Falko – German learner corpus
Theory and practice

Marc Reznicek

30.11.12

Universidad de Granada

with slides by

Hagen Hirschmann, Anke Lüdeling, Amir Zeldes, and
the whole Humboldt-University corpus linguistics team

content

- learner corpus research
- Falko - design
- annotation
  - automatic preprocessing
  - errors and target hypothesis
- research question and analysis
  - learner syntax
  - register awareness
- hands-on ANNIS query workshop

research questions

- Which linguistic structures are difficult to acquire for students of German as a foreign language?
- Do they depend on the learners native language(s)?
- Are difficulties form or function based?
- How productive are learners in their language use?
- Do learners actually lack register awareness?

methodology

Falko

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

→ find patterns in linguistic representations

→ uncover quantitative differences between learners and native speakers

→ contrast text of different L1-learner groups

Error **A**nalysis (EA) (e.g. Corder 1981, Izumi et al. 2005)

→ What kind of errors are learner specific?

→ Which depend on the learners L1?

learner corpora

Falko

- controlled and digitalized collections of learner texts
- design depends on the research question
  - spoken vs. written / task / text type
  - proficiency
  - L1 of learner
  - …
- most learner corpora in English

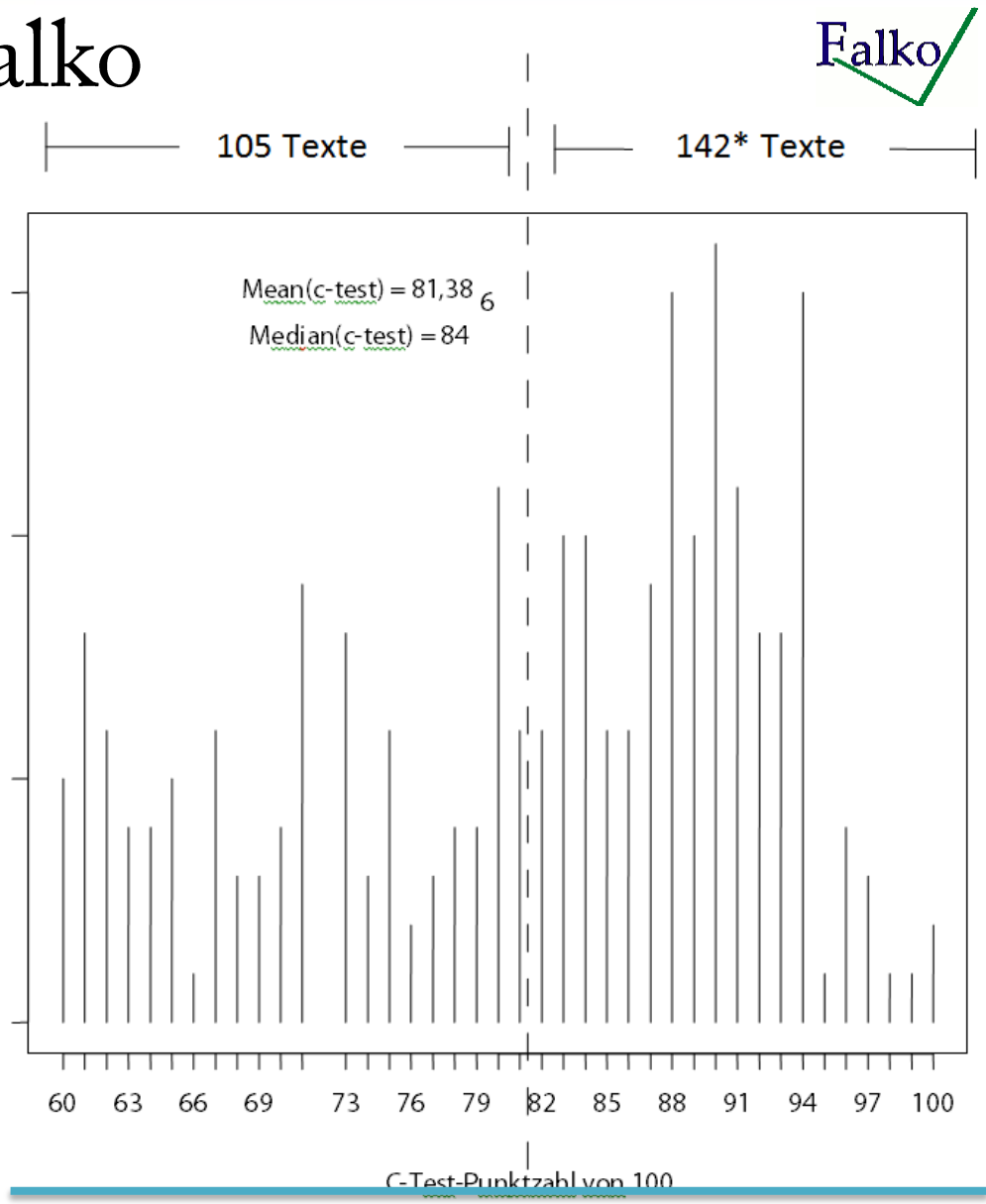
  growing amount in other languages

  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

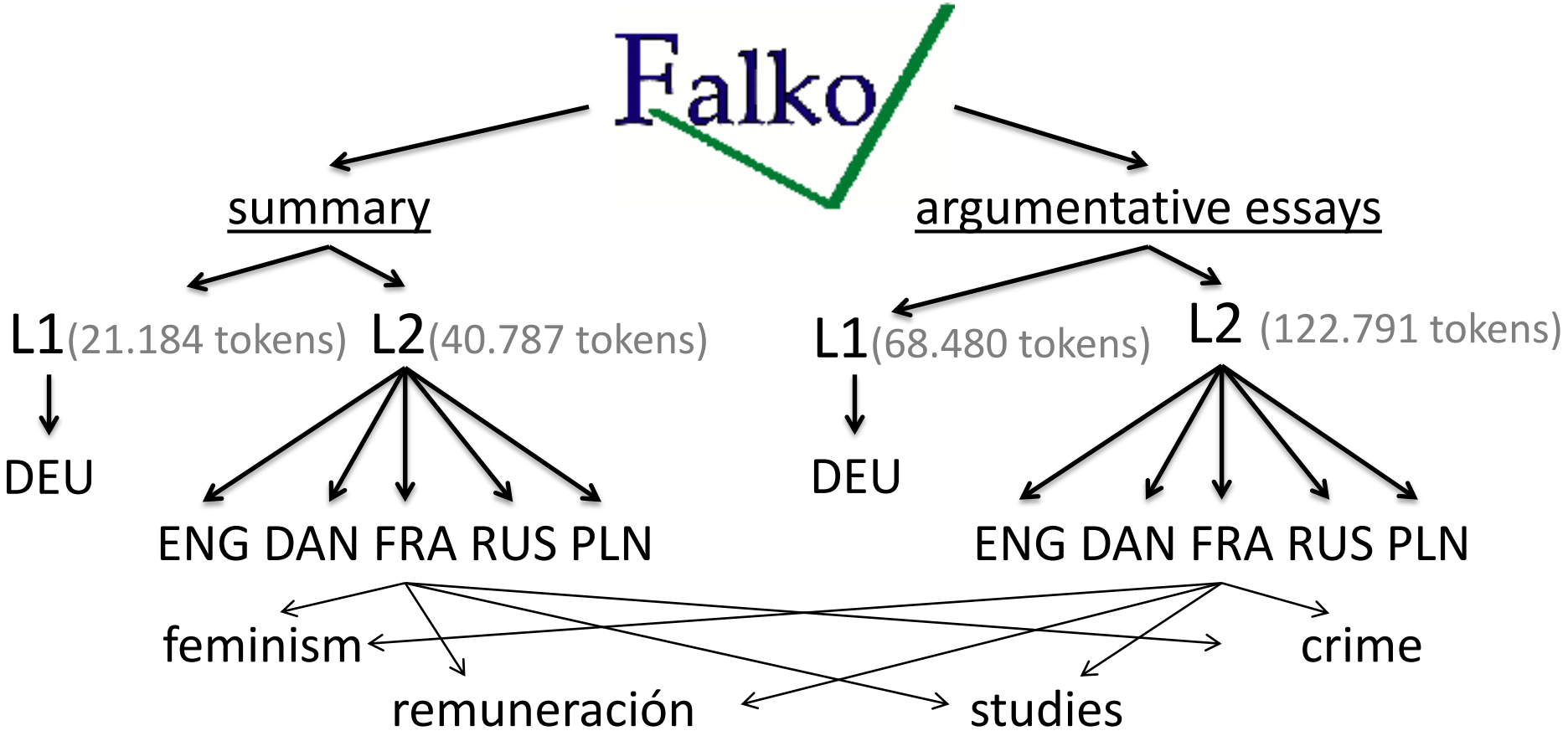
- freely available learner corpus of German as foreign language
- advanced learners

  ~B1-C1 (CEFR)



Falko

- subcorpora 2 x 2 x 4 (16)



data collection

