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

Spezifikationen des Falko-Lernerkorpus 2.0
Version 1.0

Marc Reznicek, Maik Walter, Jia Wei Chan

Stand vom:
10. September 2010

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

Übersicht

1. Falko-Korpus

Das Falko-Gesamtkorpus V2.0 setzt sich aus fünf Subkorpora zusammen. Sie unterscheiden sich in zwei Faktoren: Schreibaufgabe und Muttersprache. Für die Zusammenfassungen liegen außerdem die Originalvorlagen vor.

	Lernerkorpus	**muttersprachliches Kontrollkorpus**	**Vorlagenkorpus**	Σ
Zusammen-fassungen	FalkoSummaryL2 V1.2 (40.865 Tokens[1])	FalkoSummaryL1 V1.2 (21.184 Tokens)	FalkoSummaryVL (11.114 Tokens)	73.163
Aufsätze	FalkoEssayL2 V2.0 (122.778 Tokens)	FalkoEssayL1 V1.2 (68.491 Tokens)		191.269
Σ	163.643	89.675	11.114	264.432

Für alle Lerner wurden neben den Texten auch umfangreiche Metadaten zu Alter, Geschlecht, akademischem Hintergrund, sprachlicher Biografie und Erhebungssituation erfasst und so aufbereitet, dass sie für die Generierung individueller Subkorpora dienen können.

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

Für alle Daten wurden die Wortarten und Lemmata mit dem *Treetagger* (Schmid 1994) automatisch annotiert. Die händischen Annotationen wurden in *EXMARaLDA* (Schmidt 2004,2005) und *Microsoft Excel* vorgenommen. Mithilfe des Treetaggers wurden dann auch für die Zielhypothesen Wortarten und Lemmata automatisch hinzugefügt. 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 sind frei zugänglich und können auf der Internetseite des Instituts für deutsche Sprache und Linguistik der Humboldt-Universität nach einer Registrierung online durchsucht werden.

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

Die Subkorpora beinhalten folgende Annotationsebenen:

Subkorpus	Ebenen	Tag
FalkoSummaryVL	Lernertext :	word
	Wortart (automatisch):	pos
	Lemma (automatisch):	lemma
FalkoSummaryL1	Lernertext:	word
	Wortart (automatisch):	pos
	Lemma (automatisch):	lemma
FalkoSummaryL2	Lernertext:	word
	Wortart (automatisch):	pos
	Lemma (automatisch):	lemma

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

	Zielhypothese:[2]	target hypothesis
	korrigierte Wortart:	cpos
	Kommentar des Transkribenten:	transcriptor comment
	Topologische Felder:	matrix-satz matrix-satz_felder konstituenten-satz_1 konstituenten-satz_1_ felder konstituenten-satz_1_ felder_2 matrix-satz_2 konstituenten-satz_2 konstituenten-satz_2_felder konstituenten-satz_2_ felder_2 konstituenten-satz_3 konstituenten-satz_3_ felder konstituenten-satz_3_ felder_2
	Syntaktische Beschreibung:	[syntax_description_1] [syntax_classification_1] [syntax_classification_pos_1] [syntax_hypothesis_1] [syntax_description_2] [syntax_classification_2] [syntax_classification_pos_2] [syntax_hypothesis_2]
FalkoEssayL1	Lernertext:	word
	Wortart (automatisch):	pos
	Lemma (automatisch):	lemma
FalkoEssayL2	Basistext:	tok
	Lernertext:	word
	Wortart (automatisch):	pos
	Lemma (automatisch):	lemma
	Normalisierung:	norm
	Textstruktur:	macro
	Minimale Zielhypothese:	ZH1
	Abweichung ZH1:	ZH1Diff
	Maximale Zielhypothese:	ZH2
	Abweichung ZH2:	ZH2Diff
	Zielhypothese komplexe Verben:	ZHverb
	Abweichung ZHverb:	ZHverbDiff
	Komplexe Verben:	verbkategorie verblemma verbfehler verbform

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

1.1. Format der Metadaten:

Für alle Texte in Falko wurden Metadaten erhoben. Obwohl das volle Spektrum an Kategorien erst in zukünftigen Erhebungen aufgenommen werden, wurden auch die bereits vorhandenen Daten an das neue Format angepasst. Die Kategorien, die in der vorliegenden Version 2.0 bereits vorhanden sind, sind mit dem Index „V2.0“ gekennzeichnet.

