Zielhypothesen und Lernersprache
–
Das Falko-Lernerkorpus

Marc Reznicek

22.06.12

offene Berliner DaF-Reihe

Mit Folien des gesamten
Korpuslinguistikteams der HU-Berlin

Übersicht

- Fragestellungen der Lernerkorpusforschung
- Falko - Design
- Annotation
  - Automatische Vorverarbeitung
  - Fehler und Zielhypothesen
- Fragestellungen und Analysen
  - Lernersyntax
  - Produktivität

Fragestellungen

Falko

- Welche sprachlichen Strukturen sind für Deutschlerner schwierig?
- Sind bestimmte Fehler L1-abhängig?
- Liegt das an der Form oder an der Funktion dieser Strukturen?
- Wie produktiv sind Fremdsprachenlerner im Sprachgebrauch?

Methoden

Falko

Contrastive **I**nterlanguage **A**nalysis (CIA: Granger 2008)

- → Muster in abstrakten Repräsentationen finden
- → Quantitative Unterschiede zwischen Lernern und Muttersprachlern aufdecken
- → L2-Texte unterschiedlicher Muttersprachler kontrastieren

Fehleranalyse (EA) (Corder 1981, Izumi et al. 2005 u.v.m.)

- → Welche Fehler sind lernertypisch
- → Abhängigkeit von L1 der Lerner?

Lernerkorpora

Falko

- Kontrollierte und meist digitalisierte Sammlungen von Lernertexten
- Das Design hängt von der Fragestellung ab
  - gesprochen vs. geschrieben / Textsorte / Aufgabe
  - Fortgeschrittenheit
  - L1 der Lerner
  - …
- die meisten Korpora auf Englisch

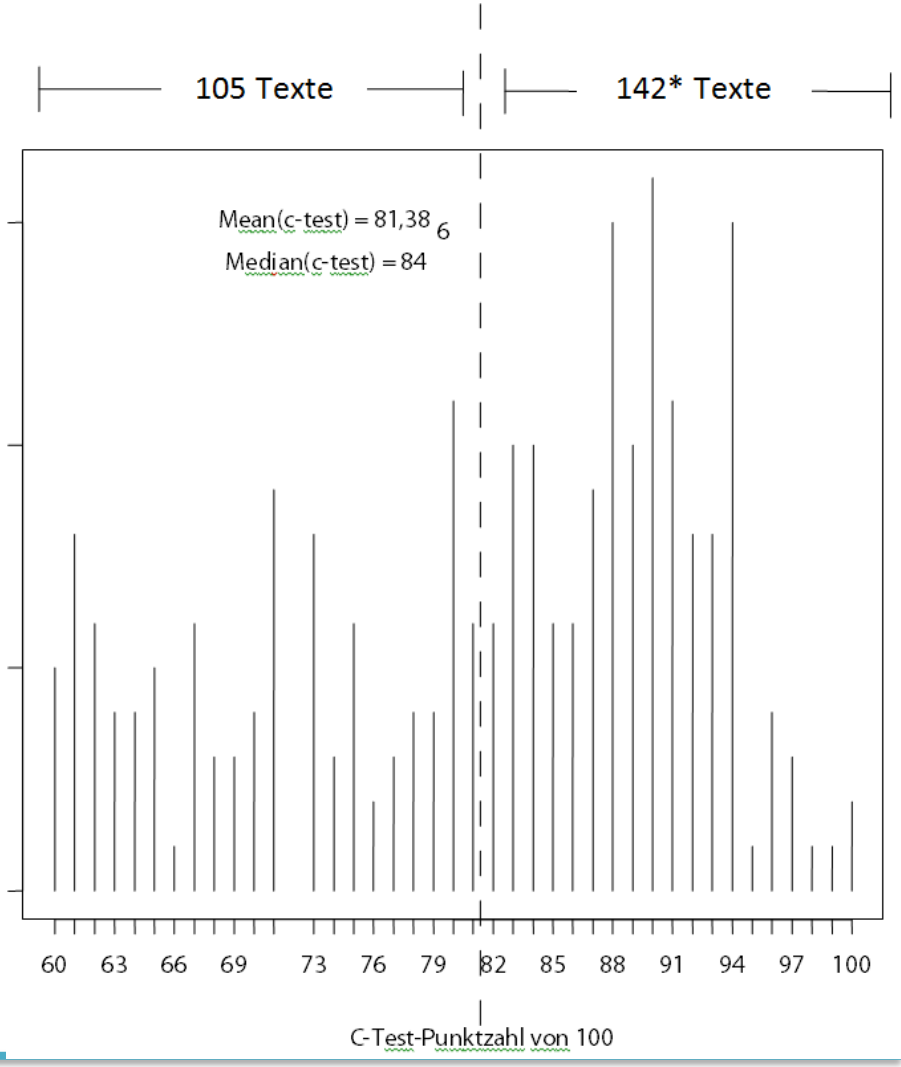
  immer mehr Lernerkorpora auch für andere Sprachen

  Granger/Hung/Petch-Tyson (2002), Cobb (2003), Tono (2003), Myles/Mitchell (2004), Nesselhauf (2004), Tenfjord/Meurer/Hofland (2004), Granger (2008), Lüdeling/Walter (2009) etc.

Falko

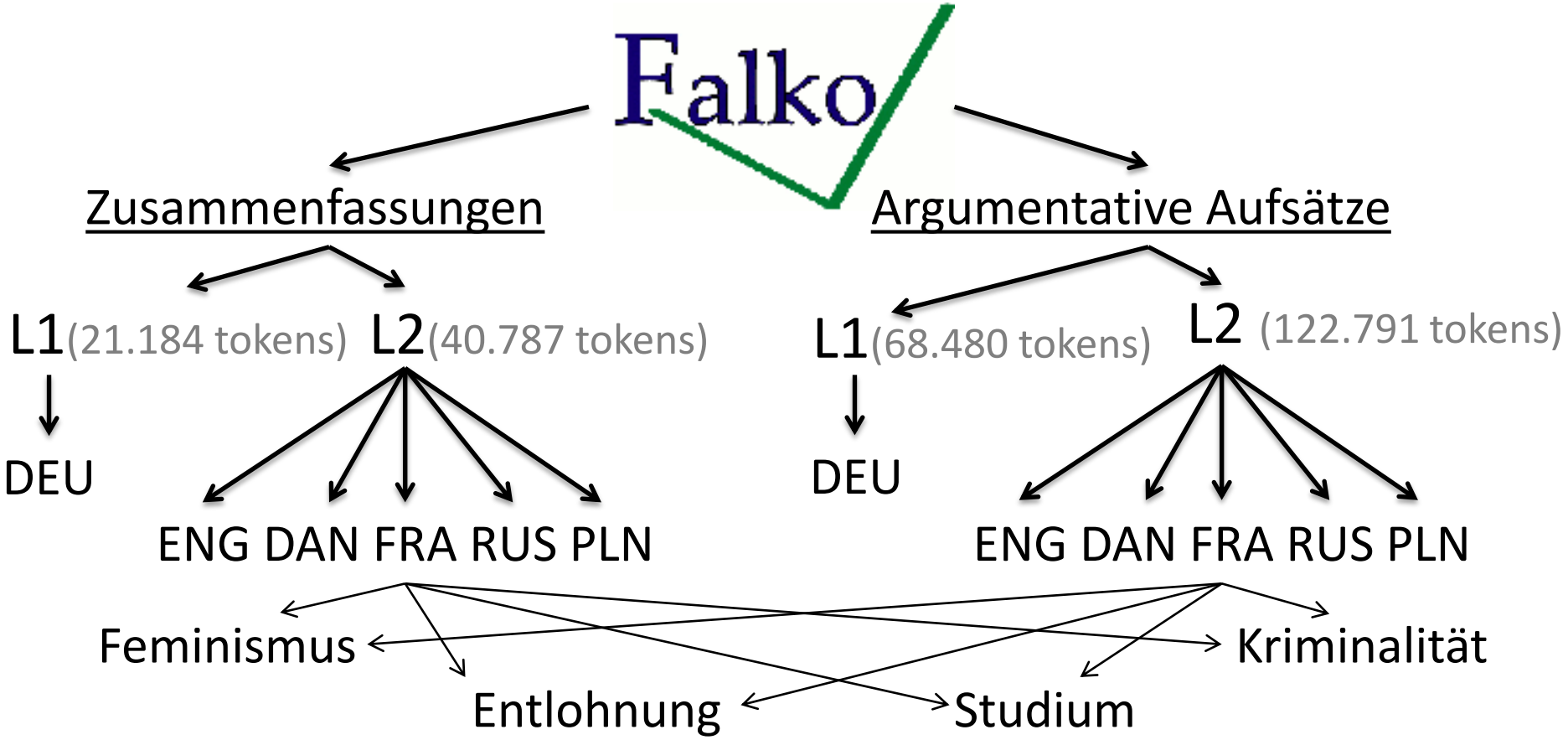
- frei verfügbares Lerner-korpus des Deutschen als Fremdsprache
- Fortgeschrittene Lerner (DHS oder C-Test: HU-Sprachenzentrum)

