Multilingual and Parallel Corpora
Non-native corpora

Amir Zeldes

amir.zeldes@georgetown.edu

Target hypotheses

- Reading:
  - Reznicek et al. (2013) Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture

Experiment (data: Gachon corpus)

For next week: formulate target hypotheses for these 5 'sentences':

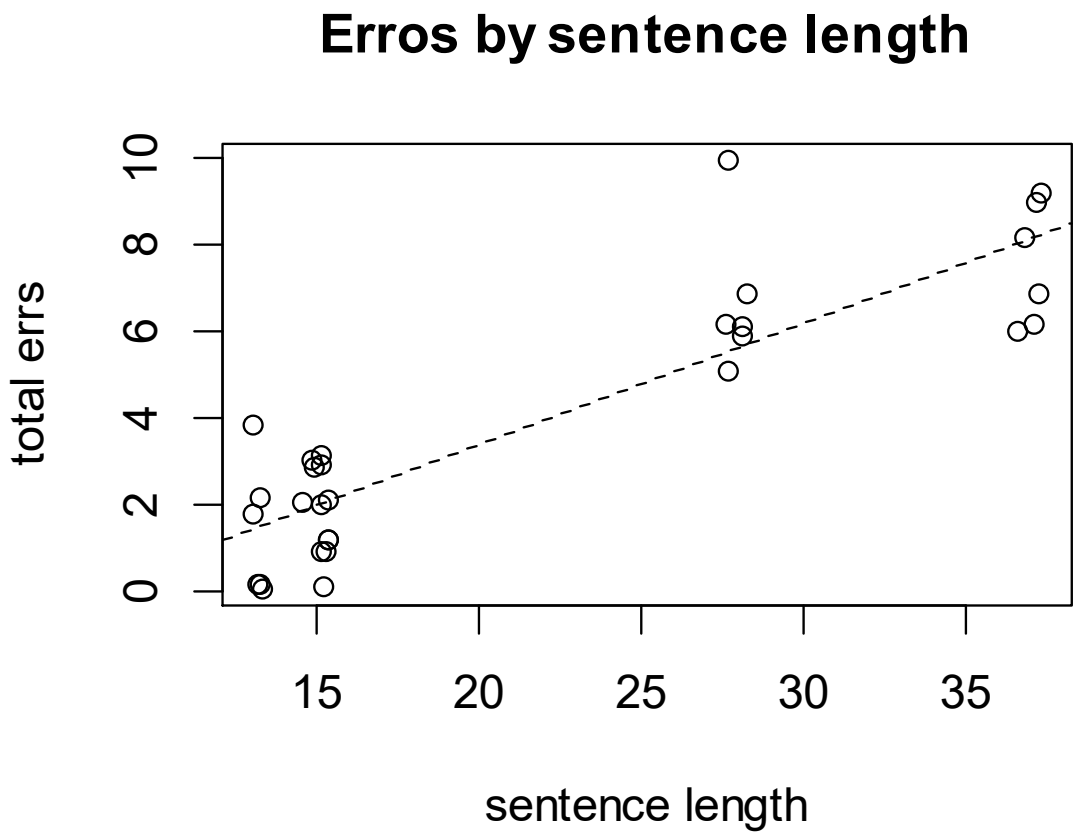
- *Nobody knows how long and how many of us work and play using computers everyday.*
- *we make calls, send email and fax, watch movies, play games, do accounting, make presentation, take training courses in front of screens instead of standing by side of all kinds of machinese in the industry times.*
- *If my computer malfunction, I will sent to repair center if i were the manager of a hotel, i want to rent ""super car"" for our guest.*
- *(super car means a very expensive car like a Lamborghini)*
- *every man hopes to drive a super car at once in their lifetime .*

Experiment (data: Gachon corpus)

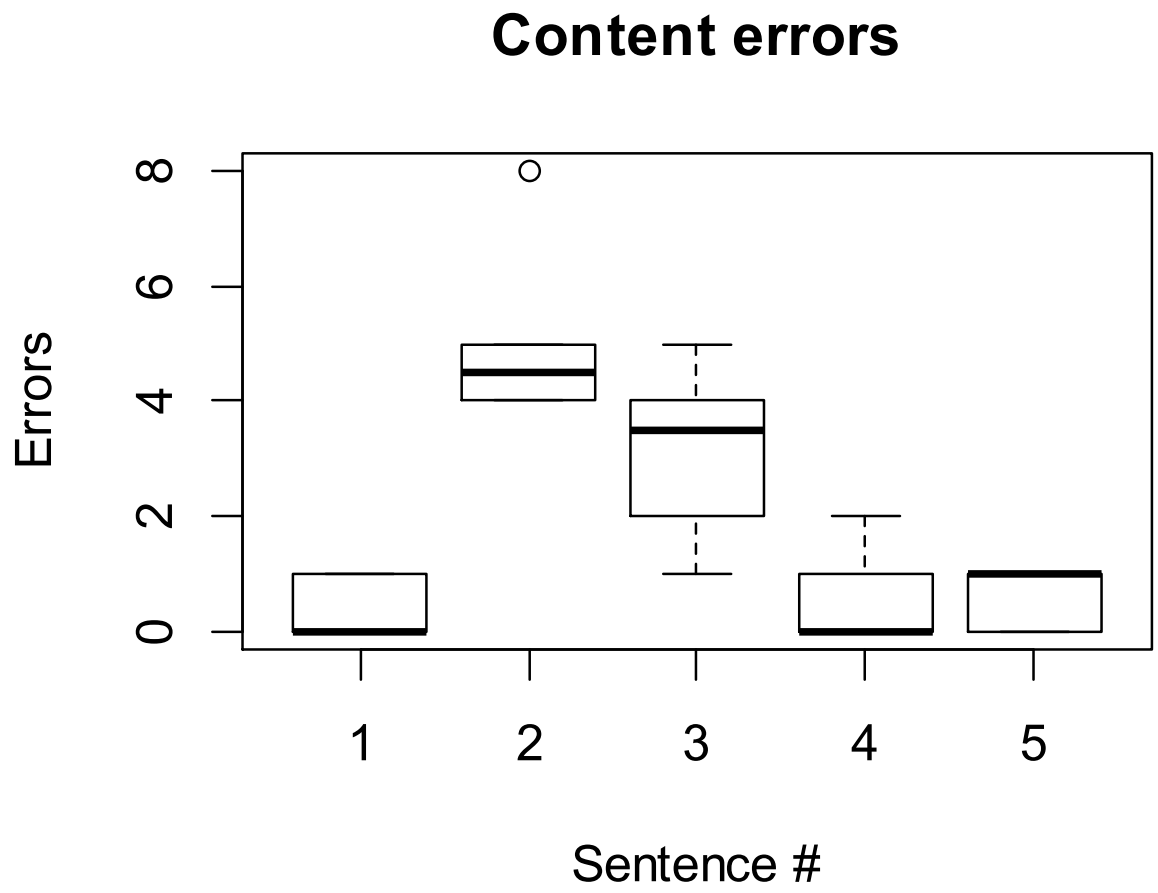
Submit your text, marking error locations and count: How many errors do you have in **function** and **content** words?

