Instructed second language acquisition and longitudinal learner corpus research: The case of lexical and syntactic complexity

Nina Vyatkina
University of Kansas

Hagen Hirschmann,
Felix Golcher
Humboldt-Universität zu Berlin

TaLC XII
Giessen, July 21, 2016

Overview

- Research goal:
  - Map development of L2 lexical complexity onto development of syntactic complexity explored in our earlier study
- Methodological question:
  - How can we describe the development of L2 writing complexity in early learners in an instructed setting?

Theoretical background

- Usage-Based Grammar
  - languages are learned primarily bottom-up: from specific examples to low-scope patterns to abstract constructions
  - inseparability of grammar and the lexicon

Bybee 2008; Ellis 2014; Flowerdew, 2011; Langacker 1987; Ortega 2015; Robinson & Ellis 2008

- Dynamic Systems Theory
  - L2 development is a dynamic process, in which regular growth stages are modulated by a complex variation within and among individuals as well as interrelated aspects of the interlanguage system

Larsen-Freeman 2006; Verspoor et al. 2008

L2 Complexity

- Measuring learner progress and proficiency – indicators employed in SLA since 1980s (Larsen-Freeman, 1983; Skehan, 1989)
- → **CAF** Measures:
  - **C**omplexity:
    - the extent to which the language produced in performing a task is elaborate and varied (Ellis, 2003)
    - the range of forms that surface in language production and the degree of sophistication of such forms (Ortega, 2003)
  - **A**ccuracy: error-free L2 production
  - **F**luency: speed of L2 production

L2 writing complexity research

- Primarily explored structural measures of syntactic and lexical complexity:
  - Syntactic complexity: length and ratios of syntactic units
    - words, clauses / sentences, T-units…
  - Lexical complexity: ratios measuring word diversity, density, and sophistication
    - type-token ratios, content words/functional words, rare words/common words, …
- Research syntheses: linear increase in some but not all measures with increasing proficiency; complex interactions between measures

  Wolfe-Quintero et al. (1998), Ortega (2003), Norris & Ortega (2009), Bulte & Housen (2012), Connor-Linton & Polio (2014)

Designs

- Many complexity studies:
  - cross-sectional or single-case longitudinal
  - manual annotation of selected features
- This study:
  - longitudinal corpus, multiple learner profiles
  - automatic corpus-based profiling (POS and lemma annotation)

    Granger & Rayson 1998; Hawkins & McCarthy 2010; Ortega & Sinicrope 2008

Data: subset of KANDEL

KANDEL is a pos-annotated, lemmatized, and error-annotated open access learner corpus
https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/korpuslinguistik/research/kandel

This study: longitudinal KANDEL subset

Metadata	
setting	Instructed SLA, large public US university
participants	12 students (5 male, 7 female)
age	18-22 (mean 19.5), 1 learner >30
languages	L1 English, L2 German (beginner to A2 CEFR proficiency)
time	4 semesters, 17 data collection points (every 3-5 weeks)
texts	185 rough drafts in-class and at-home L2 essays (personal narratives and descriptions; essays with explanatory elements; letters)
text length	100-200 words (mean 161)

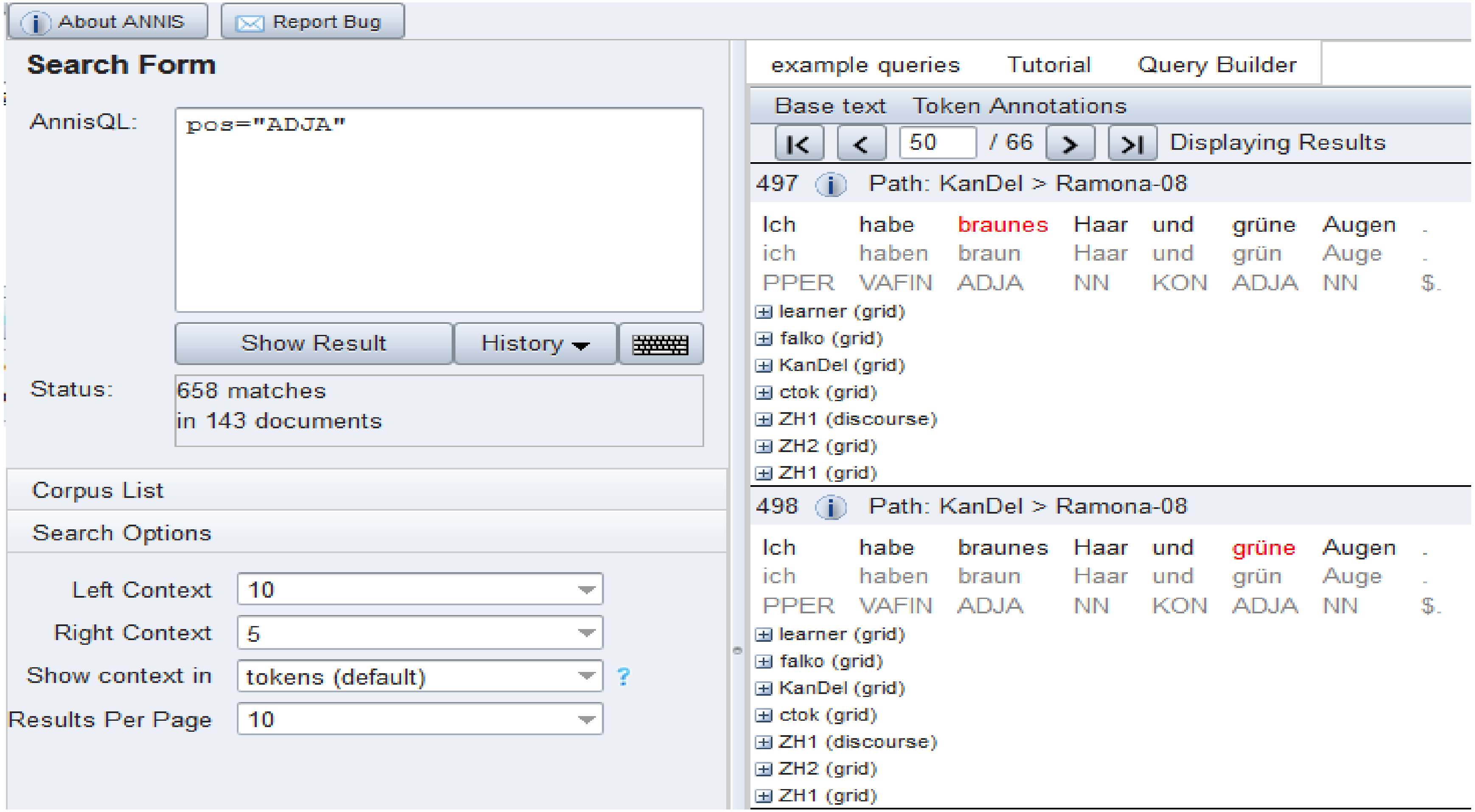
Research question and hypothesis

- RQ: Does the observed development of specific word classes (syntactic modifiers) correlate with lexical development?
- RH: Lexical richness is verifiably increasing over time, independently of growth curve of syntactic categories

Lexical complexity measures

- Structural measures:
  - ~~Lexical density~~
  - ~~Lexical sophistication~~
  - Lexical diversity (TTR and type frequency)
- Content-based measures:
  - lexical novelty (emergent words)
  - specific content words as specific syntactic modifiers (cf. Ortega & Sinicrope 2008)
  - semantic-functional aspects (cf. Ortega 2015; Brandes & Ravid 2016)

KanDeL in ANNIS – sample search



AnnisQL: `pos="ADJA"`

Status: 658 matches
in 143 documents

Left Context: 10

Right Context: 5

Show context in: tokens (default) ?