  ~B1-C1 (CEFR)



Falko

- Subkorpora 2 x 2 x 4 (16)



Metadaten

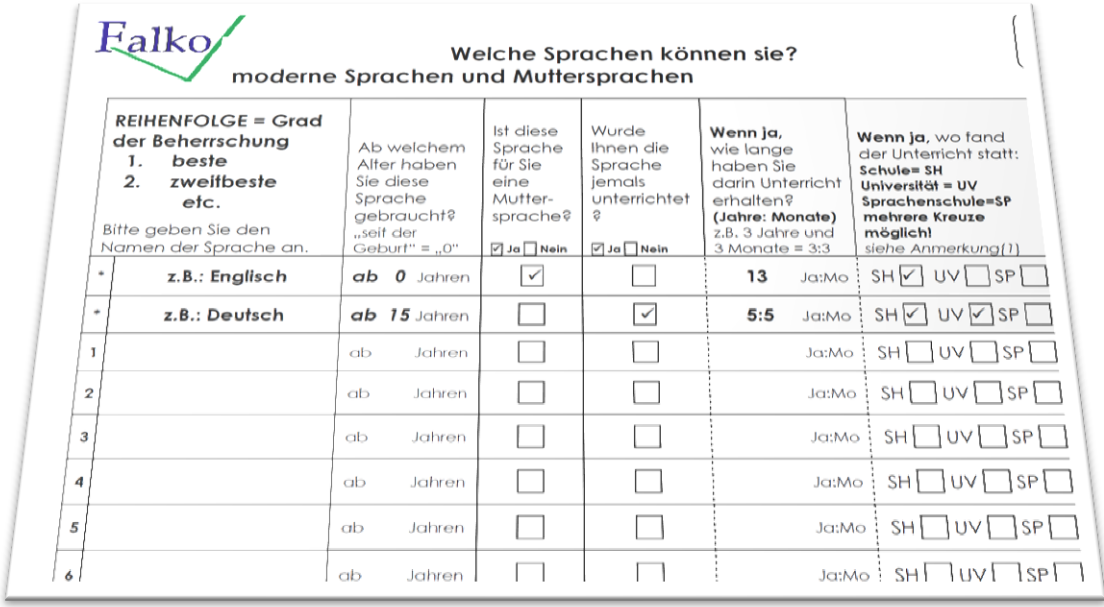
- Erhebungen
  - 90 Minuten
  - keine Hilfe (Internet, Wörterbücher, Spell-checker etc.)
  - handschriftlich(nur Zusammenfassungen) & computergeschrieben
- Dokumentation (Reznicek et al. 2010)

  http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und% 20Annotationen_v1.0.1
- Projektseite:

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

Spracherwerbsbiographie

- Lerner mit 49 Muttersprachen
- Für jede Sprache
  - Dauer des Erwerbs
  - Unterricht
  - Dauer
  - Auslandsaufenthalte
  - Beherrschungsgrad



Falko

Welche Sprachen können sie?
moderne Sprachen und Muttersprachen

	REIHENFOLGE = Grad der Beherrschung 1. ***beste*** 2. ***zweitbeste*** ***etc.*** *Bitte geben Sie den Namen der Sprache an.*	Ab welchem Alter haben Sie diese Sprache gebraucht? „seit der Geburt" = „0"	Ist diese Sprache für Sie eine Mutter-sprache? ☑ Ja ☐ Nein	Wurde Ihnen die Sprache jemals unterrichtet? ☑ Ja ☐ Nein	**Wenn ja,** wie lange haben Sie darin Unterricht erhalten? **(Jahre: Monate)** z.B. 3 Jahre und 3 Monate = 3:3	**Wenn ja,** wo fand der Unterricht statt: **Schule= SH Universität = UV Sprachenschule=SP mehrere Kreuze möglich!** *siehe Anmerkung(1)*
*	**z.B.: Englisch**	***ab 0*** Jahren	☑	☐	**13** Ja:Mo	SH ☑ UV ☐ SP ☐
*	**z.B.: Deutsch**	***ab 15*** Jahren	☐	☑	**5:5** Ja:Mo	SH ☑ UV ☑ SP ☐
1		ab Jahren	☐	☐	Ja:Mo	SH ☐ UV ☐ SP ☐
2		ab Jahren	☐	☐	Ja:Mo	SH ☐ UV ☐ SP ☐
3		ab Jahren	☐	☐	Ja:Mo	SH ☐ UV ☐ SP ☐
4		ab Jahren	☐	☐	Ja:Mo	SH ☐ UV ☐ SP ☐
5		ab Jahren	☐	☐	Ja:Mo	SH ☐ UV ☐ SP ☐
6		ab Jahren	☐	☐	Ja:Mo	SH ☐ UV ☐ SP ☐

WHIG – What is hard in German?

- Weitere Erhebungen englischer Deutschlernern
- ~120.000 tokens
- Datenerhebung und Aufbereitung nach Falko - Standard

Arts & Humanities Research Council

HU-Berlin
Anke Lüdeling
Marc Reznicek
Bangor University
Astrid Ensslin
Cedric Krummes

Annotation von Lernersprache

- Viele Lernerkorpora sind nicht annotiert
- Manche haben Fehlerannotationen
- Wenige (vor allem junge) Korpusprojekte integrieren weitere Annotationsebenen

  (ALESKO, KOBALT, BEMATAC, DALEKO, KanDel)

- Falko: Standoff-Format (jede Annotation wird unabhängig gespeichert)

→ Neue Annotationsebenen können unbeschränkt hinzugefügt werden. (Lüdeling et al. 2005)

Falko

automatische Annotation –Wortarten/Lemma

- Meiste Lernerkorpusforschung konzentrierte sich auf Oberflächenformen (Möllering 2004, Vyatkina 2007 etc.)
- Aber interessanter z.B.:
  - Welche Wortarten und Wortartenketten werden vermieden? (Oberflächensyntax: Borin/Prütz 2004)
- Automatische Wortartentagger werden immer besser: Zeitungssprache ~98% (Kübler et al. 2010)

→automatische Wortarten(POS)- und Lemma-Annotation

Fehlerannotation in Falko

An der anderen Seite, wenn da kein Feminismus wäre, stünden wir noch nur in der Küche und köchten wir. (fkb034_2008_07)

Fehlerannotationen beziehen sich immer auf eine (zumindest implizite) korrekte Entsprechung der Lerneräußerung. → Zielhypothese

Wie würden Sie korrigieren?

Zielhypothesen

An der anderen Seite, wenn da kein Feminismus wäre, stünden wir noch nur in der Küche und köchten wir.
(fkb034_2008_07)

Falko: explizite Zielhypothese

- oft konkurrierende Korrekturen möglich

ZH1: *An der anderen Seite, wenn da kein Feminismus wäre, stünden wir nur noch in der Küche und kochten.*

ZH2: *Andererseits stünden wir, wenn es keinen Feminismus gäbe, nur noch in der Küche und kochten.*