- *Nobody knows how long and how many of us work and play using computers everyday.*
- *we make calls, send email and fax, watch movies, play games, do accounting, make presentation, take training courses in front of screens instead of standing by side of all kinds of machinese in the industry times.*
- *If my computer malfunction, I will sent to repair center if i were the manager of a hotel, i want to rent ""super car"" for our guest.*
- *(super car means a very expensive car like a Lamborghini)*
- *every man hopes to drive a super car at once in their lifetime .*

Caveat – long sentences > more errors

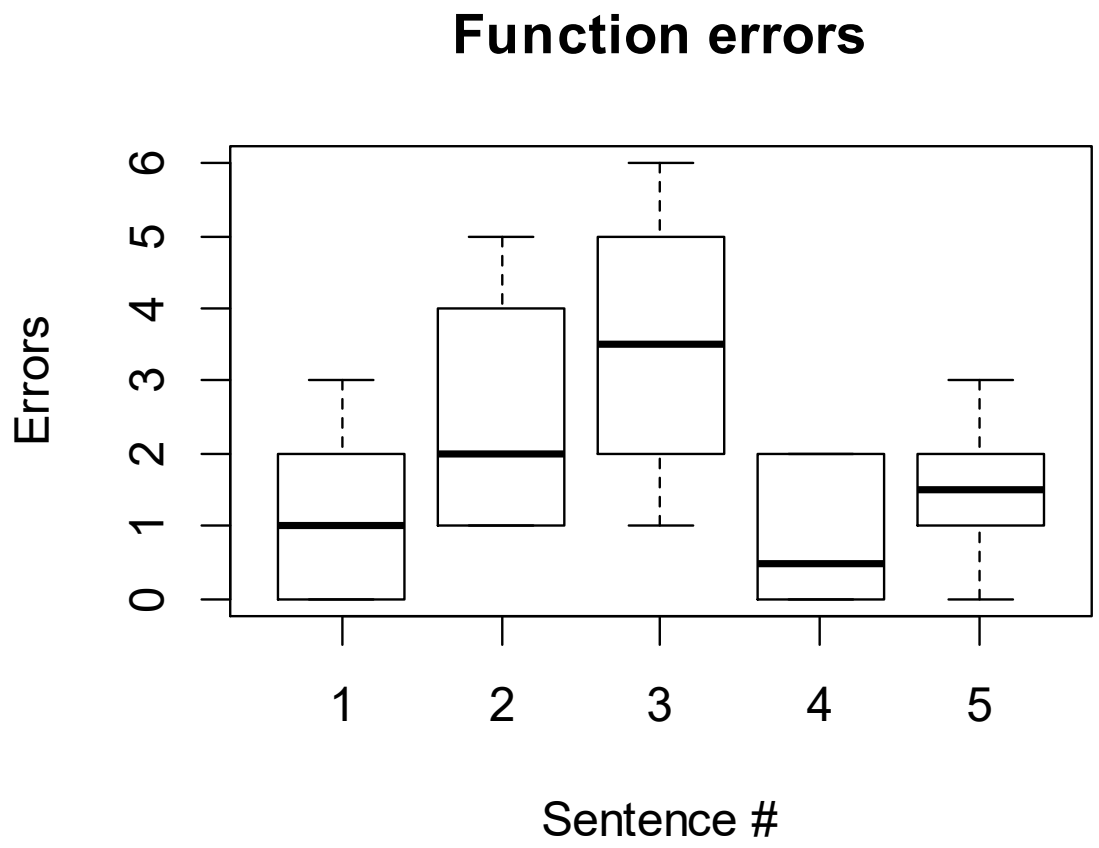


Experiment - results

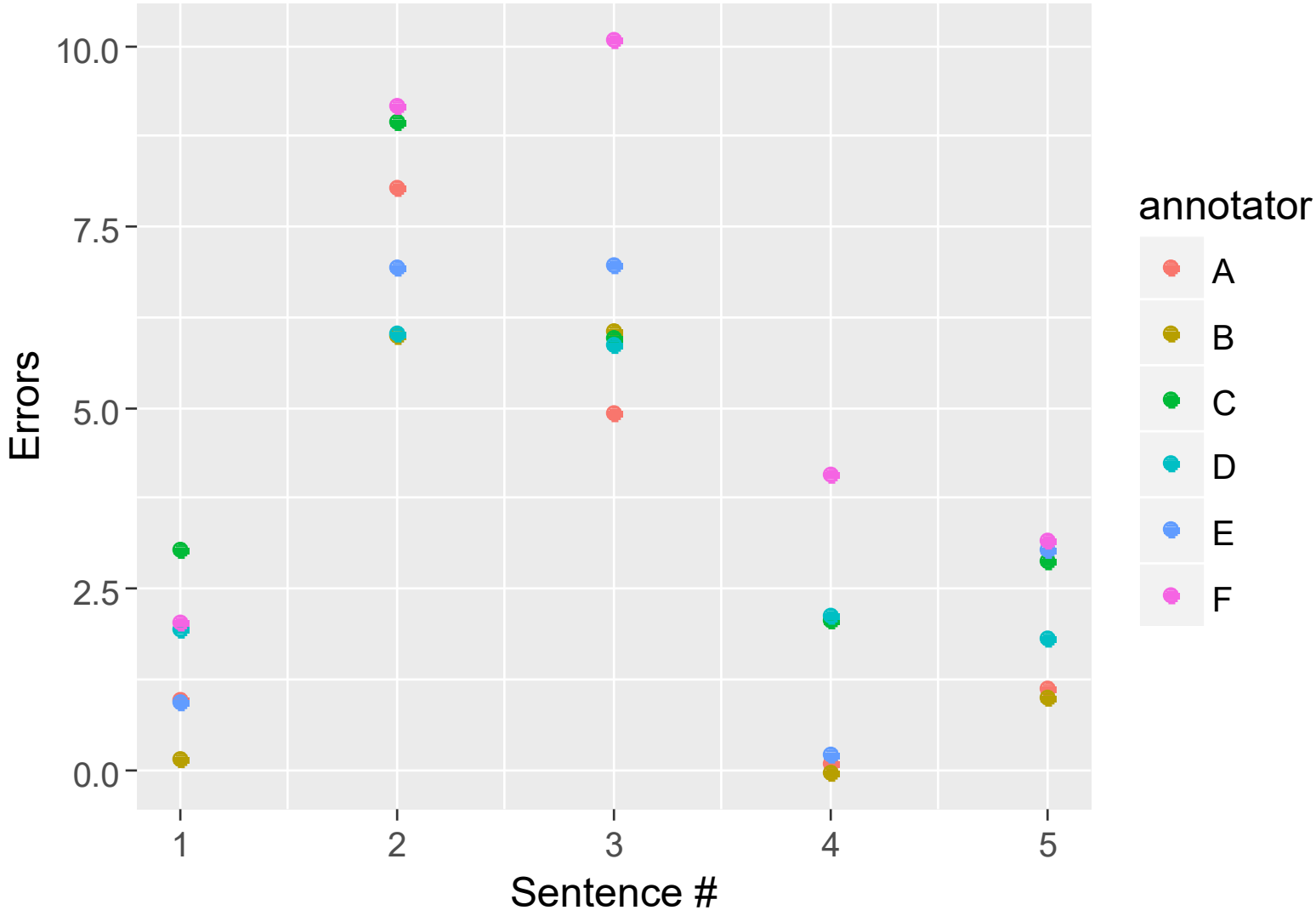


Experiment - results

- Function errors



Annotator leniency



Example – sentence 4

- (super car means a very expensive car like a Lamborghini)
  - (**A** super car **is** a very expensive car like a Lamborghini)
  - (**“**Super car**”** **means** a very expensive car**,** like a Lamborghini)
  - (**a** **“**super car**”** **is** a very expensive car like a Lamborghini)
  - (**A luxury car is** a very expensive car**,** like a Lamborghini.)
  - ***[no correction]*** (twice)
  - …

TH agreement – German example

	was		der	Novelle	oder	der	Ode	nicht	betrifft
