



Erster Tübingen-Berlin Lernerkorpus-Workshop (15.-16.10.2009)

Falko ✓

&

WHiG? ✓ ✗

Marc Reznicek

Gliederung

- 1) Falko – Design**
- 2) WHIG –What's hard in German**
- 3) Forschung an Falko**
- 4) Operationalisierung der Zielhypothese**

- Lerner
 - 132.187 Token
- Muttersprachler (Kontrollgruppe)
 - 88.730 Token
- Schriftliche Texte
 - handschriftlich
 - computergeschrieben
- Aufgaben
 - Zusammenfassungen
 - Aufsätze
- Design
 - Querschnittsdaten
 - Längsschnittdaten

Lüdeling et al. 2008

■ L1

- 49 verschiedene Muttersprachen
- eigene Einschätzung der Lerner
- häufigste Sprachen:
Dänisch, Englisch, Französisch, Polnisch, Russisch

■ Herkunftsorte

- Europa, Türkei, Kenia, Südafrika...

■ fortgeschrittene Lerner

- Externes Kriterium:
 - DSH (Summaries)
 - Zwei abgeschlossene Studienjahre Deutsch (Georgetown)
- Internes Kriterium:
 - C-Test (Essays)

- automatische Annotation

- POS
- Lemma

durch Treetagger

Schmidt 2001

- händische Annotation

- korrigierte POS
- Topologische Felder
- Korrigierte POS
- Normalisierung
- Zielhypothesen

Lüdeling et al. 2007, Doolittle 2009

- Falko Projektseite

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

- WHIG Projektseite

<http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/lernschwierigkeiten/WHIG>

- Falko CQP-Suche

<http://korpling.german.hu-berlin.de/falko/index.jsp>

- Falko Annis-Demo-Suche

<http://korpling.german.hu-berlin.de/Annis/>



Erweiterung von Falko WHIG



What's hard in German?



- Auffinden struktureller Schwierigkeiten bei Lernern des Deutschen als Fremdsprache
 - Europäisches Gemeinschaftsprojekt



Humboldt-Universität zu Berlin



Bangor University

Gefördert durch:



und



Arts & Humanities
Research Council



■ Datenerhebungen

- an acht Universitäten in Großbritannien



■ Essays

- ~ 600 Texte a 500 Token
- Lerner im zweiten und vierten Jahr „German studies“

- Contrastive Interlanguage Analysis (CIA)

- Systematische Abweichungen der Interimssprachen von der muttersprachlichen Varietät der Kontrollgruppe

