Aufbau und Nutzung des fehlerannotierten Lernerkorpus für Deutsch als Fremdsprache FALKO

Marc Reznicek

22.06.12

Universität Erlangen

Interdisziplinäres Zentrum für Lexikografie, Valenz- und Kollokationsforschung

Mit Folien des gesamten Korpuslinguistikteams der HU-Berlin

Übersicht

- Forschungsfragen
- Lernerkorpora
- Falko - Design
- Annotation
  - Automatische Vorverarbeitung
  - Fehler und Zielhypothesen
- Suchen in ANNIS2

Forschungsfragen

- 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?
- Haben Lerner Schwierigkeiten bei der Registerwahl?

Methoden

Falko

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

→ Welche Fehler sind lernertypisch?

→ Welchen Einfluss hat dabei die jeweilige Muttersprache?

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

→ Betrachtung der Lernersprache unabhängig von Norm

→ Quantitative Unterschiede zwischen Lernern und Muttersprachlern aufdecken

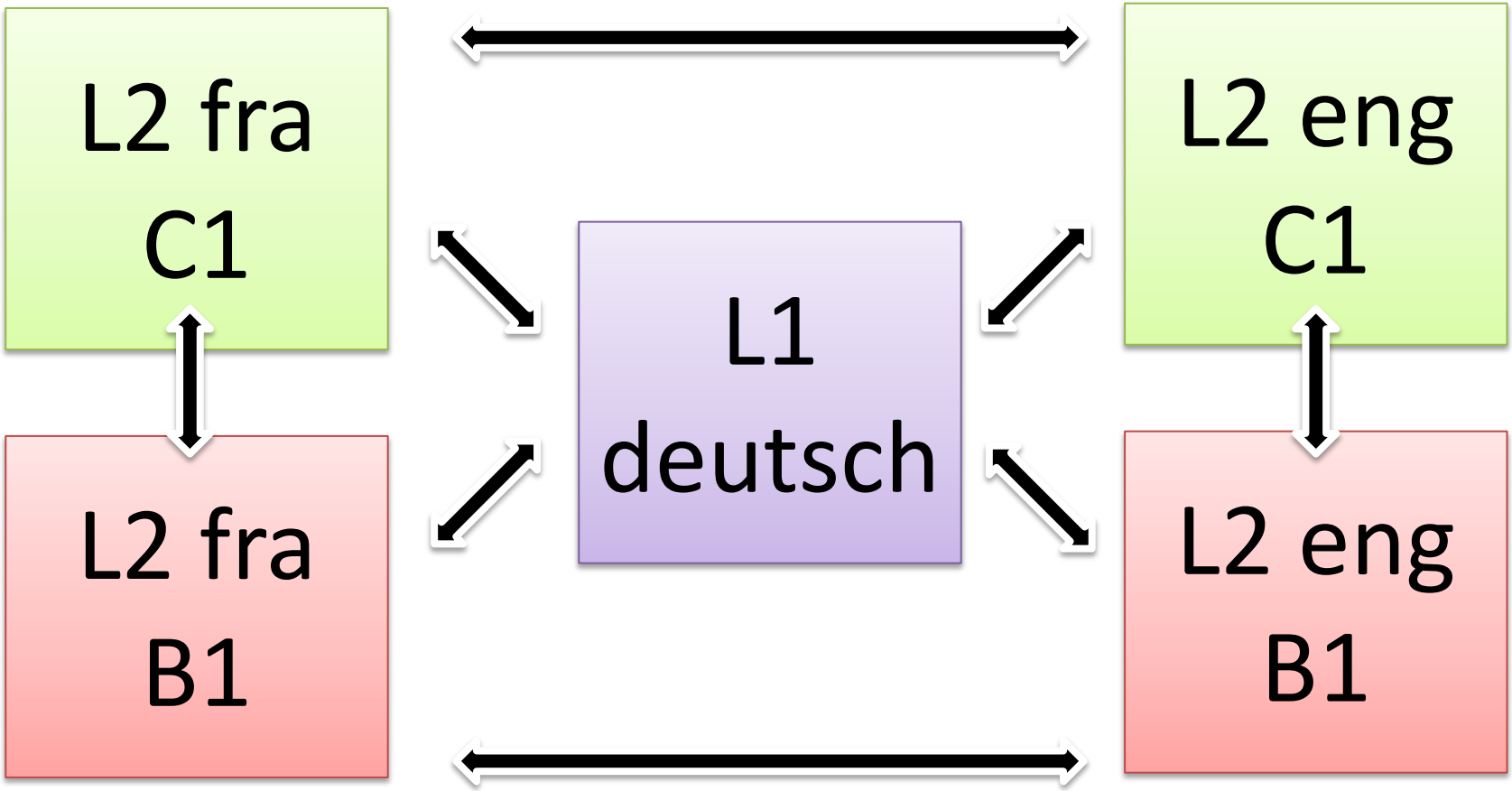
→ L2-Texte unterschiedlicher Muttersprachler kontrastieren