Zielhypothesen in Falko

ZH1: satzbasiert, nah an Lernerstruktur:
Orthographie, Morphosyntax

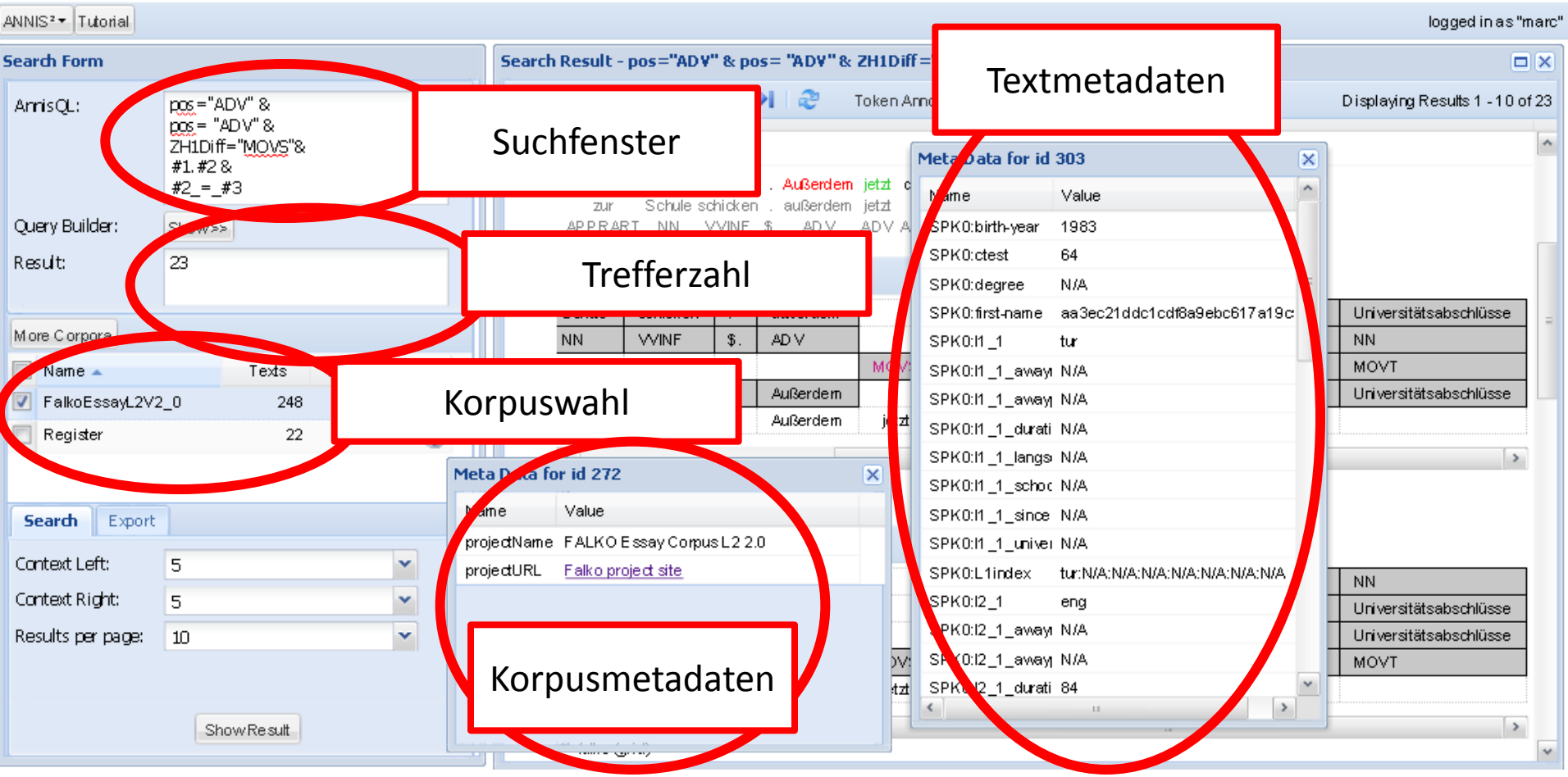
ZH2: text-basiert, nah an Lernerintention:
Semantik, Pragmatik, Stilistik

- Die Unterschiede zwischen ZH und Lernertext wird automatisch mit **edit tags** (CHAnge, INSert, DELete etc.) markiert
- Alle ZHs werden mit POS-und Lemma annotiert (TreeTagger, Schmid 1994)
- weitere manuelle Fehlertags für einige Phänomene

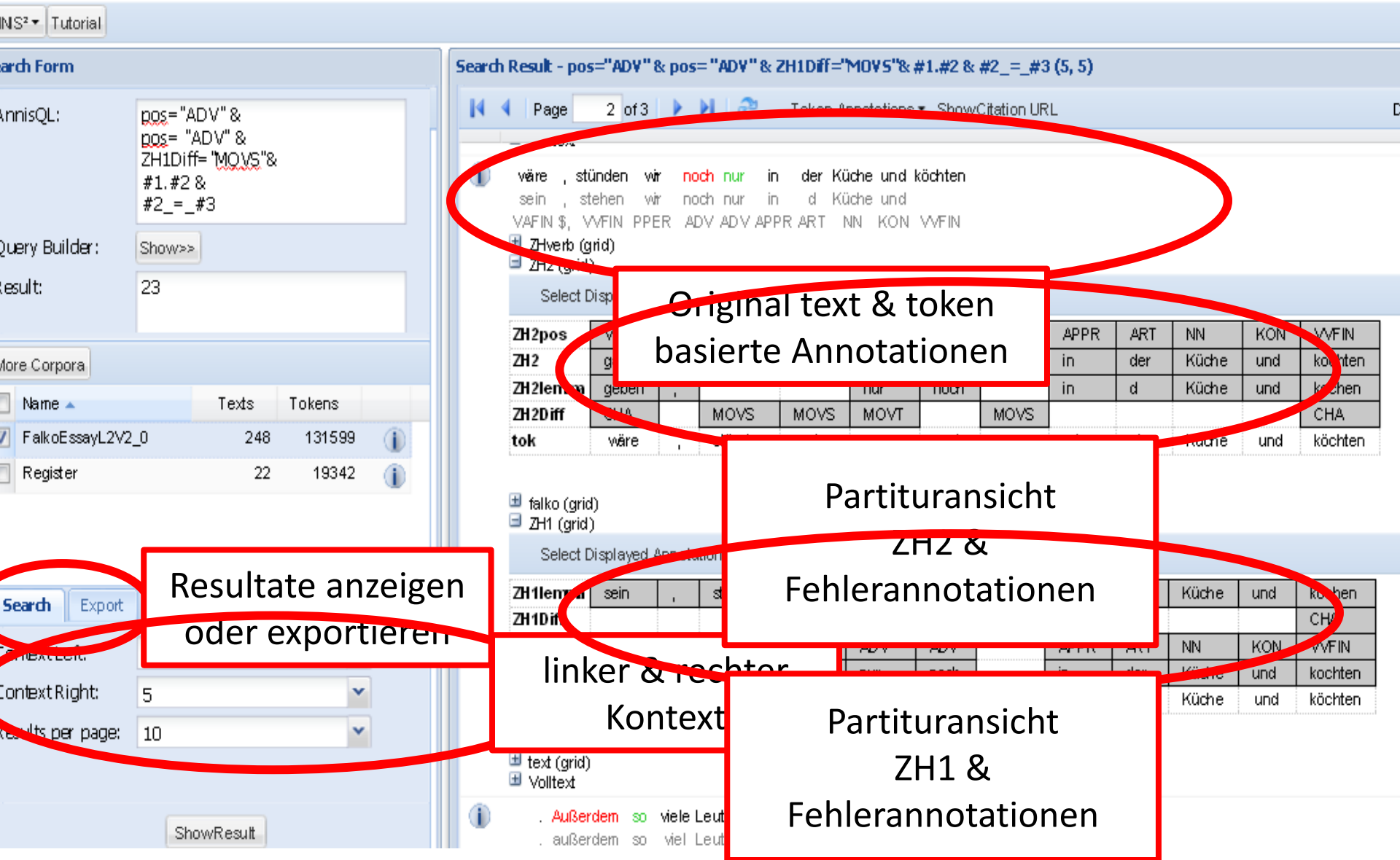
LT	TH1	TH1Diff	TH2	TH2Diff
An	Auf	CHA	Andererseits	MERGE
der	der			
anderen	anderen			
Seite	Seite			
,	,		,	
			stünden	MOVT
			wir	MOVT
			,	INS
wenn	wenn		wenn	
da	da			DEL
			es	INS
kein	kein		keinen	CHA
Feminismus	Feminismus		Feminismus	
wäre	wäre		gäbe	CHA
,	,		,	
stünden	stünden			MOVS
wir	wir			MOVS
	nur	MOVT	nur	MOVT
noch	noch		noch	
nur		MOVS		MOVS
in	in		in	

LT	pos	lemma	TH1	TH1Diff	TH1pos	TH1posDiff
An	APPR	an	**Auf**	CHA	APPR	
der	ART	d			ART	
anderen	ADJA	andere	anderen		ADJA	
Seite	NN	Seite	Seite		NN	
,	$,	,	,		$,	
wenn	KOUS	wenn	wenn		KOUS	
da	PAV	da	da		PAV	
kein	PIAT	kein	kein		PIAT	
Feminismus	NN	Feminismus	Feminismus		NN	
wäre	VAFIN	sein	wäre		VAFIN	
,	$,	,	,		$,	
stünden	VVFIN	stehen	stünden		VVFIN	
wir	PPER	wir	wir		PPER	
			nur	MOVT	ADV	MOVT
noch	ADV	noch	noch		ADV	
nur	ADV	nur		MOVS		MOVS
in	APPR	in	in		APPR	