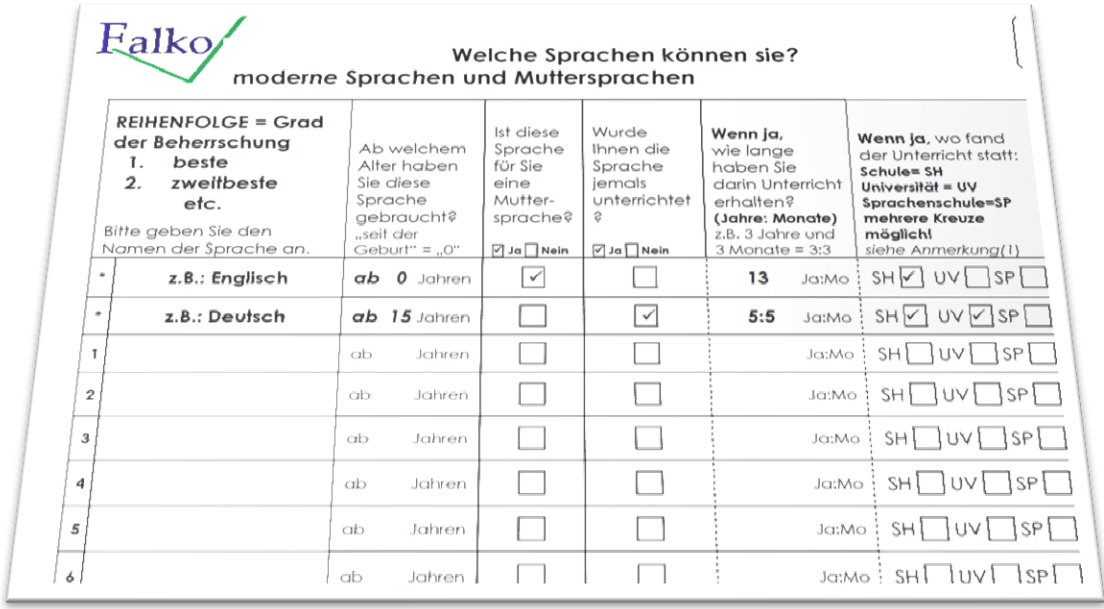
- collection
  - 90 minutes
  - no tools (internet, dictionaries, spell-checker etc.)
  - handwritten (only summaries) & typed
- documentation (Reznicek et al. 2012)

  http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20annotationen_v2.01
- project site:

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

language biography

- learners from 49 native language backgrounds
- much data for each acquired language
  - first contact
  - classes
  - duration
  - time spent in target language country
  - proficiency



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 ☐

annotation of learner language

- many corpora aren't annotated
- some include error tagging (Díaz-Negrillo & Fernández-Domínguez 2006)
- few (mainly) new learner corpora include more elaborated annotations

  (ALESKO, KOBALT, BEMATAC, DALEKO, KanDel)
- Falko: standoff-format (each annotation can be stored separately)

→ No limitation for new annotation layers (Lüdeling et al. 2005)

Falko

automatic annotation part-of-speech & lemma

- most learner corpus research on language surface (Möllering 2004, Vyatkina 2007 etc.)
- more interesting:
  - Which parts-of-speech (POS) and chains thereof are avoided? (surface syntax: Borin/Prütz 2004)
- automatic POS-taggers increase in accuracy newspaper ~98% (Kübler et al. 2010)

→Falko is automatically tagged for POS and lemma (TreeTagger & rfTagger)

error annotation in Falko

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

On other site would there no feminism be then we standed stil in the kitchen and cook.

(made up translation)

How would you correct this?

error annotation in Falko

On other site would there no feminism be then we standed stil in the kitchen and cook.