- **Overuse & Underuse**

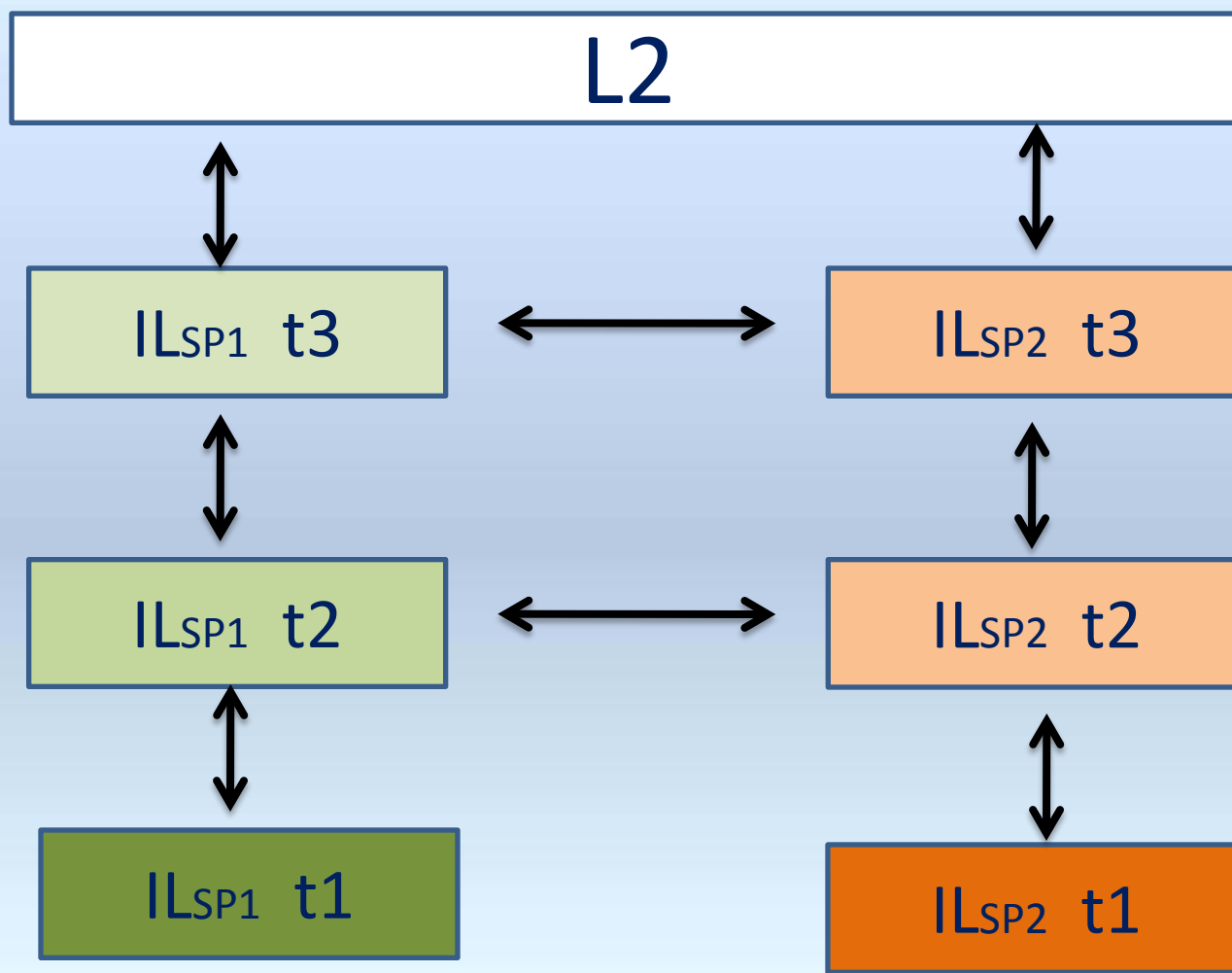
Cobb 2003, Tono 2003, Granger 2008, Walter & Grommes 2008, Mukherjee 2008