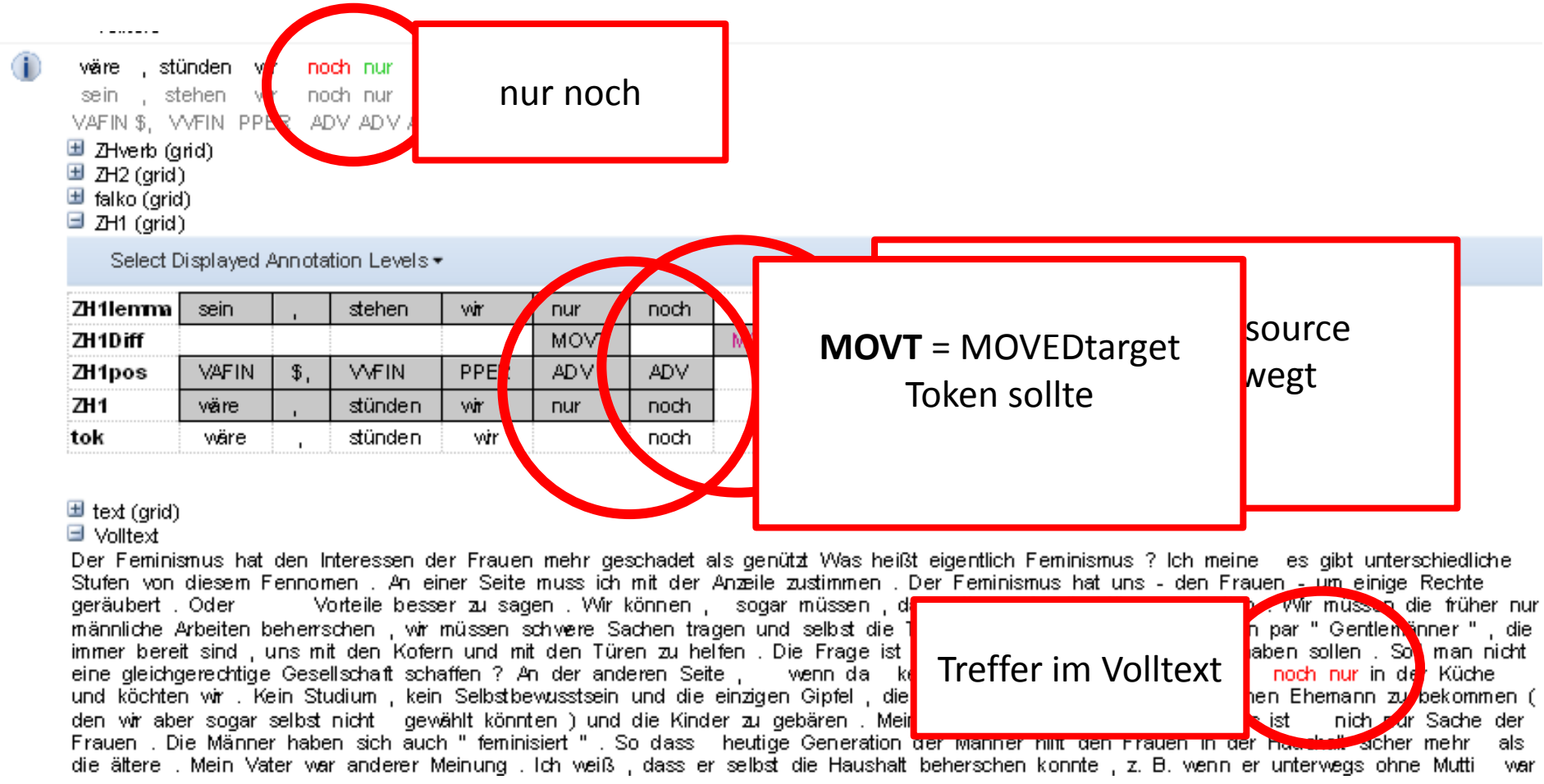
Exkurs: Suche in ANNIS



Exkurs: Suche in ANNIS



Exkurs: Suche in ANNIS



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

Mögliche Studie

- Machen romanische Deutschlerner signifikant weniger Artikelfehler als germanische?

→Finde alle Artikel in der ZH2, die mit "INS" markiert sind

→Vergleich die Ergebnisse in Texten von spanischen und italienischen Lernern mit denen in Texten von dänischen und afrikaans Lernern.

Contrastive Interlanguage Analysis

Falko

- Auffinden struktureller Lernerschwierigkeiten im L2-Deutsch
- Strukturelle Schwierigkeiten sind solche, die
  - unabhängig sind von der Lerner-L1
  - daher abhängig sind von der Grammatik der Zielsprache

DFG

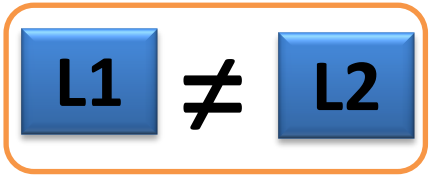
Arts & Humanities Research Council

HU-Berlin
Anke Lüdeling
Marc Reznicek
Bangor University
Astrid Ensslin
Cedric Krummes

Wie finden wir komplizierte Strukturen?

Falko

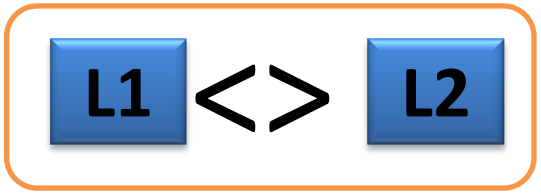
- L1-korrelierte Strukturen (Transfer)



- Von den Lernern als kompliziert eingestufte Strukturen

- Fehlerhafte Strukturen

- Unterrepräsentierte Strukturen



Wie finden wir komplizierte Strukturen?

Falko

- L1-korrelierte Strukturen (Transfer)

  grammatische Analysen

  → nur sehr beschränkte Vorhersagekraft

- von den Lernern als kompliziert eingestufte Strukturen

  Lernerintuition

  → unsystematisch, lehrformabhängig

- fehlerhafte Strukturen

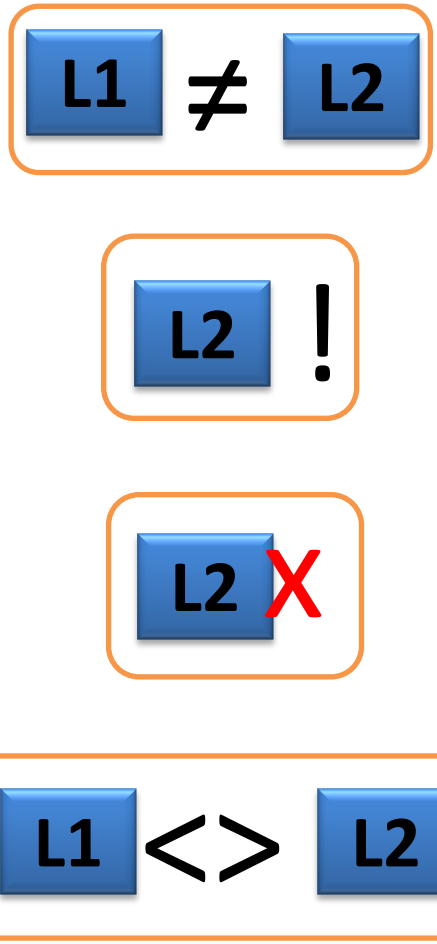
  Lehrerintuition

  → unsystematisch, normenabhängig

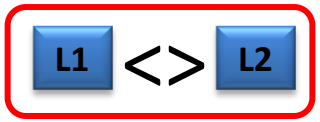
- unterrepräsentierte Strukturen

  Korpusanalyse: Frequenzvergleiche

  → **Overuse/Underuse**



Mindergebrauch (Underuse)



Falko

- Frequenzen von Strukturen im L2-Subkorpus werden verglichen mit Strukturen im L1-Subkorpus
- Overuse & underuse sind definiert als (statistisch signifikante) Unterschiede zwischen zwei Varietäten
- Eine Kategorie kann mindergebraucht sein, weil:
  - die Lerner sie nicht kennen
  - die Lerner sie zwar kennen, aber (unbewusst?!?) vermeiden

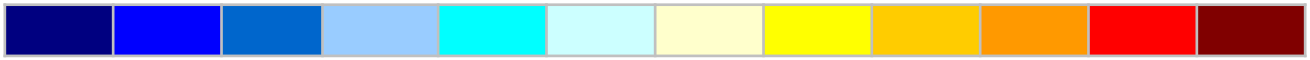
→ Diagnostik für das Auffinden schwieriger Strukturen

Overuse/Underuse-Visualisierung

- underuse: kalte Farben
- overuse: warme Farben

- Intensität der Farbe signalisiert Stärke des overuse/underuse

Underuse **Overuse**



Excel –AddIn (Amir Zeldes) verfügbar unter:
http://korpling.german.hu-berlin.de/~amir/uoaddin.htm

Overuse/Underuse-Visualisierung lexikalische Einheiten

Falko

lemma	tot_norm	deu	dan	eng	fra	pln	rus
in	0.013188	0.012261	0.014041	0.014247	0.015272	0.012135	0.009534
es	0.010897	0.011945	0.010900	0.011379	0.013347	0.008163	0.012385
sie	0.010618	0.008193	0.010643	0.008835	0.010909	0.006067	0.005613
man	0.010164	0.007900	0.012438	0.008742	0.009754	0.006950	0.007306
dass	0.009522	0.007404	0.012823	0.008789	0.009625	0.008880	0.009890
von	0.007982	0.007122	0.007309	0.006846	0.007315	0.010259	0.007930
auch	0.007028	0.008362	0.008527	0.005828	0.005775	0.005461	0.004455
für	0.006683	0.007201	0.006091	0.007216	0.006802	0.005736	0.004188
sind	0.006465	0.004271	0.008976	0.007308	0.006930	0.004964	0.005346
sich	**0.006309**	**0.011697**	**0.006283**	**0.006291**	**0.006930**	**0.007170**	**0.005435**
ich	0.006262	0.003877	0.013272	0.005366	0.003465	0.001434	0.001426
aber	0.006048	0.003347	0.007309	0.006245	0.007315	0.003365	0.003831

sich ist in allen L1-Subkorpora mindergebraucht

Stuttgart-Tübingen-Tagset (STTS)