Error annotations are always based on a (at least implicit) normalized version of the learner utterance → target hypothesis

target hypotheses

On other site would there no feminism be then we stood still in the kitchen and cooked.

Falko: explicite target hypotheses

- competing versions

TH1: *On the other site, there would be no feminism then we would still stand in the kitchen and cook.*

TH2:*On the other hand, if there was no feminism ~~then~~ we would still stand in the kitchen cooking.*

target hypotheses

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

Falko: explicite target hypotheses

- competing versions

TH1: *An der anderen Seite, wenn da kein Feminismus wäre, stünden wir nur noch in der Küche und kochten.*

TH2: *Andererseits stünden wir, wenn es keinen Feminismus gäbe, nur noch in der Küche und kochten.*

target hypotheses in Falko

TH1: sentence-based, stays close to learner language:

orthography, morpho-syntax

TH2: text-based, aproximation to learner intention:

semantics, pragmatics, style

LT	TH1	TH2
An	Auf	Andererseits
der	der	
anderen	anderen	
Seite	Seite	
,	,	
		stünden
		wir
		,
wenn	wenn	wenn
da	da	
		es
kein	kein	keinen
Feminismus	Feminismus	Feminismus
wäre	wäre	gäbe
,	,	,
stünden	stünden	
wir	wir	
	nur	nur
noch	noch	noch
nur		
in	in	in

target hypotheses in Falko

TH1: sentence-based, stays close to learner language:

orthography, morpho-syntax

TH2: text-based, aproximation to learner intention:

semantics, pragmatics, style

- Differences between the TH and the original text are automatically tagged via **edit tags** (**CHA**nge, **INS**ert, **DEL**ete etc.)

LT	TH1	TH1Diff	TH2	TH2Diff
An	Auf	CHA	Andererseits	MERGE
der	der			
anderen	anderen			
Seite	Seite			
,	,			DEL
			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	

target hypotheses in Falko

TH1: sentence-based, stays close to learner language:
orthography, morpho-syntax

TH2: text-based, aproximation to learner intention:
semantics, pragmatics, style

- Differences between the TH and the original text are automatically tagged via **edit tags** (**CHA**nge, **INS**ert, **DEL**ete etc.)
- all THs are POS-tagged and lemmatized (TreeTagger, rfTagger)

- annotation of differences in the annotations on LT and TH1 & TH2

LT	pos	lemma	TH1	TH1Diff	TH1pos	TH1posDiff
An	APPR	an	Auf	CHA	APPR	
der	ART	d	der		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			ADV	
nur	ADV	nur		MOVS		MOVS
in	APPR	in	in		APPR	

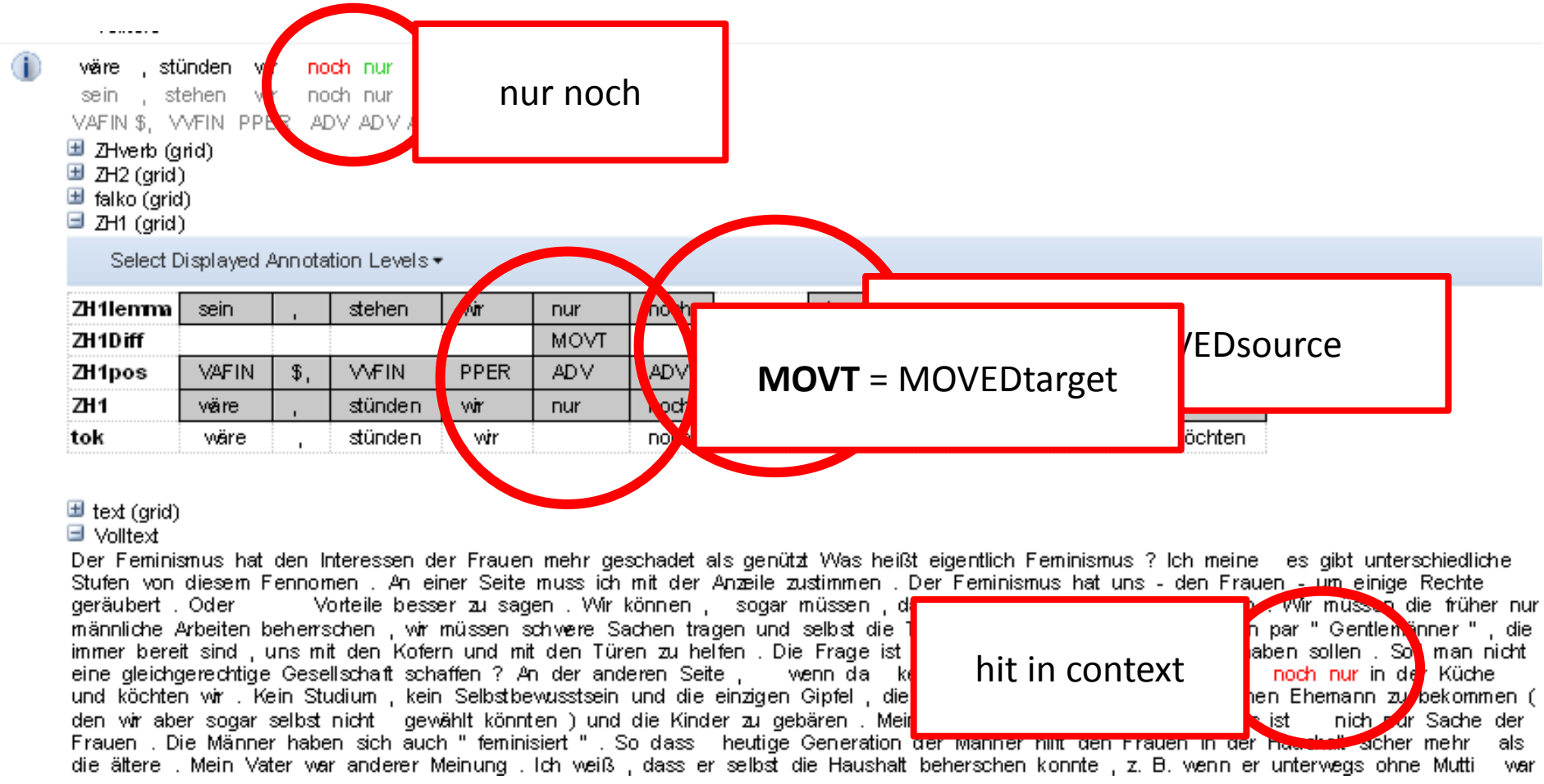
target hypotheses in Falko

TH1: sentence-based, close to learner language: orthography, morpho-syntax

TH2: text-based, close to learner intention: semantics, pragmatics, style

- Differences between the TH and the original text are automatically tagged via **edit tags** (**CHA**nge, **INS**ert, **DEL**ete etc.)
- All THs are POS-tagged and lemmatized (TreeTagger, rfTagger)
- additional **manual error tags** for some phenomena

search in ANNIS



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

Contrastive Interlanguage Analysis

Falko

- Do learners from Roman languages make more errors on articles than from Germanic languages?

