Humboldt-Universität zu Berlin
Institut für deutsche Sprache und Linguistik – Korpuslinguistik

Das Falko-Handbuch
Korpusaufbau und Annotationen
Version 2.0

Reznicek, Marc; Lüdeling, Anke; Krummes, Cedric; Schwantuschke, Franziska

Stand vom:
10. September 2012

http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko

Übersicht

1. Falko-Korpus

Das Falko-Gesamtkorpus V2.0 setzt sich aus sechs Subkorpora zusammen. Sie unterscheiden sich in zwei Faktoren: Schreibaufgabe und Muttersprache. Für die Zusammenfassungen liegen außerdem die Originalvorlagen vor.
Die Aufsatzkorpora liegen in zwei unterschiedlichen Bearbeitungsversionen vor. Basisannotationen wurden für alle Subkorpora (FalkoEssay und FalkoEssayWHIG) vorgenommen. Das FalkoEssayL2v2.3 enthält allerdings darüber hinaus Dependenzbäume für die ZH1 (siehe Abschnitt 4).

	Lernerkorpus	**muttersprachliches Kontrollkorpus**	**Vorlagenkorpus**	Σ
Zusammenfassungen	FalkoSummaryL2 V1.2 (40.787 Tokens[1])	FalkoSummaryL1 V1.2 (21.184 Tokens)	FalkoSummaryVL (11.016 Tokens)	72.987
Aufsätze	FalkoEssayL2 V2.3 (122.791 Tokens)	FalkoEssayL1 V2.3 (68.480 Tokens)		308.460
	FalkoEssayL2WHIG V2.0 (117.189 Tokens)			
Σ	280.767	89.664	11.016	381.447

Erhebungsbedingungen:

Alle Texte sind unter Prüfungsbedingungen entstanden. Die Kontrollkorpora wurden unter den gleichen Bedingungen und mit den gleichen Anforderungen erhoben wie die Lernerkorpora.

Metadaten:

Zu allen Texten wurden auch umfangreiche Metadaten zu Alter, Geschlecht, akademischem Hintergrund, sprachlicher Biografie und Erhebungssituation erfasst und so aufbereitet, dass sie für die Generierung individueller Adhoc-Subkorpora dienen können.

Basisannotation:

Für alle Daten wurden die Wortarten (POS) des STTS (Schiller et al. 1995) und Lemmata (lemma) mit dem Treetagger (Schmid 1994) automatisch annotiert. Für das EssayWHIG-Korpus liegen parallel die POS- und morphologischen Tags (morph) des rfTaggers (Schmid & Laws 2008) ebenfalls auf die STTS-Tags gemappt vor.

Die manuelle Annotation der Zielhypothesen (siehe Abschnitt Abschnitt 4) wurde anfangs mit EXMARaLDA (Schmidt 2004,2005) und später Microsoft Excel mithilfe des Falko-Excel-AddIns (Reznicek 2012) vorgenommen. In einem zweiten Schritt wurden auch die Zielhypothesen automatisch mit POS-, lemma (und morph-Annotationen versehen). Für die Essaykorpora wurden dann alle Abweichungen (ZH…Diff, siehe Abschnitt 4.4) zwischen den Zielhypothesen und der Lernerreferenzebene (ctok, siehe Abschnitt 3.6) sowie der jeweiligen Annotationsebenen automatisch annotiert. Auf Basis dieser POS-Annotationen wurden für die Essay-Korpora Satzspannen erzeugt (siehe Abschnitt 4.5). Darüber hinaus wird in einer Makrostrukturebene ("macro") der vom Schreiber wiederholte Titel (title) usw. ausgewiesen. Start und Ende eines Dokuments werden auf der Ebene

[1]Tokenanzahlen beziehen sich auf die Ebene der Lernertexte ohne die durch die Zielhypothesen verursachten Leertokens.

"TXTstructure" annotiert, um in ANNIS2 (das genuin keine Textgrenzen kennt) wiederauffindbar zu bleiben. Fremdsprachliches Material und dessen Herkunftssprache werden auf der Ebene "fm" festgehalten.

Konvertierung:

Die in Excel annotiertenen Dateien wurden ins PAULA-XML-Format (Chiarcos et al. 2008) exportiert (Exmaralda-Excel-AddIn: Zeldes 2011).
Im nächsten Schritt wurden diese Daten dann mit SaltNPepper 1.0 (Zipser 2009) nach relANNIS konvertiert, das als genuines relationales Datenbankformat für das Suchwerkzeug ANNIS2 (Zeldes et al. 2009) dient.
Die Korpora liegen in folgenden Formaten vor:

- Original-Textdokumente
- Excel-Tabellen (Blatt 1: Text & Annotationen, Blatt 2: Metadaten)
- PAULA-XML
- relAnnis

Zugang:

Die Korpora sind frei zugänglich und können auf der Internetseite des Instituts für deutsche Sprache und Linguistik der Humboldt-Universität online durchsucht werden.

http://korpling.german.hu-berlin.de/falko-suche

Darüber hinaus sind auch die Rohdaten sowie alle Annotationen für die nicht-kommerzielle Nutzung nach Unterschreiben einer Lizenzvereinbarung frei erhältlich (siehe Abschnitt XXX).

1.1. Übersicht über die Annotationsebenen:

FalkoSummaryL1 & VL

Annotation	Beispiel	Erläuterung
word	word=","	Originaltext mit aufeinanderfolgenden Tokens
pos	pos="NN"	Originaltext: Treetagger-POS-tags (STTS)
lemma	lemma="d"	Originaltext: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]

FalkoSummaryL2

Schicht	Annotation	Beispiel	Erläuterung
	word	word=","	Originaltext mit aufeinanderfolgenden Tokens
	pos	pos="NN"	Originaltext: Treetagger-POS-tags (STTS)
	lemma	lemma="d"	Originaltext: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
	target hypothesis		Ziehypothese[2]

[2] Die Zielhypothesen der Zusammenfassungen folgen nicht den Richtlinien dieses Manuals und sind beschrieben im Handbuch der Annotation der Stellungsfelder bei Falko (2006).

	cpos	manuelle POS-Tags
	transcriptor comment	Kommentare des Transkribenten
Topologische Felder	matrix-satz	Martrixsatz 1
	matrix-satz_felder	Martrixsatz 1: topologische Felder
	konstituenten-satz_1	Konstituentensatz 1
	konstituenten-satz_1_ felder	KS1: topologische Felder
	konstituenten-satz_1_ felder_2	KS1: topologische Felder
	matrix-satz_2	Martrixsatz 2
	konstituenten-satz_2	Konstituentensatz 2
	konstituenten-satz_2_felder	KS2: topologische Felder
	konstituenten-satz_2_ felder_2	KS2: topologische Felder
	konstituenten-satz_3	Konstituentensatz 3
	konstituenten-satz_3_ felder	KS3: topologische Felder
	konstituenten-satz_3_ felder_2	KS3: topologische Felder
syntaktische Beschreibung	syntax_description_1	Syntaktische Beschreibung 1
	syntax_classification_1	Syntaktische Klassifikation 1
	syntax_classification_pos_1	Syntaktische Klassifikation POS 1
	syntax_hypothesis_1	syntaktische Zielhypothese
	syntax_description_2	Syntaktische Beschreibung 2
	syntax_classification_2	Syntaktische Klassifikation 2
	syntax_classification_pos_2	Syntaktische Klassifikation POS 2

FalkoEssayL1 & L2

Schicht	Annotation	Beispiel	Erläuterung
	tok	tok =","	Basisebene mit Lücken
learner	TXTstructure	TXTstructure="start"	Text: erstes Token = "start", letztes Token = "end"
	macro	macro="title"	Text: Textabschnitte: title, subtitle, structure, comment, end
	fm	fm="eng"	Text: fremdsprachliches Material: Sprachkürzel
falko	word	word=","	Originaltext mit aufeinanderfolgenden Tokens
	pos	pos="NN"	Originaltext: Treetagger-POS-tags (STTS)
	lemma	lemma="d"	Originaltext: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
ctok	ctok	ctok=","	Kopie von *word* mit korrigierten Tokenisierungsfehlern /()-'!.? Werden richtig abgetrennt, *ctok* dient als Grundlage für die Diff-Abweichungsannotationen
	ctokpos	ctokpos="NN"	
	ctoklemma	ctoklemma="d"	

ZH0	ZH0	ZH0=","	ZH0: Zielhypothese 1, in der alle Bewegungen rückgängig gemacht wurden.
	ZH0Diff	ZH0Diff="CHA"	ZH0: Abweichungen *ZH0 - ctok*
	ZH0S	ZH0S="s5"	ZH0: Satzspannen auf Grundlage von *ZH0gposDiff*
	ZH0pos	ZH0pos="NN"	ZH0: Treetagger-POS-Tags (STTS)
	ZH0posDiff	ZH0posDiff="CHA"	ZH0: Abweichungen *ZH0pos - ctokpos*
	ZH0gpos	ZH0gpos="NN"	ZH0: manuell korrigierte POS-tags
	ZH0gposDiff	ZH0gposDiff="CHA"	ZH0: Abweichungen *ZH0gpos - ctokpos*
	ZH0lemma	ZH0lemma="d"	ZH0: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
	ZH0lemmaDiff	ZH0lemmaDiff="CHA"	ZH0: Abweichungen *ZH0lemma - ctoklemma*
ZH1	ZH1	ZH1=","	ZH1: Zielhypothese 1: minimale Normalisierungsebene: Orthografie, Morphosyntax
	ZH1Diff	ZH1Diff="CHA"	ZH1: Abweichungen ZH1 - ctok
	ZH1S	ZH1S="s4"	ZH1: Satzspannen auf Grundlage von ZH1gposDiff
	ZH1pos	ZH1pos="NN"	ZH1: Treetagger-POS-tags (STTS)
	ZH1posDiff	ZH1posDiff="CHA"	ZH1: Abweichungen ZH1pos - ctokpos
	ZH1gpos	ZH1gpos="NN"	ZH1: manuell korrigierte POS-tags
	ZH1gposDiff	ZH1gposDiff="CHA"	ZH1: Abweichungen ZH1gpos - ctokpos
	ZH1lemma	ZH1lemma="d"	ZH1: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
	ZH1lemmaDiff	ZH1lemmaDiff="CHA"	ZH1: Abweichungen ZH1lemma - ctoklemma
ZH2	ZH2	ZH2=","	ZH2: Zielhypothese 2: Semantik, Pragmatik, Lexik
	ZH2Diff	ZH2Diff="CHA"	ZH2: Abweichungen ZH2 - ctok
	ZH2S	ZH2S="s3"	ZH2: Satzspannen auf Grundlage von ZH2gposDiff
	ZH2pos	ZH2pos="NN"	ZH2: Treetagger-POS-tags (STTS)
	ZH2posDiff	ZH2posDiff="CHA"	ZH2: Abweichungen ZH2pos - ctokpos
	ZH2lemma	ZH2lemma="d"	ZH2: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
	ZH2lemmaDiff	ZH2lemmaDiff="CHA"	ZH2: Abweichungen ZH2lemma - ctoklemma
Zhverb	ZHverb	ZHverb=","	ZHverb: Zielhypothese für komplexe Verben
	ZHverbDiff	ZHverbDiff="CHA"	ZHverb: Abweichungen ZHverb - ctok
	ZHverbS	ZHverbS="s2"	ZHverb: Satzspannen auf Grundlage von ZHverbgposDiff
	ZHverbpos	ZHverbpos="NN"	ZHverb: Treetagger-POS-tags (STTS)
	ZHverbposDiff	ZHverbposDiff="CHA"	ZHverb: Abweichungen ZHverbpos - ctokpos
	ZHverblemma	ZHverblemma="d"	ZHverb: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
	ZHverblemmaDiff	ZHverblemmaDiff="CHA"	ZHverb: Abweichungen ZHverblemma - ctoklemma
	verbkategorie	verbkategorie="vpräf"	ZHverb: Unterscheidung: Präfix vs. Partikelverb
	verbform	verbform="fin"	ZHverb: morphosyntaktische Form im Satz
	verblemma	verblemma="auszahlen"	ZHverb: Lemma des komplexen Verbs

	verbfehlertyp	verbfehlertyp="orth"	ZHverb: Fehlerklassifikation für komplexe Verben, je Fehler ein Tag
	verbfehlertyp_all	verbfehlertyp_all="orth"	ZHverb: Fehlerklassifikation für komplexe Verben, alle Fehler pro Token
dep	dep	node & node & #1 ->dep #2	ZH1: Dependenzen auf der ZH1
	func	node & node & #1 ->dep[func="DET"] #2	ZH1: Grammatische Funktion nach Foth (2006)

FalkoEssayL2WHIGv2.0

Schicht	Annotation	Beispiel	Erläuterung
	tok	tok=","	Basisebene mit Lücken
learner	TXTstructure	TXTstructure="start"	Text: erstes Token = "start", letztes Token = "end"
	macro	macro="title"	Text: Textabschnitte: title, subtitle, structure, comment, end
falko	word	word=","	Originaltext mit aufeinanderfolgenden Tokens
	pos	pos="NN"	Originaltext: Treetagger-POS-tags (STTS)
	lemma	lemma="d"	Originaltext: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
ctok	ctok	ctok=","	Kopie von *word* mit korrigierten Tokenisierungsfehlern /()-'!.? Werden richtig abgetrennt, *ctok* dient als Grundlage für die Diff-Abweichungsannotationen
	ctokpos	ctokpos="NN"	
	ctoklemma	ctoklemma="d"	
	ctokrfPos	ctokrfPos= "NN"	ctok: rfTagger-POS-tags (STTS)
	ctokrfMorph	ctokrfMorph= "ADV"	ctok: rfTagger-morphologische Information
ZH0	ZH0	ZH0=","	ZH0: Zielhypothese 1, in der alle Bewegungen rückgängig gemacht wurden.
	ZH0Diff	ZH0Diff="CHA"	ZH0: Abweichungen *ZH0 - ctok*
	ZH0S	ZH0S="s5"	ZH0: Satzspannen auf Grundlage von *ZH0gposDiff*
	ZH0pos	ZH0pos="NN"	ZH0: Treetagger-POS-tags (STTS)
	ZH0posDiff	ZH0posDiff="CHA"	ZH0: Abweichungen *ZH0pos - ctokpos*
	ZH0rfPos	ZH0rfPos= "NN"	ZH0: rfTagger-POS-tags (STTS)
	ZH0rfPosDIff	ZH0rfPosDiff= "CHA"	ZH0: Abweichungen *ZH0rfPos* - ctokrfPos
	ZH0rfMorph	ZH0rfMorph= "ADV"	ZH0: rfTagger-morphologische Information
	ZH0rfMorphDiff	ZH0rfMorphDIff= "CHA"	ZH0: Abweichungen *ZH0rfMorph - ctokrfMorph*
	ZH0gpos	ZH0gpos="NN"	ZH0: manuell korrigierte POS-tags
	ZH0gposDiff	ZH0gposDiff="CHA"	ZH0: Abweichungen *ZH0gpos - ctokpos*
	ZH0lemma	ZH0lemma="d"	ZH0: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
	ZH0lemmaDiff	ZH0lemmaDiff="CHA"	ZH0: Abweichungen *ZH0lemma - ctoklemma*

ZH1	ZH1	ZH1=","	ZH1: Zielhypothese 1: minimale Normalisierungsebene: Orthografie, Morphosyntax
	ZH1Diff	ZH1Diff="CHA"	ZH1: Abweichungen ZH1 - ctok
	ZH1S	ZH1S="s4"	ZH1: Satzspannen auf Grundlage von ZH1gposDiff
	ZH1pos	ZH1pos="NN"	ZH1: Treetagger-POS-tags (STTS)
	ZH1posDiff	ZH1posDiff="CHA"	ZH1: Abweichungen ZH1pos - ctokpos
	ZH1rfPos	ZH1rfPos= "NN"	ZH1: rfTagger-POS-tags (STTS)
	ZH1rfPosDIff	ZH1rfPosDiff= "CHA"	ZH1: Abweichungen ZH1rfPos - ctokrfPos
	ZH1rfMorph	ZH1rfMorph= "ADV"	ZH1: rfTagger-morphologische Information
	ZH1rfMorphDiff	ZH1rfMorphDIff= "CHA"	ZH1: Abweichungen ZH1rfMorph - ctokrfMorph
	ZH1gpos	ZH1gpos="NN"	ZH1: manuell korrigierte POS-tags
	ZH1gposDiff	ZH1gposDiff="CHA"	ZH1: Abweichungen ZH1gpos - ctokpos
	ZH1lemma	ZH1lemma="d"	ZH1: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
	ZH1lemmaDiff	ZH1lemmaDiff="CHA"	ZH1: Abweichungen ZH1lemma - ctoklemma

ZH2	ZH2	ZH2=","	ZH2: Zielhypothese 2: Semantik, Pragmatik, Lexik
	ZH2Diff	ZH2Diff="CHA"	ZH2: Abweichungen ZH2 - ctok
	ZH2S	ZH2S="s3"	ZH2: Satzspannen auf Grundlage von ZH2gposDiff
	ZH2pos	ZH2pos="NN"	ZH2: Treetagger-POS-tags (STTS)
	ZH2posDiff	ZH2posDiff="CHA"	ZH2: Abweichungen ZH2pos - ctokpos
	ZH2rfPos	ZH2rfPos= "NN"	ZH2: rfTagger-POS-tags (STTS)
	ZH2rfPosDIff	ZH2rfPosDiff= "CHA"	ZH2: Abweichungen ZH2rfPos - ctokrfPos
	ZH2rfMorph	ZH2rfMorph= "ADV"	ZH2: rfTagger-morphologische Information
	ZH2rfMorphDiff	ZH2rfMorphDIff= "CHA"	ZH2: Abweichungen ZH2rfMorph - ctokrfMorph
	ZH2lemma	ZH2lemma="d"	ZH2: Treetagger-Lemmata, unbekannte Lemmata --> [unknown]
	ZH2lemmaDiff	ZH2lemmaDiff="CHA"	ZH2: Abweichungen ZH2lemma - ctoklemma

1.2. Format der Metadaten:

Für alle Texte in Falko wurden Metadaten erhoben. Mit der Erhebung der WHIG-Daten wurden diese erweitert. Alle Kategorien bleiben aber abwärtskompatibel.
Die Kategorien, die in der vorliegenden Version 2.0 bereits vorhanden sind, sind mit dem Index „V2.0“ gekennzeichnet.