→ Kann für alle möglichen Phänomene untersucht werden.

CIA

Falko

Wie unterscheiden sich die Frequenzen unterschiedlicher Strukturen zwischen diesen Gruppen?



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

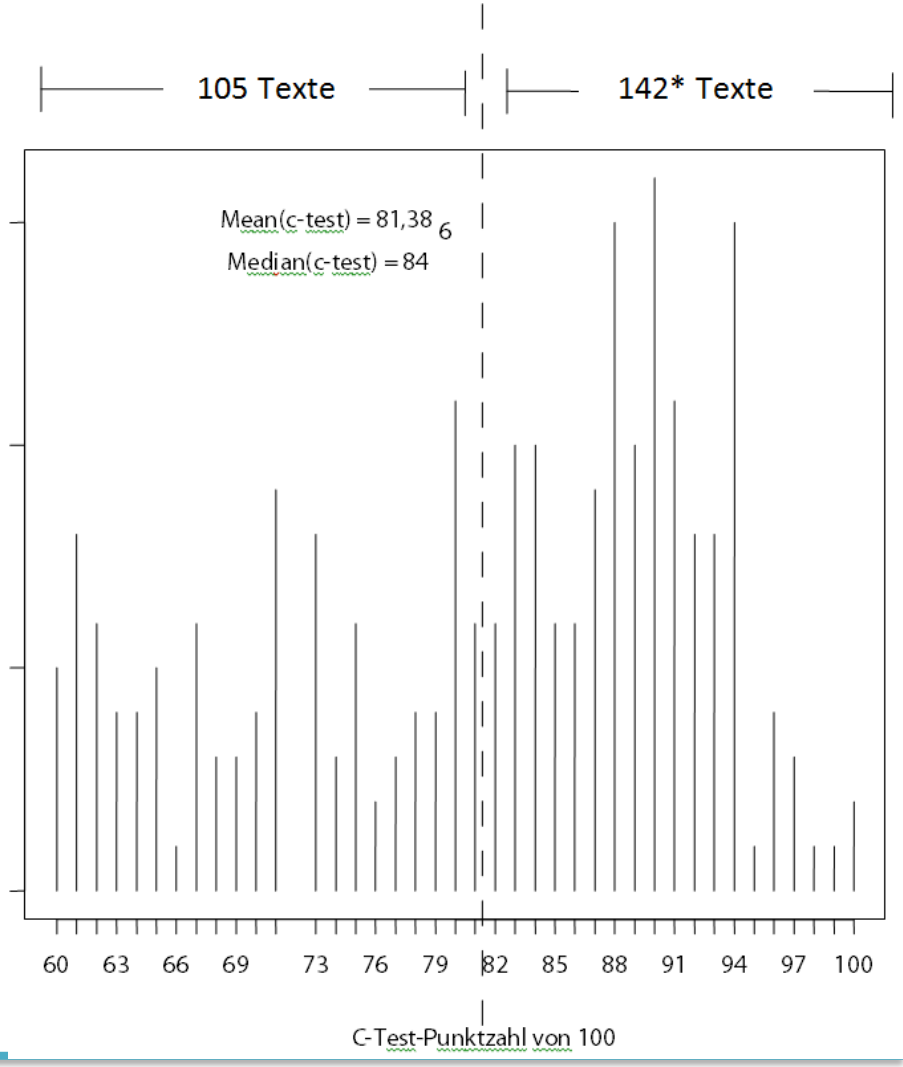
  immer mehr Lernerkorpora auch für andere Sprachen