1	was	auf		Novelle	oder		Ode	nicht	zutrifft
2	was	auf	die	Novelle	oder	die	Ode	nicht	zutrifft
3	was	bei	der	Novelle	oder	der	Ode	nicht	der Fall ist
4	was	für	die	Novelle	oder	die	Ode	nicht	zutrifft
5	das		die	Novelle	oder	die	Ode	nicht	betrifft

Lüdeling (2008)

Experiment results (Lüdeling 2008)

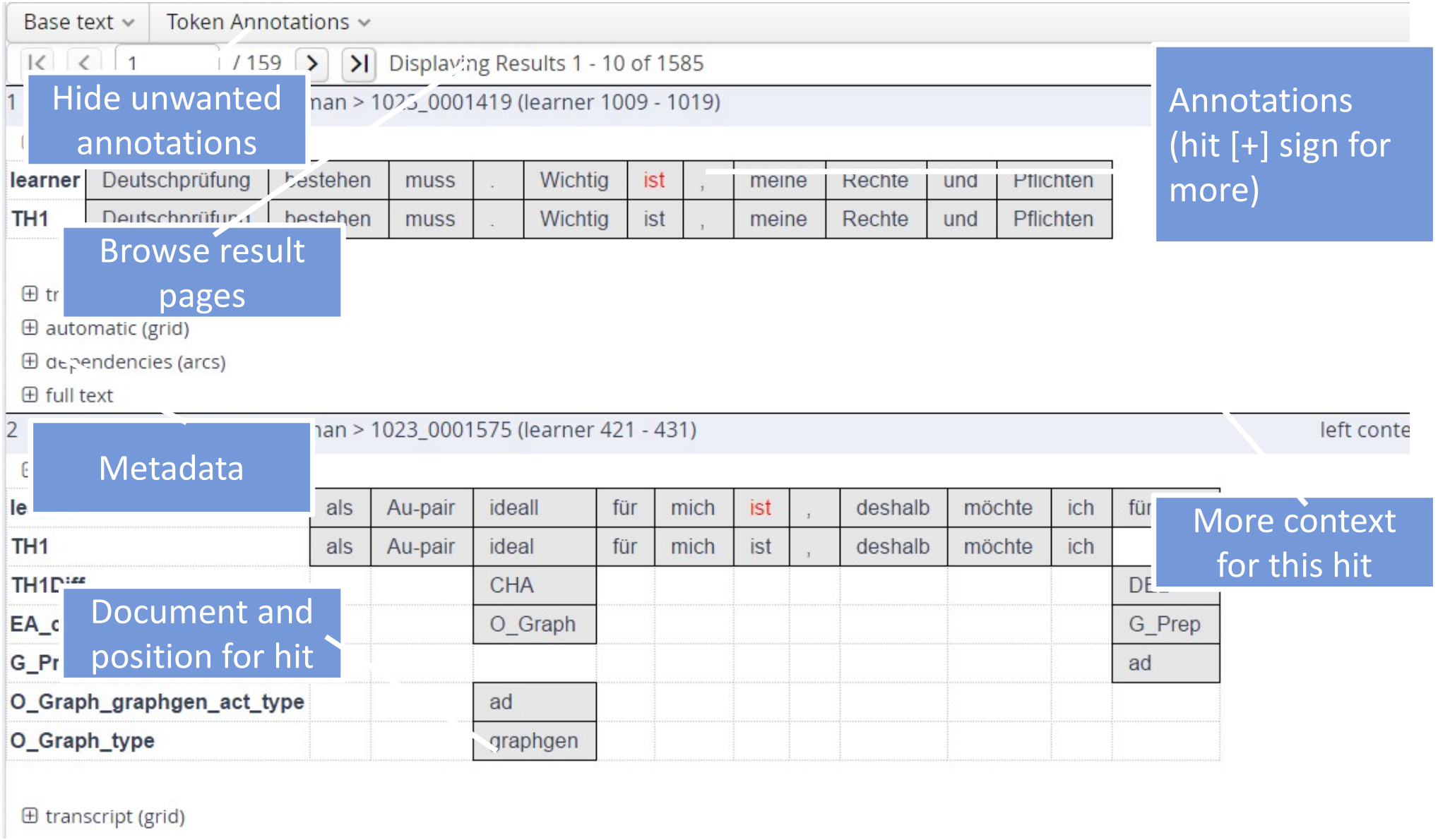
- 5 annotators
- 17 sentences (running text)

Content words	Function words
15	13
24	26
17	25
16	12
14	22

Using TH with ANNIS Query Language

- Basic AQL principle: `annotation="value"`
  - These can be actual token forms: `tok="go"`
  - Or named annotations: `pos="RP"`
  - Use regular expressions with slashes: `pos=/VV.*/`

What does the concordance look like?



Say we want “meine” only as a verb

- Example: MERLIN corpus
- We can query POS annotations
- And we can query the learner text level
  - `learner="meine"`
  - `tok_pos="VVFIN"`
- How do we combine them? How does ANNIS know whether we mean "at the same time" or "one after the other"?
- We need **operators**
  - Placed between search terms: *‘VVFIN and meine’* or *‘VVFIN then meine'*?

.	precedence	`learner="meine"` .	`tok_pos="VVFIN"`
_=_	identical coverage	`learner="meine"` _=_	`tok_pos="VVFIN"`

What about some distance away?