Schicht	Annotation	Beispiel	Erläuterung
Text	**corpus**	FalkoEssayWHIGL2v2.0	Korpusname
	subcorpus	UCL 2011	Subkorpus für Erhebung
	transcriptionName	BNG2-2011-03-201	Kürzel-Datum-Index
	collectionDate	16.03.11	Erhebungsdatum
	corrector	FS, MR, CT	Autoren der Zielhypothesen
	topic	Kriminalität	Aufgabentitel

	originalFilename	2011-03-16-10-2-05	Ursprüngliche Textdatei
Autor	**name**	C5DD1B2697720FE692C529688D3F4F8D	verschlüsselter Name
	firstName	E39E74FB4E80BA656F773669ED50315A	verschlüsselter Vorname
	birthYear	1992	Geburtsdatum
	sex	f	Geschlecht
	mayorSubject	final	Studienfach
	degree	Hochschulreife	höchster Abschluss
	ctest	65	C-Test-Ergebnis (von 100)
Sprachbiografie	**l1_1**	eng	Muttersprache 1 (am besten beherrschte Muttersprache)
	l1_1_since	0	gelernt seit (0=Geburt)
	l1_1_duration	216	Sprachanwendung seit (in Monaten)
	l1_1_school	Ja	Sprachunterricht in der Schule
	l1_1_university	Nein	Sprachunterricht in der Universität
	l1_1_langschool	Nein	Sprachunterricht in einer Sprachschule
	l1_1_awaymonths	1	Auslandsaufenthalt ins Zielsprachland (in Monaten)
	l1_1_awayplace	Vancouver/London	Auslandsaufenthaltsort (für Muttersprache = Ort des Erstspracherwerbs)
	l1_2	cmn	
	l1_2_since	0	
	l1_2_duration	216	
	l1_2_school	Nein	
	l1_2_university	Nein	
	l1_2_langschool	Ja	Muttersprache 2 (am zweibesten beherrschte Muttersprache)
	l1_2_awaymonths	0	
	l1_2_awayplace	0	
	l2_1	deu	
	l2_1_since	13	
	l2_1_duration	12	
	l2_1_school	Ja	
	l2_1_university	Ja	
	l2_1_langschool	Ja	Fremdsprache 1 (am besten beherrschte Fremdsprache)
	l2_1_awaymonths	5	
	l2_1_awayplace	Cologne	
	l2_1	fra	
	l2_1_since	10	Fremdsprache 2 (am zweitbesten beherrschte Fremdsprache)
	l2_1_duration	11	
	l2_1_school	Ja	

l2_1_university	Ja	
l2_1_langschool	Ja	
l2_1_awaymonths	5	
l2_1_awayplace	Bordeaux, Paris	
comments	fra: I never really liked french	sprachbibliographierelevante Kommentare

1.3. Falko Ländercodes nach ISO 639-3

Code	Sprache
afr	Afrikaans
cat	Katalanisch
ces	Tschechisch
cma	Maa
dan	Dänisch
ell	Neugriechisch
eng	Englisch
fin	Finnisch
fra	Französisch
hbs	Serbokroatisch
hin	Hindi
hun	Ungarisch
iii	Sichuan-Yi
ita	Italienisch
jpn	Japanisch
kik	Kikuyu
kor	Koreanisch
kua	Oshivambo
lub	Kiluba

Code	Sprache
luy	Luhya
nde	Nord-Ndebele
nld	Niederländisch
nor	Norwegisch
pol	Polnisch
ron	Rumänisch
rus	Russisch
slk	Slowakisch
sme	Nord-Sami
spa	Spanisch
sqi	Albanisch
swe	Schwedisch
tat	Tatarische
tur	Türkisch
ukr	Ukrainisch
uzb	Usbekisch
vie	Vietnamesische
zho	Chinesisch

2. Falko Zusammenfassungskorpus (Summary-Korpus)

Das Falko-Summary-Korpus besteht aus drei Subkorpora.

2.1. Lernertexte (FalkoSummaryL2):

Dieser Teil enthält Textzusammenfassungen, die von fortgeschrittenen Lernern des Deutschen erstellt wurden. Die Texte sind Zusammenfassungen von linguistischen und literaturwissenschaftlichen Fachtexten, die als Teil der obligatorischen Sprachstandsbestimmung für ausländische Studierende verfasst wurden. Die Daten wurden an der Freien Universität Berlin erhoben. Ausländische Studierende, die in einem germanistischen Hauptfach eingeschrieben sind, müssen nach dem

Grundstudium eine Sprachprüfung absolvieren, in der sie nachweisen, dass sie einen germanistischen Fachtext verstehen und sich fachsprachlich ausdrücken können. Diese Sprachstandsbestimmung ist eine Voraussetzung für die Zulassung zur Zwischenprüfung. Die Prüfung wird durch das Studiengebiet Deutsch als Fremdsprache des Instituts für Deutsche und Niederländische Philologie verantwortet. Die Textvorlagen wurden von Maik Walter (Linguistik) und Almut Hille (Literaturwissenschaft) ausgewählt. Neben dem schriftlichen Teil absolvieren die Studierenden einen mündlichen Teil. Die Verfasser der Texte haben die DSH-Prüfung erfolgreich absolviert und werden deshalb als fortgeschrittene Lerner (auf dem Niveau C1 - C2 des Europäischen Referenzrahmens) eingestuft. Der Prüfungskontext ist unten beschrieben. Die Texte wurden von Julia Kassubek, Katja Jansen und Karin Schmidt digitalisiert und mehrfach von verschiedenen Mitarbeitern und Studierenden korrigiert.

2.2. Muttersprachlertexte (FalkoSummaryL1):

Dieser Teil enthält Textzusammenfassungen, die von deutschen Muttersprachlern (Studierenden der Freien Universität Berlin und der Humboldt-Universität zu Berlin) erstellt wurden. Die Texte sind Zusammenfassungen derselben linguistischen und literaturwissenschaftlichen Fachtexte, die auch von den Lernern bearbeitet wurden. Die Rahmenbedingungen für die Erhebungen waren vergleichbar (90 Minuten, keine Hilfsmittel), allerdings wurden die Texte nicht als Prüfungsleistung erhoben. Auch hier wurden mithilfe eines Fragebogens Metadaten über die Verfasser erhoben.

2.3. Vorlagentexte (FalkoSummaryVL):

Dieser Teil enthält die linguistischen und literaturwissenschaftlichen Fachtexte, die als Vorlage für die Textzusammenfassungen in den anderen Subkorpora verwendet wurden.
Im Folgenden sind die einzelnen Erhebungszeiträume für FalkoSummaryL2 und die Zusammensetzung aller Subkorpora dokumentiert. In der Dokumentation sind die Vorlagentexte abkürzend benannt – die genauen Angaben zu jedem Vorlagentext finden sich in der Dokumentation der FalkoSummaryVL.

2.4. FalkoSummaryL2 1.2

Die folgenden 6 Datenerhebungen wurden als Grundlage für das Korpus FalkoSummaryL2 1.1 verwendet: **Datum der Erhebung**	**Anzahl der Lernertexte**		
	männlich	weiblich	Σ
09.02.2004	5	19	24
01.07.2004	7	13	20
20.01.2005	1	14	15
27.06.2005	3	20	23
06.02.2005	0	16	16
02.02.2006	3	6	9
Σ	**19**	**88**	**107**

Tokenanzahl

Lerner	98	Texte	107	Tokens	40923	Ø Text	382,46

Insgesamt haben 98 Lerner 197 Texte verfasst, von 9 Lernern sind daher zwei verschiedene Texte im Subkorpus enthalten. Die folgenden Texte wurden zusammengefasst:

Vorlagentext	**N**
Berlinromane	5
Entscheidungen	6
Epochen	5
Hermeneutik	18
Pragmatik	11
Realismus	9
Schlaf	9
Semantik	11
Syntax	4
Textgrenzen	12
Valenz	14
Volksmärchen	3
Σ	**107**

Datenerhebung vom 09.02.2004

Die Aufgabe bestand darin, einen literaturwissenschaftlichen (N=18) bzw. linguistischen (N=6) Fachtext zusammenzufassen.

Angaben zu den Ausgangstexten:

(a) Witte, Bernd (1993): Das Gericht, das Gesetz, die Schrift. Über die Grenzen der Hermeneutik am Beispiel von Kafkas Türhüter - Legende. In: Bogdal, Klaus-Michael (Hg.): Neue Literaturtheorien in der Praxis. Textanalysen von Kafkas "Vor dem Gesetz". Opladen: Westdeutscher Verlag, S. 94-97. Als Datei hermeneutik.rtf Teil des Korpus.

(b) Miller, George A. (1993): Unterscheidungen treffen. In: ders.: Wörter. Streifzüge durch die Psycholinguistik. Spektrum. Heidelberg, Berlin, New York: Akademischer Verlag, S. 223. Als Datei entscheidungen.rtf Teil des Korpus.

Aufgabenstellung

a) Beantworten Sie bitte folgende Fragen anhand des Textes.
   1. Was ist Hermeneutik?
   2. Warum ist Franz Kafkas Legende "Vor dem Gesetz" für eine hermeneutische Analyse geeignet?
   3. Was ist das "Paradoxe" in Kafkas Text?

b)

1. Fassen Sie den folgenden Text mit eigenen Worten zusammen.
2. Geben Sie ein Beispiel für eine nicht informationsübermittelnde Kommunikation (mit nicht ernsthaften Menschen).

Prüfungskontext

- keine Vorbereitungszeit
- keine Textkenntnis
- keine Hilfsmittel
- handschriftlich verfasste Klausuren unter Aufsicht
- Zeit: 90 Minuten

Datum der Erhebung	**Anzahl der Teilnehmer** männlich	weiblich	**L1**	**L2**
09.02.2004	5	19	Polnisch (11) Portugiesisch (2) Russisch (2) Georgisch (2) Koreanisch (2) Französisch (1) Bulgarisch (1) Weißrussisch (1) Englisch (1) Persisch (1)	Deutsch (24) Englisch (20) Französisch (1) Russisch (11) Ukrainisch (1) Spanisch (3) Niederländisch (4) Japanisch (1) Chinesisch (1) Italienisch (1)
Σ	24			

2.4.1. D

atenerhebung vom 01.07.2004

Die Aufgabe bestand darin, einen literaturwissenschaftlichen (N=9) bzw. linguistischen (N=11) Fachtext zusammenzufassen.

Angaben zu den Ausgangstexten:

(a) Sprengel, Peter (1998): III. Stile und Richtungen. 1. Realismus. In: ders.: Geschichte der deutschsprachigen Literatur 1870-1900. Von der Reichsgründung bis zur Jahrhundertwende. München: Verlag C.H. Beck, S. 99-101. Als Datei realismus.rtf Teil des Korpus.

(b) Meibauer, Jörg (1999): Pragmatische Erwerbsprinzipien. In: ders.: Pragmatik. Eine Einführung. Tübingen: Stauffenburg, S. 170-172. Als Datei pragmatik.rtf Teil des Korpus.

Aufgabenstellung:

3. Fassen Sie bitte den folgenden Text zusammen.
4. Fassen Sie den Text mit eigenen Worten zusammen.

Prüfungskontext

- keine Vorbereitungszeit
- keine Textkenntnis
- keine Hilfsmittel
- handschriftlich verfasste Klausuren unter Aufsicht
- Zeit: 90 Minuten

Datum der Erhebung	**Anzahl der Teilnehmer** männlich	weiblich	**L1**	**L2**
01.07.2004	7	13	Polnisch (5) Chinesisch (3) Russisch (2) Japanisch (2) Georgisch (1) Persisch (1) Slowenisch (1) Arabisch (1) Ungarisch (1) Türkisch (1) Deutsch (1) Litauisch (1) Thai (1)	Deutsch (20) Englisch (14) Russisch (4) Spanisch (3) Französisch (2) Chinesisch (1) Italienisch (1)
Σ	20			

Die Aufgabe bestand darin, einen literaturwissenschaftlichen (N=3) bzw. linguistischen (N=12) Fachtext zusammenzufassen.

Angaben zu den Ausgangstexten:

(a) Klotz, Volker (2002): Kunstmärchen: Name und Sachverhalt. In: ders.: Das europäische Kunstmärchen. Fünfundzwanzig Kapitel seiner Geschichte von der Renaissance bis zur Moderne. 3. überarbeitete und erweiterte Auflage. München: Wilhelm Fink Verlag, S. 7-8. Als Datei volksmaerchen.rtf Teil des Korpus.

(b) Linke, Angelika / Nussbaumer, Markus / Portmann, Paul R. (21994): Textgrenzen. In: dies.: Studienbuch Linguistik. Tübingen: Max Niemeyer Verlag, S. 255-256. Als Datei textgrenzen.rtf Teil des Korpus.

Aufgabenstellung:

1. Fassen Sie bitte den folgenden Text zusammen.
2. Fassen Sie den Text mit eigenen Worten zusammen.

Prüfungskontext

- keine Vorbereitungszeit
- keine Textkenntnis
- keine Hilfsmittel

• handschriftlich verfasste Klausuren unter Aufsicht
• Zeit: 90 Minuten

Datum der Erhebung	**Anzahl der Teilnehmer** männlich	weiblich	**L1**	**L2**
20.01.2005	1	14	Polnisch (4) Russisch (4) Bulgarisch (2) Ukrainisch (1) Serbo-Kroatisch (1) Japanisch (1) Armenisch (1) Englisch (1) Chinesisch (1)	Deutsch (15) Englisch (13) Russisch (3) Französisch (2) Spanisch (2) Italienisch (2) Rumänisch (1) Latein (1) Bosnisch (1)
Σ	15			

2.4.2. D

atenerhebung vom 27.06.2005

Die Aufgabe bestand darin, einen linguistischen (N=9) bzw. literaturwissenschaftlichen (N=14) Fachtext zusammenzufassen.

Angaben zu den Ausgangstexten:
(a) Eisenberg, Peter (2004): 3.2.2 Valenz und Bedeutung. Grundpositionen. In: ders.: Grundriss der deutschen Grammatik. Band 2: Der Satz. 2., überarbeitete und aktualisierte Auflage. Stuttgart, Weimar: Metzler, S. 71-72. Als Datei valenz.rtf Teil des Korpus.
(b) Alt, Peter-André (2002): Der Schlaf der Vernunft. Literatur und Traum in der Kulturgeschichte der Neuzeit. München: Beck, S. 10-12. Als Datei schlaf.rtf Teil des Korpus.

Aufgabenstellung:

3. Fassen Sie bitte den folgenden Text zusammen.
4. Fassen Sie den Text mit eigenen Worten zusammen.

Prüfungskontext

• keine Vorbereitungszeit
• keine Textkenntnis
• keine Hilfsmittel
• handschriftlich verfasste Klausuren unter Aufsicht
• Zeit: 90 Minuten

Datum der Erhebung	**Anzahl der Teilnehmer** männlich	weiblich	**L1**	**L2**
27.06.2005	3	20	Polnisch (10) Russisch (10) Weißrussisch (3) Ukrainisch (3) Portugiesisch(1) Mongolisch (1)	Deutsch (23) Englisch (20) Russisch (9) Französisch (5) Spanisch (4) Italienisch (1) Rumänisch (1) Latein (1) Polnisch (1) Niederländisch (1) Japanisch (1)
Σ	23			

2.4.3. D

atenerhebung vom 02.02.2006

Die Aufgabe bestand darin, einen literaturwissenschaftlichen (N=5) bzw. linguistischen (N=11) Fachtext zusammenzufassen.

Angaben zu den Ausgangstexten:
(a) Rosenberg, Rainer (2001): Epochen. In: Brackert, Helmut/ Stückrath, Jörn (Hg.): Literaturwissenschaft. Ein Grundkurs. Reinbek: Rowohlt Taschenbuch Verlag, S. 269-272.
Als Datei epochen.rtf Teil des Korpus.
(b) Wunderlich, Dieter (1991): Welche Verfahren gibt es zur Bedeutungsanalyse? In: ders.: Arbeitsbuch Semantik. 2., ergänzte Auflage. Frankfurt am Main: Hain, S. 124-126.
Als Datei semantik.rtf Teil des Korpus.

Aufgabenstellung:

5. Fassen Sie bitte den folgenden Text zusammen.
6. Fassen Sie den Text mit eigenen Worten zusammen.

Prüfungskontext

- keine Vorbereitungszeit
- keine Textkenntnis
- keine Hilfsmittel
- handschriftlich verfasste Klausuren unter Aufsicht
- Zeit: 90 Minuten

Datum der Erhebung	**Anzahl der Teilnehmer** männlich	weiblich	**L1**	**L2**
6	0	16	Polnisch (5) Russisch (4) Mongolisch (1) Bulgarisch (1) Kroatisch (1) Italienisch (1) Japanisch (1) Koreanisch (1) Litauisch (1)	Arabisch (1) Englisch (16) Deutsch (16) Schwedisch (2) Französisch (2) Spanisch (3) Italienisch (3) Portugiesisch (1) Russisch (3) Litauisch (1)
Σ	16			

2.4.4. D

atenerhebung vom 06.02.2007

Die Aufgabe bestand darin, einen linguistischen (N=4) bzw. literaturwissenschaftlichen (N=5) Fachtext zusammenzufassen.

Angaben zu den Ausgangstexten:
(a) Eroms, Hans-Werner (2000): Syntax der deutschen Sprache. Berlin, New York: Walter de Gruyter, S. 47-48.
Als Datei syntax.rtf Teil des Korpus.
(b) Siebenpfeiffer, Hania (2001): Topographien des Seelischen. Berlinromane der neunziger Jahre. In: Harder, Matthias (Hg.): Bestandsaufnahmen. Deutschsprachige Literatur der neunziger Jahre aus interkultureller Sicht. Würzburg: Könighausen & Neumann, S. 85-87.
Als Datei berlinromane.rtf Teil des Korpus.

Aufgabenstellung:

7. Fassen Sie bitte den folgenden Text zusammen.
8. Fassen Sie den Text mit eigenen Worten zusammen.

Prüfungskontext

- keine Vorbereitungszeit
- keine Textkenntnis
- keine Hilfsmittel
- handschriftlich verfasste Klausuren unter Aufsicht
- Zeit: 90 Minuten

Datum der Erhebung	Anzahl der Teilnehmer männlich	weiblich	L1	L2
06.02.2007	3	6	Polnisch (4) Russisch (3) Englisch (2) Baschkirisch (1)	Deutsch (9) Englisch (6) Französisch (2) Latein (2) Altgriechisch (1) Arabisch(1) Hebräisch (1) Italienisch (1) Japanisch (1) Niederländisch(1) Spanisch (1) Türkisch (1) Tschechisch (1)
Σ		9		

2.5. Annotationen in FalkoSummaryL2

2.5.1. Zielhypothesen in Summary L2

Die Zielhypothesen im Summary-Korpus entsprechen den Entwürfen in Lüdeling al. 2005. Sie weichen somit von den hier entwickelten Richtlinien ab.