Results Per Page: 10

50 / 66 Displaying Results

497 Path: KanDel > Ramona-08

Ich	habe	braunes	Haar	und	grüne	Augen	.
ich	haben	braun	Haar	und	grün	Auge	.
PPER	VAFIN	ADJA	NN	KON	ADJA	NN	$.

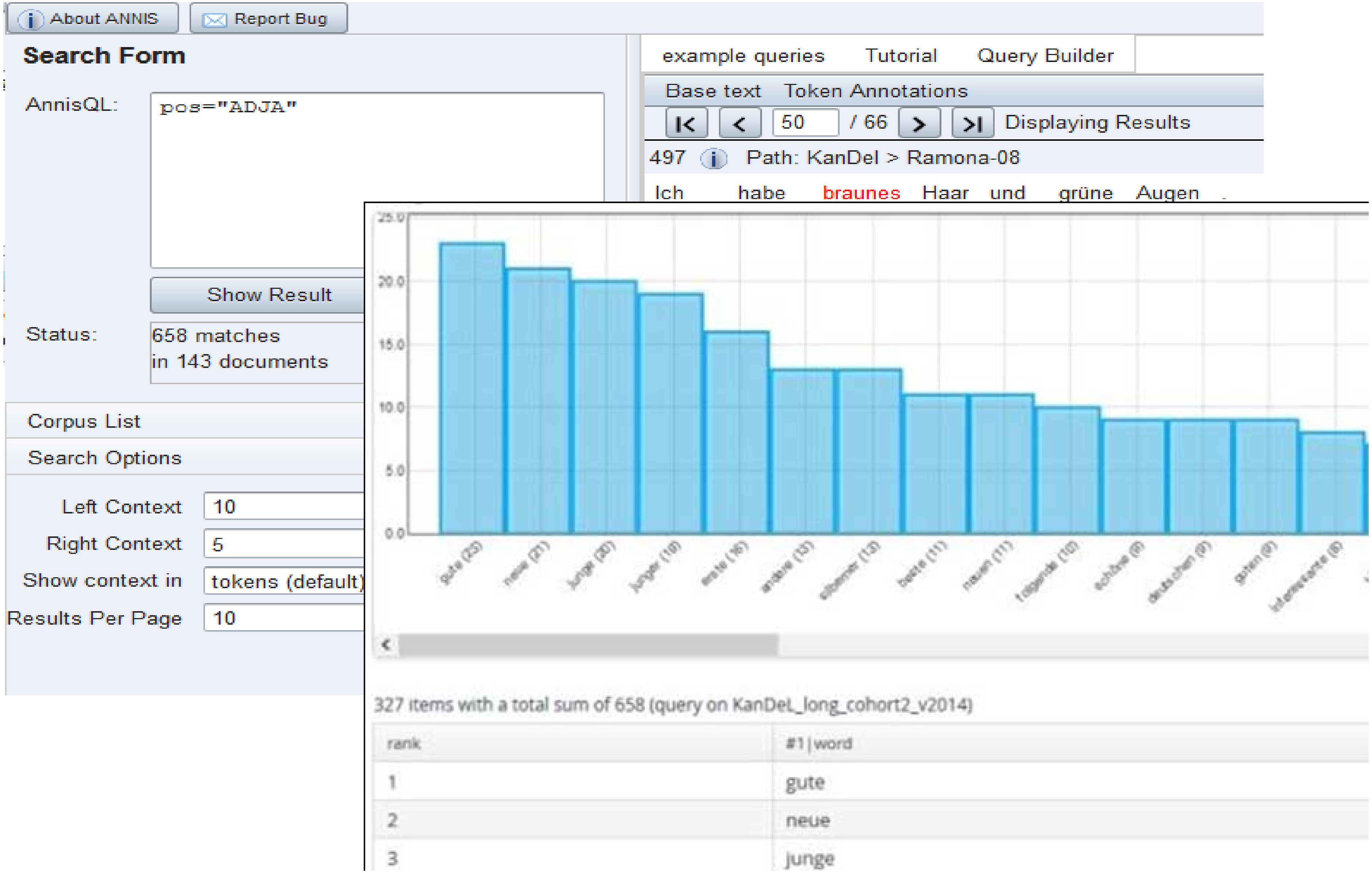
⊞ learner (grid)
⊞ falko (grid)
⊞ KanDel (grid)
⊞ ctok (grid)
⊞ ZH1 (discourse)
⊞ ZH2 (grid)
⊞ ZH1 (grid)

498 Path: KanDel > Ramona-08

Ich	habe	braunes	Haar	und	grüne	Augen	.
ich	haben	braun	Haar	und	grün	Auge	.
PPER	VAFIN	ADJA	NN	KON	ADJA	NN	$.

⊞ learner (grid)
⊞ falko (grid)
⊞ KanDel (grid)
⊞ ctok (grid)
⊞ ZH1 (discourse)
⊞ ZH2 (grid)
⊞ ZH1 (grid)

KanDeL in ANNIS – sample search



Search Form

AnnisQL: `pos="ADJA"`

Status: 658 matches in 143 documents

Left Context: 10

Right Context: 5

Show context in: tokens (default)

Results Per Page: 10

50 / 66 Displaying Results

497 Path: KanDel > Ramona-08

Ich habe braunes Haar und grüne Augen .

327 items with a total sum of 658 (query on KanDeL_long_cohort2_v2014)

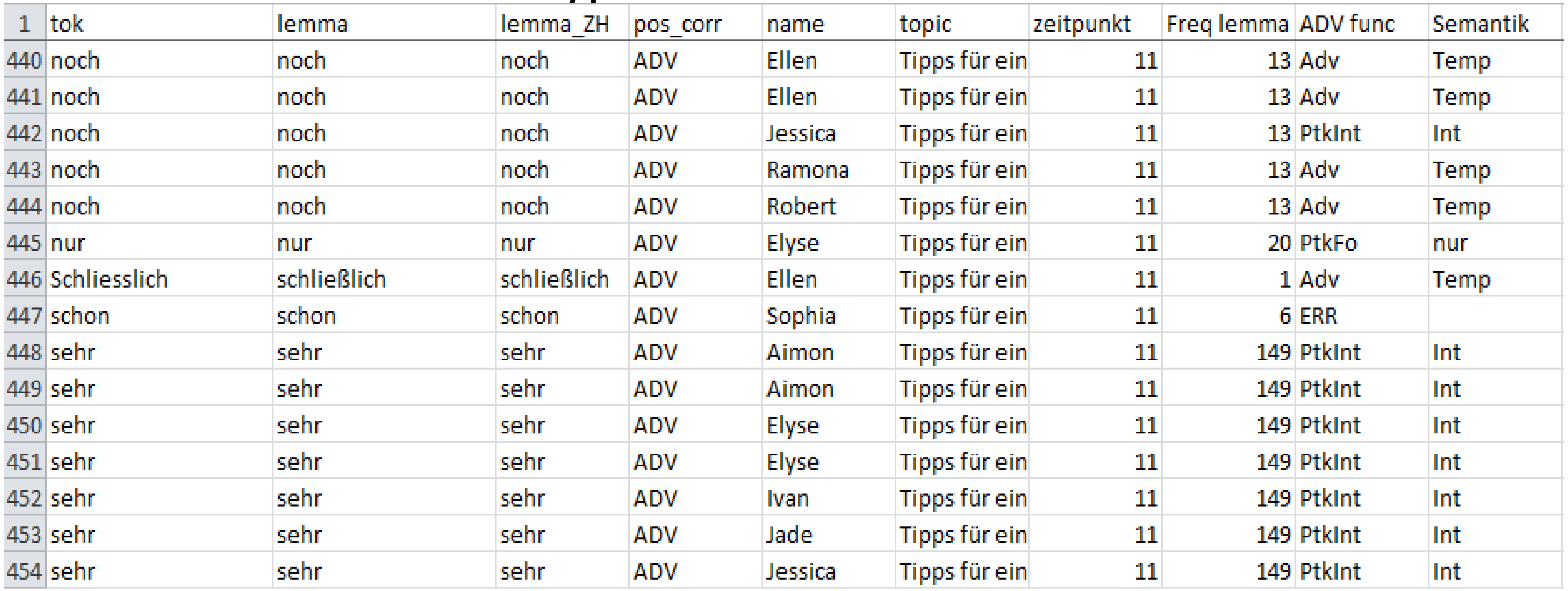
rank	#1 | word
1	gute
2	neue
3	junge

Procedure

- Focusing on modifier categories
  1. 'prenominal adjective',
  2. 'predicative adjective',
  3. 'adverb'
  - very general categories, contain different syntactic and semantic types
- Processing steps for study:
  - Export all relevant tokens with sentence contexts
  - Annotating individual tokens in MS Excel: functional syntactic and semantic categories
    - Excluding all erroneous tokens that cannot be interpreted (orthographic vs. grammatical errors)
  - Data analyses using R and MS Excel tables