  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

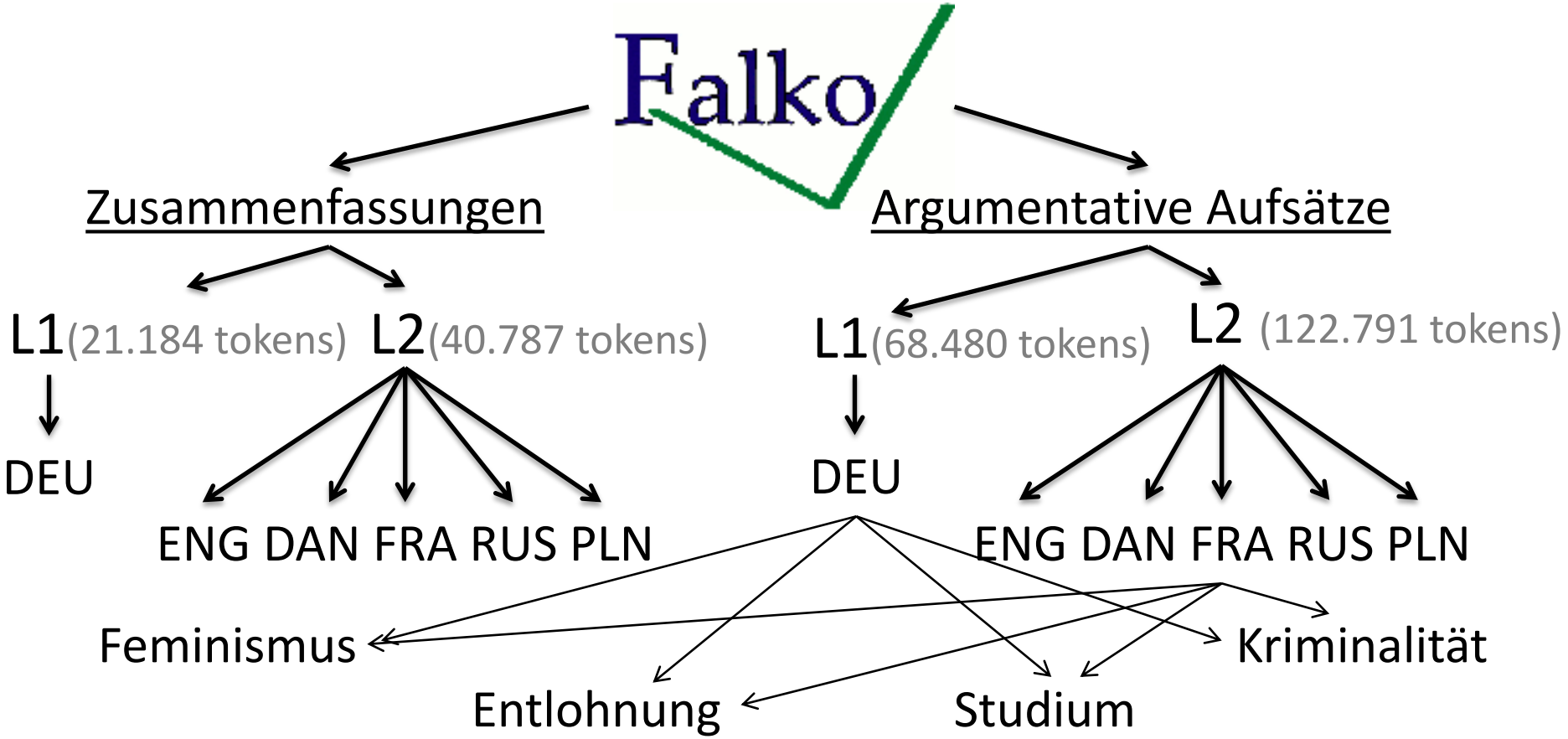
- frei verfügbares Lernerkorpus des Deutschen als Fremdsprache
- Fortgeschrittene Lerner (DHS oder C-Test: HU-Sprachenzentrum)