2.5.2. Annotation topologischer Felder und syntaktischer Beschreibung

Im Rahmen der Magisterarbeit von Seanna Doolittle (Doolittle 2009) wurden kanonische und unkanonische Sätze annotiert und für erstere die folgenden Felderannotationen vergeben

Vorfeld	Linke Satzklammer	Mittelfeld	Rechte Satzklammer	Nachfeld
VF	LSK	MF	RSK	NF

Weiterhin wurde auf der Grundlage dieser Annotationen eine syntaktische Beschreibung vorgenommen. Weitere Details zu diesen Annotationen finden Sie im **Fehler! Verweisquelle konnte nicht gefunden werden.**..

2.6. FalkoSummaryL1 1.2

Vier Datenerhebungen wurden als Grundlage für das Korpus FalkoSummaryL1 1.1 verwendet. Die Zusammenfassungen wurden an der Freien Universität Berlin und an der Humboldt-Universität zu Berlin von Studierenden eines germanistischen Faches im Hauptstudium verfasst. Ein Teil der Studierenden (N=39) absolvierte den Zusatzstudiengang Deutsch als Fremdsprache an der Freien Universität Berlin. Alle Texte wurden unter den identischen Bedingungen erhoben, das betrifft insbesondere die Aufgabenstellung und die kontrollierte Datenerhebung.

Datum (Ort) der Erhebung	**Anzahl der Texte**			
	männlich	weiblich	N/A	Σ
17.02.2005 (FU Berlin)	2 (2)	5 (5)	11(11)	18
22.05.2007 (FU Berlin)	0	10 (10)	0	10
15.07./20.07./01.08.2007 (FU Berlin)	0	11 (8)	0	11
03.05./07.06./13.06./09.07./20.07.2007 (HU Berlin)	8 (8)	10 (10)	0	18
Σ	**10**	**36**	**11**	**57**

2.6.1. Tokenanzahl

Texte	36	Lerner	33	Tokens	21184	Ø/Text	370,62

Unterschied zur Version 1.0 – proportional zum L1-Subkorpus kompiliert
Zu jedem Vorlagentext wurden (aus ökonomischen Gründen) die halbe Anzahl der Texte (der L2) in der L1 Deutsch erhoben. Bei einer ungeraden Zahl wurde aufgerundet. Nach der ersten Erhebung wurden ebenfalls die Metadaten (L1, L2, L3,…, Dauer des Erwerbs, Alter, Geschlecht) erhoben. In 11 Fällen liegen keine Metadaten vor.

Aufgabenstellung:

- identisch mit den L2-Erhebungen (s.o.)

Prüfungskontext

- keine Vorbereitungszeit
- keine Textkenntnis
- keine Hilfsmittel
- handschriftlich verfasste Klausuren unter Aufsicht
- Zeit: 90 Minuten

Vorlagentext	N
Berlinromane	5
Entscheidungen	6
Epochen	5
Hermeneutik	18
Pragmatik	11
Realismus	9
Schlaf	9
Semantik	11
Syntax	4
Textgrenzen	12
Valenz	14
Volksmärchen	3
Σ	**57**

2.7. FalkoSummaryVL 1.0

Die folgenden Texte bilden die Textbasis für das Subkorpus FalkoSummaryVL1.0:

Sigle	Quelle
Hermeneutik	Witte, Bernd (1993): Das Gericht, das Gesetz, die Schrift. Über die Gren-zen der Hermeneutik am Beispiel von Kafkas Türhüter - Legende. In: Bogdal, Klaus-Michael (Hg.): Neue Literaturtheorien in der Praxis. Text-analysen von Kafkas "Vor dem Gesetz". Opladen: Westdeutscher Verlag, S. 94-97.
Entscheidungen	Miller, George A. (1993): Unterscheidungen treffen. In: ders.: Wörter. Streifzüge durch die Psycholinguistik. Heidelberg, Berlin, New York: Spekt-rum. Akademischer Verlag, S. 223.
Pragmatik	Meibauer, Jörg (1999): Pragmatische Erwerbsprinzipien. In: ders.: Pragmatik. Eine Einführung. Tübingen: Stauffenburg, S. 170-172.
Realismus	Sprengel, Peter (1998): III. Stile und Richtungen. 1. Realismus. In: ders.: Geschichte der deutschsprachigen Literatur 1870-1900. Von der Reichs-gründung bis zur Jahrhundertwende. München: Verlag C.H. Beck, S. 99-101.
Volksmärchen	Klotz, Volker (2002): Kunstmärchen: Name und Sachverhalt. In: ders.: Das europäische Kunstmärchen. Fünfundzwanzig Kapitel seiner Ge-schichte von der Renaissance bis zur Moderne. 3. überarbeitete und erwei-terte Auflage. München: Wilhelm Fink Verlag, S.7-8.
Textgrenzen	Linke, Angelika / Nussbaumer, Markus / Portmann, Paul R. (²1994): Text-grenzen. In: dies.: Studienbuch Linguistik. Tübingen: Max Niemeyer Verlag, S. 255/-256.
Schlaf	Alt, Peter-André (2002): Der Schlaf der Vernunft. Literatur und Traum in der Kulturgeschichte der Neuzeit. München: Beck, S. 10-12.
Valenz	(a) Eisenberg, Peter (2004): 3.2.2 Valenz und Bedeutung. Grundpositio-nen. In: ders.: Grundriss der deutschen Grammatik. Band 2: Der Satz. 2., überar-

	beitete und aktualisierte Auflage. Stuttgart, Weimar: Metzler, S. 71-72.
Semantik	Wunderlich, Dieter (1991): Welche Verfahren gibt es zur Bedeutungs-analyse? In: ders.: Arbeitsbuch Semantik. 2., ergänzte Auflage. Frankfurt am Main: Hain, S. 124-126.
Epochen	Rosenberg, Rainer (2001): Epochen. In: Brackert, Helmut/ Stückrath, Jörn (Hg.): Literaturwissenschaft. Ein Grundkurs. Reinbek: Rowohlt Taschen-buch Verlag, S. 269-272.
Syntax	Eroms, Hans-Werner (2000): Syntax der deutschen Sprache. Berlin, New York: Walter de Gruyter, S. 47-48.
Berlinromane	Siebenpfeiffer, Hania (2001): Topographien des Seelischen. Berlinromane der neunziger Jahre. In: Harder, Matthias (Hg.): Bestandsaufnahmen. Deutschsprachige Literatur der neunziger Jahre aus interkultureller Sicht. Würzburg: Könighausen & Neumann, S. 85-87.

2.7.1. Tokenanzahl

Texte	12	Tokens	11114	Ø/Text	926,17

3. Falko-Aufsatzkorpus (Essay-Korpus)

Das Falko-Essay-Korpus besteht aus drei Subkorpora.

Lernertexte (FalkoEssayL2):
Dieser Teil enthält Aufsätze, die von fortgeschrittenen Lernern des Deutschen erstellt wurden. Die Texte sind argumentative Aufsätze zu einem von vier vorgegeben Themen, die aus der Gesamtmenge der im *International Corpus of Learner English* (ICLE) (Granger 1993, 2003) verwendeten Aufsatzthemen ausgewählt wurden.
Die Lernertexte stammen von Nicht-Muttersprachlern, die teilweise an Feriensprachkursen an der Freien Universität Berlin und der Humboldt-Universität zu Berlin und teilweise an ausländischen Universitäten und Goethe-Instituten erhoben wurden. Alle Lerner mussten einen Fragebogen für die Erfassung der Lernerdaten ausfüllen und in einem C-Test mindestens 60 von 100 Punkten erreichen, der vom Sprachenzentrum der Humboldt-Universität zu Berlin entwickelt wurde und dort ebenfalls eingesetzt wird.

Ergebnis im C-Test	Einstufungsniveau in Maßen des GER
60-79	B2
80-89	C1
90-100	C2

Die Texte wurden unter Aufsicht direkt in einem Texteditor geschrieben, der keine Rechtschreibkorrektur beinhaltet. Jeglicher Zugriff auf weitere Hilfsmittel bzw. das Internet wurde vorher ausgeschlossen.

Lernertexte (FalkoEssayL2WHIG):
Die Aufgabenstellung dieses Subkorpus ist mit dem des FalkoEssayL2-Subkorpus identisch. Es wurde allerdings das komplette Spektrum an Metadaten miterhoben. Die Erhebungen fanden im Rahmen des WHIG-Projektes in Kooperation zwischen der Humboldt-Universität und der Bangor University statt. (siehe WHIG-Projektseite http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/lernschwierigkeiten/WHIG)
In den folgenden Universitäten wurden die Texte erhoben:

- Aberystwyth University, German Department
- Bangor University, German Department
- Bristol University, Department of German
- University of Leeds, Department of German, Russian and Slavonic Studies
- University of Nottingham, Department of German
- Queen Mary, University of London, Department of German
- The University of Sheffield, Department of Germanic Studies

Muttersprachlertexte (FalkoEssayL1):
Dieser Teil enthält Aufsätze, die von deutschen Muttersprachlern in den Abschlussklassen dreier Gymnasien in Berlin, Eichwalde und Potsdam, sowie in einem Kurs im Studiengang „Deutsch als

Fremdsprache“ an der Freien Universität Berlin erhoben wurden. Die Texte sind Aufsätze zu denselben Themen, die auch von den Lernern bearbeitet wurden. Die Rahmenbedingungen für die Erhebungen waren vergleichbar (90 Minuten, keine Hilfsmittel). Auch hier wurden mithilfe eines Fragebogens Metadaten über die Verfasser erhoben.

Im Folgenden sind die einzelnen Erhebungszeiträume für Falko-Essay und die Zusammensetzung aller Subkorpora dokumentiert.

3.1. FalkoEssayL2 v2.3

Die Aufgabe bestand darin, zu einem der folgenden vier Themen einen argumentativen Aufsatz zu schreiben:

- Der Feminismus hat den Frauen mehr geschadet als genutzt.
- Kriminalität zahlt sich nicht aus.
- Die meisten Universitätsabschlüsse bereiten die Studenten nicht auf die wirkliche Welt vor. Sie sind deswegen von geringem Wert
- Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, den er/sie für die Gesellschaft geleistet hat.

Prüfungskontext

- keine Vorbereitungszeit
- keine Textkenntnis
- keine Hilfsmittel
- in einem Texteditor verfasste Klausuren unter Aufsicht
- Zeit: 90 Minuten

3.1.1. Übersicht über Sprache und Geschlecht der Lerner für die einzelnen Erhebungen

(Mehrfachangaben für Sprachen wurden auch mehrfach gezählt)

Erhebungsdatum	Anzahl der Teilnehmer		L1	L2
	männlich	weiblich		
02.05.2006 09.05.2006 10.05.2006	3	5	Türkisch (8)	Deutsch (8) Englisch (8) Französisch (1)
28.06.2006 11.07.2006 12.07.2006 17.07.2006	5	13	Suaheli (9) Kikuyu (5) Luo (5) Luhya (2) Meru (2)	Deutsch (18) Englisch (18) Suaheli (9) Französisch (2) Chinesisch (1)

18.07.2006 10.08.2006 12.09.2006 29.09.2006			Embu (1) Gusii (1) Kalenjin (1) Nandi (1) KiTaita (1)	Italienisch (1) Kikuyu (1) Luhya (1) Meru (1) Giryama (1) Spanisch (1) Schwedisch (1)
27.07.2006 17.08.2006	7	17	Englisch(8) Französisch(5) Neugriechisch (3) Schwedisch(3) Italienisch(2) Chinesisch(1) Dänisch(1) Hebräisch(1) Niederländisch(1) Norwegisch(1) Polnisch(1) Rumänisch(1) Russisch(1) Tschechisch(1) Ukrainisch(1)	Deutsch(24) Englisch(16) Französisch(10) Latein(10) Russisch(6) Spanisch(6) Niederländisch(2) Altenglisch(1) Indonesisch(1) Italienisch(1) Japanisch(1)
19.09.2006 26.09.2006	6	7	Englisch (5) Norwegisch (2) Griechisch (1) Finnisch (1) Französisch (1) Niederländisch (1) Polnisch (1) Spanisch (1) *(Fast alle Teilnehmer haben 2 Texte geschrieben)*	Deutsch (13) Englisch (12) Französisch (8) Italienisch (3) Spanisch (3) Griechisch (1) Latein (1) Niederländisch (2) *(Fast alle Teilnehmer hatten 2 Texte geschrieben)*
29.09.2006 01.10.2007	4	27	Dänisch(31) Schwedisch(2) Norwegisch(1)	Englisch(31) Deutsch(31) Französisch(17) Latein(8) Spanisch(4) Russisch(3) Schwedisch(3) Norwegisch(2) Italienisch(1)
04.10.2006	6	9	Usbekisch (11)	Deutsch (15)

			Russisch (4) Tadschikisch (3)	Englisch (14) Russisch (11) Usbekisch (3) Tadschikisch (2) Persisch (1) Französisch (1) Koreanisch (1)
24.10.2006 20.11.2007	5	10	Englisch (3) Japanisch (3) Chamorro (2) Meru (2) Angika (1) Embu (1) Maithili (1) Hindi (1) Koreanisch (1) Norwegisch (1) Polnisch (1) Russisch (1) Ukrainisch (1)	Deutsch (15) Englisch (11) Französisch (3) Japanisch (1) Russisch (1) Jiddisch (1)
07.12.2006	1	9	Afrikaans (8) Englisch (2)	Deutsch (10) Englisch (7) Xhosa (3) Afrikaans (2) Französisch (2) Chinesisch (1)
05.03.2007 18.05.2007 25.06.2007	1	1	Englisch (2)	Deutsch (2) Französisch (2) Tschechisch (1) Latein (1) Norwegisch (1) Chinesisch (1)
09.08.2007 20.11.2007	12	38	Russisch (13) Englisch (12) Französisch (6) Dänisch (5) Spanisch (5) Ukrainisch (4) Niederländisch (3) Polnisch (3)	Englisch (55) Deutsch (50) Französisch (28) Spanisch (18) Latein (17) Russisch (7) Italienisch (6) Niederländisch (4)

25.07.2008 06.08.2008			Rumänisch (3) Italienisch (2) Türkisch (2) Tschechisch (2) Ungarisch (2) Vietnamesisch (2) Albanisch (1) Finnisch (1) Hindi (1) Irisch (1) Katalanisch (1) Neugriechisch (1) Slovakisch (1)	Altgriechisch (2) Katalanisch (2) Schwedisch (2) Tschechisch (2) Chinesisch (3) Walisisch (2) Baskisch (1) Japanisch (1)
Σ	50	136	Dänisch(37) Englisch (32) Russisch (19) Französisch (12) Usbekisch (11) Türkisch (10) Suaheli (9) Afrikaans (8) Polnisch (6) Spanisch (6) Ukrainisch (6) Luo (5) Kikuyu (5) Niederländisch (5) Norwegisch (5) Schwedisch(5) Neugriechisch (4) Italienisch (4) Rumänisch (4) Japanisch (3) Tadschikisch (3) Tschechisch (3) Chamorro (2) Embu (2) Finnisch (2) Hindi (2) Luhya (2) Ungarisch (2) Vietnamesisch (2) Albanisch (1) Angika (1) Chinesisch(1)	Deutsch (186) Englisch (172) Französisch (74) Spanisch (32) Latein (29) Russisch (28) Italienisch (12) Suaheli (9) Niederländisch (8) Chinesisch (6) Schwedisch (6) Japanisch (3) Norwegisch (5) Tschechisch (3) Usbekisch (3) Xhosa (3) Afrikaans (2) Altgriechisch (2) Katalanisch (2) Tadschikisch (2) Tschechisch (2) Walisisch (2) Altenglisch(1) Baskisch (1) Giryama (1) Griechisch (1) Indonesisch(1) Jiddisch (1) Kikuyu (1) Koreanisch (1) Luhya (1) Meru (1)

			Griechisch (1) Gusii (1) Hebräisch(1) Irisch (1) Kalenjin (1) Katalanisch (1) KiTaita (1) Koreanisch (1) Maithili (1) Meru (4) Nandi (1) Slovakisch (1)	Persisch (1)
Σ	186			

3.1.2. Übersicht über die Orte und Textgrößen bezüglich der einzelnen Erhebungen

Erhebungsdatum	Ort	Texte	Token	Token/Text
02.05.2006 09.05.2006 10.05.2006	Cukurova University, Türkei (TRK)	2 2 4	2772	346,50
28.06.2006 11.07.2006 12.07.2006 17.07.2006 18.07.2006 10.08.2006 12.09.2006 29.09.2006	Goethe-Institut Nairobi, Kenia (KNE)	1 1 2 6 1 2 1 4	5016	278,67
27.07.2006 17.08.2006	HU Berlin (Ferienkurs), Deutschland(FK)	22 19	23883	582,51
19.09.2006 26.09.2006	Humboldt-Universität, Deutschland (HU)	13 11	12926	538,58
29.09.2006 01.10.2007	Copenhagen Business School, Dänemark (CBS)	16 15	16463	531,06
04.10.2006	National University of Uzbekistan, Usbekistan (USB)	15	6044	402,93
24.10.2006 20.11.2007	Freie Universität, Deutschland (FU)	9 14	10166	442,00
07.12.2006	Stellenbosch University, Südafrika (SA)	10	6802	680,20
05.03.2007 18.05.2007 25.06.2007	Auckland University, New Zealand (NZ)	1 1 2	2148	537
09.08.2007 20.11.2007	Humboldt-Universität Ferienkurs, Deutschland (FKB)	12 37	36559	494,04