Procedure

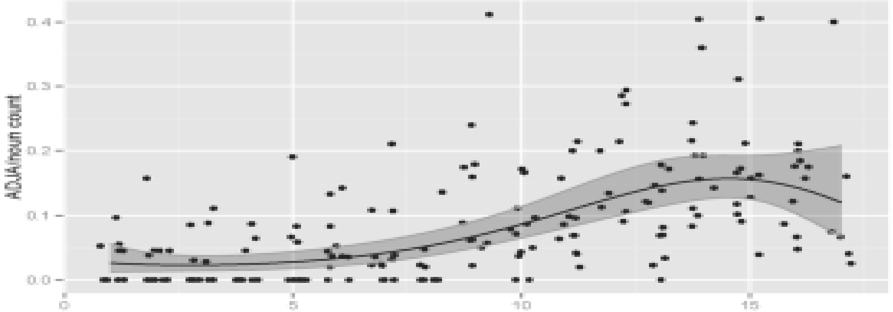
- Focusing on modifier categories
  1. 'prenominal adjective',
  2. 'predicative adjective',
  3. 'adverb'
  - very general categories, contain different syntactic and semantic types



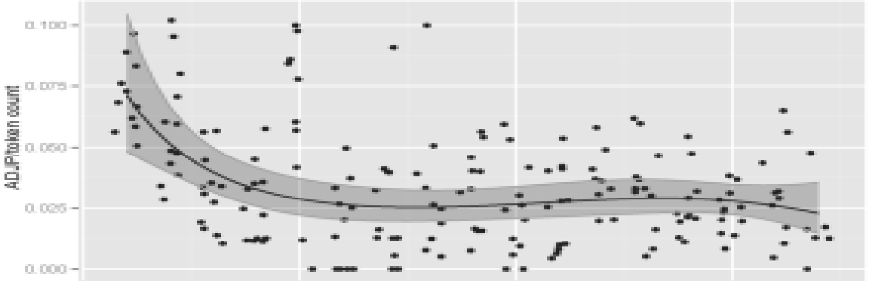
1	tok	lemma	lemma_ZH	pos_corr	name	topic	zeitpunkt	Freq lemma	ADV func	Semantik
440	noch	noch	noch	ADV	Ellen	Tipps für ein	11	13	Adv	Temp
441	noch	noch	noch	ADV	Ellen	Tipps für ein	11	13	Adv	Temp
442	noch	noch	noch	ADV	Jessica	Tipps für ein	11	13	PtkInt	Int
443	noch	noch	noch	ADV	Ramona	Tipps für ein	11	13	Adv	Temp
444	noch	noch	noch	ADV	Robert	Tipps für ein	11	13	Adv	Temp
445	nur	nur	nur	ADV	Elyse	Tipps für ein	11	20	PtkFo	nur
446	Schliesslich	schließlich	schließlich	ADV	Ellen	Tipps für ein	11	1	Adv	Temp
447	schon	schon	schon	ADV	Sophia	Tipps für ein	11	6	ERR	
448	sehr	sehr	sehr	ADV	Aimon	Tipps für ein	11	149	PtkInt	Int
449	sehr	sehr	sehr	ADV	Aimon	Tipps für ein	11	149	PtkInt	Int
450	sehr	sehr	sehr	ADV	Elyse	Tipps für ein	11	149	PtkInt	Int
451	sehr	sehr	sehr	ADV	Elyse	Tipps für ein	11	149	PtkInt	Int
452	sehr	sehr	sehr	ADV	Ivan	Tipps für ein	11	149	PtkInt	Int
453	sehr	sehr	sehr	ADV	Jade	Tipps für ein	11	149	PtkInt	Int
454	sehr	sehr	sehr	ADV	Jessica	Tipps für ein	11	149	PtkInt	Int

Use of modifier categories…

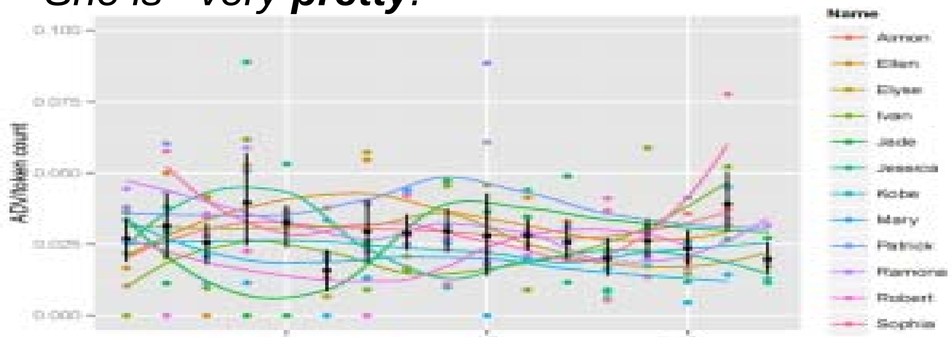
(Vyatkina&Hirschmann&Golcher 2015, no lexical perspective)



*Ich habe die **beste** Familie in der Welt.* (Aimon 03)
*I have the **best** family in the world.*



*Sie ist sehr **schön**.* (Aimon 03)
*She is very **pretty**.*

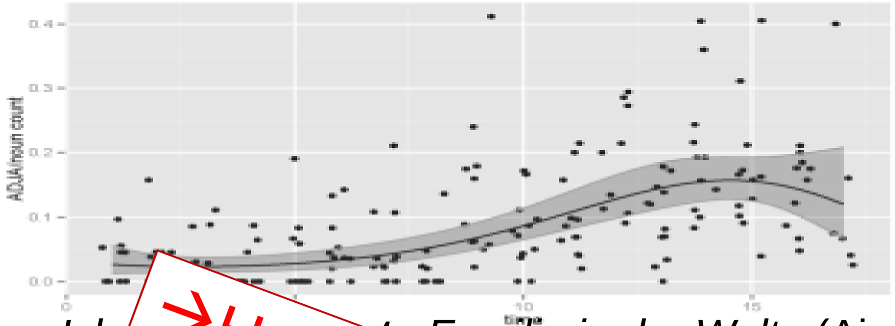


***Gestern** kam Julchen zu mir.* (Patrick 15)
***Yesterday** came Julchen to me.*

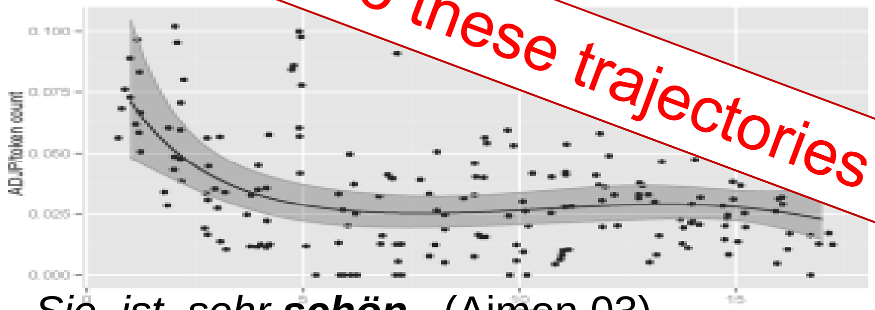
- **Prenominal adjectives** significantly increasing over time (despite great variation)
- **Predicative adjectives** significantly decreasing over time (despite great variation)
- **Adverbs** show no significant trend