(Schiller et al. 1995)

ADJektiv

- ADJA
- ADJD

ADVerb

Nomen

- NN
- NE

Pronomen

- PDS
- PDAT
- PIS
- PIAT
- PIDAT
- PPER
- PPOSS
- PPOSAT
- PRELS
- PRELAT
- PRF
- PWS
- PWAT
- PWAV

Verb

- VVFIN
- VVIMP
- VVINF
- VVIZU
- VVPP
- VAFIN
- VAIMP
- VAINF
- VAPP
- VMFIN
- VMINF
- VMPP

ParTiKel

- PTKZU
- PTKNEG
- PTKVZ
- PTKANT
- PTKA

KOnjunktion

- KOUI
- KOUS
- KON
- KOKOM

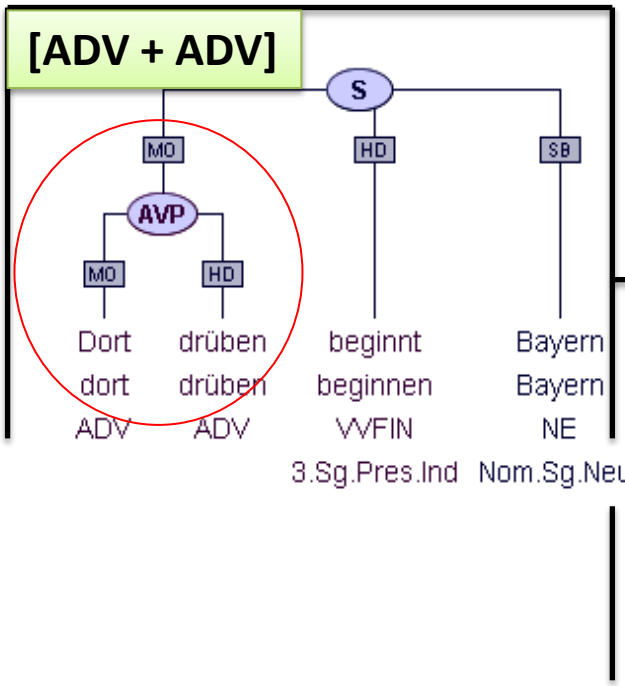
Overuse/Underuse-Visualisierung POS-Bigramme

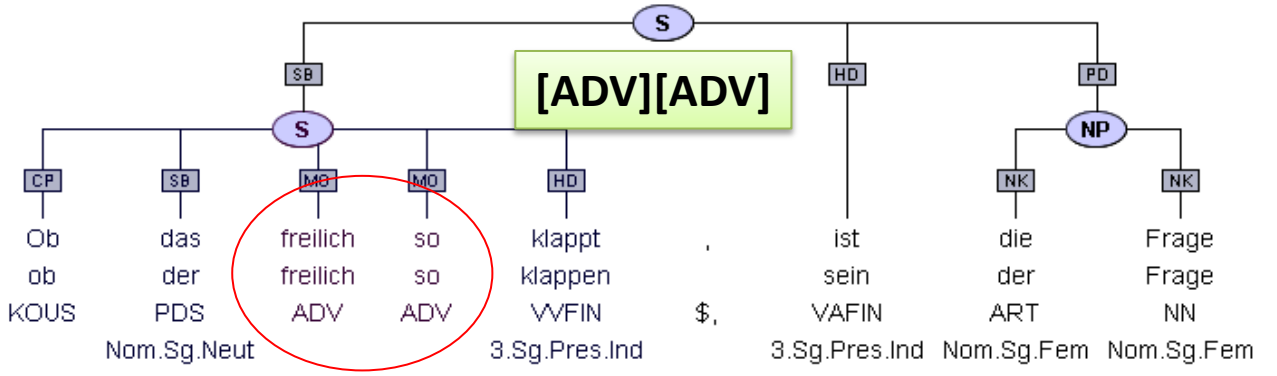
bigram	tot_norm	de	da	en	fr	pl	ru
$.-PPER	0.042384	0.005297	0.009748	0.007963	0.006166	0.005801	0.007409
VVFIN-$,	0.042131	0.006457	0.00776	0.006343	0.006937	0.006243	0.008391
PPOSAT-NN	0.041739	0.008058	0.007247	0.007269	0.007066	0.006298	0.005802
ADV-ADV	**0.041604**	**0.012858**	**0.010518**	**0.006111**	**0.006166**	**0.003094**	**0.002856**
ADV-APPR	0.039742	0.009117	0.008016	0.005324	0.007837	0.004807	0.004642
PDAT-NN	0.03956	0.005409	0.004233	0.005509	0.007837	0.007735	0.008837
ADV-ART	0.037125	0.007629	0.006349	0.006898	0.005653	0.006133	0.004463

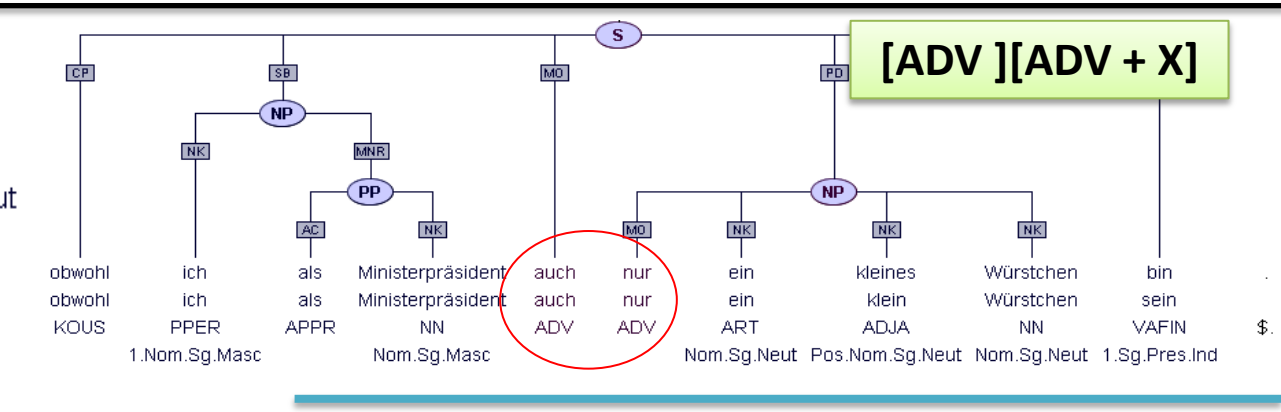
Adverbketten sind in allen L1-Subkorpora mindergebraucht

Fazit

- Adverbketten werden von allen Lernern unabhängig von der Muttersprache vermieden.







Fazit

- Adverbketten werden von allen Lernern unabhängig von der Muttersprache vermieden.

[ADV + ADV] **[ADV][ADV]** **[ADV][ADV + X]**

→Strukturen mit sehr variabler Tiefenstruktur werden besonders häufig vermieden.

→ Werden die Formen vermieden oder die Funktionen?

Hypothese: Modifikation wird generell vermieden.

CIA: syntaktische Funktion

Falko

Untersuchung zu Underuse feinerer Adverbklassen (Hirschmann erscheint) zeigt einen Unterschied zwischen adverbialen Funktionen.

→Wie sieht der Vergleich rein syntaktischer Funktionen unabhängig von der Füllerkategorie aus?

→Methode: tiefe syntaktische Annotation der Lernerdaten

Falko – syntaktische Vorverarbeitung

- Direkte Verarbeitung von Lernersprache ist problematisch

  - Äußerungen gehorchen nicht der beschreibenden Grammatik

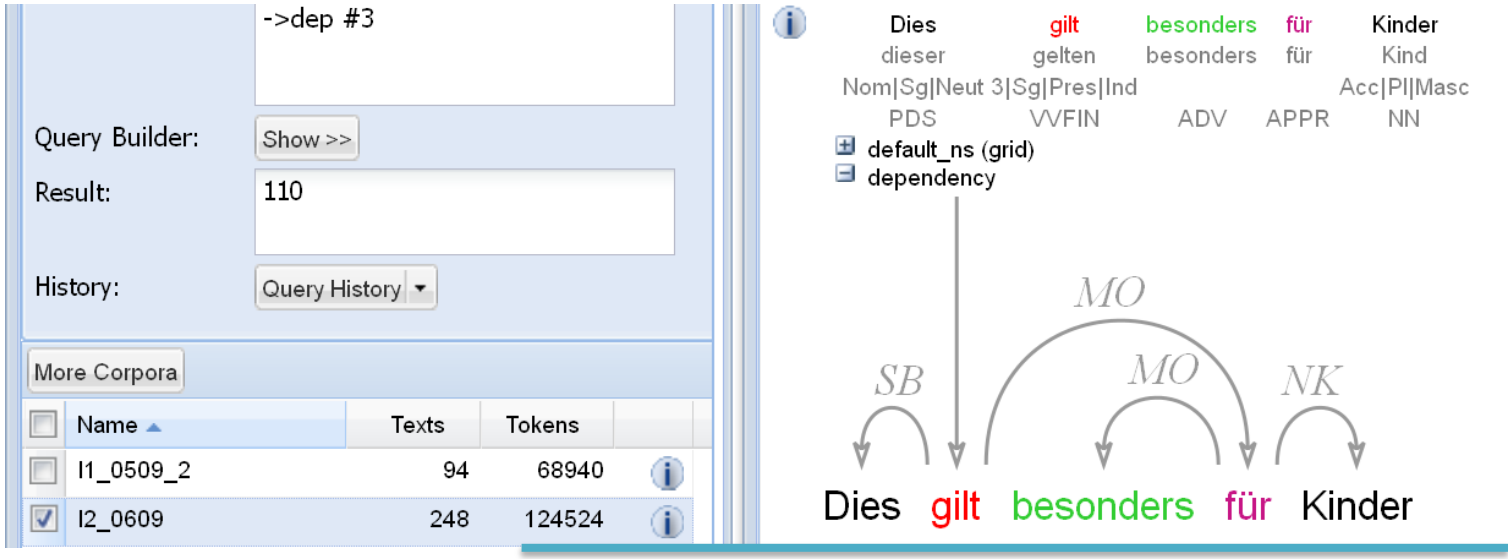
  (siehe Hirschmann et al. 2009, Díaz-Negrillo et al. 2010)