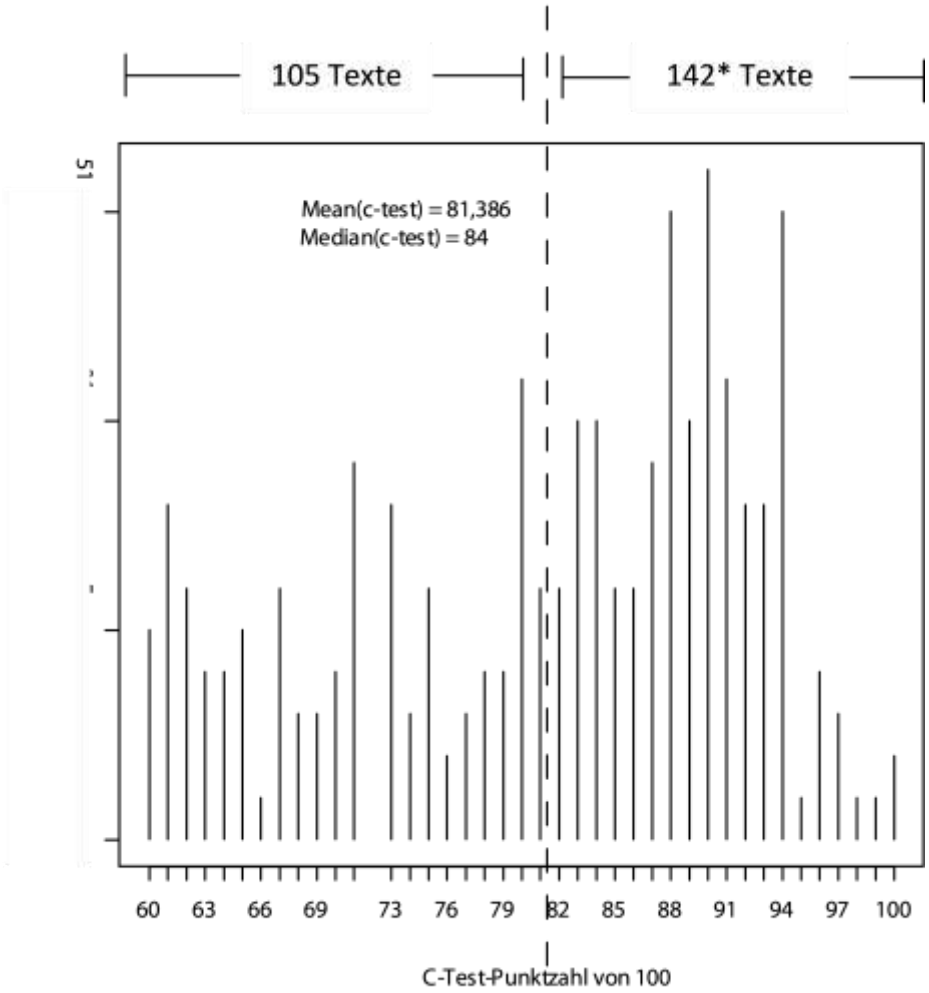
25.07.2008		1		
06.08.2008		24		
Σ		**248**	**122791**[3]	**495,08**

3.1.3. Tokenanzahl

Texte	248	Lerner	186	Tokens	122.778	Ø/Text	495,08

3.1.4. Verteilung der C-Test-Ergebnisse in FalkoEssayL2 v2.3

Insgesamt überwiegt die Zahl der Texte von sehr fortgeschrittenen Lernern (C-Test >= 80).



* Daten für einen Lerner fehlen

[3] Diese Tokenanzahl bezieht sich auf die Originaltexte und stimmt nicht mit der im Annis-Interface angezeigten Zahl überein. Dort werden die durch die Zielhypothesen entstandenen Leertokens mitgezählt.

3.2 FalkoEssayL2WHIG v2.0

Die Aufgabe bestand darin, zu einem der folgenden vier Themen einen argumentativen Aufsatz zu schreiben:

- Der Feminismus hat den Frauen mehr geschadet als genutzt.
- Kriminalität zahlt sich nicht aus.
- Die meisten Universitätsabschlüsse bereiten die Studenten nicht auf die wirkliche Welt vor. Sie sind deswegen von geringem Wert
- Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, den er/sie für die Gesellschaft geleistet hat.

Prüfungskontext

- keine Vorbereitungszeit
- keine Textkenntnis
- keine Hilfsmittel
- in einem Texteditor verfasste Klausuren unter Aufsicht
- Zeit: 90 Minuten

3.3.1. Übersicht über Sprache und Geschlecht der Lerner

(Mehrfachangaben für Sprachen wurden auch mehrfach gezählt)

Geschlecht	
männlich	58
weiblich	138

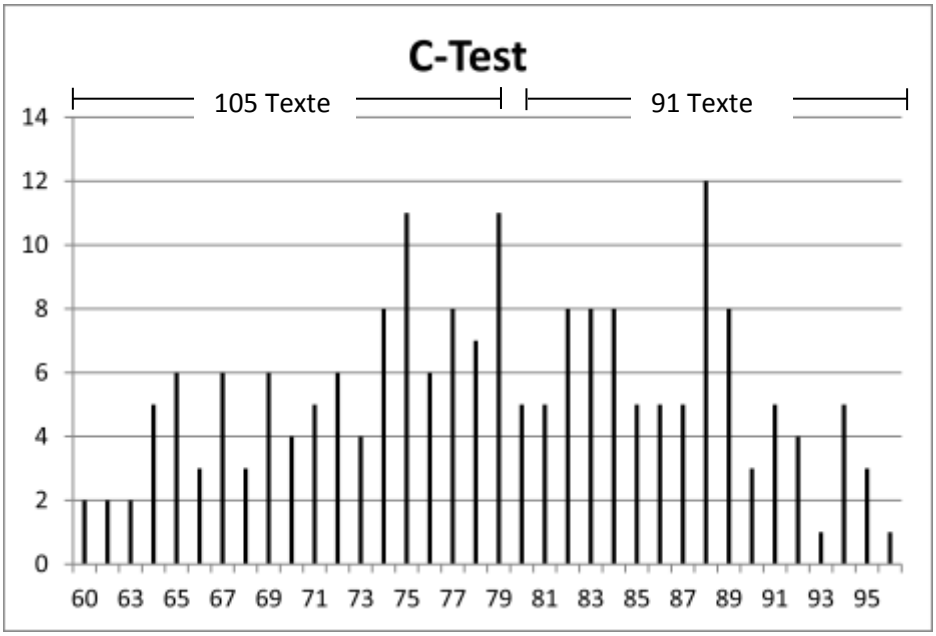
Aufgabenstellung	
Entlohnung	23
Feminismus	56
Studium	59
Kriminalität	58

L1	
eng	184
pol	2
rus	2
fra	2
nld	1
fin	1
cmn	1
ces	1
tam	1
lit	1
ita	1
Gesamtergebnis	**197**

L2	
deu	195
fra	124
spa	55
lat	45
ita	27
rus	16
eng	12
jpn	8
cym	7
nld	7
por	5
swe	5
cmn	4
ara	3
ces	3
non	3
pol	3
heb	3
gmh	3
cat	2
hbo	2
bfi	1
gle	1
sna	1
fro	1
ang	1
nap	1
ben	1
Gesamtergebnis	**539**

3.3.2. Verteilung der C-Test-Ergebnisse in FalkoEssayL2WHIG v2.0

Insgesamt überwiegt die Zahl der Texte von weniger fortgeschrittenen Lernern (C-Test < 80).



3.3 FalkoEssayL1 v2.3

Auch für die Muttersprachler bestand die Aufgabe darin, zu einem der folgenden vier Themen einen argumentativen Aufsatz zu schreiben:

- Der Feminismus hat den Frauen mehr geschadet als genutzt.
- Kriminalität zahlt sich nicht aus.
- Die meisten Universitätsabschlüsse bereiten die Studenten nicht auf die wirkliche Welt vor. Sie sind deswegen von geringem Wert
- Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, der er/sie für die Gesellschaft geleistet hat.

Prüfungskontext

- keine Vorbereitungszeit
- keine Textkenntnis
- keine Hilfsmittel
- in einem Texteditor verfasste Klausuren unter Aufsicht
- Zeit: 90 Minuten

3.3.3. Übersicht über Sprache und Geschlecht der Lerner für die einzelnen Erhebungen

(Mehrfachangaben für Sprachen wurden auch mehrfach gezählt)

Datum der Erhebung	Anzahl der Teilnehmer männlich	weiblich	L1	L2
25.10.06	5	2	Deutsch(7)	Englisch(7) Französisch(5) Latein(3) Spanisch(2) Schwedisch(1) Altgriechisch(1) Russisch(1)
15.06.2007	8	31	Deutsch(39)	Englisch(39) Französisch(34) Latein(19) Russisch(3) Spanisch(2) Altgriechisch (2) Chinesisch(1)
11.09.2007	5	10	Deutsch(15)	Englisch(15) Latein(12) Französisch(10) Russisch(2) Spanisch(1)
13.09.07	2	11	Deutsch(13)	Englisch(13) Französisch(12) Latein(8) Spanisch(2) Russisch(1)
09.10.2007	7	8	Deutsch(15) Thailändisch(1)	Englisch(15) Französisch(15) Latein(7) Spanisch(1) Japanisch(1)
23.10.07	2	4	Deutsch(6)	Englisch(6) Französisch(6) Spanisch(4) Latein(3) Russisch(1) Jiddisch(1)
Σ	29	66	Deutsch(95) Thailändisch(1)	Englisch(95) Französisch(85) Latein(54) Spanisch(12)

			Russisch(8) Altgriechisch (3) Jiddisch(1) Japanisch(1) Chinesisch(1)
Σ	95		

3.3.4. Übersicht über die Orte und Textgrößen bezüglich der einzelnen Erhebungen

Erhebungsdatum	Ort	Texte	Token	Token/Text
25.10.06	Freie Universität Berlin, Deutschland (FUD)	7	4670	778,33
15.06.2007	Evangelisches Gymnasium Hermannswerder, Deutschland (DHW)	39	34502	884,67
11.09.2007	Humboldt-Gymnasium Eichwalde, Deutschland (DEW),	15	8828	588,53
13.09.07	Humboldt-Gymnasium Eichwalde, Deutschland (DEW),	13	6399	492,23
09.10.2007	Carl-Siemens-Schule Berlin, Deutschland (DCS)	15	9302	620,13
23.10.07	Freie Universität Berlin, Deutschland (FUD)	6	4790	684,29
Σ		**95**	**68491**	**720,96**

3.3.5. Tokenanzahl

Texte	95	Muttersprachler	95	Tokens	68491	Ø/Text	720,96

4. Richtlinien für die Annotation im Falko-Essay-Korpus v2.3

Seit der Version 2.0 kann Falko mithilfe des Suchwerkzeugs ANNIS2 (Zeldes et al. 2009) durchsucht werden kann, das speziell für Abfragen tiefannotierter Korpora wie Falko entwickelt wurde.
Ältere Versionen werden nach der Publikation nicht mehr verändert, um die auf ihr basierenden Ergebnisse replizierbar zu halten. Die Daten wurden zwar alle von mindestens zwei Annotatoren kontrolliert, allerdings ist damit zu rechnen, dass auch weiterhin vereinzelt Dinge übersehen wurden.

Wir möchten daher alle Nutzer dazu auffordern, uns Fehler und Verbesserungsvorschläge immer sofort zu melden, damit wir zeitnah verbesserte Versionen herausgeben können.

4.1. Textgrenzen: [TXTstructure]

Um in ANNIS2 die Anzahl von Texten suchen zu können, in der eine bestimmte Struktur auf-

taucht, wurde auf dieser Ebene jedes erste Token mit dem Tag „start“ und jedes letzte mit dem Tag „end“ versehen.

Erstes Token im Text
Das erste Token erhält das Tag „start“.

tok	Der	Feminismus	hat	den
TXTstructure	start			

Letztes Token im Text
Das letzte Token erhält das Tag „end“.

tok	müssen	berücksichtigt	werden	.
TXTstructure				end

4.2. Korrigierte Tokenebene: [ctok]

Die Korrigierte Tokenebene stellt die Grundlage für die Berechnung jeglicher Abweichungen der Zielhypothesen dar. Hier werden lediglich Fehler in der automatischen Tokenisierung verbessert. Die Korrigierte Tokenebene besteht aus einem kompletten Text, der die Änderungen enthält. Sie erlaubt darüber hinaus die Simulation von Subtokens, also einem einzelnen Token auf der ctok-Ebene, das mehreren Tokens auf einer Zielhypothesenebene entspricht.

Tokens zusammenfassen
Durch die Tokenisierung fälschlich getrennte Zeichen werden auf der ctok-Ebene in einer Spanne zusammengeführt.

tok	das	20	.	Jhd
ctok	das	20.		Jhd

pos-Konstanz
Werden Tokens getrennt, wird darauf geachtet, dass gleiche Wortarten untereinander stehen.

tok	ob	er/		sie	kommt
pos	KOUS	<unknown>		PPER	VVFIN
ctok	ob	er	/	sie	kommt
	KOUS	PPER	&(	PPER	VVFIN

Basis für Subtokenisierung auf einer Zielhypothese
Werden in den Zielhypothesen Tokens aus dem Lernertext getrennt, so wird die ctok-Ebene zur Spanne über alle „Subtokens“ der Zielhypothese.

ctok	oder	auf	der	Jungfrau	
ZH2	oder	auf	eine	junge	Frau

4.3. Makrostrukturebene: [macro]

Auf der Makrostrukturebene werden Textabschnitte gekennzeichnet, die nicht zum eigentlichen Textverlauf beitragen, um sie in einer späteren Suche gesondert behandeln zu können.

start
Beginnt ein Essay mit der Beschreibung der Umstände, dem Namen oder Datum, so wird dies mit „start“ gekennzeichnet.

ctok	Im	Zuge	des	DaF-Seminars	soll	ff.	These	diskutiert	werden	:
macro	start									

(fu081d_2007_10)

title
Viele Essays beginnen mit der Wiederholung der Fragestellung. Diese wird als „title“ gekennzeichnet.

ctok	Kriminalität	zahlt	sich	nicht	aus	.	Die	Frage
macro	title							

subtitle
Tauchen in den Texten Zwischenüberschriften auf, so werden sie als „subtitle“ markiert, damit sie beispielsweise für syntaktische Auswertungen ignoriert werden können.

ctok	nicht	ändern	wollen	.	Der	Feminismus	Feminismus	hat	schon
macro					subtitle				

structure
In Texten auftretende Strukturierungszeichen wie Listenzeichen (* - • → etc.) werden als „structure“ gekennzeichnet, um gezielt nach ihnen suchen zu können.

ctok	geht	.	-	Alle	Leute	wissen	viel	von
macro			structure					

citation
Zitate im Lenertext werden als solche mit „citation“ markiert.

ctok	"	Does	feminsit	mean	ugly	,	loud	woman	or	does
macro	citation									

ctok	it	mean	someone	who	stands	up	for	their	own	rights
macro	citation									

ctok	.	I	believe	it	is	is	the	latter	.	"	Margaret	Atwood
macro	citation											

(fk010_2006_07)

comment
Kommentare, die sich auf die Schreibsituation oder die Erhebungssituation beziehen, werden als „comment“ gekennzeichnet. Diese können bei einer Analyse des Textaufbaus ignoriert werden.

ctok	hat	.	Ach	so	!	Also	,
macro			comment				

alternative
Bietet der Lerner in Klammern alternative syntaktische Varianten an, so werden diese mit dem Label (alternative) gekennzeichnet, um für eine spätere syntaktische Verarbeitung besser ausgeschlossen werden zu können.

ctok	jeder	sollte	(	=	jedem	ist		erlaubt	)	so	viel	heraushohlen	als	möglich	ist
macro			alternative												

end
Metatextuelle Zusätze am Ende des Textes, die sich keinen Einfluss mehr auf den Textverlauf haben, werden mit „end“ gekennzeichnet, um sie später gesondert behandeln zu können.

ctok	Vorname	Name	Studierende	und	Sachbearbeiterin
macro	end				

4.4. Fremdsprachliches Material: [fm]

Auf dieser Ebene werden Tokens annotiert, die aus einer Fremdsprache stammen, jedoch noch nicht als Teil des deutschen Fremdwortschatzes betrachtet werden können. Die Entscheidung unterlag dem subjektiven Empfinden des Annotators und kann nur als Hinweis auf mögliche Kandidaten dienen.

[Sprache]
Das passende dreistellige Kürzel der Herkunftssprache.

ctok	learning	by	doing
macro	fm:eng		

5. Zielhypothesen

Gegenstand:

Eine wichtige Komponente eines Lernerkorpus ist die Fehlerannotation. Um Fehler annotieren zu können, muss man sie identifizieren. Laut Corder(1986:37) müssen dafür die Lerneräußerungen mit den „korrekt" rekonstruierten Äußerungen verglichen werden.

We identify errors by comparing original utterances with what I shall call reconstructed utterances, that is, correct utterances having the meaning intended by the learner.

Was man unter der rekonstruierten Äußerung zu verstehen hat, wird erkennbar, wenn man sich Lennons(1991:182) Definition für Fehler ansieht. Unter einem Lernerfehler versteht er „ *a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterparts.* Dabei führt der Begriff der "rekonstruierten Äußerung" leicht in die Irre, denn eine direkte Verbindung zwischen Lerneräußerung und muttersprachlicher Rekonstruktion existiert nicht. Jede Lerneräußerung erlaubt eine Vielzahl gleichwertiger Äußerungen in der Muttersprache. Lüdeling(2008) zeigt in einer Studie mit verschiedenen professionellen Deutschlehrern, dass die Abweichung zwischen „Fehlerannotatoren" sehr groß ist. Aus diesem Grund spricht sie von einer „Zielhypothese", welche eine Interpretation der Lerneräußerung durch einen geschulten Annotator darstellt. Drei Konsequenzen erwachsen daraus. Zum Ersten muss eine Zielhypothese explizit sein, damit darauf basierende Fehlerannotationen transparent werden. Zweitens muss für Benutzer des Korpus die Möglichkeit bestehen, eigene, konkurrierende Zielhypothesen zu entwickeln und einzubinden. Drittens wird deutlich, dass die Erarbeitung einer Zielhypothese einer Operationalisierung mit mehr oder weniger engen Grenzen der Interpretation bedarf. Diese Richtlinien liegen hiermit vor.