- Fehleranalyse

- Abweichung der Interimssprachen von der Norm der Zielsprache

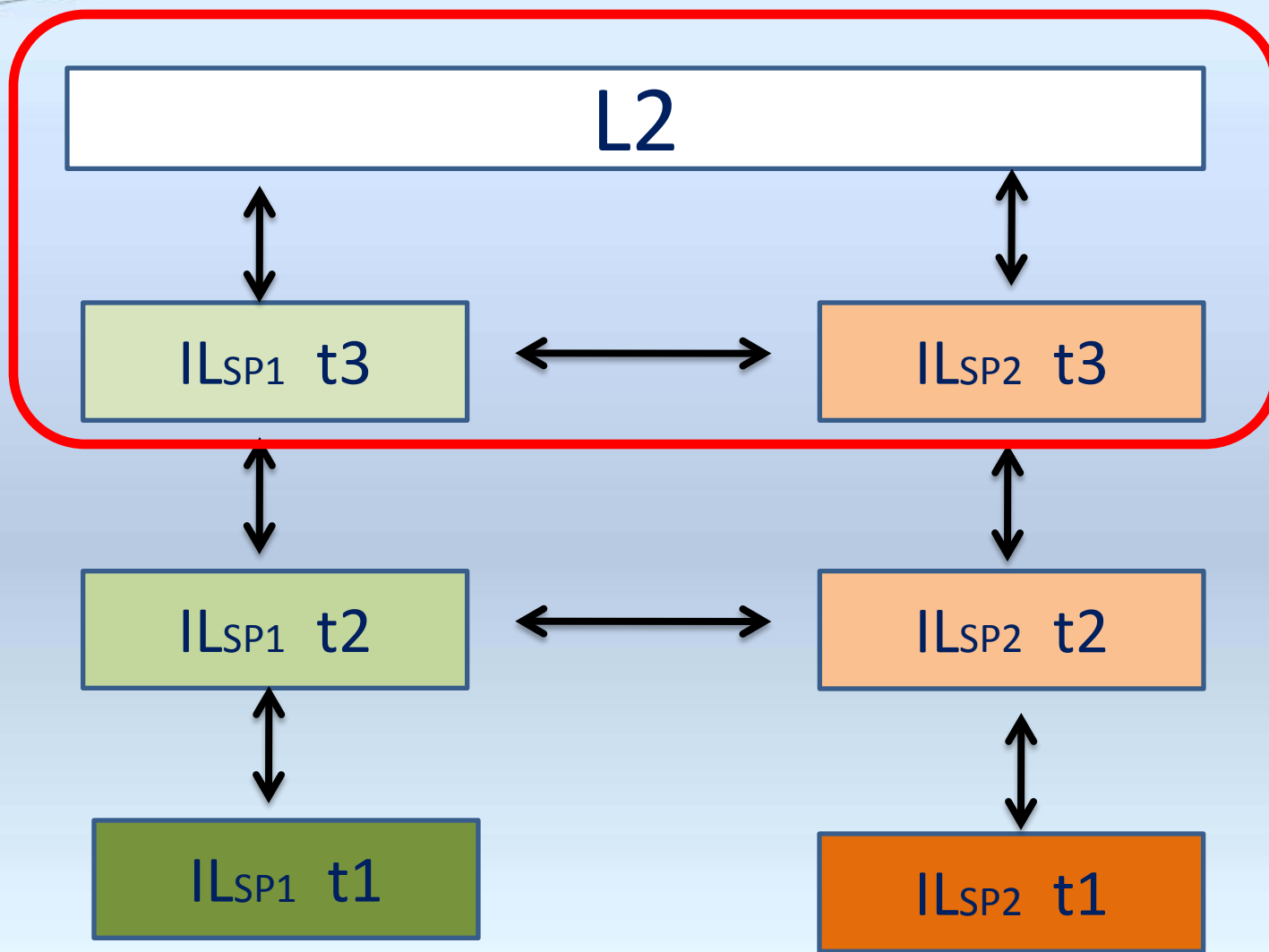
- **Fehlerannotation**

Corder 1981 u.a.