→Find al articles marked with "INS" on ZH1Diff
(note: we look for something missing in LT)

→Compare results in texts written by Spanish and Italian learners with Danish and Afrikaans

Contrastive Interlanguage Analysis

Falko

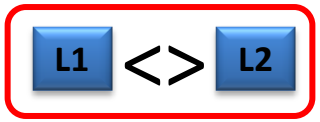
- What are difficult structures in learner German?
- structural difficulties are
  - independent of the learners L1
  - dependent of grammar of L2

DFG

Arts & Humanities Research Council

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

underuse



- Comparison of frequencies in L1 and L2 subcorpora
- overuse & underuse defined as **statistically significant differences** between two varieties
- structure may be underused because …

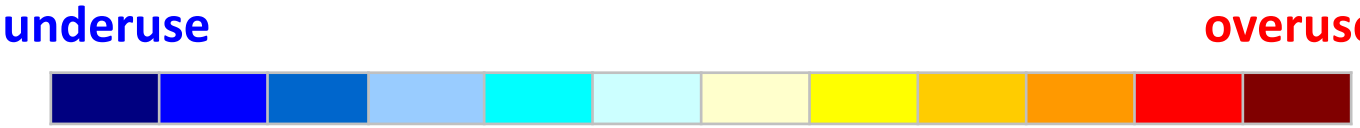
  … the learners did not acquire it yet.

  … the learners know it, but (unconsciously?!?) avoid it.

→ Diagnostics for finding difficult structures

overuse/underuse-visualization

- underuse: cold colors
- overuse: warm colors
- intensity of color signals strength of overuse/underuse



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

overuse/underuse visualization lexical items

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 underused in all L1 subcorpora

Stuttgart-Tübingen-Tagset (STTS)

(Schiller et al. 1995)

ADJective	Noun	Pronoun	Verb	ParTiKel	KOnjunction
ADJA	NN	PDS	VVFIN	PTKZU	KOUI
ADJD	NE	PDAT	VVIMP	PTKNEG	KOUS
		PIS	VVINF	PTKVZ	KON
ADVerb		PIAT	VVIZU	PTKANT	KOKOM
		PIDAT	VVPP	PTKA	
		PPER	VAFIN		
		PPOSS	VAIMP		
		PPOSAT	VAINF		
		PRELS	VAPP		
		PRELAT	VMFIN		
		PRF	VMINF		
		PWS	VMPP		
		PWAT			
		PWAV			

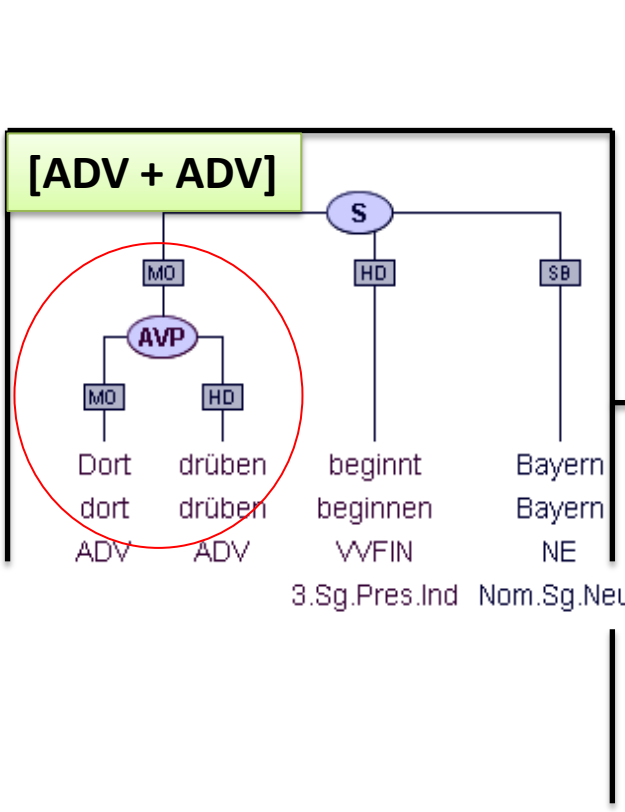
overuse/underuse visualization POS-bigrams