Use of modifier categories…

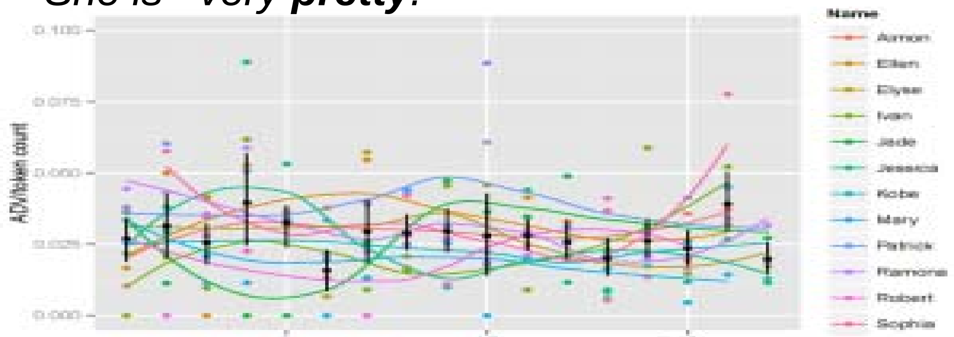
(Vyatkina&Hirschmann&Golcher 2015, no lexical perspective)



Ich [illegible] ***te** Familie in der Welt.* (Aimon 03)
I have th[illegible]*ily in the world.*



*Sie ist sehr **schön**.* (Aimon 03)
*She is very **pretty**.*

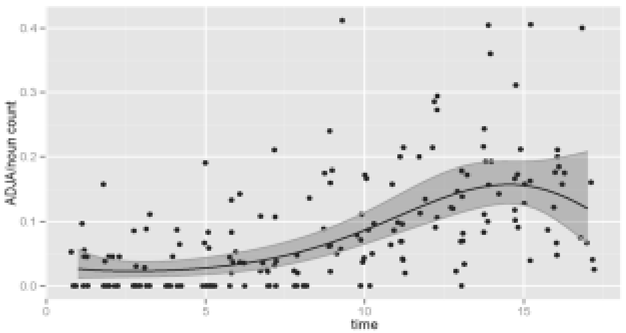


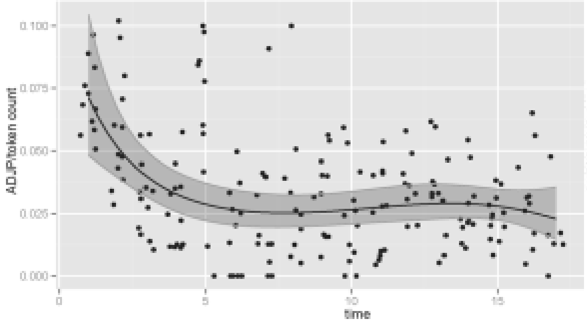
***Gestern** kam Julchen zu mir.* (Patrick 15)
***Yesterday** came Julchen to me.*

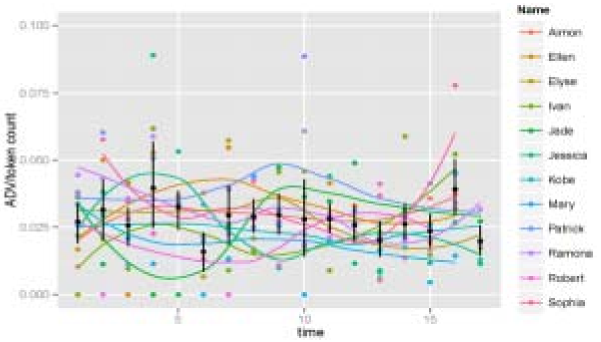
- **Prenominal adjectives** significantly increasing over time (despite great variation)
- **Predicative adjectives** significantly decreasing over time [illegible] great variation)
- **Adverbs** show no significant trend

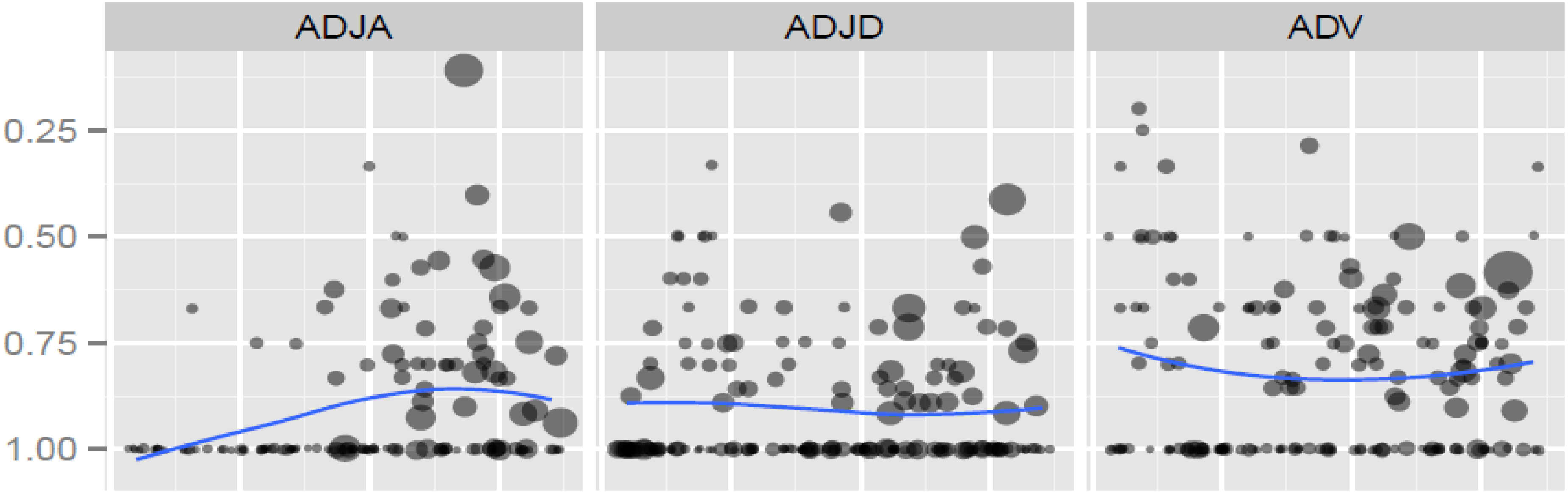
→How do these trajectories correspond with lexical use?