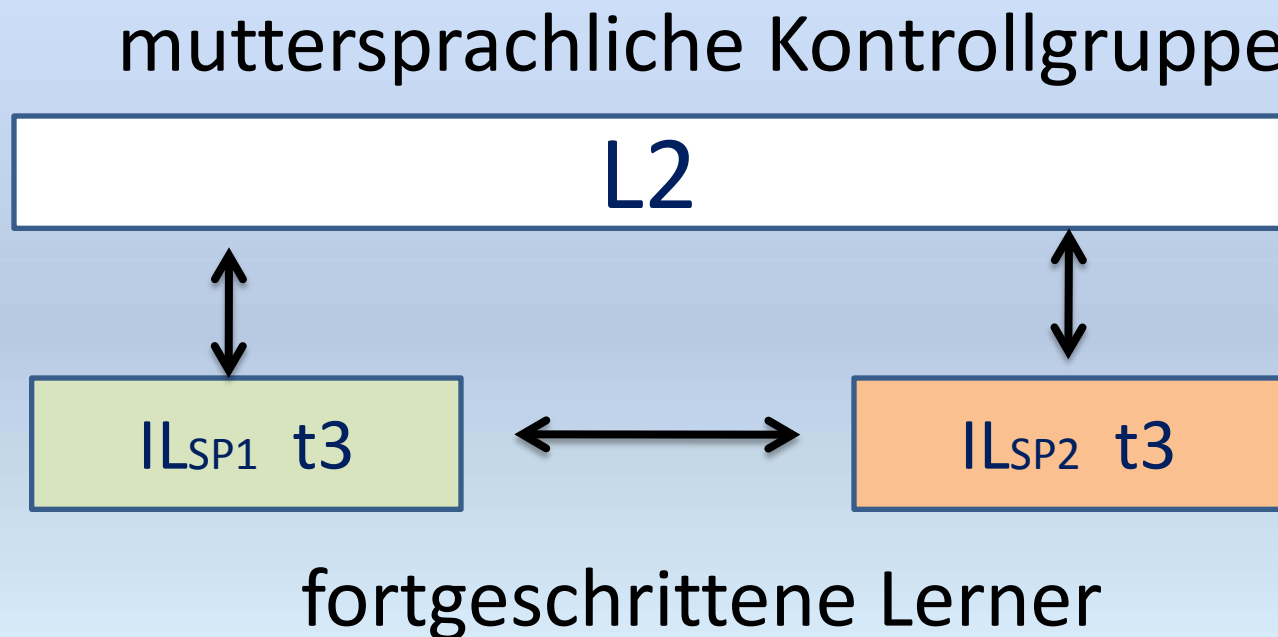


Forschung an Falko

Contrastive Interlanguage Analysis



■ Underuse



- Annahme 1: Vermeidung
 - schwierige Strukturen werden häufiger vermieden als einfache.
- Folgerung: Underuse
 - Vermeidung schwieriger Strukturen führt zum Underuse dieser Strukturen in den Lernertexten

schwierige
Struktur

Vermeidungs-
strategie

Underuse

- Problem 1:

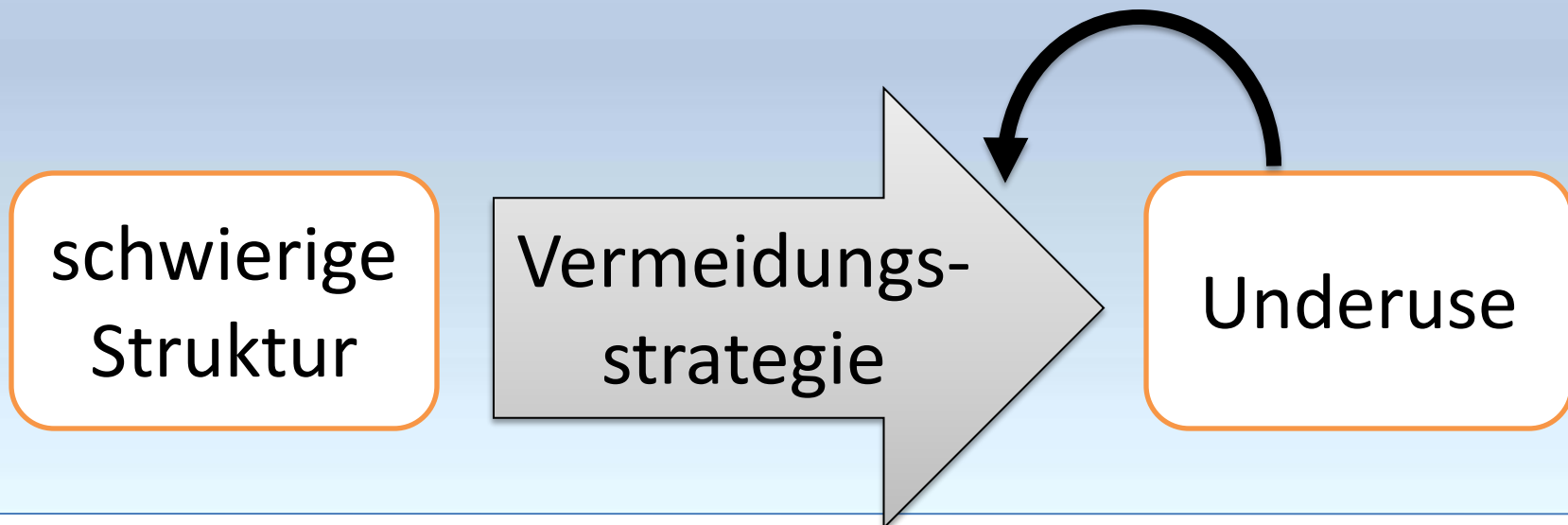
Underuse kann unterschiedliche Gründe haben

- a) Struktur ist nicht bekannt
beisp. Lexem „behände“
- b) Lerner weiß um seine Wissenslücke und benutzt
Alternativstruktur
- c) Lerner kennt die Struktur, vermeidet sie aber
unbewusst

■ Lösung:

Es werden nur solche Strukturen untersucht, die

- a) in den Lernertexten frequent sind.
- b) in den Lernertexten unterrepräsentiert sind.