→ZH1 als Grundlage für das Parsing

→Rückübetragung der Annotationen auf den Lernertext (ZH0)

Falko - Dependenzbäume

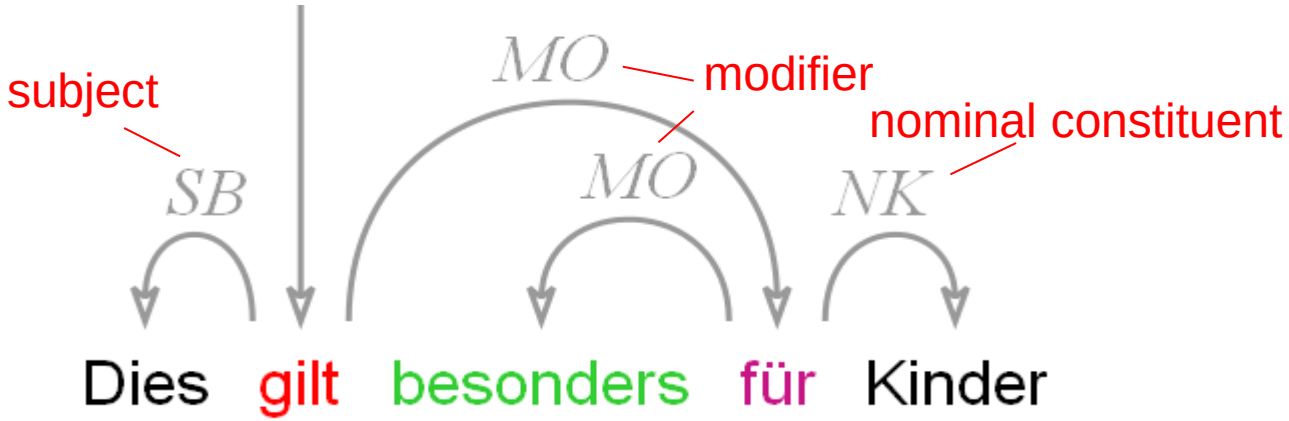
- Dependenzparser (Bernd Bohnet 2010; Syntactic Analyser)
- Trainingsdaten: TiGer dependency bank
  - (abgeleitet aus ~50000 trees of the TiGer treebank)
- Ergebnis: sehr genaue Dependenzbäume
  - mit syntaktischen Funktionen



Syntaxschema (Kurzformat)

Falko

- Jedes Wort ist mit den von ihm abhängigen Worten verbunden.
- Pfeile zeigen zu hierarchisch niedrigeren Dependenten.
- Jeder Pfeil (Dependenz) trägt eine Funktionsbezeichnung.

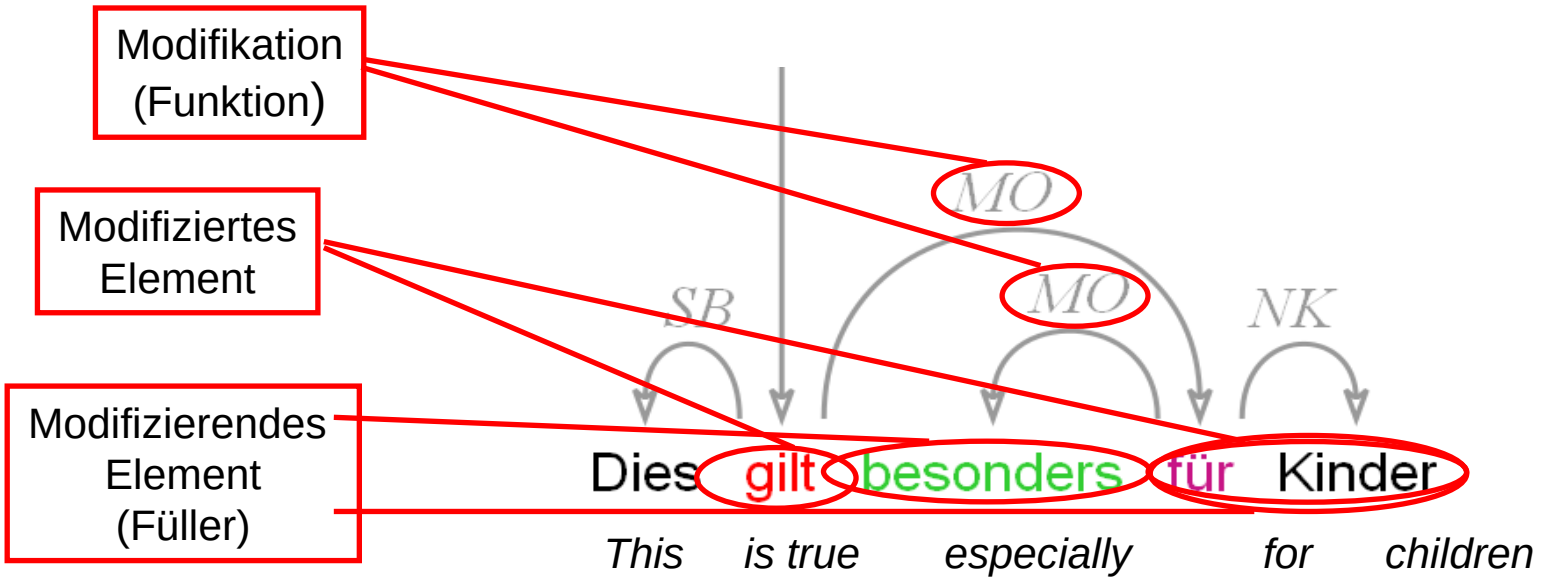


Suche nach Modifikationen

Falko

Verschiedene Aspekte des Problems

- Wird die syntaktische Funktion vermieden?
- Werden bestimmte Ziele der Modifikation vermieden?
- Werden bestimmte modifizierende Kategorien vermieden?



Underuse/Overuse von Funktionen

Falko

label	de	da	en	fr	ru	usb
NK	0,264067	0,278546	0,284881	0,303271	0,29552	0,295136
HD	0,156192					
MO	0,141968	0,12789	0,113704	0,110112	0,112513	0,108707
SB	0,07398	0,078506			0,078852	0,085512
CJ	0,059604	0,053397	0,056411	0,050632		0,072183
AC	0,057051					0,04916
OC	0,050335					0,040679
OA	0,044213					
CD	0,026549			0,022156		
CP	0,017653	0,021732	0,020325			
PD	0,014435		0,015943		0,016947	0,018002
NG	0,011065					
MNR	0,010995	0,013707	0,013429	0,013383		
RC	0,010051				0,006268	0,005366

MO (modification) ist signifikant mindergebraucht unabhängig von L1

Modifikationen in Falko

- Alle Kategorien werden oft modifiziert in L1- und L2-Daten
- aber ***alle* syntaktischen Modifikationsrelationen** zeigen einen Underuse
- Adverbmodifizierer zeigen den stärksten Underuse

- Das gilt unabhängig von der L1 der Lerner

Falko

Produktivität bei Lernern

- Viele Studien zeigen, dass Muttersprachler oft Worte neubilden
  - sind sich (implizit) Wortbildungsregeln bewusst
  - wahrscheinlich aus Input abgeleitet

  (Baayen 1992, 2001 etc., Plag 1998, Bauer 2001, Lüdeling & Evert 2005, Kiss 2007, Zeldes to appear etc.)