Kategorie			Beispiel	Tag	Kommentar
Allgemeine Informationen zum Lerner		**Dateiname** V2.0	HUB_001_2005_10.txt	Subkorpus + *laufende Nummer*	dreistelliges Kürzel der erhebenden Institution (wird vor der Erhebung mit dem Falkoteam besprochen), dreistellige laufende Nummer, 2005 und 10 stehen für den Erhebungszeitraum im Oktober 2005 – alles durch Unterstriche getrennt
		Name V2.0	Mustermann	name	werden gelöscht und durch MD5-Hashwerte ersetzt
		Vorname V2.0	Max	first-name	
		Geburtsjahr V2.0	1980	birth-year	in vierstelliger Form: JJJJ
		Geschlecht V2.0	m	sex	männlich = m, weiblich = f
		Bildungsgrad	M.A.	degree	höchster erreichter akademischer Abschluss
		Thema V2.0	Studium	topic	eines der vier vorgegebenen Themen (Entlohnung, Studium, Feminismus, Kriminalität)
		Fach V2.0	Deutsch	major-subject	Studienfach
		Transkribent V2.0	RM	transcriptor	Kürzel des Transkribenten, der die Daten digitalisiert
		Erhebungsdatum V2.0	25.05.2009	transcription-date	Datum, an dem die Erhebung stattfand.
		C-Test V2.0	85	ctest	Ergebnis aus dem C-Test
		Textsorte V2.0	Essay	production-modality	Textsorte
Sprache		**Sprachbezeichnung** V2.0	eng	l1_x l2_x	Kürzel aus der Sprachliste l1 = Muttersprache, l2 = Fremd-/Zweitsprache
		gesprochen seit	0	l1_x_since l2_x_since	ab welchem Alter gesprochen/gelernt, 0 = ab Geburt
	Unterricht	**Monate** V2.0	6	l1_x_duration l2_x_duration	Dauer des Unterrichts der Sprache in MOonaten

	Institution	**Schule**	Ja	l1_x_school l2_x_school	Sprache in einer Schule gelernt?
		Uni	Ja	l1_x_university l2_x_university	Sprache an einer Universität gelernt?
		Sprachschule	Ja	l1_x_langschool l2_x_langschool	Sprache an einer Sprachschule gelernt?
	Auslands-ufenthalt	**Monate**	26	l1_x_awayMonths l2_x_awayMonths	Aufenthalt in einem Land der Zielsprache in Monaten
		Ort	Frankfurt	l1_x_awayPlace l2_x_awayPlace	Ort des Aufenthalts
		Indexfelder zur Suche mit regulären Ausdrücken [V2.0]	dan:N/A:0:Ja:Ja:Nein: N/A:N/A	l1index	Abfolge aller Werte für die Muttersprachen *Sprache1:Wert1$_{Sprache1}$:Wert2$_{Sprache1}$; Sprache2: Wert1$_{Sprache2}$…*
			deu:48:12:Nein:Ja:Nein:12:Berlin	l2index	Abfolge aller Werte für die Fremdsprachen *Sprache1:Wert1$_{Sprache1}$:Wert2$_{Sprache1}$; Sprache2: Wert1$_{Sprache2}$…*
			l1:dan,l1_duration:N/A,l1_since:N/A,l1_school:N/A,l1_university:N/A,l1_langschool:N/A,l1_awayMonths:N/A,l1_awayPlace:N/A;l2:eng,l2_duration:N/A,l2_since:N/A,l2_school:N/A,l2_university:N/A,l2_langschool:N/A,l2_awayMonths:N/A,l2_awayPlace:N/A	reg	Abfolge aller Sprachen und ihrer Werte als Variablen-Wert-Paare *Variable$_1$: Wert$_{Sprache1}$, Variable$_2$: Wert$_{Sprache1}$; Variable$_1$: Wert$_{Sprache2}$, Variable$_2$: Wert$_{Sprache2}$; …*

2. Falko Zusammenfassungskorpus(Summary-Korpus)

Das Falko-Summary-Korpus besteht aus drei Subkorpora.