Forschung an Falko Underuse

Falko

- Visualisierung von Overuse & Underuse

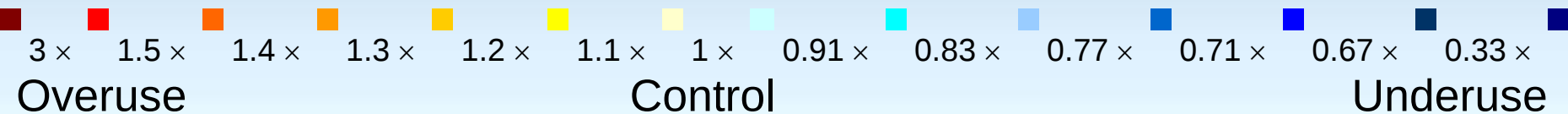
Excel-Plugin

Amir Zeldes

Forschung an Falko

Underuse

1	word	de	da	en	fr	pl	ru
2	,	0.055799	0.062897	0.064648	0.070923	0.050833	0.080987
61	dann	0.002754	0.002885	0.001459	0.001736		0.002165
62	einer	0.002666	0.001603	0.002465	0.002671	0.001412	0.003465
63	schon	0.002503	0.001154	0.000704	0.000801	0.001977	0.003032
64	da	0.002488	0.001154	0.000503	0.000801	0.000282	
65	Frau	0.002458	0.000769	0.001157	0.00187	0.001694	0.005197
66	muss	0.002384	0.002372	0.002515	0.001469	0.003389	0.007796
67	Geld	0.002369	0.000769	0.002515	0.000534		0.002599
68	viele	0.002281	0.003847	0.002214	0.00187	0.005083	0.002599
69	zum	0.002147	0.001282	0.002666	0.002938	0.000565	0.000866
70	des	0.002132	0.001731	0.003471	0.004408	0.001694	0.001299
71	bei	0.002103	0.001539	0.000755	0.000134	0.000282	0.000433
72	können	0.002088	0.003526	0.002616	0.003206	0.001977	0.001732
73	denn	0.002073	0.001154	0.001006	0.001336	0.00113	
74	Leben	0.002058	0.001411	0.001811	0.002003	0.000847	0.003465
75	sollte	0.00197	0.001026	0.000956	0.001469	0.000847	
76	einem	0.001925	0.001731	0.001459	0.000935	0.001694	0.000866



- Visualisierung von Overuse & Underuse

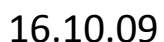
Excel-Plugin

Amir Zeldes

- transferunabhängige Lernerschwierigkeiten

a) unterrepräsentierte Strukturen in allen L1-Gruppen

‚denn‘ ist in allen L1-
Interlanguages unterrepräsentiert



- Visualisierung von Overuse & Underuse

Excel-Plugin

Amir Zeldes

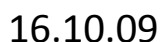
- Transferunabhängige Lernerschwierigkeiten

a) unterrepräsentierte Strukturen in allen L1-Gruppen

- Ebenen des Underuses

a) Lexeme

‚denn‘ ist in allen L1-
Interlanguages unterrepräsentiert



- Visualisierung von Overuse & Underuse

Excel-Plugin

Amir Zeldes

- transferunabhängige Lernerschwierigkeiten

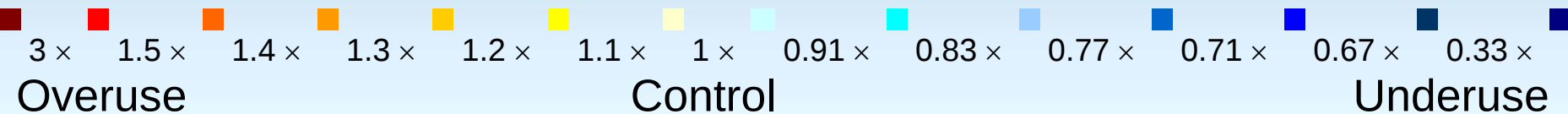
- a) unterrepräsentierte Strukturen in allen L1-Gruppen