Mehrere Zielhypothesen

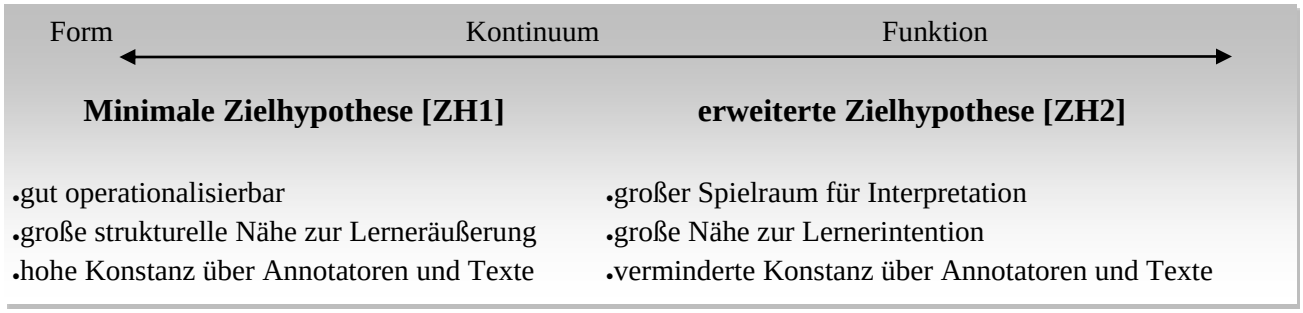
Während der Erarbeitung der Operationalisierung der Falko-Zielhypothese sind zwei widerstrebende Tendenzen deutlich geworden. Je stärker man die Interpretation der Lerneräußerung lenkt und beschränkt, desto weniger Spielraum bleibt dem Annotatoren, die eigentliche Intention, des Lerners sinngemäß zu rekonstruieren. Je mehr von dieser Intention er jedoch versucht in der Zielhypothese widerzuspiegeln, desto weniger Einheitlichkeit verspricht das Resultat. Aus diesem Grund wurde für das Falko-Essay-Korpus entschieden, beide Strategien zu verfolgen, indem zwei verschiedene Zielhypothesen für den gleichen Text erstellt wurden.

Eine **minimale Zielhypothese** dient als Normalisierungsebene und hat zum Ziel eine Ebene zu erzeugen, die für die automatische Verarbeitung dienen zu können. Die Anpassung des Originaltextes an eine parsbare Struktur ist somit höchste Priorität bei der Erstellung der minimalen Zielhypothese.
Gleichzeitig wird versucht, die Anzahl der Abweichungen von der Originaläußerungen zu minimieren und nimmt dafür in Kauf, sich vor allem auf Fehler niedriger sprachlicher Ebenen (Orthografie, Morphologie, Syntax) zu beschränken.

Die **erweiterte Zielhypothese** versucht eine große Bandbreite von Fehlern auch auf höheren Ebenen (Semantik, Lexik, Pragmatik, Stilistik) einzufangen auf Kosten der Operationalisierbarkeit und Konstanz über Annotatoren und Texte. Auf diese Weise kann eine geschickt ausgenutzte Kombination

beider Zielhypothesen bereits ohne weitere explizite Fehlerannotation ein großes Spektrum von Abweichungen der Lernertexte erfassen.

Zusammengefasst handelt es sich bei den Zielhypothesen NICHT um explizite nach Fehlerkategorien unterteile Fehlerannotationen, sondern um eine implizite Annotation von Abweichungen der Lernertexte von einer postulierten Standardvariante, vor deren Hintergrund eine Fehlerannotation möglich wird und die auf dieser aufbauen kann. Für eine kompakte Darstellung auf Englisch siehe Reznicek et al. (im Druck).



Die Richtlinien für die Annotation der Zielhypothesen sind in vier Teile unterteilt:

- Allgemeine Information über das technische Vorgehen bei der Annotation der Daten mit MS Excel oder EXMaRALDA[4].
- Kriterien für die Annotation der minimalen Zielhypothese [ZH1]
- Kriterien für die Annotation der maximalen Zielhypothese [ZH2]
- Richtlinien für die Annotation der Differenzmarkierungen zwischen den Zielhypothesen und der ctok-Ebene

5.1. Technische Vorgaben für die Erstellung der Zielhypothesen

Geringe Korrekturen
Was nicht geändert werden muss, um eine Version zu erzeugen, die den oben beschriebenen Kriterien entspricht, wird nicht verändert. Je weniger Tokens verändert werden müssen, desto besser.

ctok	Frauen	konnten		solchen		geselschaftlichen	Zustand	verändern
ZH1	Frauen	konnten	einen	solchen		gesellschaftlichen	Zustand	verändern
¬ZH1	Frauen	konnten		solch	einen	gesellschaftlichen	Zustand	verändern

(fk024_2006_07)

Bewegen statt Tauschen
Zielposition für eine bewegte Konstituente in der Zielhypothese ist immer ein leeres Token. Konstituenten werden nicht getauscht.

[4] siehe Schmidt 2001

ctok	Man	hat		ihr	es	geglaubt	.
ZH1	Man	hat	es	ihr		geglaubt	.
¬ZH1	Man	hat		es	ihr	geglaubt	.

Bewegung nach links
Spricht keine der vorangegangenen Regeln dagegen, werden Bewegungen nach links denen nach rechts vorgezogen.

ctok	Man	hat		ihr	es		geglaubt	.
ZH1	Man	hat	es	ihr			geglaubt	.
¬ZH1	Man	hat			es	ihr	geglaubt	.

Kurze Bewegungen
Weicht die Position eines Tokens in der Zielhypothese von ihrem Pendant auf der ctok-Ebene ab, so soll der Abstand zwischen beiden Positionen minimal bleiben.

ctok		Die	Frauen	haben		die	Macht	auch	.
ZH1		Die	Frauen	haben	auch	die	Macht		.
¬ZH1	Auch	die	Frauen	haben		die	Macht		.

Bewegung leichter Konstituenten
Tokens werden so in einem Satz verschoben, dass möglichst wenige Konstituenten und Tokens davon betroffen sind.

ctok	Es	hat					Geld	seiner	Firma	in	Polen		gebracht	.
ZH2	Es	hat						seiner	Firma	in	Polen	Geld	gebracht	.
¬ZH2	Es	hat	seiner	Firma	in	Polen	Geld						gebracht	.

Interpunktionszeichen werden für Regel 6a nicht berücksichtigt.
Wird ein Token an einer Stelle im Satz eingefügt, an der ein Interpunktionszeichen gelöscht wird, so kann das Token links oder rechts eingefügt werden, um eine Zielstruktur zu erhalten, die der Ausgangsstruktur ähnelt.

ctok	nach	der	Meinung	der	Feministinnen		,	als	altmodisch
ZH2		Der	Meinung	der	Feministinnen	nach		als	altmodisch

Token verschmelzen
Werden zwei Token miteinander verschmolzen, werden auch die Zellen zu einer Spanne verschmolzen.

ctok	an	dem	Wohlstand
ZH1	am		Wohlstand

¬ZH1		am	
¬ZH1	am		Wohlstand

Token aufteilen
Werden zwei Token miteinander verschmolzen, werden auch die Zellen zu einer Spanne verschmolzen.

ctok	zum		Frau
ZH1	zu	der	Frau

Doppelte Sequenzen löschen
Fälschlicherweise verdoppelte Sequenzen werden rechts gelöscht.

ctok	es	nicht	es	nicht
ZH1	es	nicht		
¬ZH1			es	nicht

Gleiche Funktion mit anderen Mitteln
Werden in der Zielhypothese bestimmte Satzfunktionen durch andere Phrasentypen erfüllt als im Originaltext, so werden beide Phrasen (wenn möglich) als überlappende Spannen konstruiert.

ctok	ob	Forschung	frei	oder	mit	einem	gegebenen	Ziel	sein	solle	.
ZH2	ob	Forschung	frei	oder	zielgerichtet				sein	sollte	.
ZH2Diff					MERGE					CHA	

ctok: Zu dieser Frage von Praxis versus Teori hört zusätzlich auch eine Diskution zwischen Gruppen von Forscher und Gruppen von Politiker und Unternehmer über die Thema ob Forschung frei oder **mit einem gegebenen Ziel** sein solle.
ZH2: Zu dieser Kontroverse zwischen Praxis und Theorie gehört zusätzlich auch eine Diskussion zwischen Gruppen von Forschern und Gruppen von Politikern und Unternehmern über das Thema, ob Forschung frei oder **zielgerichtet** sein sollte. (fk021_2006_08)

ctok	über	sich	selbst	und	ihre	Erwachsenwerdenprobleme				schreiben
ZH2	über	sich	selbst	und	ihre	Probleme	mit	dem	Erwachsenwerden	schreiben
ZH2Diff						SPLIT				

ctok: Plötzlich können sie, über sich selbst und ihre **Erwachsenwerdenprobleme** schreiben - und es ist interessant für die Gesellschaft.
ZH2: Plötzlich können sie über sich selbst und ihre **Probleme mit dem Erwachsenwerden** schreiben und es interessiert die Gesellschaft. (fk022_2006_07)

5.2. Minimale Zielhypothese (Satzebene, Orthografie, Morphosyntax) [ZH1]

Das Ziel der minimalen Zielhypothese ist es, durch minimale Änderungen einen Satz zu konstruieren, der den Regeln der deutschen **Morphosyntax** und **Orthografie** gehorcht. Sie soll vor allem als Normalisierung für die weitere automatische Verarbeitung dienen. Abweichungen werden gegenüber der ctok-Ebene vorgenommen. Die Zielhypothese besteht aus einem kompletten Text, der die Änderungen enthält. Semantik, Lexik, zielsprachliche Kollokationen und Pragmatik werden ignoriert.

Orthografie

Betonung durch Großbuchstaben
In Großbuchstaben geschriebene Silben oder Wörter werden der deutschen Orthografie angepasst.

ctok	Vater	UND	Mutter	arbeiten
ZH1	Vater	und	Mutter	arbeiten

Movierung akzeptiert
Großschreibung in Wörtern, die eine Movierung ausdrückt, wird nicht korrigiert.

ctok	AbsolverntIn
ZH1	AbsolverntIn

Umlaute und Sonderzeichen
Ersatzschreibungen für Umlaute werden korrigiert.

ctok	zurueck
ZH1	zurück

Morphologie

Pluralitanta & Singularitanta
Im Singular gebrauchte Plurariatanta und im Plural gebrauchte Singulariatanta werden nur dann korrigiert, wenn sie nicht im Duden stehen und im DEWAC nicht mindestens 10 Belege gefunden werden können.

ctok: Obwohl die Bibel und unsere Moralen uns immer gesagt haben, dass Kriminalität immer schlecht ist, ist es aber auch erforderlich zu fragen, ob sie auch gut sein kann.

ZH1: Obwohl die Bibel und unsere Moralen uns immer gesagt haben, dass Kriminalität immer schlecht ist, ist es aber auch erforderlich zu fragen, ob sie auch gut sein kann. (BNG2-2011-03-186)

http://www.duden.de/suchen/dudenonline/Moralen

Verschmelzungen aus Präposition und Artikel
Im mündlichen Sprachgebrauch übliche Verschmelzungen von Präposition und Artikel werden nicht korrigiert.

ctok: Heutzutage ist es selbstverständlich, sich mit dem sogenannten" Ellbogen-Prinzip" **durchs** Leben zu schlagen.
ZH1: Heutzutage ist es selbstverständlich, sich mit dem sogenannten" Ellbogen-Prinzip" **durchs** Leben zu schlagen.(dhw_010_2007_06)

Syntax

Abweichung von minimaler Bewegung
Führt eine minimale Bewegung zu einer Struktur, die einerseits sehr stark von der Originalsyntax abweicht, die intendierte Zielstruktur aber andererseits nicht gut abbildet, so darf die minimale Bewegung verletzt werden.

ctok		In	Neu	Seeland	auch	haben		wir	eine	Prime	Minister
ZH1	Auch	in	Neuseeland			haben		wir	eine	Preministerin	
¬ZH1		In	Neuseeland			haben	auch	wir	eine	Preministerin	

Konnektoren/Subjunktoren und Wortstellung
Findet sich in einem Nebensatz eine falsche Kombination aus Konnektorvorgabe und Wortstellung, so wird die Wortstellung auch dann angepasst, wenn eine Veränderung des Konnektors weniger Abweichung insgesamt bedeuten würde.

ctok	,	weil	man		findet	keine	Arbeit		.
ZH1	,	weil	man			keine	Arbeit	findet	.
¬ZH1	,	denn	man		findet	keine	Arbeit		.

ctok: Und ich kann es nicht verstehen, dass viele Leute glauben an diese Wesen(Terroristen), die ich als Menschen nicht nennen kann.
ZH1: Und ich kann es nicht verstehen, dass viele Leute an diese Wesen (Terroristen) glauben, die ich nicht Menschen nennen kann. (usb015_2006_10)

Konkurrierendes „dass“ und Infinitiv mit „zu“
Tauchen in einem Satz der „dass“-Subordinierer und ein Infinitiv mit „zu“ auf, wird die Variante mit der kleineren Abweichung gewählt. Dies ist eine Ergänzung zur vorhergehenden Regel.

ctok: Aber... Einige streben danach, **dass** noch reicher **zu sein**, noch gewaltiger;

ZH1: Aber: Einige streben danach, noch reicher **zu sein** und noch gewaltiger;
(usb015_2006_10)

Verbstellung subordinierten Nebensätzen
In subordinierten Nebensätzen wird Verbzweitstellung korrigiert. Stellungsvarianten mit eingebettetem V2-Satz nach „dass“, wie sie etwa Freywald (2008,2009) auch bei Muttersprachlern findet, werden hier nicht berücksichtigt.

ctok: Die reiche Studenten möchten nicht studieren" - hatte ich gesagt, aber ist das falsche Meinung, weil viele reichen Studenten **erwerben** in diesem Moment in großem Ergebnisse als arme Studenten.
ZH1: Die reichen Studenten möchten nicht studieren", hatte ich gesagt, aber das ist eine falsche Meinung, weil viele reiche Studenten in diesem Moment größere Ergebnisse als arme Studenten **erwerben**.(usb007_2006_10)

Linke Satzklammer beibehalten
In Hauptsätzen wird das finite Verb als linke Satzklammer behandelt. Diese wird nicht bewegt. Diese Regel widerspricht der generellen Bewegung nach links und hat höhere Priorität.

ctok	Heute		Männer	gehen		nicht	oft	einkaufen	.
ZH1	Heute			gehen	Männer	nicht	oft	einkaufen	.
¬ZH1	Heute	gehen	Männer			nicht	oft	einkaufen	.

Linke Satzklammer erzeugen
Wird aus einem Originalsatz mit besetzter linker (LSK) und rechter Satzklammer (RSK) eine Zielhypothese erstellt, in der lediglich die LSK besetzt ist, wird der Inhalt der ehemaligen LSK gelöscht und das Token der ehemaligen RSK rechts von der Löschung bewegt.

ctok: Je mehr ein Job belohnt, desto weniger der Arbeitnehmer mit die Gesellschaft zu tun **hat**.
ZH1: Je mehr ein Job belohnt, desto weniger **hat** der Arbeitnehmer mit der Gesellschaft zu tun. (BNG2-2010-11-138)

Rechte Satzklammer
Die rechte Satzklammer hat keinen besonderen Status in Bezug auf die Annotation, sodass das finite Verb in diesem Fall bewegt werden darf.

ctok	was	ist	darüber	gemeint		.
ZH1	was		damit	gemeint	ist	.
ZH1Diff		MOVS	CHA		MOVT	

ctok: Wenn wir über" Feminismus" oder" die Interessen der Frauen" reden, sollen wir erklären, was **ist** darüber gemeint.

ZH1: Wenn wir über" Feminismus" oder" die Interessen der Frauen" reden, sollen wir erklären, was damit gemeint **ist**. (fkb031_2008_07)

Verb als Kern

Stimmen das Verb und seine Argumente nicht überein, wird das Verb beibehalten und die Argumente angepasst. Für fehlende obligatorische Objekte werden Dummys eingefügt. Gleiches gilt für vom Verb abhängige Präpositionalobjekte.
Numerus und Kasus richten sich nach dem Verb.

ctok	dass	unsere	Zivilisation	sich	aus	Gruppen	besteht	.
ZH1	dass	unsere	Zivilisation		aus	Gruppen	besteht	.
¬ZH1	dass	unsere	Zivilisation	sich	aus	Gruppen	zusammensetzt	.

ctok: In diesem Fall kann ich nur über mein Land sagen, aber vermutlich sieht diese Situation auch in anderen Ländern ähnlich.
ZH1: In diesem Fall kann ich nur **etwas** über mein Land sagen, aber vermutlich sieht diese Situation auch in anderen Ländern ähnlich. (fkb028_20)

ctok: Wenn wir über "Feminismus" oder "die Interessen der Frauen" reden, sollen wir erklären, was ist **darüber** gemeint.
ZH1: Wenn wir über "Feminismus" oder "die Interessen der Frauen" reden, sollen wir erklären, was **damit** gemeint ist. (fkb031_2008_07)

ctok: Die Universitäten sollten praxisorientirt sein, damit die Studenten, wenn sie auf dem Arbeitsmarkt kommen, **auf** die Herauforderungen gut gewapnet sind.
ZH1: Die Universitäten sollten praxisorientiert sein, damit die Studenten, wenn sie auf den Arbeitsmarkt kommen, **für** die Herausforderungen gut gewappnet sind. (cbs006_2007_10.)

ctok: Hauptsätslich denken Leute , dass sie über die Ideologien einer Gruppe wissen.
ZH1: Hauptsächlich denken Leute, dass sie **etwas** über die Ideologien einer Gruppe wissen. (BNG2-2010-11-101)

ctok: Es ist natürlich sehr schwierig eine Stelle zu bekommen, wenn es **der** Arbeitgeber klar wird.
ZH1: Es ist natürlich sehr schwierig, eine Stelle zu bekommen, wenn es **dem** Arbeitgeber klar wird.

Änderungen am Verb

Können dem Verb keine Argumente hinzugefügt werden, wird das Verb präfigiert oder mit Partikel versehen.

ctok: Man sagt, dass die Arbetsbedingungen fuer Arbeitseinsteiger konnen nicht verbessert werden, es sei denn der Staat **passt** neue Arbeitsmarkt regelnde Gesetze.

ZH1: Man sagt, dass die Arbeitsbedingungen für Arbeitseinsteiger nicht verbessert werden können, es sei denn, der Staat **passt** neue arbeitsmarktregelnde Gesetze **an**. (fk016_2006_08)

Kasusrektion durch Präpositionen
Kasusrektion durch Präpositionen wird gemäß der Darstellung in Grammis (Breindl 2000) behandelt, hierbei werden auchprogressive Entwicklungen hin zum Dativ akzeptiert

ctok: Während den sechziger Jahren, gab es eine grosse Debatte über Porn.
ZH1: Während den Sechzigerjahren gab es eine große Debatte über einen Porno. (BNG2-2010-11-120)

Korrelate
Es wird vermieden, Korrelate hinzuzufügen und zu entfernen, wenn dies noch grammatisch möglich ist.

ctok: Wenn man die Argumente abwägt, wird es klar, dass Kriminalität oft sich nicht auszahlt.
ZH1: Wenn man die Argumente abwägt, wird klar, dass Kriminalität sich oft nicht auszahlt.