1.2 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.

1.3 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.

1.4 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.

3.1. FalkoSummaryL2 1.1

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		L1	L2
	männlich	weiblich		
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			

Datenerhebung 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 text-grenzen.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			

Datenerhebung 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			

Datenerhebung 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			

Datenerhebung 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)	Deutsch (9) Englisch (6) Französisch (2)

		Baschkirisch (1)	Latein (2) Altgriechisch (1) Arabisch(1) Hebräisch (1) Italienisch (1) Japanisch (1) Niederländisch(1) Spanisch (1) Türkisch (1) Tschechisch (1)
Σ	9		

3.1.1. Annotationen in FalkoSummaryL2

Anotationsebene	Kürzel
Lernertext:	word
Wortart (automatisch):	pos
Lemma (automatisch):	lemma
einfache Zielhypothese:	target hypothesis
korrigierte Wortart:	cpos
Kommentar des Transkribenten:	transcriptor comment
Topologische Felder:	matrix-satz matrix-satz_felder konstituenten-satz_1 konstituenten-satz_1_ felder konstituenten-satz_1_ felder_2 matrix-satz_2 konstituenten-satz_2 konstituenten-satz_2_felder konstituenten-satz_2_ felder_2 konstituenten-satz_3 konstituenten-satz_3_ felder konstituenten-satz_3_ felder_2
Syntaktische Beschreibung:	[syntax_description_1] [syntax_classification_1] [syntax_classification_pos_1] [syntax_hypothesis_1] [syntax_description_2] [syntax_classification_2] [syntax_classification_pos_2] [syntax_hypothesis_2]

2.2.1.1 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.

2.2.1.2 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 Handbuch der Annotation der Stellungsfelder bei Falko..

3.2. FalkoSummaryL1 1.1

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**

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**

3.3. 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. Streif-züge durch die Psycholinguistik. Heidelberg, Berlin, New York: Spektrum. 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., über-arbeitete 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.

Tokenanzahl

Texte	12	Tokens	11114	Ø/Text	926,17

3. Falko-Aufsatzkorpus (Essay-Korpus)

Das Falko-Essay-Korpus besteht aus zwei 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 Gesamt-menge der im *International Corpus of Learner English* (ICLE) (Granger 1993, 2003) ver-wendeten 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.

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.

4.1. FalkoEssayL2 2.0

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, 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

3.1.1 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 18.07.2006 10.08.2006 12.09.2006 29.09.2006	5	13	Suaheli (9) Kikuyu (5) Luo (5) Luhya (2) Meru (2) Embu (1) Gusii (1) Kalenjin (1) Nandi (1) KiTaita (1)	Deutsch (18) Englisch (18) Suaheli (9) Französisch (2) Chinesisch (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 Teil-nehmer haben 2 Texte geschrieben)*	Deutsch (13) Englisch (12) Französisch (8) Italienisch (3) Spanisch (3) Griechisch (1) Latein (1) Niederländisch (2) *(Fast alle Teil-nehmer 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) Russisch (4) Tadschikisch (3)	Deutsch (15) 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 25.07.2008 06.08.2008	12	38	Russisch (13) Englisch (12) Französisch (6) Dänisch (5) Spanisch (5) Ukrainisch (4) Niederländisch (3) Polnisch (3) 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)	Englisch (55) Deutsch (50) Französisch (28) Spanisch (18) Latein (17) Russisch (7) Italienisch (6) Niederländisch (4) 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)	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)