- Ebenen des Underuses

- a) Lexeme

- b) Wortfolgen (n-Gramme)

c_pos_bi	de	da	en	fr	pl	ru
VVINP-\$.	0,00714479	0,00808055	0,0100463	0,01014902	0,00679558	0,00803356
PTKZU-VVINP	0,00722367	0,00775989	0,01148148	0,01014902	0,00662983	0,00696242
VVINP-\$,	0,0070321	0,00872186	0,00967593	0,01258993	0,00519337	0,00696242
\$. -PRELS	0,00776461	0,00705445	0,00675926	0,01220452	0,00883978	0,00723021
NN-ADV	0,0106383	0,00885013	0,00703704	0,00668037	0,00718232	0,00642685
APPRART-NN	0,00798999	0,00647727	0,00689815	0,00642343	0,0081768	0,00883692
ADV-ADJD	0,00815903	0,00878599	0,0087963	0,00796506	0,00530387	0,00473088
\$. -PPER	0,00529661	0,00974796	0,00796296	0,0061665	0,0058011	0,00740873
VVFIN-\$,	0,00645736	0,00775989	0,00634259	0,00630731	0,00624309	0,00839061
PROGAT-NN	0,00805761	0,00724684	0,00726852	0,00706578	0,00629834	0,00580292
ADV-ADV	0,01285837	0,01051754	0,00611111	0,0061665	0,00309392	0,00285638
ADV-APPR	0,00911693	0,00801642	0,00532407	0,00783659	0,00480663	0,00434161
PDAT-NN	0,0054093	0,00423267	0,00550926	0,00783659	0,00773481	0,00883692
ADV-ART	0,00762937	0,006349	0,00689815	0,00565262	0,0061326	0,00446309
VAFIN-\$.	0,0052628	0,00724684	0,00597222	0,00539568	0,00635359	0,00669464
VVFIN-\$.	0,00578119	0,00468159	0,00490741	0,00513875	0,00883978	0,00749799



- Visualisierung von Overuse & Underuse

Excel-Plugin

Amir Zeldes

- transferunabhängige Lernerschwierigkeiten

- a) unterrepräsentierte Strukturen in allen L1-Gruppen

- Ebenen des Underuses

- a) Lexeme

- b) Wortfolgen (n-Gramme)

- c) Syntaktische Einheiten (minimal trees)

- d) Funktionale Einheiten

- offene Fragen

- Welche unterrepräsentierten Strukturen lassen sich auf den verschiedenen Ebenen finden
- Finden sich diese Strukturen auch bei Lernern mit anderer L1?
- Was macht die unterschiedlichen Strukturen schwierig?

- Hypothese

Lerner haben Probleme mit Strukturen hoher Variabilität

Zeldes et al. 2008, Hirschmann (dieser Workshop)

- Fehlererkennung
- Fehlerannotation (sehr ausführlich)

- Fehlerebenen:
 - linguistische Ebene
 - Fehlertyp
 - Prinzipienverletzung
 - beteiligte Wortart

Lüdeling et al 2008, Granger 2009

- Fehleranalyse

- Fehlerannotation beruht IMMER auf impliziter korrekter Vorlage (Zielhypothese)
 - Beisp.: ICLE
Dagneaux/Denness/Granger/Meunier 1996
 - Die Zielhypothese zu finden, ist nicht trivial. Lüdeling 2008
- Explizite Zielhypothese in Falko
- Konkurrierende Zielhypothesen

■ tokenbasierte Zielhypothese

[word]	Er	tatsächlich	war		sehr	wohlhabend
ZH	Er		war	tatsächlich	sehr	wohlhabend
NICHT		war tatsächlich				

■ automatische Annotation der Abweichungen

- Löschung: DE
 - Modifikation: MOD
 - Einsetzung: INS
- } OP Operation

[word]	Er	tatsächlich	sein		sehr	wohlhabend
ZH	Er		war	tatsächlich	sehr	wohlhabend
OP		DE	MOD	INS		

Zielhypothesen

Operationalisierung

- minimale Abweichungen
 - Je weniger OPs, desto besser

word	dass	es	so	groß	wie	Polen	ist
------	------	----	----	------	-----	-------	-----