Nominale Kongruenzhierarchie
Bei fehlender Kongruenz zwischen Determinierer, Attribut und Nominalkern wird vom Kern ausgegangen. Danach wird zugunsten des Determinierers entschieden.

höchste Priorität — syntaktischer Kern (lexikalischer Kopf)
↑ — Determinierer (syntaktischer Kopf)
Attribute
niedrigste Priorität

ctok	keine	Arbeit	bekommen	,	die	ihrer	Qualifikationen	entspricht	.
ZH1	keine	Arbeit	bekommen	,	die	ihren	Qualifikationen	entspricht	.
¬ZH1	keine	Arbeit	bekommen	,	die	ihrer	Qualifikation	entspricht	.

ctok: Das Problem entsteht auch, weil man zwischen **öffentliche** und **private** Unternehmen entscheiden muss.
ZH1: Das Problem entsteht auch, weil man zwischen **öffentlichen** und **privaten** Unternehmen entscheiden muss. (fkb010_2008_07)

Fehlende Artikel
Für fehlende Artikel wird ein indefiniter Artikel eingesetzt.

ctok: Deswegen muss Verbrecher bestraft werden.
ZH1: Deswegen muss **ein** Verbrecher bestraft werden.

ctok: Nächstes Jahr werde ich mein Studium enden und ich neide meinen Freundinnen aus Gymnasium, die sich für Jura- oder Medizinstudium entschieden.
ZH1: Nächstes Jahr werde ich mit meinem Studium enden und ich beneide meine Freundinnen aus **dem** Gymnasium, die sich für ein Jura- oder Medizinstudium entschieden. (fk015_2006_08)

ABER:

ctok: Seit dem Anfang des zwanzigsten Jahrhunderts ist Feminismus einer der umstrittensten Themen gewesen, am wenigstens in demokratischen Länder.
ZH1: Seit dem Anfang des zwanzigsten Jahrhunderts ist Feminismus eines der umstrittensten Themen gewesen, wenigstens in demokratischen Ländern. (fkb037_2008_07)

Wortstellung im Mittelfeld
Abfolgen von Konstituenten im Mittelfeld werden nur dann korrigiert, wenn dadurch ungrammatische Strukturen entstehen. Wortstellungstendenzen werden ignoriert (vgl. Eisenberg 2006:405ff.).

ctok	Er	hat		meiner	Firma	es	gesagt	.
ZH1	Er	hat	es	meiner	Firma		gesagt	.
¬ZH1	Er	hat		meiner	Firma	es	gesagt	.

ctok: Aber vielleicht ist das Problem auch, dass man **nicht** auf eine Universität alles lernen kann.
ZH1: Aber vielleicht ist das Problem auch, dass man auf einer Universität **nicht** alles lernen kann. (cbs003_2007_10)

ctok: Einerseits muss ich sagen, dass jeder Student muss unbedingt gute theoretische Grundkenntnisse haben, weil ohne diesen **er** keine Chance um ein(egal ob Mathematisch- oder soziales) Problem zu Lösen hat.
ZH1: Einerseits muss ich sagen, dass jeder Student unbedingt gute theoretische Grundkenntnisse haben muss, weil **er** ohne diese keine Chance hat, ein(egal ob mathematisches oder soziales) Problem zu lösen. (fkb011_2007_09)

Koordinierte Strukturen nicht teilen
Führt eine kurze Bewegung zur Teilung koordinierter Strukturen, wird eine längere Bewegung bevorzugt.

ctok	obwohl	ich	erkläre	meine	Meinung		und		von	meinen	Freunden	
ZH1	obwohl	ich		meine	Meinung		und	die	von	meinen	Freunden	erkläre
¬ZH1	obwohl	ich		meine	Meinung	erkläre	und	die	von	meinen	Freunden	

Semantisch konfluierendes Tempus & Modus
Auch in Fällen, in denen Tempus oder Modus semantisch oder pragmatisch durch eine Äuße-

rung gefordert wird, wird **NICHT** korrigiert.

ctok	Erst	verliert	er	das	Geld	und	dann	floh	er
ZH1	Erst	verliert	er	das	Geld	und	dann	floh	er
¬ZH1	Erst	verlor	er	das	Geld	und	dann	floh	er

ctok: So **verhält** es sich jedenfalls früher in Dänemark
ZH1: So **verhält** es sich jedenfalls früher in Dänemark (cbs005_2007_10)

Lexik

Fremdsprachliches Material
Fremdsprachliches Material außerhalb von Klammern wird, wenn möglich, übersetzt.

ctok: Die Stunden an den dänischen Unis, sollten nicht zu lange Vorträge von das gelesene beinhalten, und mehr Übungen in den Stunden einbauen lassen in Form von z. B **Cases**.
ZH1: Die Stunden an den dänischen Unis sollten nicht zu lange Vorträge des Gelesenen beinhalten und mehr Übungen in die Stunden einbauen lassen z. B. in Form von **Fällen**. (cbs006_2007_10)

5.2.1. A

uf der Ebene der minimalen Zielhypothese [ZH1] NICHT annotierte Fehler

Syntax

Verberstdeklarativsätze ohne Subjekt
Verberstdeklarativsätze ohne Subjekt , in denen Topikdrop stattgefunden haben kann, werden nicht korrigiert, indem ein Subjekt eingesetzt wird.

ctok: Sie stossen auf Situationen wo sie nuicht genau wissen was sie tun sollten. **Fühlen** sich etwas orientirungslos.
ZH1: Sie stoßen auf Situationen, in denen sie nicht genau wissen, was sie tun sollten. **Fühlen** sich etwas orientierungslos. (cbs006_2007_10)

Verblose Sätze
Verblose Sätze, die nicht als Paraphrase an den vorhergehenden oder den nachfolgenden Satz angehängt werden können, werden nicht korrigiert.

ctok: Ins Besondere die sprachliche Gewinne (natürlich), aber nicht nur die!
ZH1: Insbesondere die sprachlichen Gewinne (natürlich), aber nicht nur die! (cbs006_2006_09)

Uneingeleitete Nebensätze
Uneingeleitete Nebensätze werden nicht korrigiert.

ctok: Ich meine, die staatlichen Arbeiter leisten der Gesellschaft auf keinen Fall einen geringeren Beitrag.
ZH1: Ich meine, die staatlichen Arbeiter leisten der Gesellschaft auf keinen Fall einen geringeren Beitrag.(fkb028_2008_07)

Isoliert stehende Nebensätze
Isoliert stehende Nebensätze werden nicht korrigiert.

ctok: Deswegen sollen sie ein bisschen besser vom Staat unterstützt werden.
ZH1: Deswegen sollen sie ein bisschen besser vom Staat unterstützt werden. (fkb028_2008_07)

Uneingeleitete Nebensätze: Zweiteilige Konnektoren
Zweiteilige Konnektoren wie „zwar… , aber“, „je…desto“, „nicht nur…, sondern auch“ müssen in einem Matrixsatz verbunden werden. Diese Regel hat höhere Priorität als "Isoliert stehende Nebensätze".

ctok: Zwar ist es wahr, dass noch viele Frauen zu Hause bleiben und sich mit wenig mehr als Kinder und Haushalt beschäftigen**. Es** ist aber die Entscheidung die sie selbst gretroffen haben - im Prinzip könnten sie auch etwas Anderes mit ihrem Leben tun.
ZH1: Zwar ist es wahr, dass noch viele Frauen zu Hause bleiben und sich mit wenig mehr als Kindern und Haushalt beschäftigen**, es** ist aber die Entscheidung, die sie selbst getroffen haben - im Prinzip könnten sie auch etwas Anderes mit ihrem Leben tun. (fkb037_2008_07)

Lexik

Neologismen und falsche Bedeutungen
Morphosyntaktisch korrekte Neologismen und semantisch falsch verwendete Ausdrücke werden **nicht** korrigiert.

ctok: Ich finde diesen Zitat sehr schön, weil er erlaubt ein Raum für eine kreative Bildung des **Frauentums**, die nicht von einer patriarchalischen Gesellschaft diktiert ist und die nicht von einen einseitigen Sicht an Feminismus diktiert ist.
ZH1: Ich finde dieses Zitat sehr schön, weil es einen Raum für eine kreative Bildung des **Frauentums** erlaubt, die nicht von einer patriarchalischen Gesellschaft diktiert ist und die nicht aus einer einseitigen Sicht vom Feminismus diktiert ist. (fk010_2006_07)

ctok: JOurnalisten haben photographien von Frauen gemacht während sie **Gefängnisse** brutalischerweise schlagten.
ZH1: Journalisten haben Fotografien von Frauen gemacht, während sie **Gefängnisse** brutalerweise schlugen.(fk010_2006_07)

ctok: Sie meinten, dass Männer die mit der kultur beschäftigt sind sind **Gewaltiger** und Unrechter als Frauen (weil Kultur kann auch zum Krieg führen), die mit der" schöne Natur" verbunden sind.
ZH1: Sie meinten, dass Männer, die mit der Kultur beschäftigt sind, **gewaltiger** und ungerechter sind als Frauen (weil Kultur auch zum Krieg führen kann), die mit der" schönen Natur" verbunden sind. (fk010_2006_07)

wo als Relativpronomen
"Wo" wird als Relativpronomen nur dann korrigiert, wenn es sich sich nicht auf Orte oder Situationen bezieht.

ctok: Ein Semester, **wo** die Studierenden in relevanten Unternehmen arbeiten konnten.
ZH1: Ein Semester, **in denen** die Studierenden in relevanten Unternehmen arbeiten konnten. (cbs006_2006_09)

ctok: In einer Gesellschaft wie die dänische, **wo** alle in unserem Sozialstaat teilnehmen, wird es übel aufgenommen, wenn einige Leute versucht das System zu entkommen.
ZH1: In einer Gesellschaft wie der dänischen, **wo** alle an unserem Sozialstaat teilnehmen, wird es übel aufgenommen, wenn einige Leute versuchen, dem System zu entkommen. (cbs008_2006_09)

ctok: das Studiums sollte Studenten für eine Welt bereit machen, **wo** Kapitalismus und Wirtschaft unseres Leben reguliert.
ZH1: Das Studium sollte Studenten für eine Welt bereit machen, **wo** Kapitalismus und Wirtschaft unser Leben regulieren. (fk004_2006_08)

Lexematisch geblockte morphologisch zulässige Varianten werden nicht korrigiert
Fremdsprachliches Material außerhalb von Klammern wird wenn möglich übersetzt.

ctok: Oder man probiert aus **Leichtsinnigkeit** illegale Drogen aus.
ZH1: Oder man probiert aus **Leichtsinnigkeit** illegale Drogen aus.
(dew09_2007_09_ta)

Pragmatik

Sprechereinstellungen
Fehlende Mittel zum korrekten Ausdruck von Sprechereinstellungen werden nicht verbessert.

ctok: Wir sollen, also, die Qualität und Quantität unterscheiden.
ZH1: Wir sollen also die Qualität und Quantität unterscheiden. (usb001_2006_10)

5.3. Erweiterte Zielhypothese [ZH2] (Textebene, Semantik, Pragmatik, Referenz, informationsstrukturelle Gliederung, Stil)

Das Ziel der erweiterten Zielhypothese ist es, ein möglichst weites Spektrum dessen zu erfassen, was in der Fehlerliteratur unter dem Begriff "Akzeptanz" gefasst wird. Die erweiterte Zielhypothese soll den Originaltext so wenig wie möglich verändern und dennoch einer muttersprachlichen Äußerung möglichst ähnlich zu sein. Hierzu gehört, dass sowohl der semantische als auch der pragmatische Gehalt sowie die informationsstrukturelle Gliederung des Satzes im Kontext des gesamten Textes korrekt bzw. angemessen ist. Nicht verbessert wird eine aus Sicht der Annotatoren falsche oder widersinnige Interpretation oder Einstellung bezüglich des besprochenen Themas im Aufsatz. Dass hier teilweise Überschneidungen auftauchen können, ist uns bewusst. Die Vielzahl möglicher, notwendiger und unvermeidlich subjektiver Abweichungen vom Originaltext, die in die erweiterte Zielhypothese mit einbezogen wird, macht eine vollständige Liste von Annotationsregeln für diese unmöglich. Aus diesem Grund sollen hier lediglich eine grobe Leitlinie und einige exemplarische Entscheidungen für die Annotation aufgeführt werden. Bei der Arbeit mit dieser Zielhypothese muss man sich der Subjektivität und nur teilweise Reproduzierbarkeit durch andere Annotatoren bewusst sein.

Syntax

Verberstdeklarativsätze ohne Subjekt

Verberstdeklarativsätze ohne werden korrigiert, indem ein Subjekt eingesetzt wird.

ctok: Sie stossen auf Situationen wo sie nuicht genau wissen was sie tun sollten. **Fühlen** sich etwas orientirungslos.
ZH2: Sie stoßen auf Situationen, in denen sie nicht genau wissen, was sie tun sollten. **Sie** fühlen sich etwas orientierungslos. (cbs006_2007_10)

Uneingeleitete Nebensätze

Uneingeleitete Nebensätze werden nicht korrigiert, wenn sie nicht selbst mit einem eingebetteten Nebensatz beginnen, wenn Verbletztstellung vorliegt oder wenn das Verb typischerweise mit eingeleiteten Nebensatz vorkommt.

ctok: Ich meine, es lohnt sich nicht zu denken, dass Terror der einzige Ausgang ist.
ZH2: Ich meine, es lohnt sich nicht zu denken, dass Terror eine Lösung ist. (usb015_2006_10)

ctok: Mein Fazit lautet deshalb, wenn man im Leben Schwein hat, kann sich auch Kriminalität lohnen,
ZH2: Mein Fazit lautet deshalb, **dass** sich, wenn man im Leben Glück hat, auch Kriminalität lohnen kann. (cbs002_2007_10)

ctok: Deshalb finde ich nicht, dass man sagen kann, die Universitätsabschlüsse von einem geringeren Wert sind.

ZH2: Deshalb finde ich nicht, dass man sagen kann, **dass** die Universitätsabschlüsse von einem geringeren Wert sind (cbs003_2007_10)

ctok: Man soll auch anerkennen, ein Universitätsabschluss **bereitet** die Jugend auf die Welt **vor**, weil die meisten einflussreichen Menschen der Welt schön einen Abschluss haben.
ZH2: Man sollte auch anerkennen, **dass** ein Universitätsabschluss die Jugend auf die Welt **vorbereitet**, weil die meisten einflussreichen Menschen der Welt schon einen Abschluss haben. (fu120_2006_10a)

ctok: Beim Abschluss erkennt der Student, er **kann** schwierige Ziele erreichen und sein Selbstwert nimmt zu.
ZH2: Beim Abschluss erkennt der Student, **dass** er schwierige Ziele erreichen **kann** und sein Selbstwert nimmt zu. (fkb055_2008_08)

Besetzung des Nachfeldes
Umfangreiche Präpositionalphrasen, Vergleichssätze, Relativsätze und Appositionen werden ins Nachfeld ausgelagert.

ctok: Oder Vorteile besser zu sagen. Wir können, sogar müssen, das selbe **wie die Männer** schaffen. (fkb034_2008_07)
ZH2: Oder, um eher die Vorteile zu nennen: Wir können und müssen sogar, dasselbe schaffen **wie die Männer**.(fkb034_2008_07)

„dass"-Sätze und Infinitivsätze mit „zu"
Ein Objektsatz, in dem das Subjekt einen Referenten des Hauptsatzes wiederaufnimmt, wird durch Infinitvsatz mit „zu" konstruiert.

ctok: Sie suchen nach der Lösung und anklagen die Universitäten, dass die keine gute Spezialisten **bereiten können**.
ZH2: Sie suchen nach der Lösung und klagen die Universitäten an, keine guten Spezialisten **auszubilden**. (fkb044_2008_08)

Semantik

Verb ersetzen
Anders als in der ZH1 hat das Verb in der ZH2 keinen besonderen Status. In Funktionsverbgefügen bildet der nominale Teil den Kern, um den korrigiert werden soll.

ctok: Vielleicht auch manche Frauen würden gerne diese Meinung vertreten, vor allem die, die anspruchsvolle Arbeit ausüben müssen um Geld zu **gewinnen.**
ZH2: Vielleicht würden auch manche Frauen gerne diese Meinung vertreten, vor allem die, die anspruchsvolle Arbeit ausüben müssen, um Geld zu **verdienen.**
(fkb030_2008_07)

ctok: **Oder Vorteile besser zu sagen**. Wir können, sogar müssen, das selbe wie die Männer schaffen. (fkb034_2008_07)
ZH2: Oder, **um eher die** Vorteile zu **nennen**: Wir können und müssen sogar, dasselbe schaffen wie die Männer.(fkb034_2008_07)

Definitheit
Der fehlerhafte Ausdruck einer kontextabhängigen Definitheit von Referenten wird korrigiert.

ctok: Das ist wahr, dass es nicht so für die Frauen in allen Ländern ist, aber das ist keine Folge **von** Feminismus.
ZH2: Es ist wahr, dass es nicht für die Frauen in allen Ländern so ist, aber das ist keine Folge **des** Feminismus. (fkb031_2008_07
ctok: Ich finde, dass **eine** finanzielle Entlohnung eines Menschen dem Beitrag entsprechen soll, den er / sie für das Unternehmen leistet.
ZH2: Ich finde, dass **die** finanzielle Entlohnung eines Menschen dem Beitrag entsprechen sollte, den er / sie für ein Unternehmen leistet. (fkb010_2008_07)

Tempus & Modus
Tempus und Modus werden dem Kontext angepasst.

ctok: Wie die Berliner haben von Barrack Obama gerade gestern gehört, unsere Sozietät **glaube**, und muss glauben bleiben, der Idee der Ebenheit und der Gleichkeit aller Personen, Mann oder Frau.
ZH2: Wie die Berliner von Barrack Obama gerade gestern gehört **haben**, **glaubt** unsere Gesellschaft und muss weiter glauben an die Idee der Gleichheit aller Personen, ob Mann oder Frau.
(fkb031_2008_07)

ctok: Wegen des Obenstehendes ist es also nicht in Ordnung zu äußern, dass die Universitätsabschlüsse von geringem Wert **sind**.
ZH2: Deshalb ist es also nicht in Ordnung zu sagen, dass die Universitätsabschlüsse von geringem Wert **seien**. (cbs007_2006_09.)