			Finnisch (2) Hindi (2) Luhya (2) Ungarisch (2) Vietnamesisch (2) Albanisch (1) Angika (1) Chinesisch(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)	Giryama (1) Griechisch (1) Indonesisch(1) Jiddisch (1) Kikuyu (1) Koreanisch (1) Luhya (1) Meru (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	Freie Universität, Deutschland (FU)	9	10166	442,00

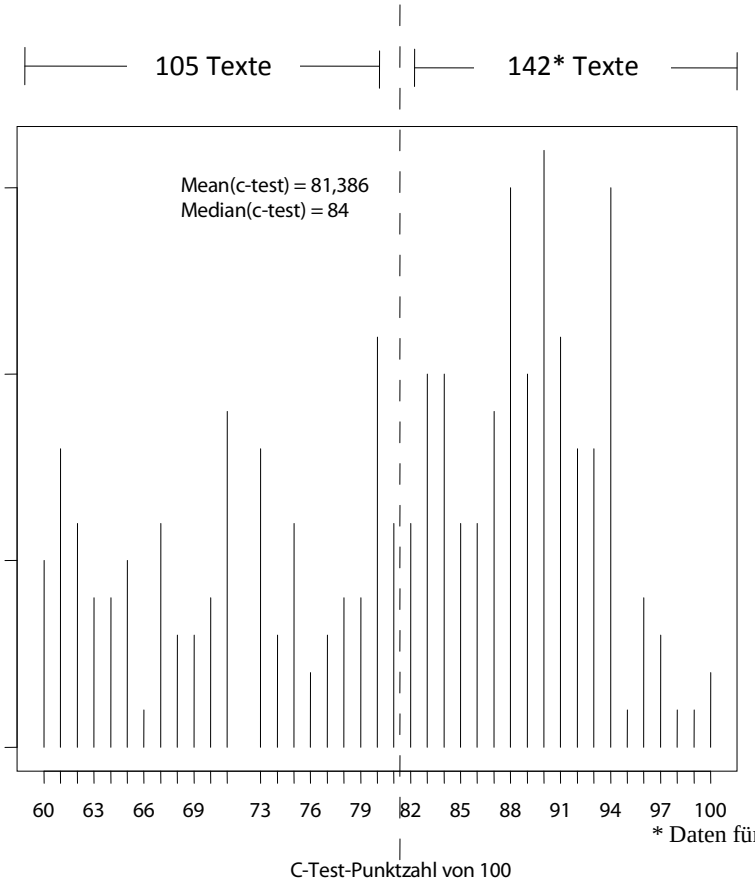
20.11.2007		14		
07.12.2006	Stellenbosch University, Südafrika (SA)	10	6802	680,20
05.03.2007 18.05.2007 25.06.2007	Auckland University, New Zealand (NZ)	1 1 2	2148	537
09.08.2007 20.11.2007 25.07.2008 06.08.2008	Humboldt-Universität Ferienkurs, Deutschland (FKB)	12 37 1 24	36559	494,04
Σ		**248**	**122778**[3]	**495,08**

Tokenanzahl

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

3.1.3 Verteilung der C-Test-Ergebnisse in FalkoEssayL2 2.0

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 von 131510 Tokens überein. Dort werden die durch die Zielhypothesen entstandenen Leertokens mitgezählt.

4.2. FalkoEssayL11.2

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

3.2.1 Übersicht über Sprache und Geschlecht der Lerner für die einzelnen Erhebungen

(Mehrfachangaben für Sprachen wurden auch mehrfach gezählt)

Datum der Erhebung	Anzahl der Teilnehmer		L1	L2
	männlich	weiblich		
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			

3.2.2 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**

Tokenanzahl

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

4. Literatur:

(2006): Handbuch der Annotation der Stellungsfelder bei Falko.

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

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.

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.

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.

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.

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

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.

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.

5. Kontakt

Marc Reznicek
Wissenschaftlicher Mitarbeiter
Institut für deutsche Sprache und Linguistik
Korpuslinguistik und Morphologie
Dorotheenstraße 24
Raum 3.339
Tel: +49 (30) 2093-9720
Marc.Reznicek@staff.hu-berlin.de