- You can qualify the dot (precedence) operator with a range of token numbers:
  - `TH1Diff="CHA" .1,5 TH1Diff="CHA"`

für	neuen	Aufgaben	
für	neue	Aufgaben	
	CHA		
für	neue	Aufgaben	

- Use .* for 1-50 tokens:
  - `tok_lemma="fragen" .* learner="was"`
- If you don't care about order you can use **the near operator ^ (with the same options)**

kein	mensch	hat	gefragt		wa
kein	Mensch	hat	gefragt	,	wa
	CHA			INS	

  - `tok_lemma="je" ^1,10 tok_lemma="desto"`
  - `tok_lemma="je" ^* tok_lemma="desto"`

Searching for things inside things

- Let's look for verbs inside essay or letter closings
- Closings (**closing**), verbs (any pos starting with **V**) and a relationship between them
- We can use the **_i_ includes** operator to combine the two:
  - `closing _i_ tok_pos=/V.*/`
- You could also search for verbs at…
  - the left edge (start) of a closing **_l_**
    `closing _l_ tok_pos=/V.*/`
  - the right edge (end) of a heading with **_r_**
    `closing _r_ tok_pos=/V.*/`

Contrastive analysis

- Contrastive Interlanguage Analysis (CIA, Selinker 1972, Granger et al. 2002):
  - Theoretical framework for SLA studies
  - Each L2 (and level, and L1, and …) is an independent language, not just an imperfect 'imitation' of L1
  - Studying the properties (syntactic, semantic, ...) of each L2
  - Contrastive analysis of L2s and native L1
  - Error Analysis (EA)

Aligned hypotheses for CIA

- How can we use THs for contrastive analysis?
- What kinds of differences can we find?

Case study – L2 German compounds

Research questions:

- Do advanced learners of German use compounds differently from natives?
- What constructions do they acquire?
- Which ones are difficult and why?
- Are L2 compounds as productive or more memorized?
- Which patterns are extensible for L1/L2?
- How does compounding relate to proficiency level?

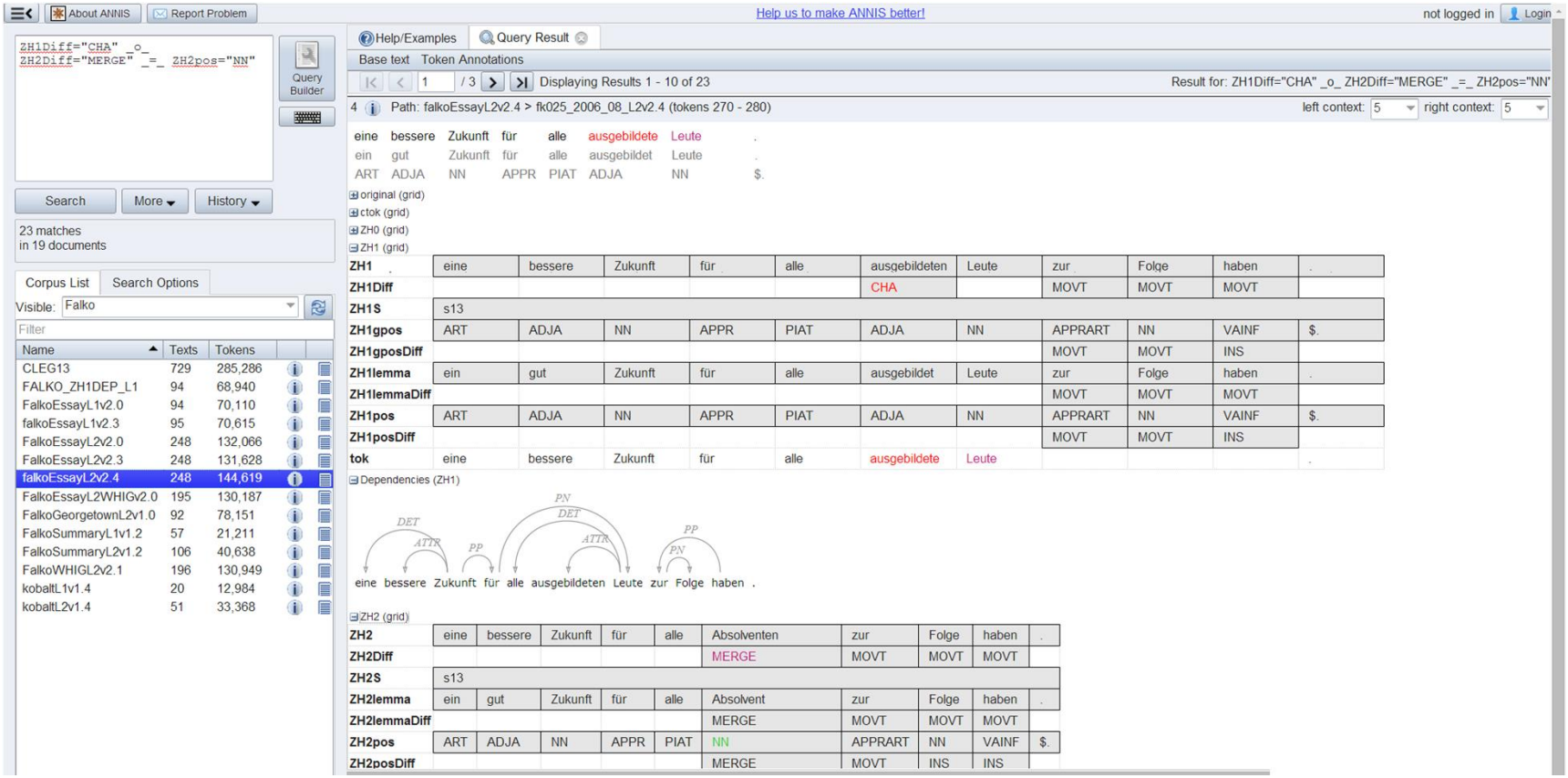
The data

- The Falko corpus (Reznicek et al. 2010)
  - Written language, designed to match ICLE (Granger et al. 2009)
  - Comparable native corpus with exactly the same tasks
  - "Advanced" learners (c-test > 60)
  - Multi-layer corpus: POS, lemma, syntactic dependencies, complex verb analysis, 3 target hypotheses + edit annotation
- **Essay** sub-corpus:
  - Argumentative texts on four topics, currently containing:
    - L1: 95 texts / 68,480 tokens (≠ word forms)
    - L2: 443 texts / 274,806 tokens

Target hypotheses

- Multiple hypotheses specify what we think the learner is trying to say on different levels:
  - Original: *alle ausgebildete Leute* ~all educates people
  - TH1 - Form: *alle ausgebildeten Leute* all educated people
  - TH2 - Content: *alle **Absolventen*** all graduates
  - TH0 - same as TH1 but word order not corrected