Ausdruck von Modus und Aspekt
Für das deutsche ungebräuchliche Formen des Ausdrucks von Modus und Aspekt werden durch gebräuchliche ersetzt.

ctok: Wie die Berliner haben von Barrack Obama gerade gestern gehört, unsere Sozietät glaube, und muss glauben **bleiben**, der Idee der Ebenheit und der Gleichkeit aller Personen, Mann oder Frau.
ZH2: Wie die Berliner von Barrack Obama gerade gestern gehört haben, glaubt unsere Gesellschaft und muss **weiter** glauben an die Idee der Gleichheit aller Personen, ob Mann oder Frau.
(fkb031_2008_07)

Modalverben
Eine dem deutschen untypische Verwendung von Modalverben, wird angepasst.

ctok: Obwohl die Lehren so wichtig ist, **können** die Lehrer nicht gut **verdienen**.
ZH2: Obwohl die Lehrer so wichtig sind, **verdienen** sie nicht gut. (trk008_2006_05)

Konnektoren
Drücken die verwendeten Konnektoren eine im Kontext unstimmige Relation aus, werden sie ersetzt.

ctok: **Weil** für meisten Menschen das Geld als die beste Entlohung gilt, wollen die andere andere Entlohungen, schätzen die andere Werte, glauben an die unsiegbare Macht des Gelds nicht.
ZH2: **Während** für die meisten Menschen das Geld als die beste Entlohnung gilt, wollen andere eine andere Entlohnung, schätzen andere Werte und glauben nicht an die unbesiegbare Macht des Gelds. (fu127_2006_10c)

ctok: müssen uns auch fragen, ob die These des oberliegenden Titels heutzutage wirklich erreichbar sei, **weil** in diesem politischen Klima, wer hat die Macht, solche Entscheidungen über Lohne und entsprechende gesellschaftliche Beitrage eigentlich zu treffen?
ZH2: Wir müssen uns auch fragen, ob die These des obigen Titels heutzutage wirklich erreichbar ist. **Denn** wer hat in diesem politischen Klima eigentlich die Macht, solche Entscheidungen über Löhne und entsprechende gesellschaftliche Beiträge zu treffen? (hu006_2006_09)

Fehlende Objekte
Semantisch fehlende Objekte werden ergänzt. Syntaktisch fehlende bereits auf ZH1.

ctok: Nach seine Universitätsabschluss kennt er wahrscheindlich jeder Wirkung von eine besondere Maschine aber er hat noch nie früher die Gelegenheit physisch mit dieser Maschine um zu gehen und sich genauer zu beschauen.
ZH2: Nach seinem Universitätsabschluss kennt er wahrscheinlich jede Wirkung einer besonderen Maschine, aber er hatte früher noch nie die Gelegenheit, physisch mit dieser Maschine umzugehen und **sie** sich genauer zu beschauen. (sa001_2006_09)

Numerus
Ein unplausibler Numerus wird korrigiert.

ctok: . **Unsere Krankenschwester** und andere Mitarbeiter des öffenlichen Unternehmens haben ihre Meinungen hierzu geäussert.
ZH2: . **Unsere Krankenschwestern** und andere Mitarbeiter in öffentlichen Unternehmen haben ihre Meinung hierzu geäußert. (fkb010_2008_07.)

Referenz innerhalb eines Matrixsatzes
Wird innerhalb eines Satzes durch Anaphern und Kataphern falsch verwiesen, wird dies korrigiert.

ctok: Wenn man denkt, dass Kriminalität auszahlt, sollten sie über die Folgen auch denken.
ZH2: Wenn man denkt, dass Kriminalität sich auszahlt, sollte man auch über die Folgen nachdenken. (BNG2-2010-11-116)

Referenz über den Matrixsatz hinaus
Die Referenz wird nur dann über den Matrixsatz hinaus angeglichen, wenn innerhalb eines Matrixsatzes zwei mögliche Versionen denkbar sind.

ctok: Es kann passieren dass die Leute verändern sich und auch seine Meinungen im Knast.
Aber bekommen sie noch eine andere Chance vom Leben und vom Bekannten? Wird **ihm** jemand hilfen, oder werden **sie** kein Chance haben sondern nur eins - noch einmal ausprobieren ein besserer und kluger Kriminalist zu werden?
ZH1: Es kann passieren, dass die Leute im Knast sich und auch ihre Meinungen verändern.
Aber bekommen sie noch eine andere Chance vom Leben und von Bekannten? Wird **ihnen** jemand helfen oder werden **sie** keine Chance haben, sondern nur eines : noch einmal ausprobieren, ein besserer und klügerer Krimineller zu werden? (fkb005_2007_09)

Morphologie

Nichtstandardsprachliche Reduktion von Partikelverben
Eine nichtstandardsprachliche Reduktion von Partikeln in Partikelverben werden korrigiert.

ctok: Und ich bin der Meinung, dass es dann schwerer ist da wieder **raus**zukommen.
ZH2: Und ich bin der Meinung, dass es dann schwerer ist, da wieder **heraus**zukommen. (dew18_2007_09)

Lexik

Feste Wendungen
Gibt es für eine unübliche Formulierung eine feste Wendung, dann wird diese bevorzugt.

ctok: In der dänischen Gesellschaft **werden** diese Ausbildungen sehr **nachgefragt**, weil es zur Zeit ein Mangel an diesen Fachkräften besteht.
ZH2: In der dänischen Gesellschaft **sind** diese Ausbildungen sehr **gefragt**, weil zur Zeit ein Mangel an diesen Fachkräften besteht. (fkb049_2008_08)

ctok: Europa war im 20. Jh zu viel **in den** Kriegen **tätig**.
ZH2: Europa war im 20. Jh. zu viel **mit** Kriegen **beschäftigt**. (fkb051_2008_08)

ctok: **Wegen des Obenstehendes** ist es also nicht in Ordnung zu äußern, dass die Universitätsabschlüsse von geringem Wert sind.
ZH2: **Deshalb** ist es also nicht in Ordnung zu sagen, dass die Universitätsabschlüsse von geringem Wert seien. (cbs007_2006_09.)

Konstruktionen: Modalpartikel
Fehlt in bestimmten Konstruktionen eine sonst übliche Modalpartikel, wird diese ergänzt.

ctok: Sei die im Titel benannte These provozierend, und auf ersten Blick völlig unbegründet scheinen, auf jeden Fall ist sie eine Diskussion wert.
ZH2: Sei die im Titel benannte These **auch** provozierend und mag sie auf den ersten Blick völlig unbegründet scheinen, auf jeden Fall ist sie eine Diskussion wert. (fkb058_2008_08)

ctok: Das heißt, dass sie versuchen, ihren Studenten so viel wie möglich die universelle Werte der Menschheit beizubringen. Das kann nur dann passieren, wenn sie sich mehr mit dem Geist des Menschen beschäftigen.
ZH2: Das heißt, dass sie versucht ihren Studenten so viele universelle menschliche Werte beizubringen wie möglich. Das kann **aber** nur dann passieren, wenn sie sich mehr mit dem menschlichen Geist beschäftigen. (fk006_2006_08.)

Übliche Bezeichnungen
Steht im Lernertext anstatt einer üblichen Bezeichnung für einen bestimmten Referenten ein anderer Begriff, bei dem nicht erkennbar ist, dass der Lerner bewusst auf die üblichere Variante verzichtet hat, wird letztere durch erstere ersetzt. Die teilweise sehr strenge Korrektur soll dazu dienen, die Zahl auf Kosten der "false positives" zugunsten einer geringen Menge von "false negatives" zu verschieben.

ctok: Die Errungenschaften der deutschen Philosophen, der deutschen Historiker des 19. Jhs bewertet man sich sehr hoch in der **Weltwissenschaft**.
ZH2: Die Errungenschaften der deutschen Philosophen und der deutschen Historiker des 19. Jhs. bewertet man in der **globalen Wissenschaftsgemeinde** sehr hoch. (fkb051_2008_08)

ctok: Die wenige Schicksallose, denen gelang es ihre **Fatum** zu verändern und finanziell zu wachsen erreichten das wieder durch Macht, Schönheit, **Schlauheit**, aber nur in einzelne Außnahmefälle durch Arbeit oder durch seinen Kopf, die könnten manchmal die Annerkennung des Menschen in der Gesellschaft zu gewinnen, aber wenig Geld zu bringen.

ZH2: Die wenigen Schicksalslosen, denen es gelänge, ihr **Schicksal** zu verändern und finanziell zu wachsen, erreichten das wieder entweder durch Macht, Schönheit und **Intelligenz**, aber nur in einzelnen Ausnahmefällen durch Arbeit, die manchmal die Anerkennung der Menschen in der Gesellschaft gewinnen könnte, aber wenig Geld brächte, oder durch ihren Kopf. (fu127_2006_10c)

ctok: Es gibt gewaltsam Kriminalität wie Mord, Kriminalität gegen **humanität**
ZH2: Es gibt gewaltsame Kriminalität wie Mord, Kriminalität gegen die **Menschlichkeit** (kne20_2006_07)

Pragmatik

Referenzen anpassen
Referenzen werden dem satzübergreifenden Kontext angepasst.

ctok: Sie hatten keine Lust mehr sich nur um **eigene Kinder** zu kümmern, Essen zu kochen, Wäsche zu waschen, kurz zusammengefasst: zu Hause wie in einem Käfig zu sitzen.
ZH2: Sie hatten keine Lust mehr, sich nur um **die eigenen Kinder** zu kümmern, Essen zu kochen, Wäsche zu waschen, kurz gesagt, zu Hause wie in einem Käfig zu sitzen. (fkb030_2008_07)

ctok: Die wenige Schicksallose, denen gelang es ihre Fatum zu verändern und finanziell zu wachsen erreichten das wieder durch Macht, Schönheit, Schlauheit, aber nur in einzelne Außnahmefälle durch Arbeit oder durch **seinen** Kopf, die könnten manchmal die Annerkennung des Menschen in der Gesellschaft zu gewinnen, aber wenig Geld zu bringen.
ZH2: Die wenigen Schicksalslosen, denen es gelänge, ihr Schicksal zu verändern und finanziell zu wachsen, erreichten das wieder entweder durch Macht, Schönheit und Intelligenz, aber nur in einzelnen Ausnahmefällen durch Arbeit, die manchmal die Anerkennung der Menschen in der Gesellschaft gewinnen könnte, aber wenig Geld brächte oder durch **ihren** Kopf. (fu127_2006_10c)

Informationsstruktur: Informationsstatus
Eine Nominalphrase, die auf einen im gleichen Satz bereits erwähnte Referenten verweist, wird pronominalisiert.

ctok: Die Unis sollten auch besser den Studenten erklären, wofür sie etwas lernen. Das, was **man** lernt sollte **der Student** zu einer bestimmten Anwendungssituation verbinden können.
ZH2: Die Unis sollten den Studenten auch besser erklären, wozu sie etwas lernen. Das, was **sie** lernen, sollten **sie** mit einer bestimmten Anwendungssituation verbinden können. (cbs006_2007_10)

ctok: Die meisten Frauen möchten, damit sich die Männer mit ihnen manchmal wie Gentlemen umgehen, **ihnen** Hilfe leisten.
ZH2: Die meisten Frauen möchten, dass die Männer ihnen wie Gentlemen Hilfe leisten. (fk019_2006_07)

ctok: Der Mann spielte eine zentrale Rolle in der Gesellschaft. **Der Mann** war nicht der Frau untertan
ZH2: Der Mann spielte eine zentrale Rolle in der Gesellschaft. **Er** war der Frau nicht untertan (fk024_2006_07)

Informationsstruktur: Fokus
Lerneräußerungen mit einer Fokusgliederung, die dem Kontext nicht entspricht, wird korrigiert.

ctok: **Die besonders schwere** Lage war in Deutschland.
ZH2: Besonders schwer war die Lage in Deutschland.
(fkb051_2008_08)

ctok: Es gibt aber Häuser, wo die Mutter zu Hause nur faulenzt.
Es ist **so** beispielweise in der Familie meines Freundes

ZH2: Es gibt aber Haushalte, in denen die Mutter zu Hause nur faulenzt.
So ist **es** beispielweise in der Familie meines Freundes (fk015_2006_07)

Informationsstruktur: Fokuspartikel
Die Stellung der Fokuspartikel wird an den intendierten Skopus im Satz angepasst.

ctok: Wir müssen uns auch fragen, ob die These des oberliegenden Titels heutzutage wirklich erreichbar sei, weil in diesem politischen Klima, wer hat die Macht, solche Entscheidungen über Lohne und entsprechende gesellschaftliche Beitrage **eigentlich** zu treffen?
ZH2: Wir müssen uns auch fragen, ob die These des obigen Titels heutzutage wirklich erreichbar ist, denn wer hat in diesem politischen Klima **eigentlich** die Macht, solche Entscheidungen über Löhne und entsprechende gesellschaftliche Beiträge zu treffen? (hu006_2006_09)

ctok: Dies Erkentniss ist auch **wesentlich** in finanziellen Berufen, und in Berufen, in den man muss Argumenten machen und schützen, zum Beispiel politische Berufen.
ZH2: Diese Erkenntnis ist auch in der Finanzbranche **wesentlich** und in Berufen, in denen man Argumente bringen und unterstützen muss, zum Beispiel in der Politik. (fk013_2006_08.)

Informationsstruktur: Topik
Topikausdrücke stehen vor Fokusausdrücken, wenn letztere nicht im Kontext als topikalisiert angesehen werden können.

ctok: Z. B. Leibniz hatte keine Möglichkeit an der Leipziger Universität zu promovieren, denn **die grosse Korruption** herrschte **dort**, die Professoren halteten die schlechten Vorlesungen, nicht immer kamen zu den Vorlesungen u s. w.
ZH2: Z. B. hatte Leibniz keine Möglichkeit, an der Leipziger Universität zu promovieren, denn **dort** herrschte **große Korruption**, die Professoren hielten schlechte Vorlesungen, sie kamen nicht immer zu den Vorlesungen usw. (fkb051_2008_08)

Stil

Direkte Rede
Direkte Rede wird nur korrigiert, wenn sie fehlerhaft ist.

ctok: Wozu nutzt es, wenn einer beim Bewerbungsgespräch sagt: "Ja, ich habe die Theorien von Marx, Weber und Durkheim im Griff und ich bin ein netter und pünktlicher Mensch."?
ZH2: Wozu nutzt es, wenn einer beim Bewerbungsgespräch sagt: "Ja, ich habe die Theorien von Marx, Weber und Durkheim im Griff und ich bin ein netter und pünktlicher Mensch."? (cbs011_2006_09)

ctok: Was soll die Lösung sein? Sicherlich nicht die Satz "Universitätsabschlüsse sind von geringem Wert".
ZH2: Was soll die Lösung sein? Sicherlich nicht der Satz: "Universitätsabschlüsse sind von geringem Wert." (fk004_2006_08.)

Einleitungen im Zeitungsstil werden nicht korrigiert
Uneingeleitete Nebensätze werden nicht korrigierte, wenn sie nicht selbst mit einem eingebetteten Nebensatz beginnen oder Verbletztstellung vorliegt.

ctok: Das Unterschied: das Praktikum muss völlig anhand des Studenten ausgehen und wird nur bis zum 50 Prozent als Teil des Studiums anerkannt.
ZH2: Der Unterschied: Das Praktikum muss völlig von dem Studenten ausgehen und wird nur bis zu 50 Prozent als Teil des Studiums anerkannt. (cbs005_2006_09)

Relativpronomen
Nichtstandardsprachliche Relativpronomen werden ersetzt

ctok: Die Schule, **wo** man streng kontrolliert wird, ist schon in der Vergangenheit geblieben.
ZH2: Die Schule, **in der** man streng kontrolliert wurde, ist schon Vergangenheit. (fkb044_2008_08)

Umgangssprachliche Lexik
Umgangssprachliche Lexik wird durch im Kontext angemessenere ersetzt. Im Zweifelsfall siehe Duden.

ctok: Wegen der methodischen / theoretischen Vorgangsweise **kriegt** eine Wissenschaftliche Arbeit auch eine besondere Form.
ZH2: Wegen der methodischen bzw. theoretischen Vorgehensweise **bekommt** eine wissenschaftliche Arbeit auch eine besondere Form. (cbs013_2006_09)

5.4. Zielhypothese für die Annotation der komplexen Verben [ZHverb]

Diese Zielhypothese stimmt fast komplett mit der maximalen Zielhypothese überein und unterscheidet sich nur an einigen Stellen, an denen die Annotatorin der komplexen Verben eine Zielstruktur gewählt hat, die entweder der minimalen Zielhypothese oder einer dritten Variante entspricht.

5.5. Abweichungen der Zielhypothesen von der ctok-Ebene

Für jede Zielhypothese wird automatisch die Abweichung gegenüber der ctok-Ebene berechnet und in einer ZH1Diff, ZH2Diff, ZHverbDiff etc. eingetragen.
Diese Abweichungen erlauben eine gezielte Suche nach fehlerhaften Strukturen in den Lernerdaten.