+

ZH	dass	es	so	groß	ist	wie	Polen	
OP					INS			DEL

-

ZH	dass	es	so	groß			ist	wie	Polen
OP					DEL	DEL		INS	INS

Zielhypothesen

Operationalisierung

- minimale Abweichungen
 - Je kürzer die Abstände zwischen Löschung und Einsetzung desto besser

word	Oft	sie	spielen	mit	uns	Schach
------	-----	-----	---------	-----	-----	--------

+

ZH	Oft		spielen	sie	mit	uns	Schach
OP		DEL	1	INS			

-

ZH		sie	spielen	oft	mit	uns	Schach
OP	DEL	1	2	INS			

Zielhypothesen

Operationalisierung

- topologische Felder herstellen
 - In Hauptsätzen wird das finite Verb nicht bewegt.
 - Felderfehlbesetzungen werden so besser erkannt

word	Oft	sie	spielen	mit	uns	Schach
------	-----	-----	---------	-----	-----	--------

+

ZH	Oft		spielen	sie	mit	uns	Schach
OP		DEL	1	INS			

-

ZH	Oft	spielen	sie		mit	uns	Schach
OP		INS		DEL			

- Vollverb bestimmt seine Argumente
 - falsche Argumente werden an das Vollverb angepasst.

word	Sie	wissen	,	was	Ihnen	erwartet
------	-----	--------	---	-----	-------	----------

+

ZH	Sie	wissen	,	was	sie	erwartet
OP					MOD	

-

ZH	Sie	wissen	,	was	ihnen	blüht
OP						MOD

Zielhypothesen

Operationalisierung

- zwei Zielhypothesen

[word]	das	ich	als	für	wertlos	im	Praktischen	schätze
--------	-----	-----	-----	-----	---------	----	-------------	---------

Zielhypothesen

Operationalisierung

- zwei Zielhypothesen
 - ZH1: Rechtschreibung, Morphosyntax, Valenz
→ strukturell grammatischer deutscher Satz

[word]	das	ich	als	für	wertlos	im	Praktischen	schätze
--------	-----	-----	-----	-----	---------	----	-------------	---------

- zwei Zielhypothesen
 - ZH1: Rechtschreibung, Morphosyntax, Valenz
→ strukturell grammatischer deutscher Satz

[word]	das	ich	als	für	wertlos	im	Praktischen	schätze
--------	-----	-----	-----	-----	---------	----	-------------	---------

[ZH1]	das	ich	im	Praktischen	als		wertlos		einschätze
-------	-----	-----	----	-------------	-----	--	---------	--	------------



- zwei Zielhypothesen

- ZH1: Rechtschreibung, Morphosyntax, Valenz
→ strukturell grammatischer deutscher Satz

- ZH2: Lexik, Semantik, Pragmatik
→ sinnvoller (-getreuer) akzeptabler deutscher Satz

[word]	das	ich	als	für	wertlos	im	Praktischen	schätze
--------	-----	-----	-----	-----	---------	----	-------------	---------

[ZH1]	das	ich	im	Praktischen	als		wertlos		einschätze
-------	-----	-----	----	-------------	-----	--	---------	--	------------



- zwei Zielhypothesen

- ZH1: Rechtschreibung, Morphosyntax, Valenz
→ strukturell grammatischer deutscher Satz

- ZH2: Lexik, Semantik, Pragmatik
→ sinnvoller (-getreuer) akzeptabler deutscher Satz

[word]	das	ich	als	für	wertlos	im	Praktischen	schätze
---------------	-----	-----	-----	-----	---------	----	-------------	---------

[ZH1]	das	ich	im	Praktischen	als		wertlos		einschätze
--------------	-----	-----	----	-------------	-----	--	---------	--	------------

[ZH2]	das	ich	in	der	Praxis		für	wertlos		halte
--------------	-----	-----	----	-----	--------	--	-----	---------	--	-------

Ausblick

- automatische Annotation der Zielhypothesen
 - POS-Tagging
 - Lemmatisierung
- syntaktische Annotation der Zielhypothese
 - Chunks
 - Trees
 - topologische Felder
- Qualitätsprüfung automatisch geparster Lernaltersprache
 - Abweichung der Bäume: Rohtext vs. Zielhypothese
- Fehlerannotation
 - Kategorisierung der Abweichungen v.d. Zielhypothese

Vielen DANK!