Falko in ANNIS



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

Extracting nominal compounds

- AQL search for nouns: pos="NN"
- Identify and analyze compounds automatically:
  - Lexicon from SMOR morphological analyzer (Schmid et al. 2004)
  - List of simplex lemmas from TuePP (~300M tokens newspaper) http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tuepp-dz.html
  - Extract longest possible head
  - Rule out impossible modifiers (morphophonology)
- Use target hypotheses to locate errors:
  - pos="NN" _=_ ZH1Diff

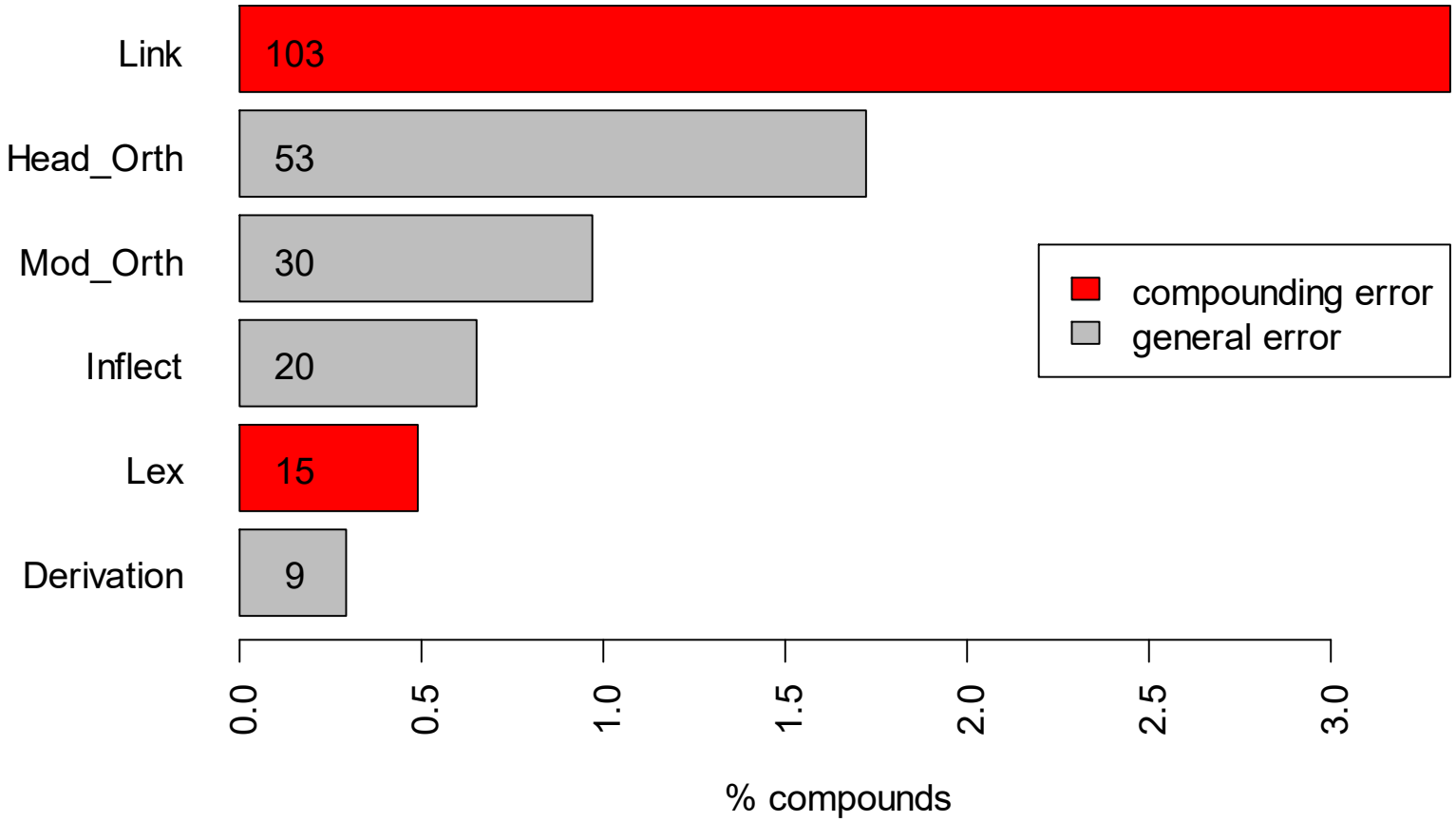
Errors in compounds

- 2113 compounds
- 218 errors (~10.3%) using target hypothesis [TH] (TH0Diff="CHA")
- Compounding removed by annotator in TH0: **1 time**
  - ***Studienlesesahl*** **'study-read-hall'** > TH0: ***Lesesaal*** **'read-hall'**
- Simplex corrected to compound in TH0: **13 times**
  - ***Engineering*** > TH0: ***Ingenieurswissenschaft*** **'eng. science'**
  - ***Tabloiden*** > TH0: ***Boulevardzeitung*** **'tabloid'**
  - ***Punkten*** **'points'** > TH0: ***Punktezahl*** **'point-number'**

➢ Learners use compounds too rarely? → CIA

Error Analysis

- Error classes:
  (multiple classes possible for same compound)



Compounding errors - lexical

- Errors are rare:
  - Lexeme error: ***Heimstadt*** > TH0: ***Heimatstadt*** **'hometown'**
  - Blocking: ***Strafschein*** > TH0: ***Strafzettel*** **'penalty ticket'**
  - Wrong part of speech: ***Finanziellewelt*** **'financial-world'** > TH0: ***Finanzwelt*** **'finance world'**
- Almost no inappropriate constructions by advanced learners, e.g.:
  - # ***Umgebungswald*** **'surroundings-forest'** > TH2: ***der umliegende Wald*** **'the surrounding forest'**

Contrastive Interlanguage Analysis

- Overuse/Underuse Methodology:
  Examine overuse and underuse of tokens/annotations/constructions compared to L1 controls
- Example from Zeldes et al. (2008): Underuse of adverb chains in L2 data