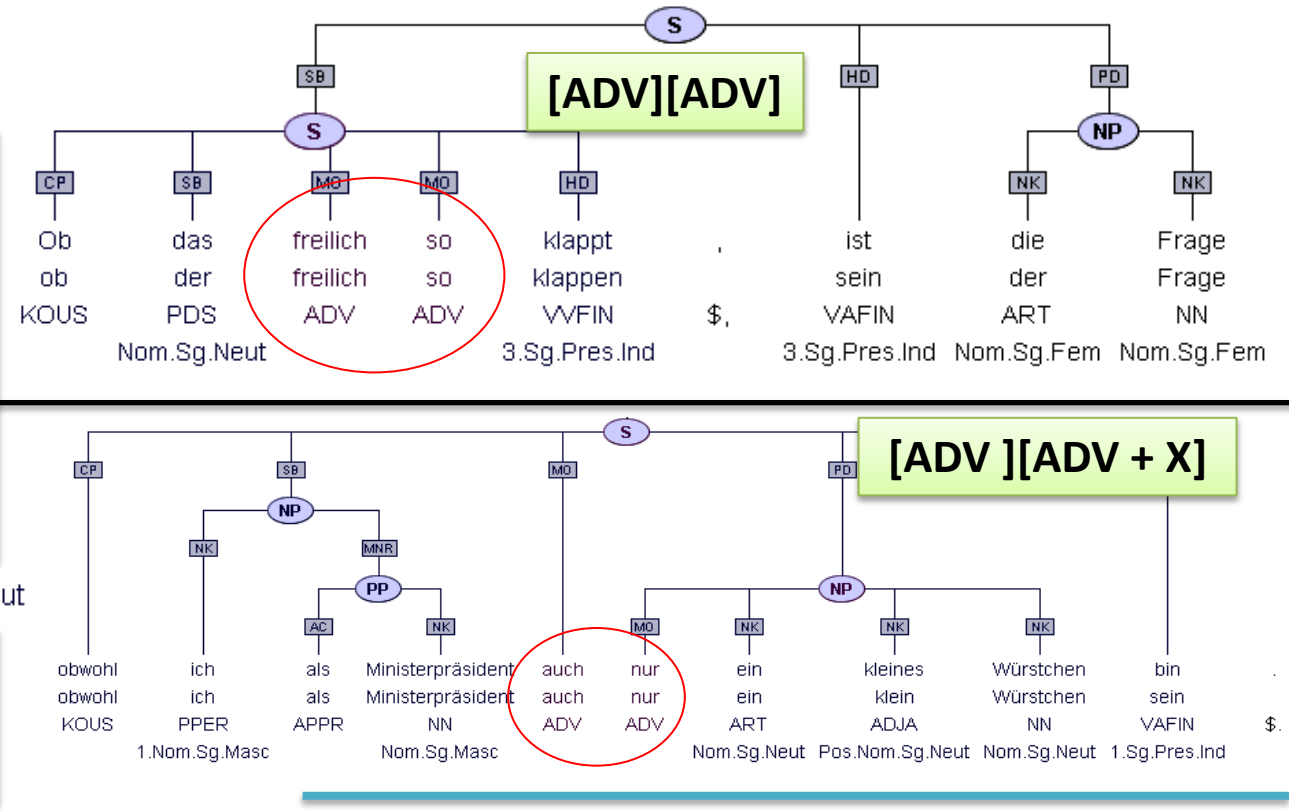
bigram	tot_norm	de	da	en	fr	pl	ru
$.-PPER	0.042384	0.005297	0.009748	0.007963	0.006166	0.005801	0.007409
VVFIN-$,	0.042131	0.006457	0.00776	0.006343	0.006937	0.006243	0.008391
PPOSAT-NN	0.041739	0.008058	0.007247	0.007269	0.007066	0.006298	0.005802
ADV-ADV	**0.041604**	**0.012858**	**0.010518**	**0.006111**	**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

Adverb chains underused in all L1 subcorpora

first conclusion

- adverb chains are avoided by all L1 groups





first conclusion

Falko

- adverb chains are avoided by all L1 groups

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

→structures with variable deep syntactic structure are avoided.

→ What is avoided?

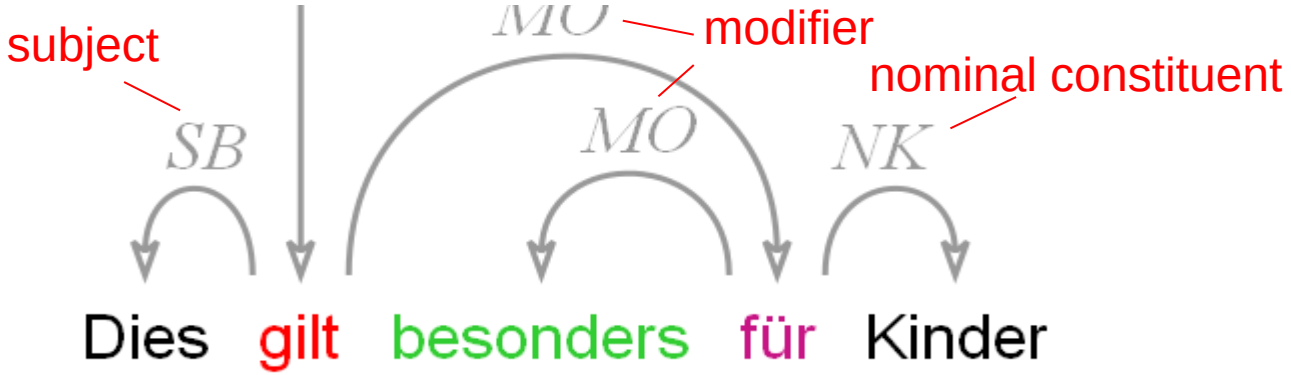
→difficult forms or functionen?

hypothesis: modification is avoided in general

syntactical annotation (dependencies)

Falko

- Every word (except sentence root) is governed by another word
- Arrows point to hierachical lower items
- Each depedency arrow carries a **grammatical function** label.