- **an Falko arbeite(te)n:**

Anke Lüdeling, Maik Walter, Karin Schmidt, Hagen Hirschmann, Seanna Doolittle, Victoria Oketch, Emil Kroymann, Amir Zeldes, Cedric Krummes, Jia Wei Chan, Thomas Krause, Florian Zipser

- Cobb, T. (2003):** Analyzing late interlanguage with learner corpora. Quebec replications of three European studies. In: *The Canadian Modern Language Review* 59 (3), S. 393–423.
- Dagneaux, Estelle/Denness, Sharon/Granger, Sylviane (1998):** Computer-aided error analysis. In: *System* 26 (2), S. 163–174.
- Dagneaux, Estelle/Denness, Sharon/Granger, Sylviane/Meunier, Fanny (1996):** Error Tagging Manual Version 1.1. **Université Catholique de Louvain (Hrsg.).** Louvain-la-Neuve: Centre for English Corpus Linguistics.
- Dagneaux, Estelle/Denness, Sharon/Meunier, Fanny/Neff, J./Thewissen, J. (2005):** Error Tagging Manual Version 1.2. Louvain-la-Neuve: Centre for English Corpus Linguistics. Université Catholique de Louvain.
- 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.
- Doolittle, Seanna/Hirschmann, Hagen/Lüdeling, Anke (2008):** Syntactic annotation of non-canonical linguistic structures. In: *Proceedings of Corpus Linguistics 2008*.
- Granger, Sylviane (2008):** Learner corpora. In: Lüdeling, Anke et al. (Hrsg.): *Corpus linguistics. An International Handbook*. Berlin [u.a.]: Mouton de Gruyter (= Handbucher zur Sprach- und Kommunikationswissenschaft; 1), S. 259–275.
- Granger, Sylviane (2009):** The contribution of learner corpora to second language acquisition and foreign language teaching. A critical evaluation. In: Aijmer, Karin (Hrsg.): *Corpora and language teaching*. University of Gothenburg. Amsterdam, Philadelphia: John Benjamins (= Studies in corpus linguistics; 33), S. 13–32.

- 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.
- Lüdeling, Anke/Doolittle, Seanna/Hirschmann, Hagen/Schmidt, Karin/Walter, Maik (2008):** Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache* (2), S. 67–73.
- Lüdeling, Anke/Ensslin, Astrid:** What's Hard in German? Research Grant Proposal (Standard Route) relating to the Memorandum of Understanding (MoU) between the AHRC and DFG.
- Lüdeling, Anke/Walter, Maik:** Korpuslinguistik für Deutsch als Fremdsprache. Sprachvermittlung und Spracherwerbsforschung.
- Lüdeling, Anke/Walter, Maik/Kroymann, Emil/Adolphs, Peter (2005a):** Falko. Ein fehlerannotiertes Lernerkorpus des Deutschen. Poster auf der 25. Jahrestagung Deutschen Gesellschaft für Sprachwissenschaften. Köln.
- Lüdeling, Anke/Walter, Maik/Kroymann, Emil/Adolphs, Peter (2005b):** Multi-level error annotation in learner corpora. In: *Proceedings of Corpus Linguistics 2005*. Birmingham.
- Mukherjee, Joybrato (2008):** Anglistische Korpuslinguistik und Fremdsprachenforschung. Entwicklungslinien und Perspektiven. In: *Zeitschrift für Fremdsprachenforschung* 19 (1), S. 31–60.
- 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.

- Tono, Yukio (2003):** Learner corpora. Design, development, and applications. Lancaster: Pre-conference workshop on learner corpora, Corpus Linguistics 2003.
- Walter, Maik/Grommes, Patrick (Hrsg.) (2008):** Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung. Tübingen: Niemeyer (= Linguistische Arbeiten; 520).
- Zeldes, Amir (2008):** Underuse / Overuse Excel Add-In (V1.2). URL: <http://korpling.german.humboldt-berlin.de/~amir/uoaddin.htm> [Stand: 26. November 2009].
- Zeldes, Amir/Lüdeling, Anke/Hirschmann, Hagen (2008):** What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data. Helsinki, Finland: Quantitative Investigations in Theoretical Linguistics 3.