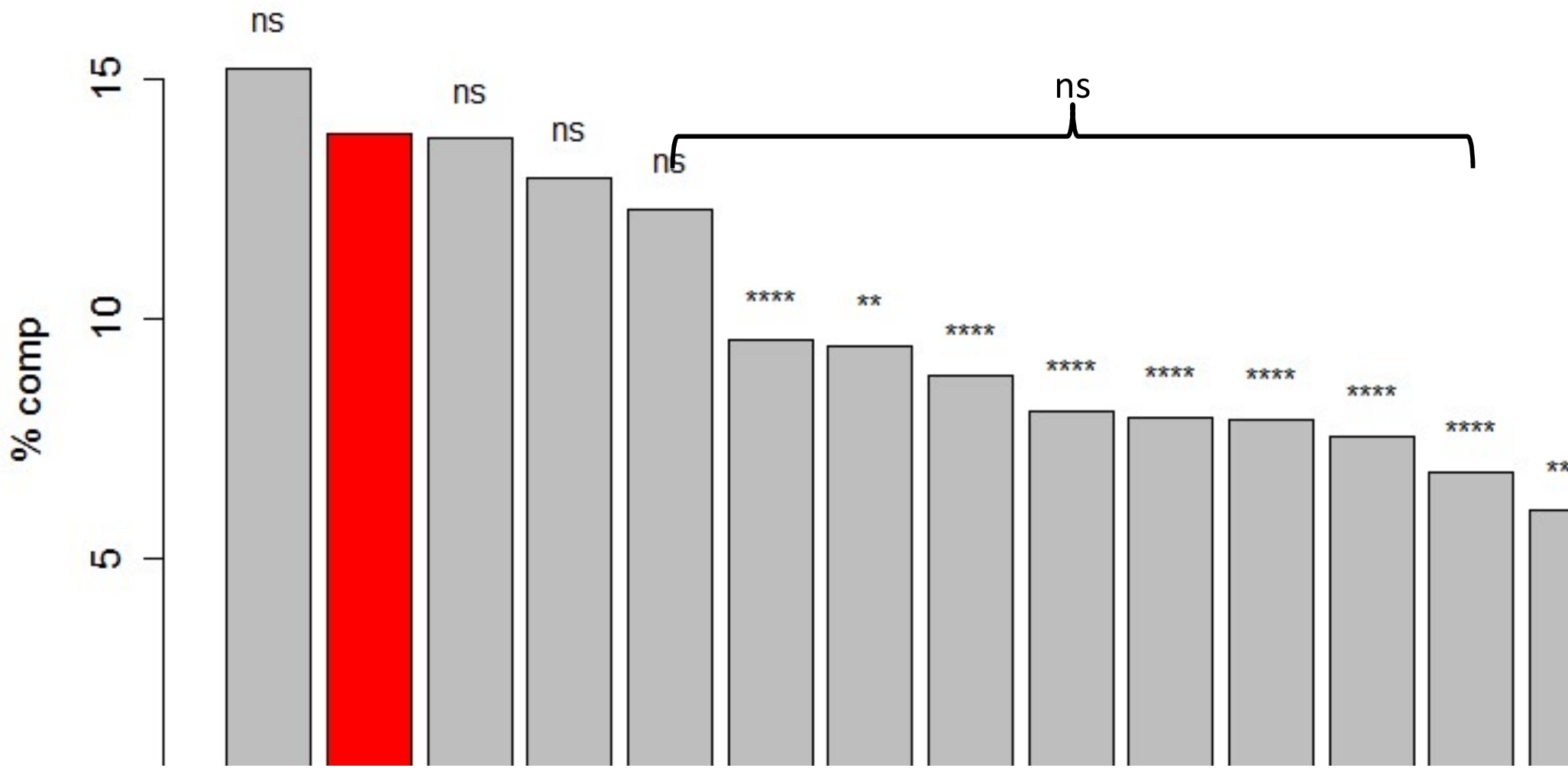
bigram	tot_norm	de	da	en	fr	pl	ru
$.-PPER	0.042384	0.005297	0.009748	0.007963	0.006166	0.005801	0.007409
ADV-ADV	**0.041604**	**0.012858**	**0.010518**	**0.006111**	**0.006166**	**0.003094**	**0.002856**
ADV-APPR	0.039742	0.009117	0.008016	0.005324	0.007837	0.004807	0.004642
PDAT-NN	0.03956	0.005409	0.004233	0.005509	0.007837	0.007735	0.008837

Qualitative analysis

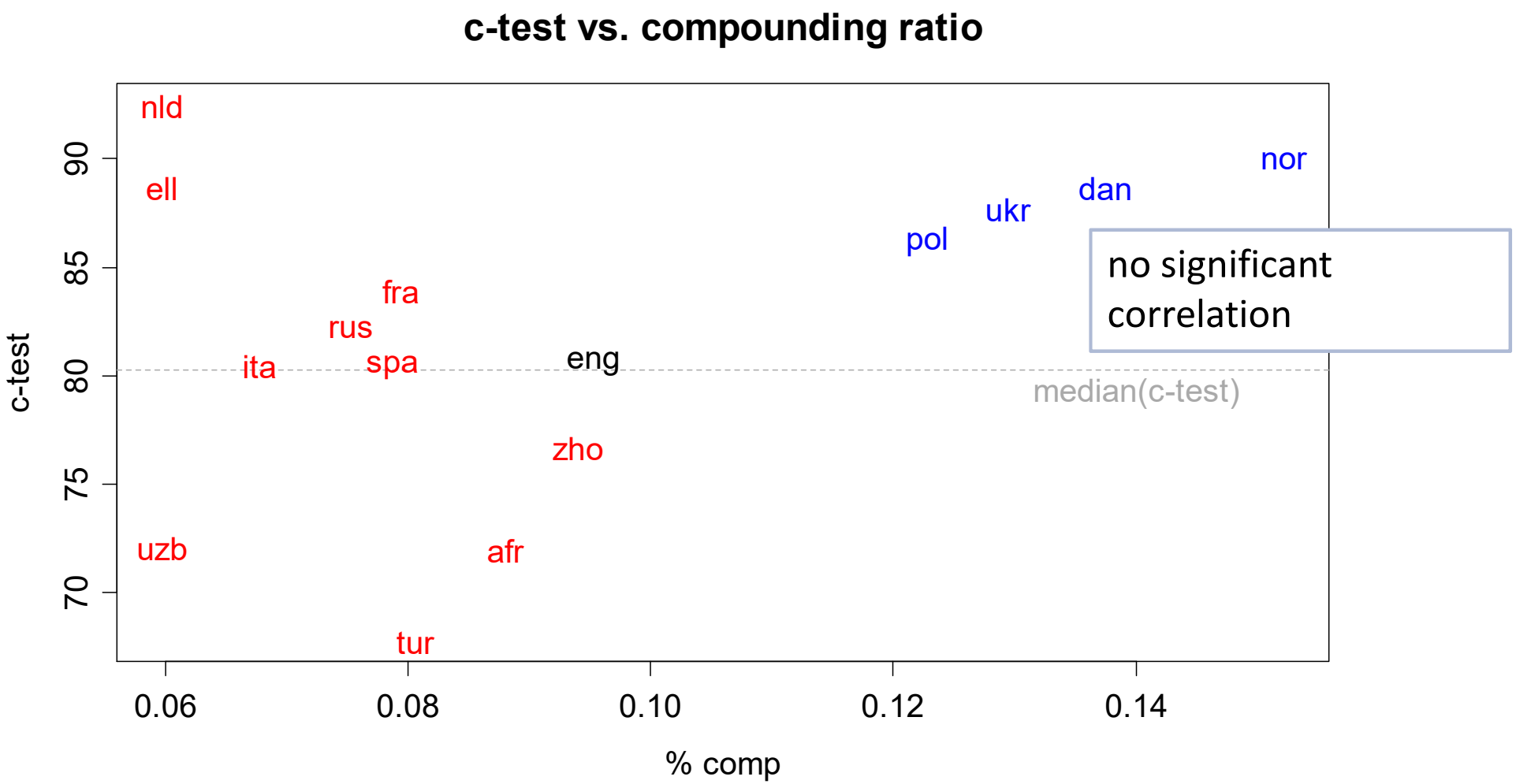
- Every quantitative analysis is based on categorization = qualitative analysis
- What are 'adverb chains' like?