Results: TTR



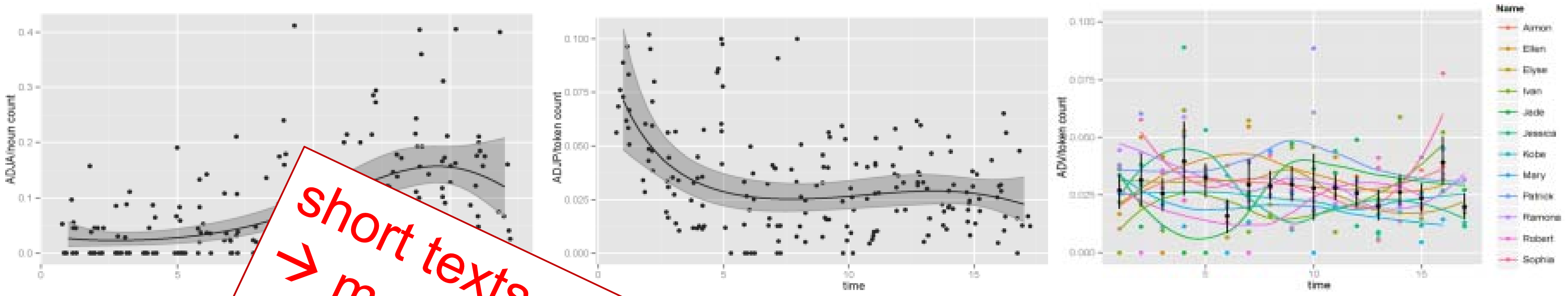


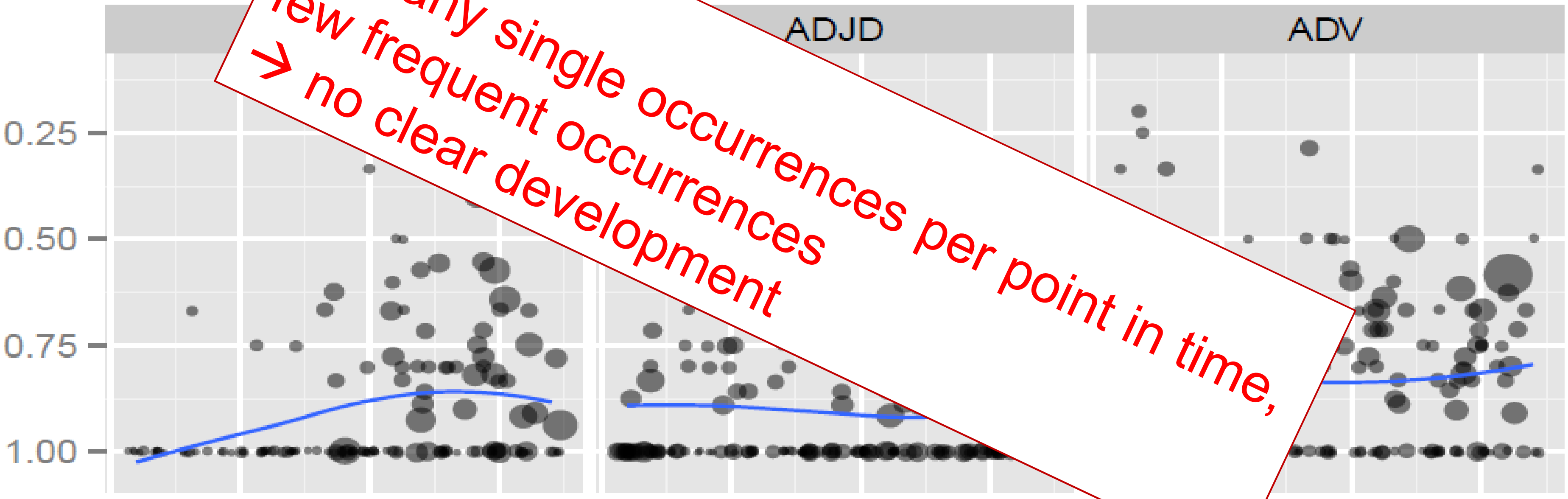




- TTR over time: black dots → TTR per text and point in time (bigger dots symbolize longer texts)

Results: TTR

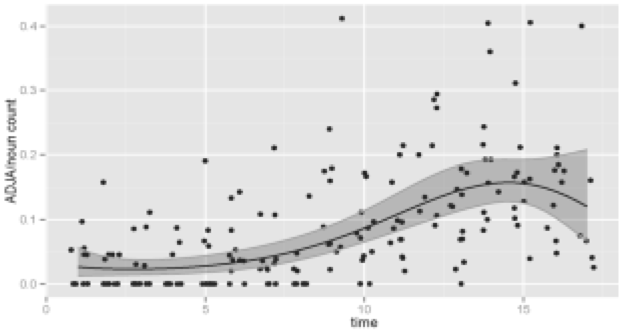


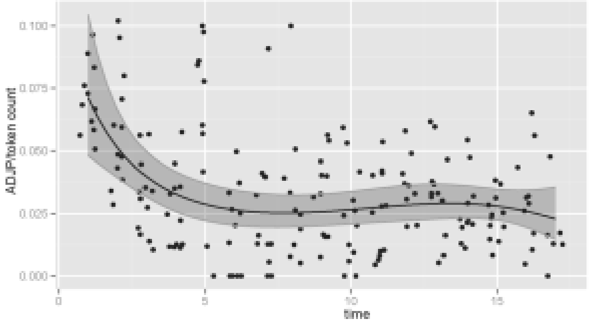


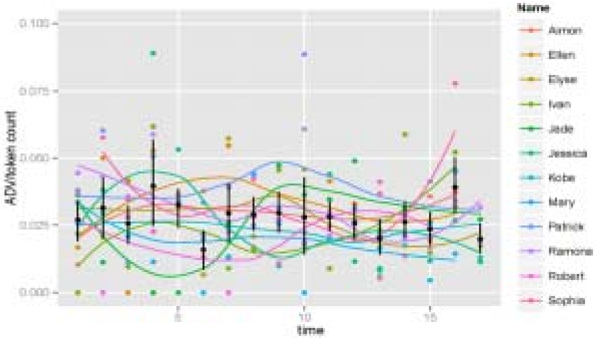
short texts
→ many single occurrences per point in time,
few frequent occurrences
→ no clear development

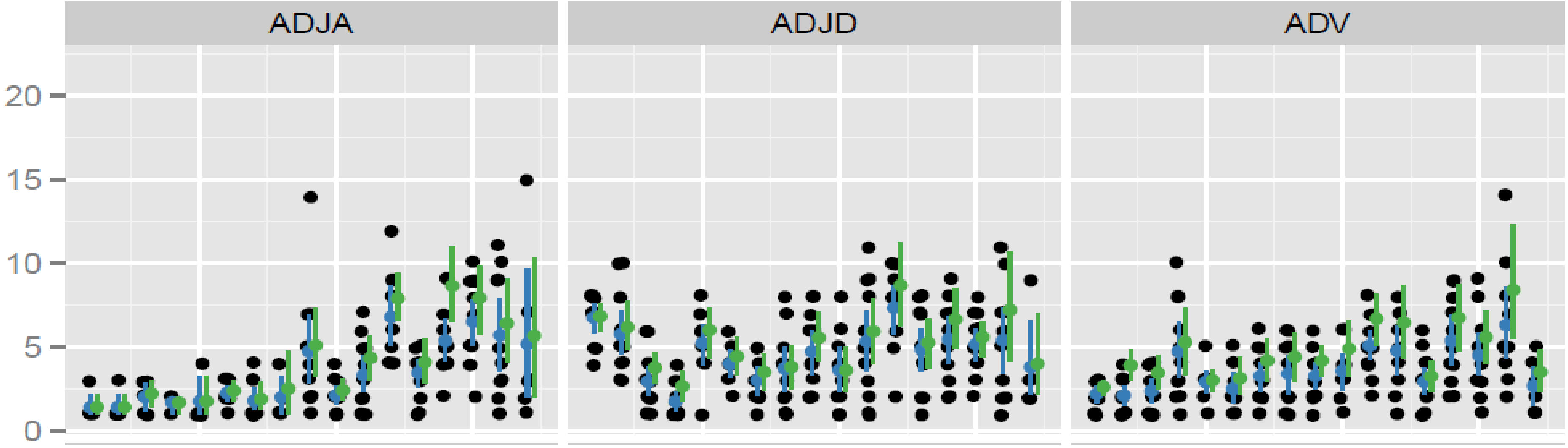
- TTR over time: black dots → TTR per text and point in time (bigger dots symbolize longer texts)