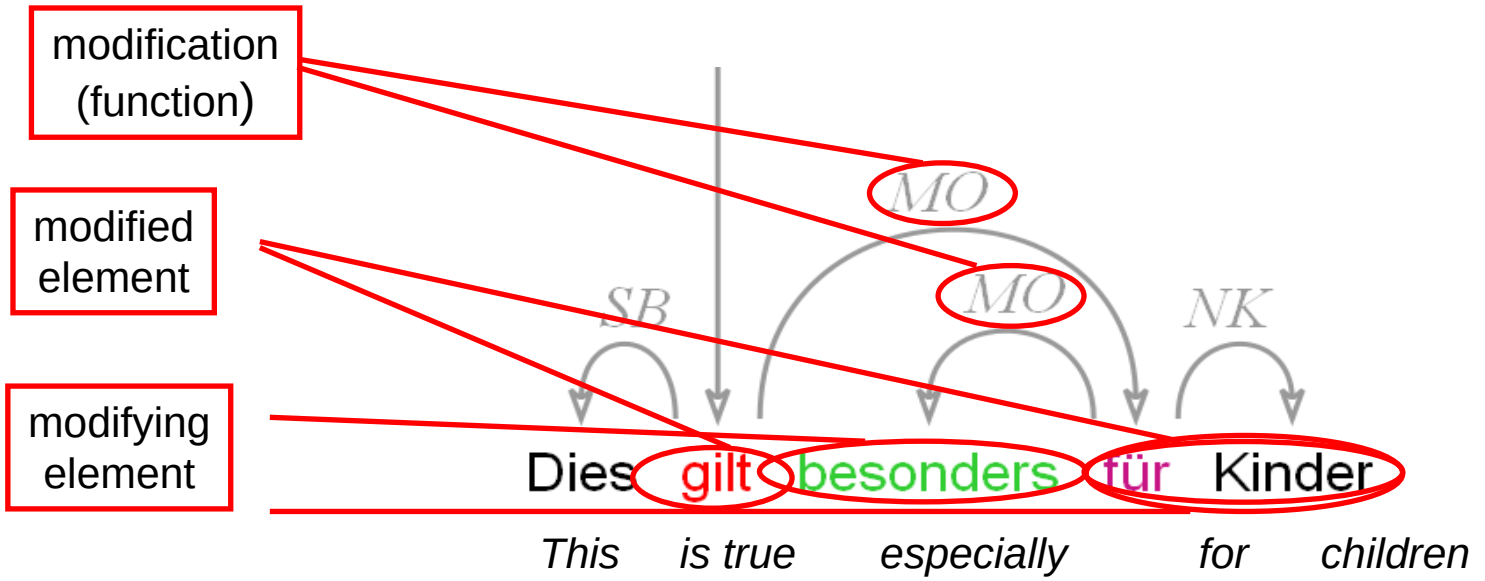


search for modification

Falko

different aspects

- Are grammatical functions avoided?
- Are special targets of modification avoided?
- Are special modifying categories avoided?



Falko

underuse/overuse of functions

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 significantly underused by all L1

modificationen in Falko

- all categories are often modified by all L1

(Hirschmann et al 2012)

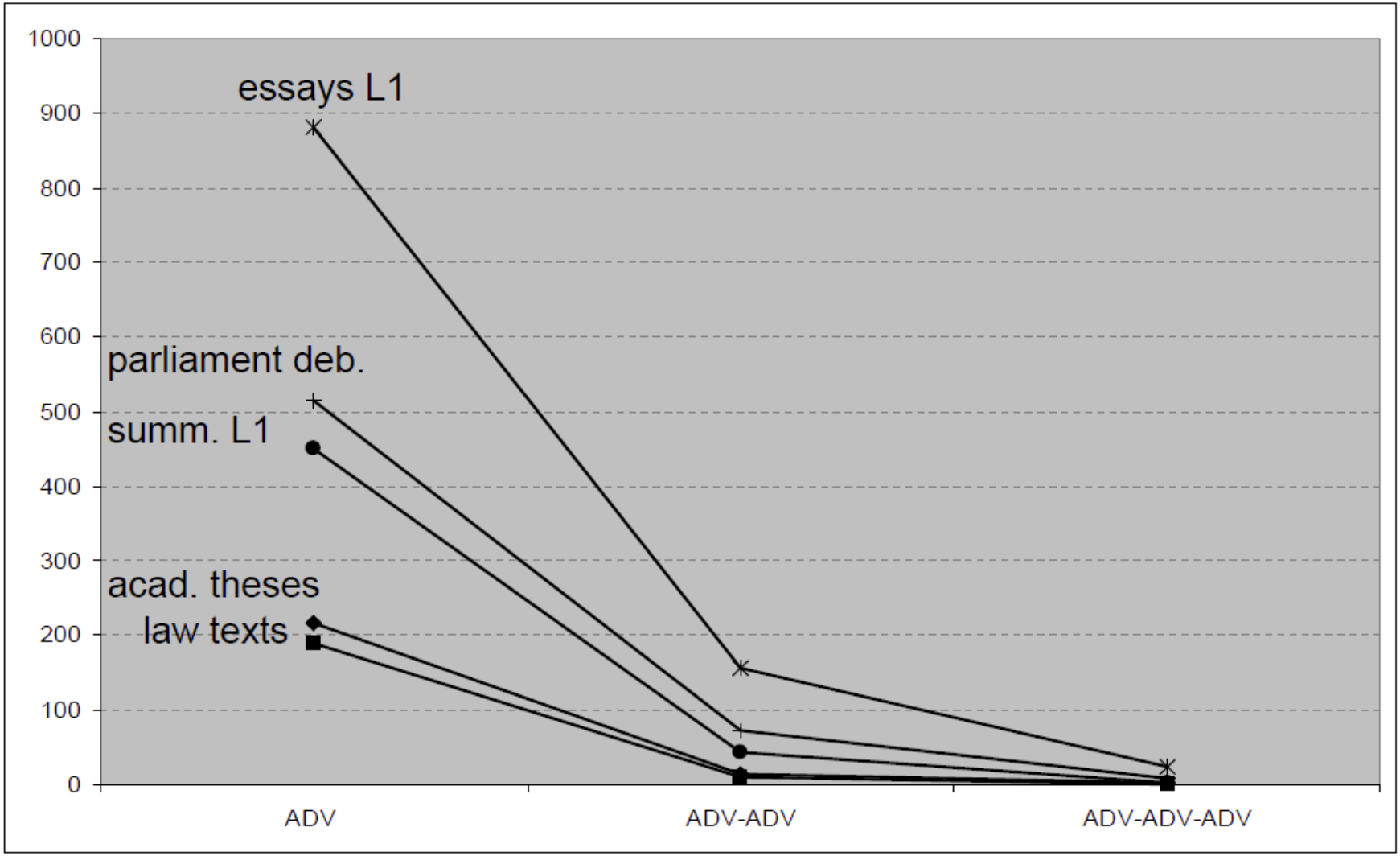
- **all modification relations** show **underuse**
- **adverb modifiers** show **strongest underuse**
- independent of L1