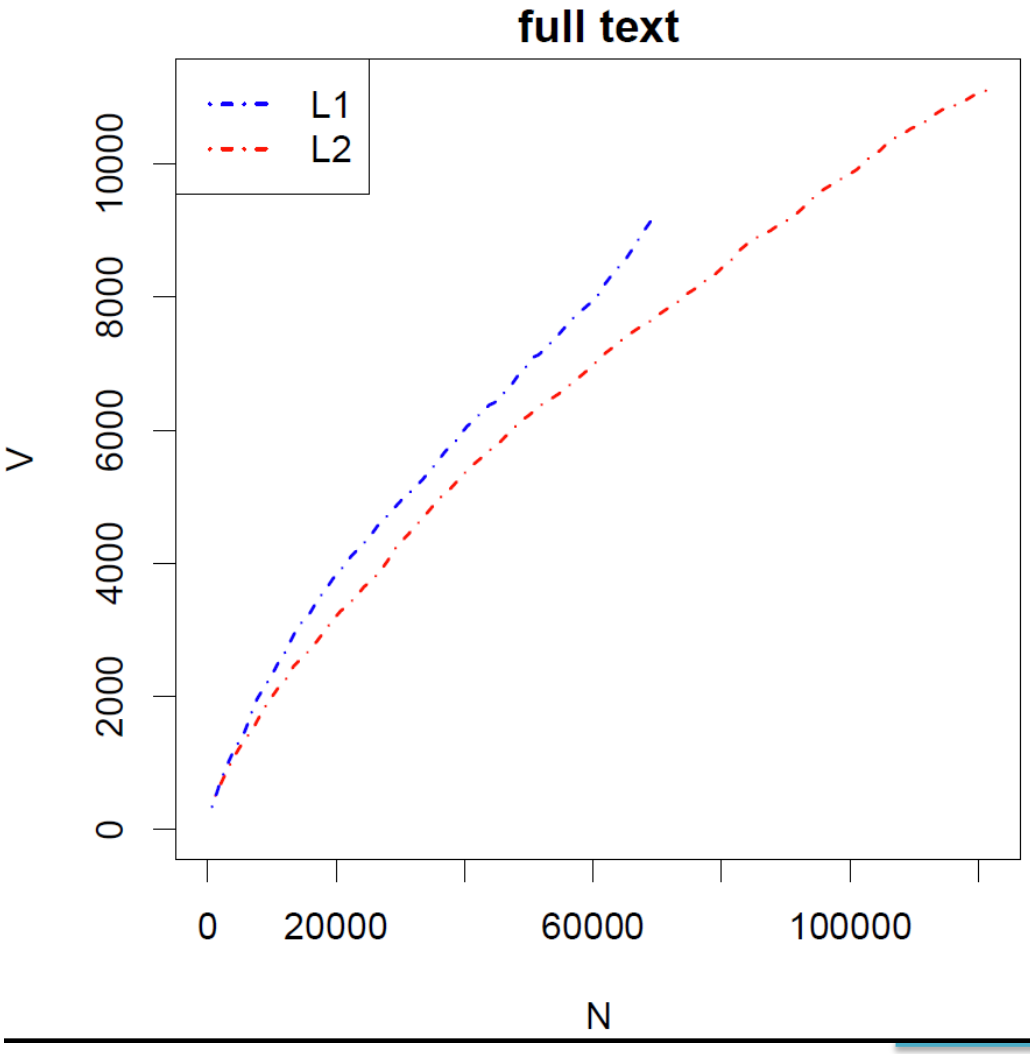
- Gilt das auch für Lerner?
- Wie können Lernerkorpora uns bei der Lösung helfen?

Vokabelwachstumskurven

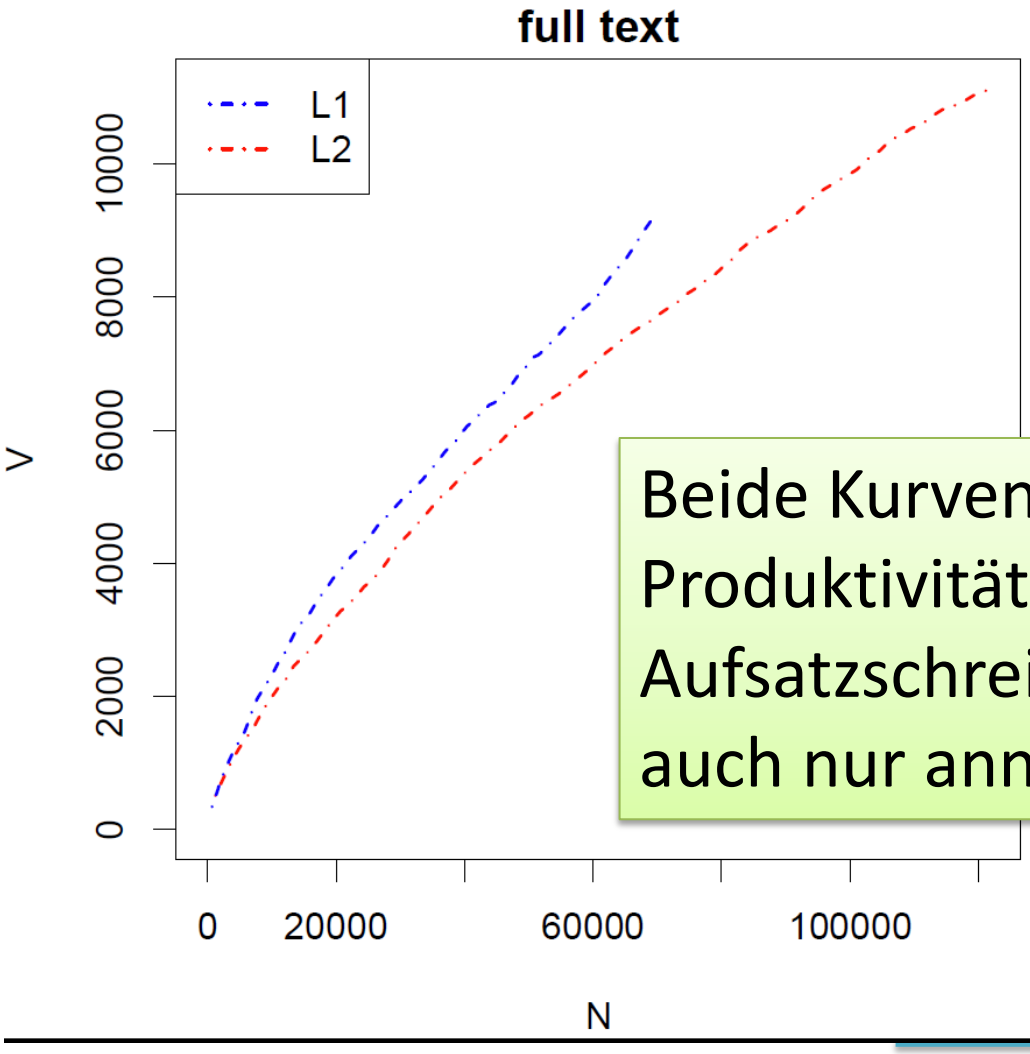
- Viele korpusbasierte Maße für morphologische Produktivität (Baayen 2001, 2009 etc.)
- Kurz: Vokabelwachstumskurven
  - **N** = Token vs. **V** = Typen

- Zeigen ob die Stichprobe (Korpus) das gesamte Vokabular ausreizt, oder noch viele (mögliche) Worte in der Gesamtpopulation gibt.

Vokabelwachstumskurve für Falko L1 und L2



Vokabelwachstumskurve für Falko L1 und L2



Beide Kurven zeigen eine hohe Produktivität – weder L1 noch L2 Aufsatzschreiber kommen erreichen auch nur annähernd die Vokabelgrenze

Komplexe Verben im Deutschen

Falko

Präfixverben

- *ver•kaufen*
- *[…] dass Peter Schokolade verkauft.*
- *Peter verkauft Schokolade.*
- Infinitive: *zu verkaufen*
- Partizip: *verkauft*

Partikelverben

- *auf•essen*
- *[…] dass Peter die Schokolade aufisst.*
- *Peter isst die Schokolade auf.*
- Infinitive: *aufzuessen*
- Partizip: *aufgegessen*

ZHverb: komplexe Verben in Falko

- Grundlage ist ZHverb (konzentriert sich auf komplexe Verben)

- manuelle Annotation
  - Typ
    - Partikel vs. Präfix
  - Lemma
  - Fehlerklasse
    - Orthographie, Flexion, Argumentstruktur, Trennbarkeit etc.
  - Form
    - finit, infinit, Partizip, getrennt etc.