Results: Types per text



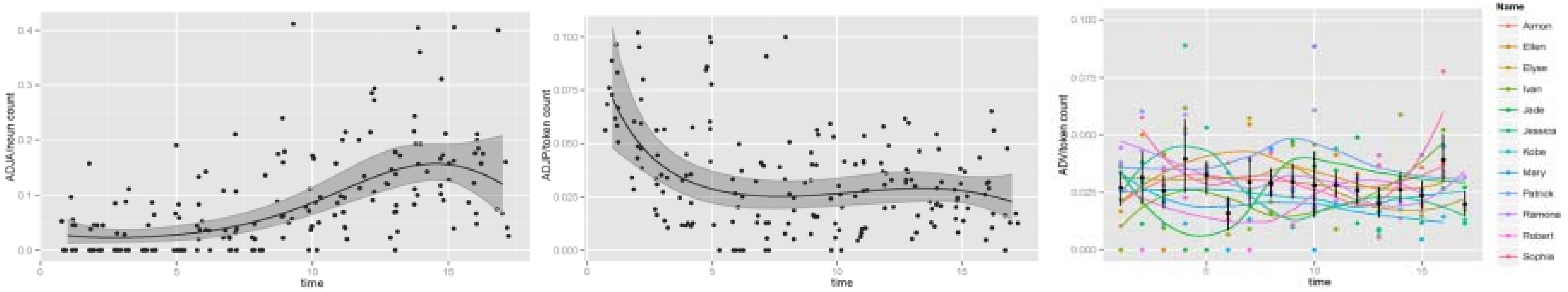


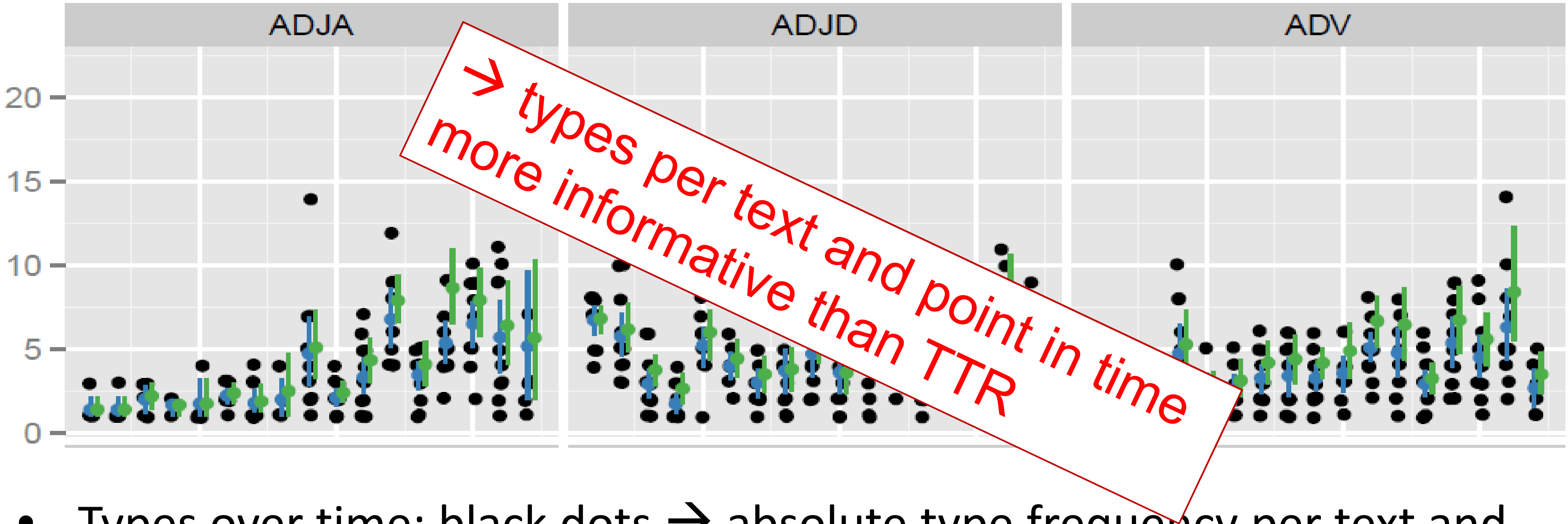




- Types over time: black dots → absolute type frequency per text and point in time, **black** dots: individual texts, blue dots: mean type values for group, green dots: mean token values