learner register awareness

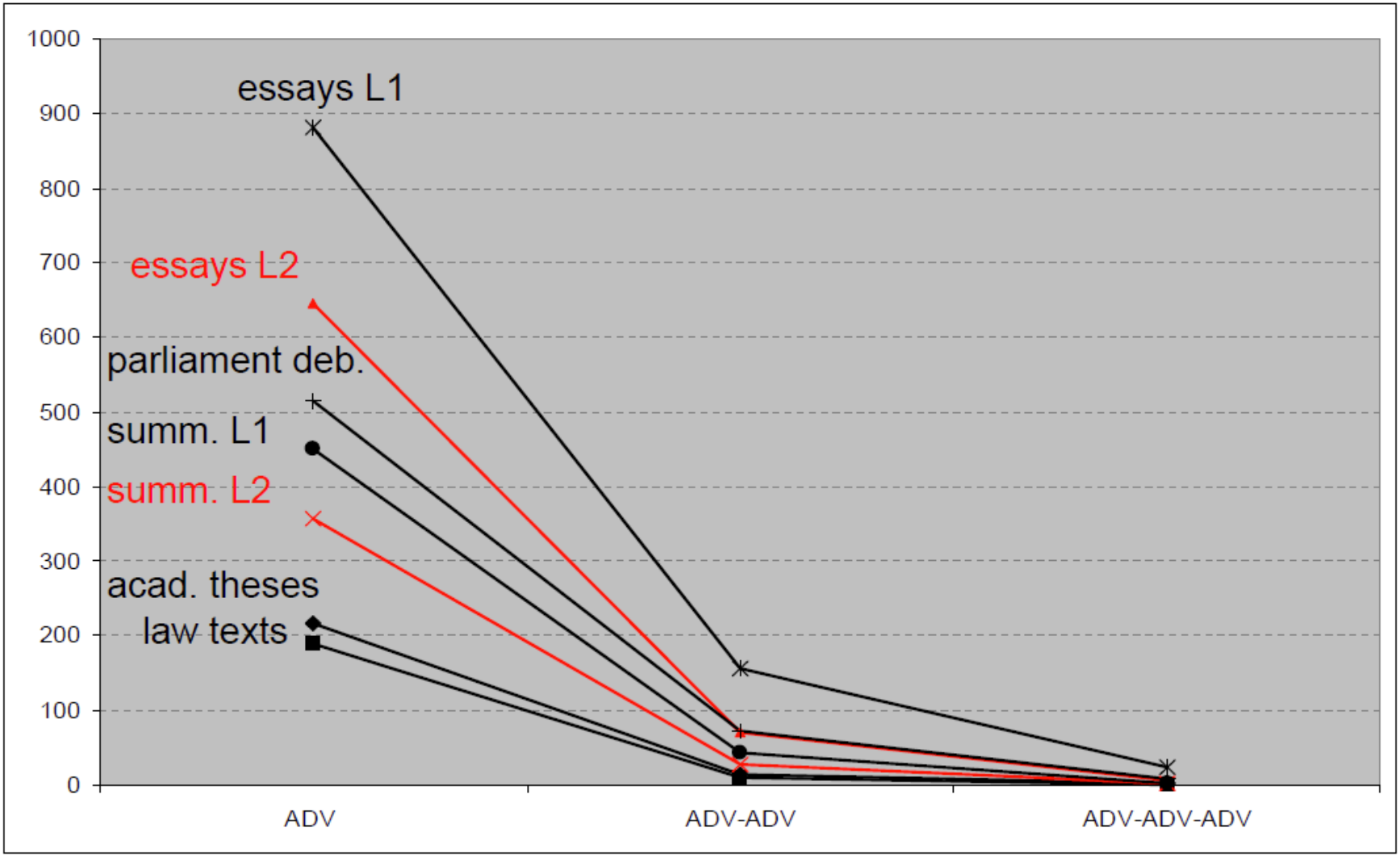
- "lack of register awareness" (Gilquin/Paquot 2007)
- complex structures show less style shift (Sato 1985)
- production of L2 ADV-ADV-chains depends on syntactic complexity (Zeldes, Hirschmann & Lüdeling 2008)

So, do learners know how to adjust language to different registers?

study 1: ADV-ADV L1 (Hirschmann et al. 2009)



study 1: ADV-ADV L1 & L2 (Hirschmann et al. 2009)



study 2: register factor analysis

- factor analysis:
  students annoted learner and native speaker texts
- 2 topics:
  - remuneration (impersonal topic)
  - studies (personal topic)

→ **Are there bundles of features which occur together in a certain register?**

- principle component analysis (explorative)