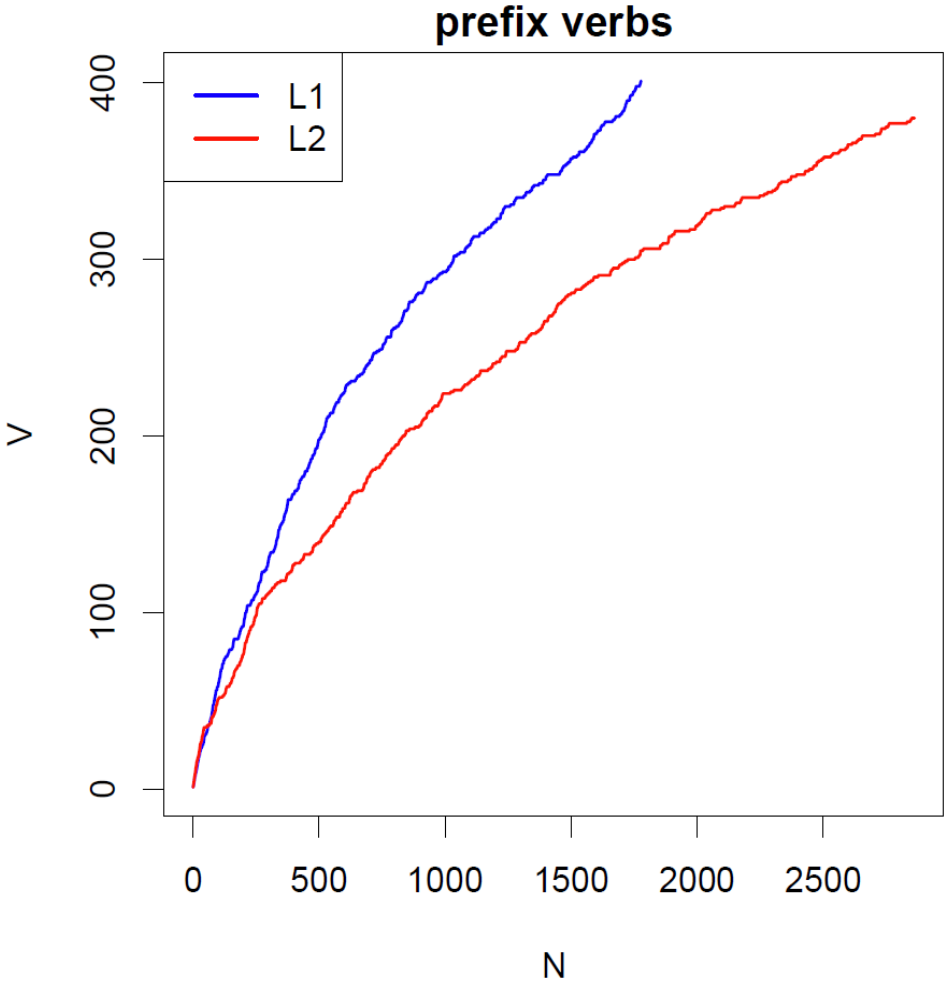
Exkurs: Fehler

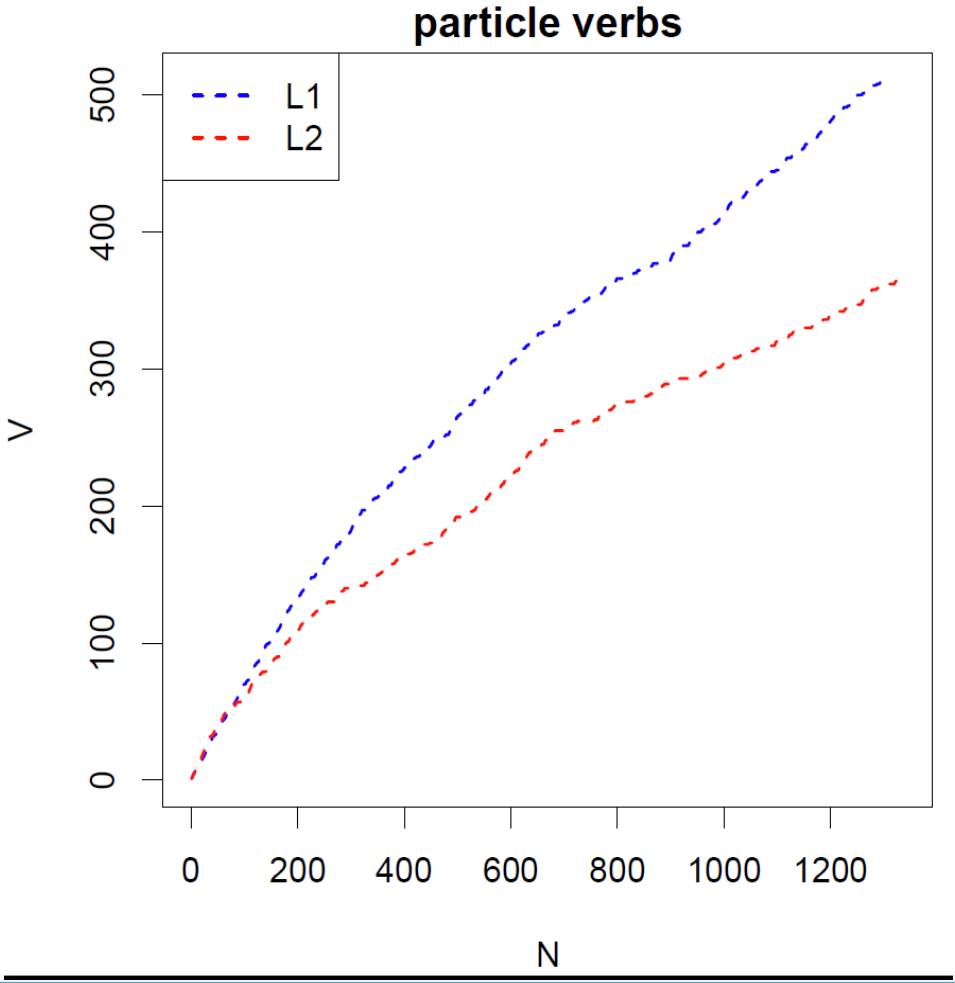
- Alle Fehlerdifferenzen sind signifikant

→Lerner machen signifikant mehr strukturelle und semantische Fehler

→Muttersprachler machen signifikant mehr Orthographiefehler

Vokabelwachstum





Komplexe Verben

- Komplexe Verben sind für L1 und L2 produktiv
- L1 benutzen mehr produktive Verben

- Kennen L1-Sprecher mehr komplexe Verben?
- Wissen L1-Sprecher besser wie man komplexe Verben bildet?

→ a) Analyse produktiver Muster

→ b) Analyse klarer Fälle von Wortneubildung

Komplexe Verben

Falko

a) Produktives Muster : hinein + Verb

hineinbewegen – move into
hineingehen – go into
hineintreten – step into

hineinbringen – bring into
hineinfallen – fall into
hineingeraten – get into
hineinpassen – fit into
hineinversetzen – put yourself in someones position
hineinwachsen – grow into
hineinziehen – pull into

- für viele mögliche Muster zeigt L1 mehr Typen
- L1 verwendet mehr Muster

!!! Nicht klar, welche davon produktiv gebildet!!!

Komplexe Verben

Falko

b) 'unbekannte' Worte

L2

34 'neue' Form (uns unbekannt, Token)

… was eine sozialere oder gerechtigere Entlohnung benachträchtigt. (hu_005_2006_09)

274 neue Lesarten für bekannte Formen (Token)

Viele Leute sich bewundern, ob … (hu_006_2006_10)

L1

23 'neue' Formen (uns unbekannte, Token)

Sie vollrichten in ihrer Arbeitszeit fast doppelt so viel … (dhw_031_2007-06)

34 neue Lesarten für bekannte Formen (Token)

Die Regierung kann dieses große Netz der Machenschaften meist nicht entspinnen. (dew_03_2007_09)

Komplexe Verben

- Lerner bilden 'neue' Wörter signifikant häufiger als benutzen 'neue' Lesarten häufiger als Muttersprachler

- machen Fehler beim Anwenden der Muster

Zusammenfassung

Tief annotierte Lernerkorpora erlauben die Beantwortung neuer Forschungsfragen.

- Abweichungen in der Frequenz bestimmter Muster auf unterschiedlichen Abstraktionsebenen
  - Lemmata, Wortarten, Wortartenketten, syntaktische Funktionen
- Unterschiede in der graduellen Beschreibung von Produktivität Produktionsmustern
- Auffinden von in den Lernertexten NICHT vorhandenen Elementen (unterlassene Artikel)

Zusammenfassung

Sowohl für Fehleranalysen als auch für rein kontrastive Analysen ist eine explizite Zielhypothese nötig

Fehler: Grundlage für Abweichungen

Kontrast: Beschreibungsgrundlage für zielsprachliche Annotationsschemata

Falko

Einladung

There is no data like more data.

Wir suchen fleißige Kollaborateure bei der Vergrößerung des Korpus.

Jeder ist eingeladen, mitzumachen

Danke

Literatur:

Borin, Lars; Prütz, Klas (2004): New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In: Guy Aston, Silvia Bernardini, Dominic Stewart (Eds.): Corpora and language learners. Amsterdam, Philadelphia: John Benjamins (Studies in corpus linguistics, 17), 67–87.

Granger, Sylviane (2008): Learner Corpora. In: Anke Lüdeling, Merja Kytö (Eds.): Corpus linguistics. Berlin, New York: Mouton de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science, 29,1), 259–275.

Kübler, Sandra; Scheutz, Matthias; Baucom, Eric; Israel, Ross (2010): Adding Context Information to Part of Speech Tagging for Dialogues. In: Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories (TLT 10). Tartu, Estonia.

Schiller, Anne; Teufel, Simone; Thielen, Christine (1995): The Stuttgart-Tübingen Tagset (STTS).