Results: Types per text

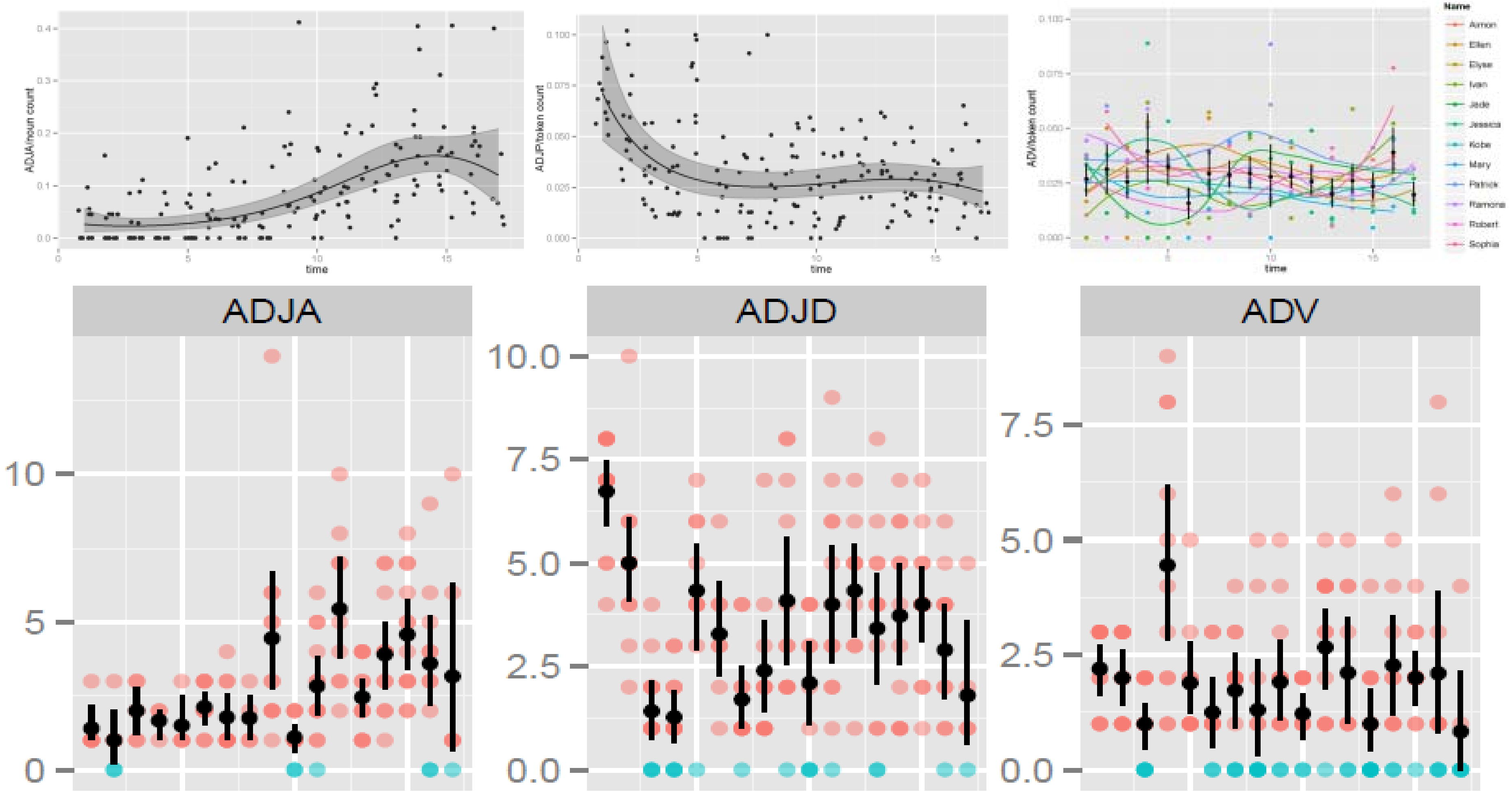




→ types per text and point in time more informative than TTR

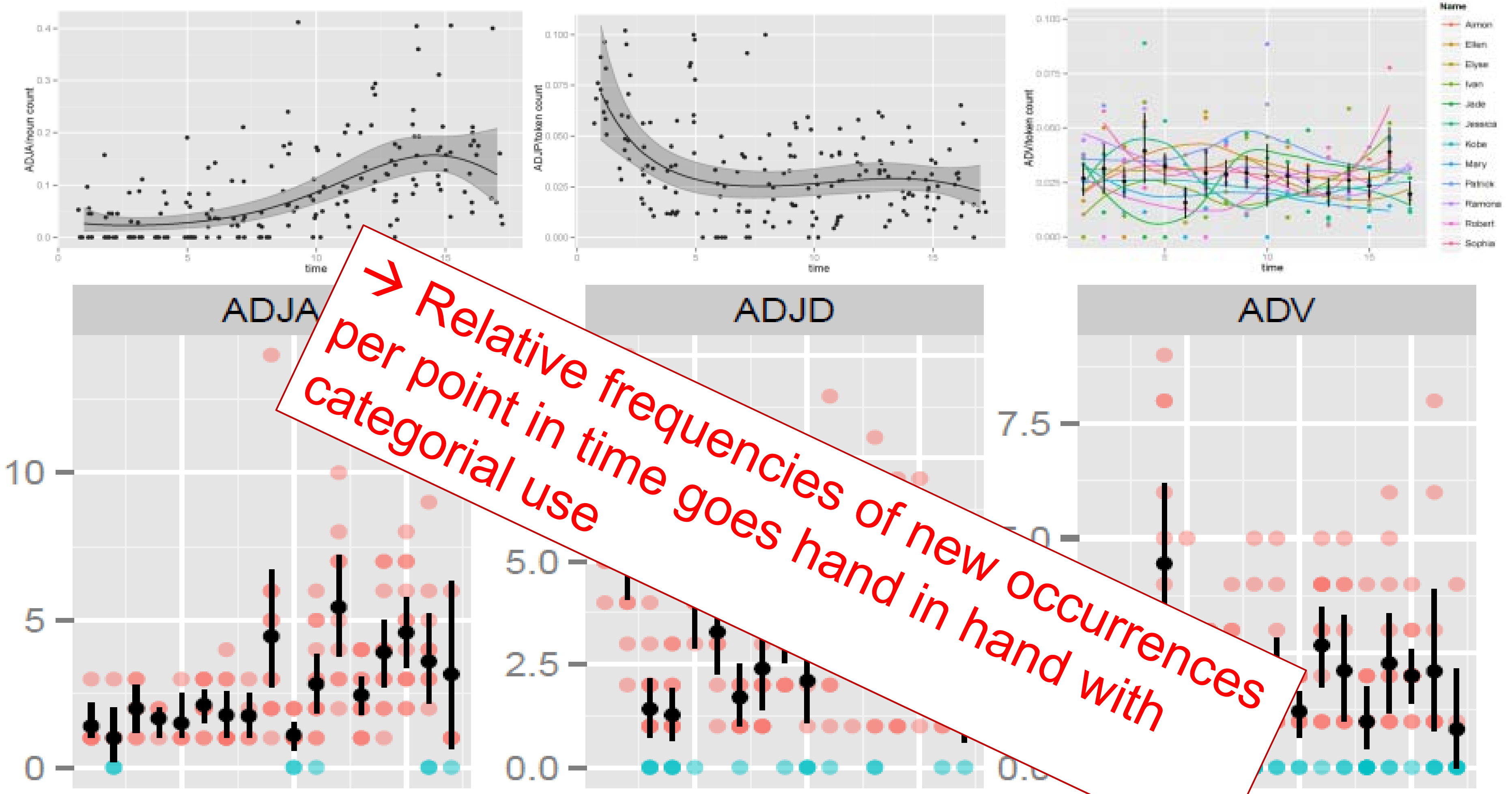
- Types over time: black dots → absolute type frequency per text and point in time, **black** dots: individual texts, blue dots: mean type values for group, green dots: mean token values

Results: New types



- New types per point in time and individual person. Red and blue dots: single texts, black dots: mean values for group with bootstrapped confidence intervals

Results: New types

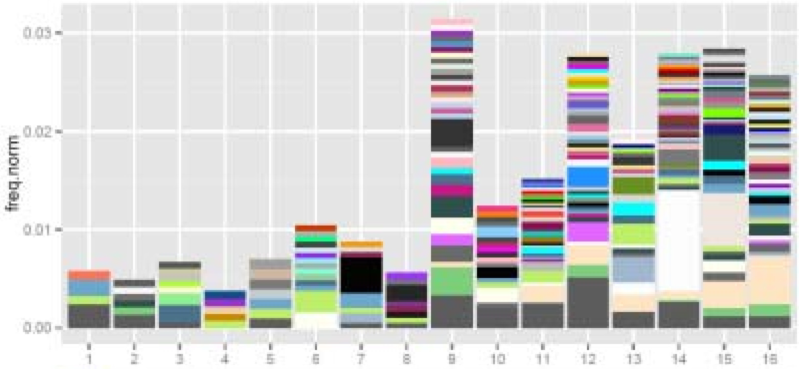


→ Relative frequencies of new occurrences per point in time goes hand in hand with categorial use

- New types per point in time and individual person. Red and b e dots: single texts, black dots: mean values for group with bootstrapped confidence intervals

Individual lexemes per point in time (lexical diversity for whole group)

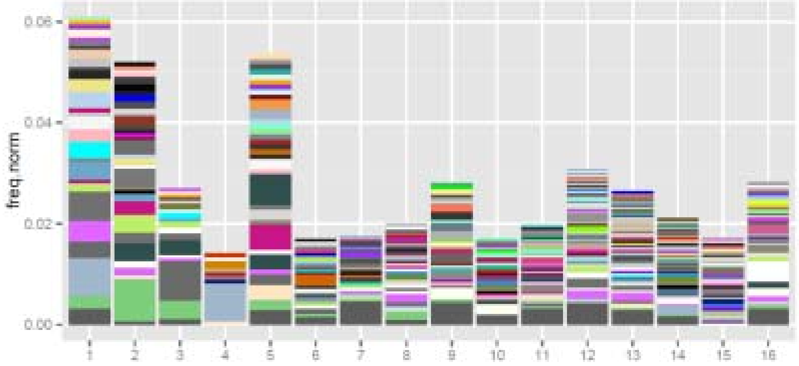
- ADJA



most frequent:

gut (good)	10,5 (65)
neu (new)	6,5 (40)
jung (young)	6,5 (40)
erst (first)	4,1 (25)
silbern (silver)	3,0 (18)