Fehlendes Token (INS)
Für ein Token, das in der Zielhypothese gefüllt ist, in der ctok-Ebene dagegen nicht, wird dieses Token mit „INS“ gekennzeichnet.

ctok	Kriminalität	zahlt	sich	nicht		.
ZH1	Kriminalität	zahlt	sich	nicht	aus	.
ZH1Diff					INS	

(cbs002_2007_10)

Überflüssiges Token (DEL)
Für ein Token, das in der ctok-Ebene gefüllt ist, in der Zielhypothese dagegen nicht, wird dieses Token mit „DEL“ gekennzeichnet.

ctok	Sich	ständig	in	der	kriminellen	Welt	aufzuhalten	lohnt	sich	auch	nicht	aus	,
ZH1	Sich	ständig	in	der	kriminellen	Welt	aufzuhalten	lohnt	sich	auch	nicht		,
ZH1Diff												DEL	

(cbs002_2007_10)

Verändertes Token (CHA)

Ein Token, das in der Zielhypothese einen anderen Wert hat als in der ctok-Ebene, das aber beide Male nicht leer ist, wird mit „CHA“ markiert.

ctok	die	Unterricht	an	der	Uni	geht	auch	so	.
ZH1	der	Unterricht	an	der	Uni	geht	auch	so	.
ZH1Diff	CHA								

(usb007_2006_10)

Geteiltes Token (SPLIT)
Ein Token der ctok-Ebene, das zwei Tokens in der Zielhypothese entspricht, wird mit „SPLIT“ markiert.

ctok	,	die	er	im	Universität	verbracht	hat
ZH1	,	die	er	in der	Universität	verbracht	hat
ZH1Diff				SPLIT			

(usb008_2006_10)

Token mit anderer Position (MOVS / MOVT)
Taucht in einem Satz das gleiche Token an unterschiedlichen Positionen in der ctok-Ebene und der Zielhypothese auf, so werden sie koindiziert[5] und als Start (MOVS) und Ziel (MOVT) einer „Bewegung“ markiert.

ctok	Und	meiner	Meinung	nach	,	die	Frau	muss			selbst	entscheiden
ZH1	Und	meiner	Meinung	nach				muss	die	Frau	selbst	entscheiden
ZH1Diff					DEL	MOVS	MOVS		MOVT	MOVT		

(usb006_2006_10)

ctok	sich	wohl	draußen	von	der	akademischen	Umgebung		zu	fühlen	,
ZH1	sich		außerhalb	von	der	akademischen	Umgebung	wohlzufühlen			,
ZH1Diff		MOVS	CHA					MOVT			

(fkb057_2008_08)

Token zusammenführen (MERGE)
Sollen in einer Zielhypothese mehrere Token des Ausgangstextes verbunden werden, so wird in der ZHDiff-Ebene mit dem Tag „MERGE“ versehen.

ctok	Man	hat	so	zu	sagen	die	Welt	vor	sich	.
ZH1	Man	hat	sozusagen			die	Welt	vor	sich	.
ZH1Diff			MERGE							

6. Satzspannen

Auf Grundlage der vorhandenen POS-Annotationen wurden für alle Ebenen Satzspannen annotiert. In bestimmten Fällen werden komplexe Sätze allerdings in unterschiedlichen Satzspannen

[5] Dieser Index kann in der aktuellen Version noch nicht dargestellt werden, soll aber in Zukunft die Grundlage für die Realisation als Parallelkorpus dienen.

repräsentiert. Abschnitte, in in der *macro*-Ebene als *start, title, subtitle, comment* oder *end* annotiert sind, werden ausgespart.

Ganze Sätze nach : und ;
Folgt auf einen Doppelpunkt oder ein Semikolon ein Satz mit finitem Verb, wird eine neue Satzspanne begonnen. Dies ist nicht der Fall, wenn mehrere direkte Redebeiträge im komplexen Satz aufeinanderfolgen.

ctok	Wir müssen aber die Frage stellen:	Ist Praxiserfahrung wirklich nötig bei allen Studiengängen?
ctokS	s16	s17

7. Dependenzen

Für die Annotation tiefer syntaktischer Relationen lassen sich unterschiedliche Modelle (Konstituenten, Depenzenzen, hybride Ansätze) verwenden. Konstituentenbäume eignen sich besonders gut für die Beschreibung von Sprachen mit wenig flexibler Wortstellung (siehe Kübler et al. 2009). Lernersprache zeichnet sich unter Anderem aber gerade durch eine große Variation in der Wortstellung aus. Daher haben wir uns dazu entschieden, die Daten auf Dependenzen zu annotieren. Hierbei liegt das Annotationsschema aus Foth (2006) zugrunde.

Lernersprache kann in vielen Bereichen nicht durch gängige Grammatiken beschrieben werden. Um dennoch gängige Parser benutzen zu können, wurde nicht der Originaltext, sondern eine Normalisierungsebene (ZH1) als Grundlage für die syntaktische Annotation ausgewählt. Die Kombination mit den bereits vorhandenen Abweichungsannotationen für die Zielhypothesen kann dann auf die Syntax der Originaltexte zurückgeschlossen werden.

Für die Dependenzanalysen wurden zuerst die automatischen POS-Tags dreier Tagger (TreeTagger, rfTagger,Standford-Tagger) verglichen und manuell korigiert (für genauere Erläuterung siehe Rehbein et al. 2012). Diese POS-Repräsentation bildete den Input für den MALT-Parser (Nivre 2006).

8. Komplexe Verben

Annotatorin: Anke Lüdeling

Gegenstand:
Annotiert werden Präfixverben und Partikelverben. Andere komplexe Verben (z.B. neoklassische) werden nicht betrachtet. Wir wollen möglichst viele Phänomene bzgl. komplexer Verben erfassen, daher sind Zweifelsfälle mit erfasst.

Was zählt dazu?

alle finiten und infiniten zusammengeschriebenen Präfix- oder Partikelverben, auch solche, die auch auseinandergeschrieben werden könnten, zB auch:

> deren Mitglieder sich irgendwo **festgeketten** haben (cbs001_2007_10)[6]

Da es keine klare Definition von Partikelverben und Präfixverben gibt, haben wir hier vieles eingeschlossen, also auch komplexe Verben mit Infinitiv (*kennenlernen*) oder Nomen (*klavierspielen*) als Erstglied. Bei Präfixverben haben wir auch ‚halbverdunkelte' Formen wie *empfinden* aufgeführt. Semantische Transparenz war kein Kriterium.
syntaktisch getrennte Partikelverben.
morphologisch getrennte Partikelverben
Partizipien in Zustandspassivkonstruktionen:

> ctok: …, dass es nicht selten für den Erfolg der Kommunikation **entscheidend** ist, ob …
> ZHverb: …, dass es nicht selten für den Erfolg der Kommunikation **entscheidend** ist, ob … (cbs002_2006_09)

aber nicht Partizipien in pränominaler Adjektivposition
Wir haben auch solche Adj/Adv/Part+V-Kombinationen aufgenommen, die getrennt geschrieben wurden, aber zusammen geschrieben werden können (*schiefgehen, ernstnehmen* etc.). Wir haben hier die zusammengeschriebene Form in die Zielhypothese genommen und einen Orthographiefehler annotiert (orth, siehe unten). .

> ctok: Kriminalitäten wo etwas schief geht,
> ZHverb: Kriminalität, bei der etwas schiefgeht (cbs002_2007_10)

8.1. Annotationsebenen für die komplexen Verben

Die komplexen Verben wurden immer am Verb der Lerneräußerung annotiert. Bei korrekten Verben (keine Änderung in der ZH-Verb) wurden die Ebenen **verbkategorie**, **verblemma** und **verbform** annotiert. Bei Fehlern zusätzlich die Ebene **verbfehlertyp**.

8.1.1. verbkategorie

Annotationswert	*Beschreibung*
vpart	bei Partikelverben, immer am Verb
ppart	bei getrennt stehenden Partikeln
vpräf	bei Präfixverben
ppräf	bei getrennt stehenden Präfixen (das sind dann Fehler) *während die Arbeitenden den Steuer **be zahlen** müssen*

[6] Zu den Fehlern bei komplexen Verben siehe weiter unten. Alle anderen Fehler in den Beispielen sind auf anderen Ebenen zu behandeln und werden hier nicht betrachtet. Die betrachtete Form ist jeweils fett markiert.

	(fkb008_2008_07) (ZH: bezahlen)
vpartx	an Stellen, bei denen ein Partikelverb hätte stehen müssen (ZHVerb), aber keines stand *Das Leben auf die Universität führt zu dem zunehmenden Selbstbewusstsein und bereitet man in diese Weise auf die wirkliche Welt.* (fk008_2006_08) ZH-Verb: *Das Leben an der Universität führt zu einem zunehmenden Selbstbewusstsein und **bereitet** einen in dieser Weise auf die wirkliche Welt **vor**.*
ppartx	an Stellen, an denen eine Partikel hätte stehen müssen (ZHVerb), aber keine stand
vpräfx	an Stellen, an denen ein Präfixverb hätte stehen müssen (ZHVerb), aber keines stand *Diese Frage, nämlich, ob die finanzielle Entlohnung eines Menschend dem Beitrag entsprechen sollte, den er für die Gesellschaft geleistet hat, ist nicht einfach zu antworten* (fk013_2006_07) (ZHVerb: beantworten)

Es gibt (einige wenige) Lerneräußerungen, in denen klar ein Verb fehlt, das dann in ZH-Verb eingesetzt wurde. Solche Verben, auch wenn sie komplex sind, werden gar nicht annotiert, da hier der Lerner keinen Wortbildungsfehler gemacht hat.

8.1.2. verblemma

Das Verblemma wird beim finiten Verb notiert. Üblicherweise steht hier das Verb aus der Lerneräußerung, auch wenn es ein ‚unmögliches' Verb ist.
In den x-Fällen wird hier das Lemma aus der Zielhypothese eingesetzt.

8.1.3. verbfehlertyp

Die Fehlertypen beziehen sich auf die ZHverb

Annotationswert	*Beschreibung*
sem	Verblemma existiert, ist aber falsch verwendet (hier muss in der ZHVerb ein anderes Verb angegeben sein) *Infolgedessen kann man nicht **vernichten**, dass der in der XXten Jahrhundert entwickelte Feminismus […] eine entscheidende Rolle gehabt hat.* (fkb009_2008_07) (ZH: *verneinen*)
orth	orthographischer Fehler *Vielleicht können die dann anschliesend auch versuchen zu **erklähren** was […]* (hu007_2006_10) (ZH: *erklären*) Dazu zählen auch Getrennt- und Zusammenschreibungsfehler

	(dann in der ZH anders). *Der Arbeiter, der nicht **weiterausgebildet** wird, wird bald entkündigt.* (fkb008_2008_07) (ZH: *weiter ausgebildet*) Es gibt ‚Partikelverben' wie z.B. *ernstnehmen*, die getrennt oder zusammen geschrieben werden können. Hier wird eine einheitliche (zusammengeschriebene) Form in der ZH angenommen. Daher wird hier bei Getrenntschreibung ein Fehler angemerkt, der nach Orthographieprinzipien keiner ist.	
lex	Verblemma ist nicht gebräuchlich (hier muss in der ZH ein anderes Verb angegeben sein) *Der Arbeiter, der nicht weiterausgebildet wird, wird bald **entkündigt**.* (fkb008_2008_07) (ZH: *gekündigt*)	
part	Partikel zu viel *Sich ständig in der kriminellen Welt aufzuhalten lohnt sich auch nicht aus.* (cbs002_2007_10) (ZH-Verb: *… lohnt sich auch nicht*.)	
as	Argumentstrukturfehler. Hier wird Argumentstruktur sehr weit gefasst – interessant ist, ob die Lerner die korrekte Verwendung des betreffenden Verbs kennen. In *Dies **verhält** sich auch mit der Praxisorientierung und Praxiserfahrung* (cbs003_2006_09) fehlt z.B. ein *so*. Hierunter werden auch falsche Präpositionen (bei subkategorisierten Ps) gefasst.	
flex	Flexionsform falsch[7] *[…] haben sie trotzdem viel gelernt und erfahrt* (fkb010_2007_09) (ZH: *erfahren*) Hier sind auch falsche Flexionsformen an Auxiliaren markiert (*sein* statt *haben*, *sein* statt *werden* etc.) *[…] oder dass noch eine Frau Vergewaltigt ist* (sa008_2006_09) (ZH *vergewaltigt wurde*)	
ws	Wortstellung (betrachtet wird hier nur die Verbstellung) falsch (vor allem V2-Fehler). Wortstellung innerhalb des Mittelfelds (die mit dem Verb nichts zu tun hat) wird nicht betrachtet.	
Besondere Flexionsprobleme bei komplexen Verben	ge	*ge* steht falsch *Aber auf jeden Fall wird es nicht **rechtgefertigt*** (usb015_2006_10) oder *ge* fehlt *Die Rollen können gelengtlich **austauscht** werden* (fkb040_2008_08)

[7] Manchmal ist es schwierig zwischen flex und orth zu unterscheiden, wie z. B. in *und man **wachst** einfach auf* (fkb010_2007_09, von *aufwachsen*). Wir haben in solchen Fällen flex markiert.

	zu	*zu* steht falsch *Als ich an dem Hochschulsommerkurs **zu teilnehmen** begann* (usb012_2006_10)
Besonderes Wort-stellungsprobleme bei komplexen Verben	sep	Trennung falsch *und nun **anbieten** die meisten davon eine praxisorien-tierte Ausbildung auf dem Niveau BA* (fkb007_2007_09)

8.1.4. verbform

Hier wird die Flexionsform angegeben. Bei Fehlern wird hier die korrigierte Form aus der ZHVerb angegeben.

Annotationswert	*Beschreibung*
fin	finit
finsep	finit und syntaktisch getrennt
inf	Infinitiv ohne *zu*
infzu	Infinitiv mit *zu*
p2	Partizip II
p1	Partizip I
nn	In einigen (ganz wenigen) Fällen steht ein Verb, wo eigentlich ein Nomen stehen müsste (ZH). In solchen Fällen wurde als Fehler flex angegeben und bei Verbform nn.

9. Literatur:

ANNIS2: http://www.sfb632.uni-potsdam.de/d1/annis/, 26.5.2010.

Breindl, Eva et al. (2000): GRAMMIS - ein Projekt stellt sich vor. In: Sprachreport. Informationen und Meinungen zur deutschen Sprache 2000 (1), S. 19–24.

Chiarcos, Christian; Dipper, Stefanie; Götze Michael; Ritz, Julia; Stede, Manfred (2008): A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. In: Proceeding of the Conference on Global Interoperability for Language Resources, Hong Kong, January 2008.

Corder, Stephen Pit (1986): The role of interpretation in the study. In: Corder, Stephen P. (Hrsg.): *Error analysis and interlanguage.* 4. impr. Oxford: Oxford University Press, S. 35–44.

Das elektronische Valenzwörterbuch deutscher Verben (E-VALBU): http://hypermedia2.ids-mannheim.de/evalbu/index.html, 26.5.2010.

Doolittle, Seanna (2006): Handbuch der Annotation der Stellungsfelder bei Falko.

Doolittle, Seanna (2009): Entwicklung und Evaluierung eines auf dem Stellungsfeldermodell basierenden syntaktischen Annotationsverfahrens für Lernerkorpora innerhalb einer Mehrebenen-Architektur mit Schwerpunkt auf schriftlichen Texten fortgeschrittener Deutschlerner. Magisterarbeit HU-Berlin.

Eisenberg, Peter (2006): Der Satz. 3., durchges. Aufl. Stuttgart, Weimar: Metzler (= Grundriss der deutschen Grammatik; Bd. 2).

Engel, Ulrich (1996): Deutsche Grammatik. 3., korr. Aufl. Heidelberg: Groos.

Foth, Kilian A. (2006): Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Universität Hamburg.

Freywald, Ulrike (2008): Zur Syntax und Funktion von dass-Sätzen mit Verbzweitstellung. In: Deutsche Sprache 36), S. 246–285.

Freywald, Ulrike (2009): Kontexte für nicht-kanonische Verbzweitstellung. V2 nach dass und Verwandtes. In: Ehrich, Veronika et al. (Hrsg.): *Koordination und Subordination im Deutschen.* Hamburg: Buske (= Linguistische Berichte; Sonderheft 16), S. 113–134.

Granger, Sylviane (1993): The International Corpus of Learner English. In: Aarts, Jan/Haan, P. de/Oostdijk, Nelleke (Hrsg.): *English Language Corpora. Design, Analysis and Exploitation.* Amsterdam: Rodopi, S. 57–69.

Granger, Sylviane et al. (2009): The International Corpus of Learner English. Version 2. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.

Kübler, Sandra; McDonald, Ryan; Nivre, Joakim (2009): Dependency Parsing. In: Graeme Hirst (Ed.): Synthesis Lectures on Human Language Technologies: Morgan Kaufmann Publishers Inc.; Claypool Publishers.

Lennon, Paul (1991): Error. Some Problems of Definition, Identification, and Distinction. In: Applied Linguistics 12 (2), S. 180–196.

Lüdeling, Anke; Walter, Maik; Kroymann, Emil; Adolphs, Peter (2005): Multi-level error annotation in learner corpora. In: *Proceedings of Corpus Linguistics 2005.* Birmingham.

Lüdeling, Anke (2008): Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora. In: Walter, Maik/Grommes, Patrick (Hrsg.): *Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung.* Deutsche Gesellschaft für Sprachwissenschaft. Tübingen: Niemeyer (= Linguistische Arbeiten; 520), S. 119–140.

Nivre, Joakim (2006): Inductive Dependency Parsing. Dordrecht: Springer (Text, Speech and Language Technology, 34).

Rehbein, Ines; Hirschmann, Hagen; Lüdeling, Anke; Reznicek, Marc (2012): Better Tags Give Better trees or Do They? In: LiLT 7.

Reznicek, Marc (2012): Falko-Excel-AddIn. Version 1.5. MS-Office 2010. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/marc/Falko_0.1.5.xla.

Reznicek, Marc; Lüdeling, Anke; Hirschmann, Hagen (im Druck): Competing Target Hypotheses in the Falko Corpus. A Flexible Multi-Layer Corpus Architecture. In: Ana Díaz-Negrillo (Ed.): Automatic Treatment and Analysis of Learner Corpus Data: John Benjamins.

Schiller, Anne; Teufel, Simone; Thielen, Christine (1995): The Stuttgart-Tübingen Tagset (STTS). http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html.

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing.

Schmid, Helmut; Laws, Florian (2008): Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: Donia Scott (Ed.): 22nd International Conference on Computational Linguistics. Coling 2008. Manchester, United Kingdom., 777–784. http://dl.acm.org/citation.cfm?id=1599081.1599179.

Schmidt, Thomas (2001): The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse. In: *Proceedings of the IRCS Workshop On Linguistic Databases, 11-13 December 2001.* Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania, S. 219–227. URL: http://www.exmaralda.org/files/IRCS_Paper.pdf.

Schneider, Roman (2008): E-VALBU. Advanced SQL/XML processing of dictionary data using an object-relational XML database. In: Sprache und Datenverarbeitung, International Journal for Language Data Processing 32 (1), S. 33–44.

Zeldes, Amir et al. (2009): ANNIS. A Search Tool for Multi-Layer Annotated Corpora. In: *Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009.*

Zipser, Florian (2009): Entwicklung eines Konverterframeworks für linguistisch annotierte Daten auf Basis eines gemeinsamen (Meta-)modells. Diplomarbeit. Institut für Informatik. Berlin.

10.Kontakt

Marc Reznicek
Wissenschaftlicher Mitarbeiter
Institut für deutsche Sprache und Linguistik
Korpuslinguistik und Morphologie
Dorotheenstraße 24
Raum 3.310
Tel: +49 (30) 2093-9720
Marc.Reznicek@staff.hu-berlin.de