  - *Es ist* [*doch*] [*auch*] *statistisch belegt*
    it is [indeed] [also] statistically attested
  - *ein Kampf, dass bis* [***heute*** *noch*] *andauert*
    a battle which until [today still] continues
  - [[*viel* ***mehr***] *Arbeitsplätze*]
    [much more] work places
  - [*immer noch*] *kann man eine unzufriedenheit spüren*
    [always still] can one sense a dissatisfaction

CIA – Do all learners use as many compounds? (from Zeldes, to appear)

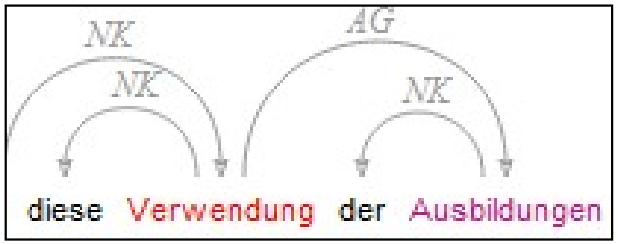


Relationship with proficiency?



Competitors?

- If compounds are underused, what replaces them?
- Several nominal paraphrases:
  - Prepositions: pos="APPR"
  - Attributive adjectives: pos="ADJA"
  - Genitives attributes: ->dep[func="AG"]
  - Relative clauses: ->dep[func="RC"]
- Are these overused?



ZH0	Definitionen	von	Feminismus
ZH0Diff	CHA		
ZH0lemma	Definition	von	Feminismus
ZH0pos	NN	APPR	NN
tok	definitionen	von	Feminismus

Analytic adnominal constructions

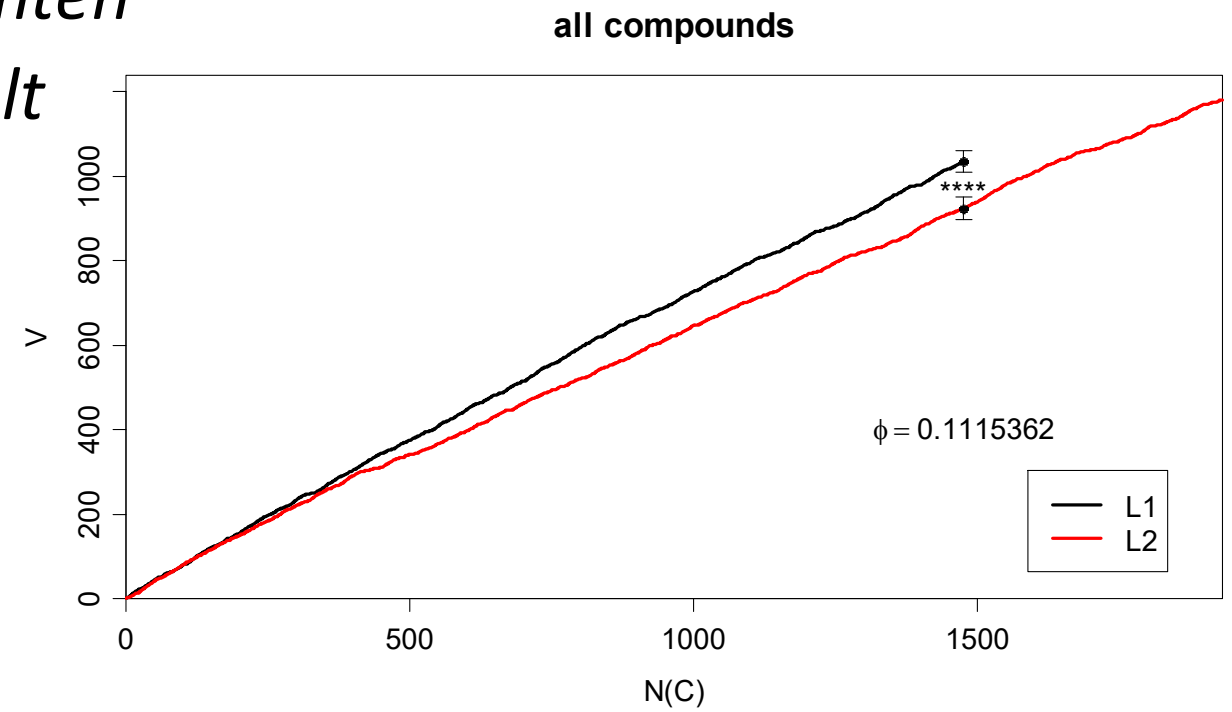
- **ADJA** and **AG** not significant
- Underuse of **RC** and **APPR** but overuse of ***von*** 'of'
  - *Oft in **Zeit von Kriegen*** 'time of wars' > TH2: ***Kriegszeiten*** 'Wartimes'
  - *eine **Position von Autorität*** > TH2: ***Führungsposition*** 'leadership-position'
  - *Schlechte Benützung **von** Zeit und Geld* 'bad use of time and money' > TH2: ***Zeit-** und **Geldverschwendung*** 'time- & money-waste'

	f(L1)	norm(L1)	f(L2)	norm(L2)	±use	p-val
pos=ADJA	2759	0.04029	5068	0.0412801	2.45%	ns
func=AG	771	0.011259	1325	0.0107925	-4.20%	ns
func=RC	843	0.012311	1292	0.0105237	-14.52%	0.00039
pos=APPR	4173	0.060939	7014	0.0571308	-6.25%	0.00069
von	408	0.005958	836	0.0068094	14.29%	0.02848
¬von	3765	0.054981	6178	0.0503213	-8.48%	0.00001

Simply **less modification** (cf. Hirschmann et al. 2013)