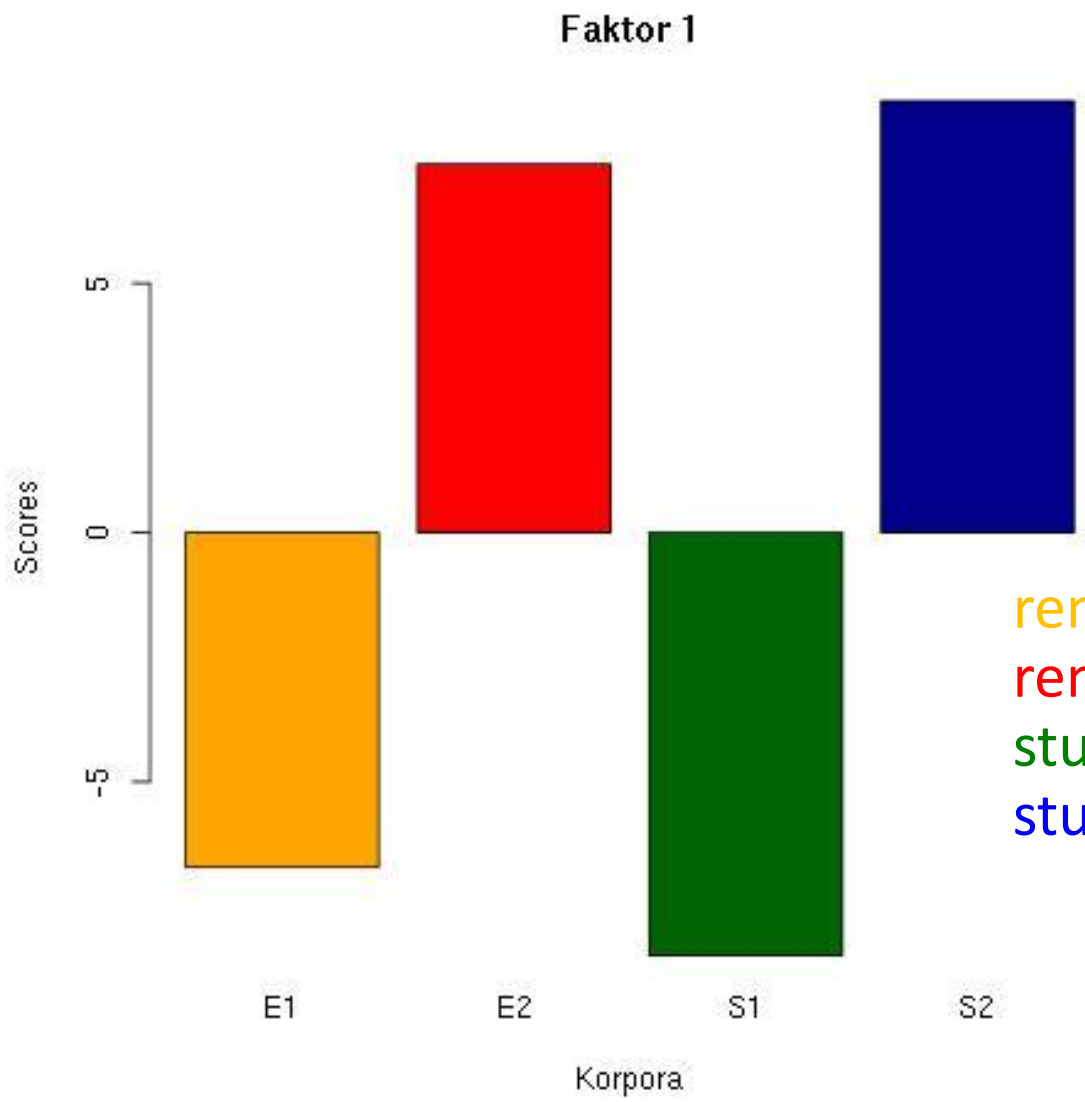
→ factors

→ **Are there differences between L1 & L2 texts?**

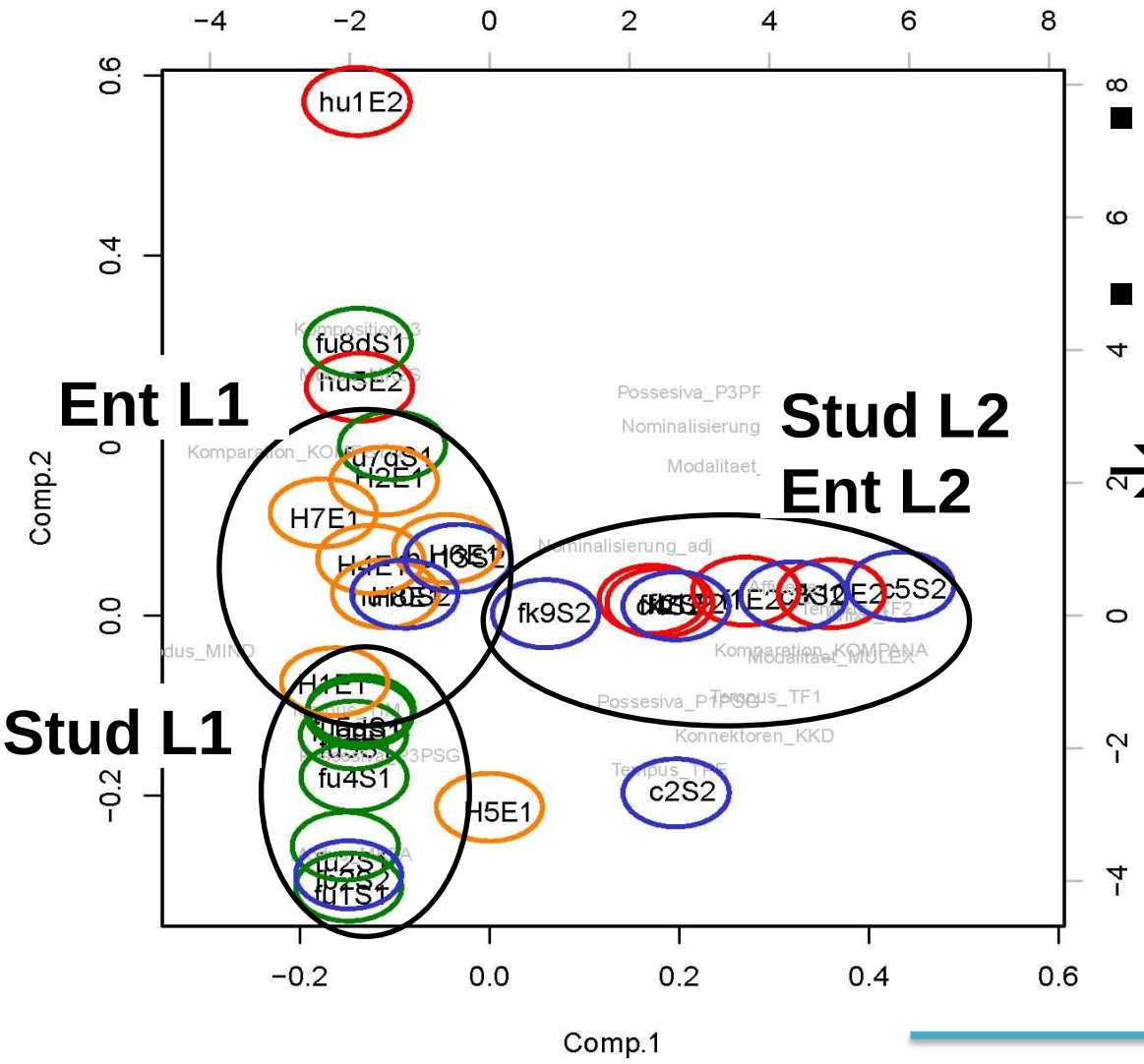
Factor 1 features

- **Affixe_t**
- **Affixe_e**
- **Komparation_KOMPANA**
- **Konnektoren_KKD**
- **Modalitaet_MULEX**
- **Modalitaet_MUZAM**
- **Nominalisierung_adj**
- **Nominalisierung_conversion**
- **Possesiva_P1PSG**
- **Possesiva_P3PPL**
- **Tempus_TF1**
- **Tempus_TF2**
- **Tempus_TPE**

- - Possesiva_P3PSG
- - Tempus_TIM
- - Komposition_3
- - Komparation_KOMPSY
- - Modus_MIND
- - Modus_MK2A
- - Modus_MK2S



remuneration L1
remuneratio L2
studies L1
studies L2



- Factor 1 distinguishes topics for L1 texts
- L2-texts are very similar in factor 1

→ Learners seem to show a (partial) lack in register awareness

remuneration L1
remuneratio L2
studies L1
studies L2

- Annotated learner corpora allow for a wide variety of quantitative and explorative methods
- Basing error analysis on an explicit target hypothesis brings the advantage of …
  - provide a high transparency and reproducibility of error analyses
  - allows for competing analyses
  - provides a reference text as basis for automatic annotation and thus contrasting of those annotations
  - elaborated statistical methods (like principle component analysis) boosts explorative investigations

Gracias!

Thanks!

Danke!