  ~B1-C1 (CEFR)



Falko

- Subkorpora 2 x 2 x 4 (16)



Metadaten

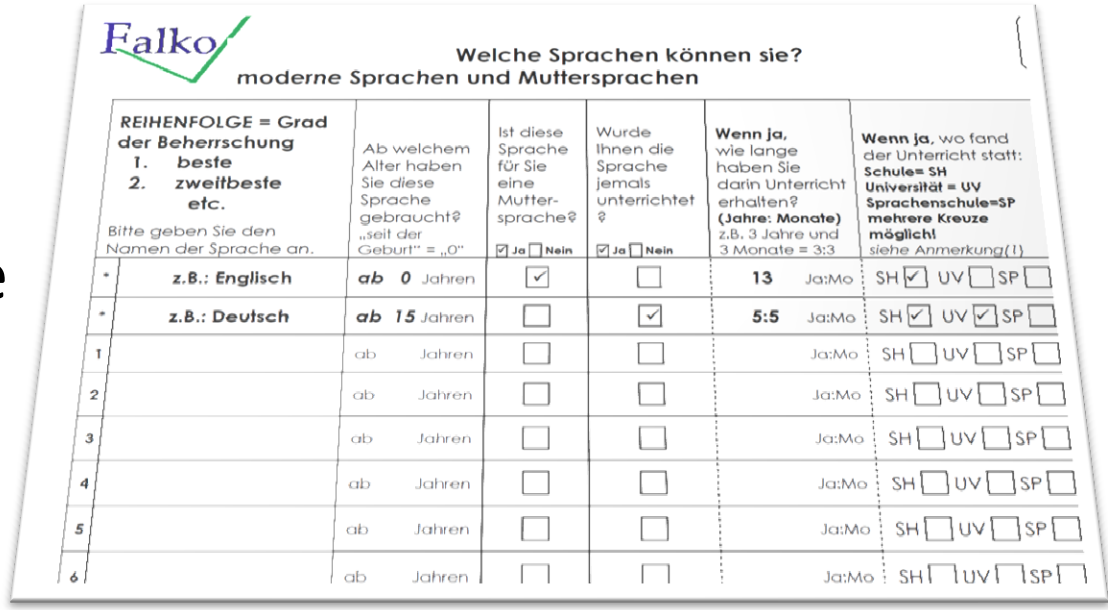
- Erhebungen
  - 90 Minuten
  - keine Hilfe (Internet, Wörterbücher, Spell-Checker etc.)
  - handschriftlich (nur Zusammenfassungen) & computergeschrieben
- Dokumentation (Reznicek et al. 2010)

  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
- 195 neue Texte
- ~130.000 Tokens (~ Wordformen)
- 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 Abfolgen von Wortarten tauchen auf? (Oberflächensyntax: Borin/Prütz 2004)
- Automatische Wortartentagger werden immer besser: Zeitungssprache ~98% (Kübler et al. 2010)

→automatische Wortarten(POS)- und Lemma-Annotation

Fehlerannotation in Falko

Ich merke, dass die Menschen ein schlechter gedenken über die Globalisierung haben.

(Ihr Beispieltext)

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

Zielhypothese

Wie würden Sie korrigieren?

Zielhypothesen

Ich merke, dass die Menschen ein schlechter gedenken über die Globalisierung haben.

(fkb034_2008_07)

- oft konkurrierende Korrekturen möglich

→ Falko: explizite Zielhypothese

ZH1: *Ich merke, dass die Menschen einen schlechten Gedanken über die Globalisierung haben.*

ZH2: *Ich merke, dass die Menschen eine schlechte~~r~~ Meinung von der Globalisierung haben.*

Zielhypothesen in Falko

ZH1: satzbasiert, nah an Lernerstruktur: Orthographie, Morphosyntax

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

- 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		einen	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	ZH1	ZH1Diff	ZH1pos	ZH1posDiff
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	

Mögliche Studie

- Machen russische und germanischen Muttersprachler im Deutschen unterschiedliche Artikelfehler?

→Finde alle Artikel in der ZH1, die mit "INS" markiert sind. (obligatorische Kontexte)

→Ziehe diese Zahl von der aller Artikel ab (100%)

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