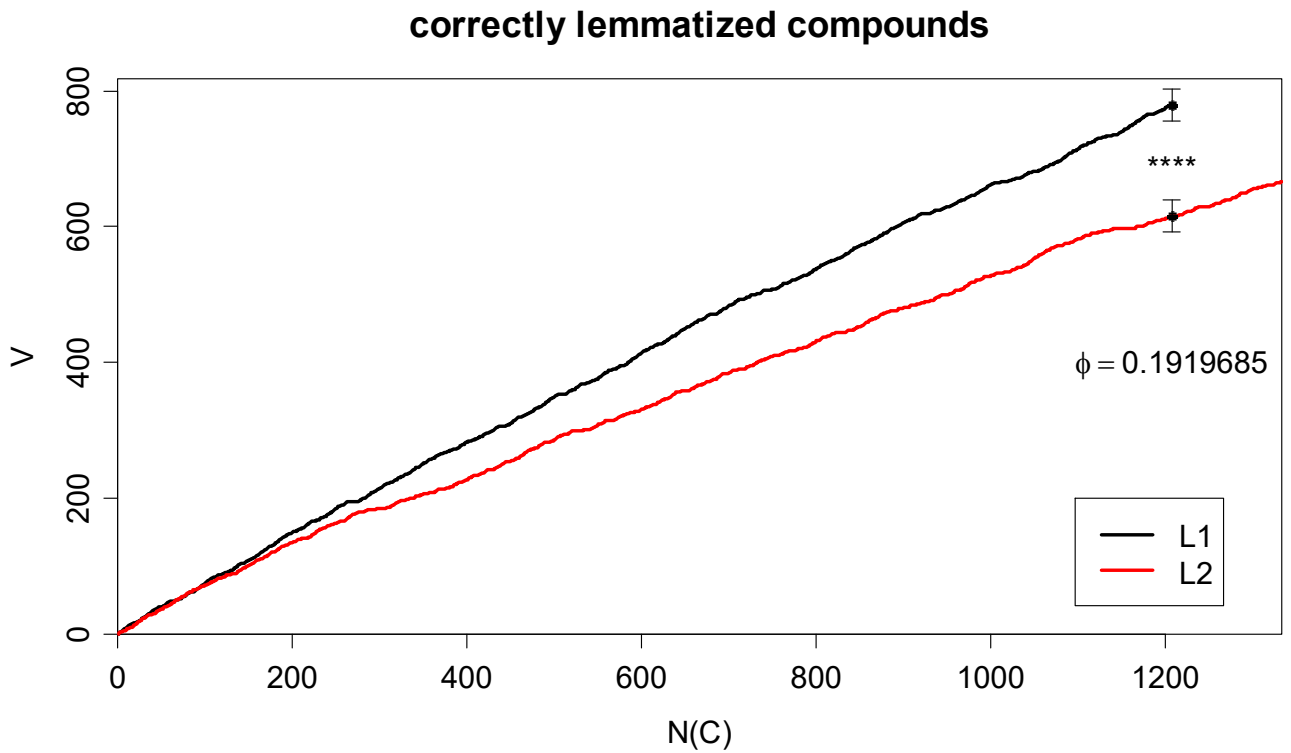
Are L2 compounds productive?

- Use **comparable** L1 corpus
- Significant differences – natives are more productive
- Does an error effect skews the data?
  - *Universitä**a**t**s**tudenten*
  - *Auslandsauf**t**enhalt*
  - *Wahl**en**rechten*
  - …



Productivity – TH compounds

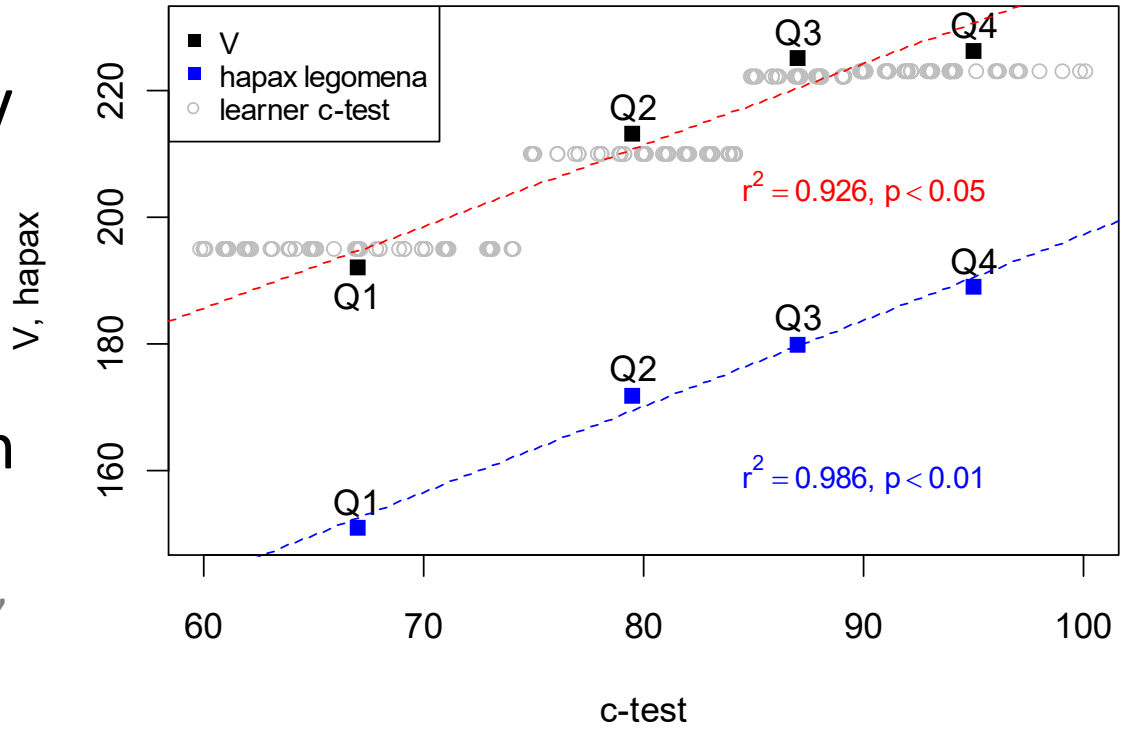
- Effect size substantially bigger (almost double)
- Does this correlate with proficiency too?



Does productivity correlate with proficiency?

- Texts too short to measure vocabulary per learner
  - Solution: bin learners into c-test score quartiles (bottom 25%, lower mid, upper mid, top 25%)
- Strong correlations
- Advanced learners have larger vocabulary
- Proportion of hapax legomena even more striking
  - Skewed distribution in advanced L2

    (cf. Ellis & Ferreira-Junior 2009, Wulff et al. 2009)



Conclusion

- Corpus and annotation architecture determine what we can do with non-standard data
- Error analysis
  - Error categories
  - Theoretical issues: target hypotheses
  - Empirical problems: textual reuse, digitization, task propoerties…
- Many pitfalls, but great potential for analyzing different forms of ‘deviations’

Next

- We've seen fine-grained word alignment and sentence alignment
- Now we will look at full hierarchical alignment:
  - all sentence constituents
  - annotation of alignment types
- Work with the Stockholm TreeAligner:
  - http://kitt.cl.uzh.ch/kitt/treealigner/wiki/TreeAlignerDownload