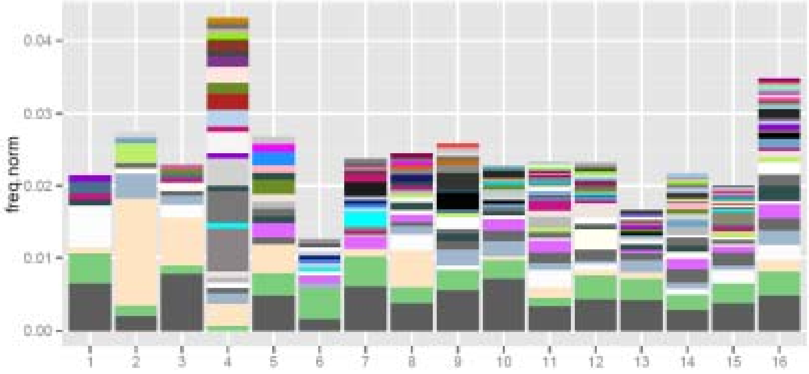
- ADJD



most frequent:

gut (good)	9,7 (83)
groß (big)	4,0 (34)
deutsch (German)	3,4 (29)
interessant (interesting)	3,0 (26)

- ADV

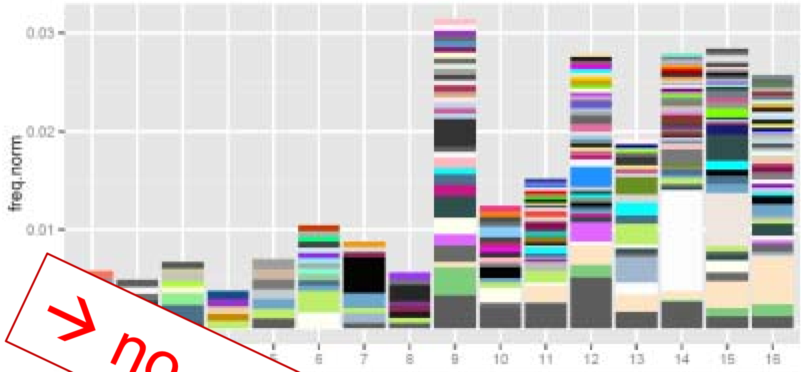


most frequent:

sehr (very)	15,6 (149)
auch (also)	9,6 (92)
gern (with pleasure)	7,0 (67)
jetzt (now)	4, 2 (40)
aber (however)	4,0 (38)

Individual lexemes per point in time (lexical diversity for whole group)

- ADJA

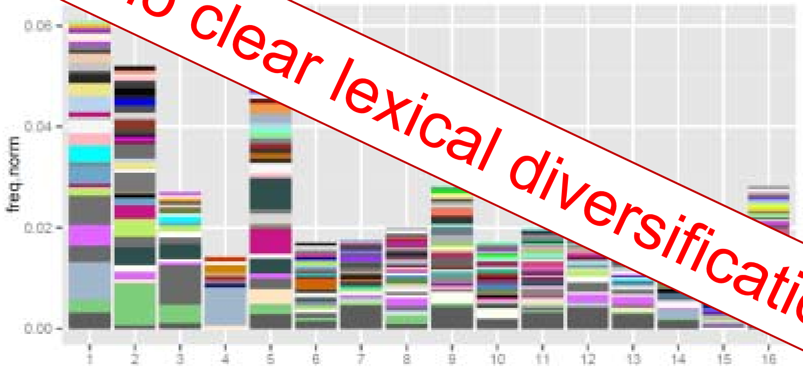


most frequent:

gut (good)	10,5 (65)
neu (new)	6,5 (40)
jung (young)	6,5 (40)
erst (first)	4,1 (25)
silbern (silver)	3,0 (18)

→ no clear lexical diversification over time

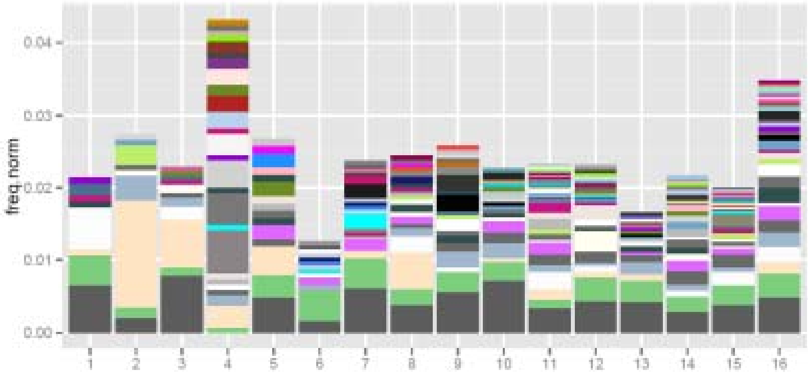
- ADJD



most frequent:

gut (good)	9,7 (83)
groß (big)	4,0 (34)
deutsch (German)	3,4 (29)
interessant (interesting)	3,0 (26)

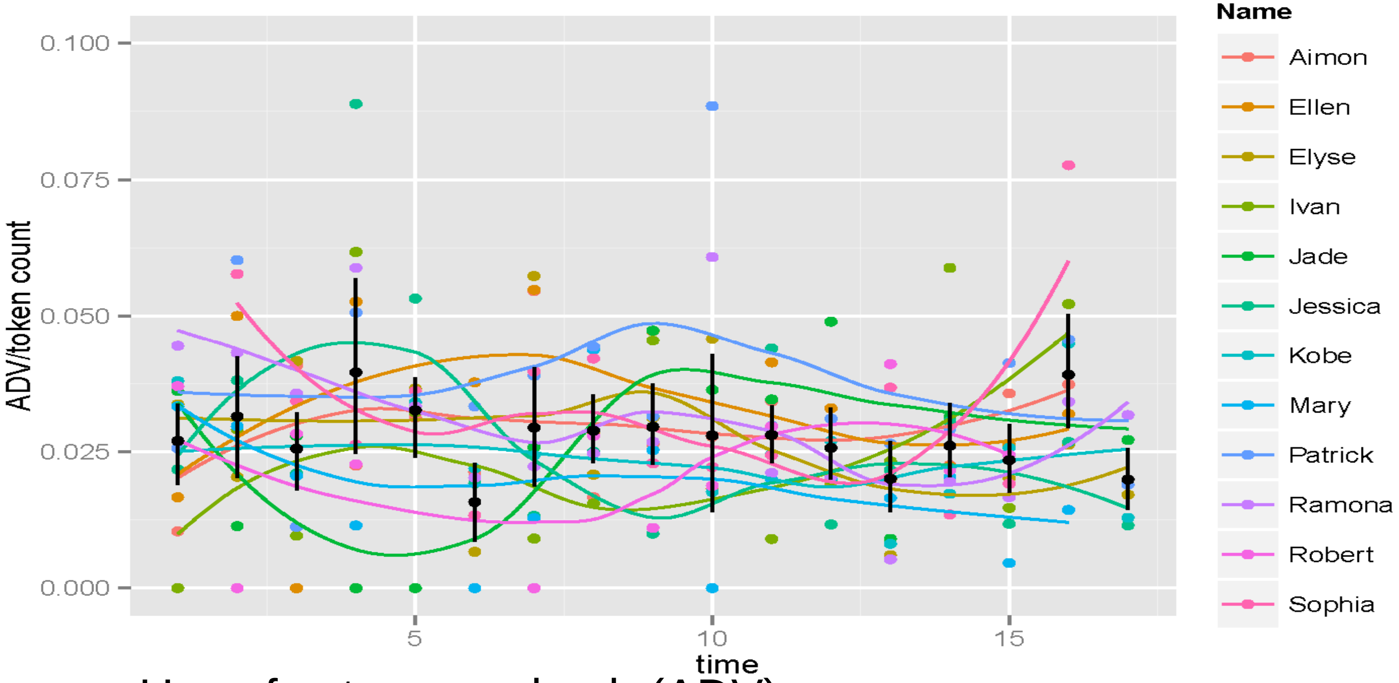
- ADV



most frequent:

sehr (very)	15,6 (149)
auch (also)	9,6 (92)
gern (with pleasure)	7,0 (67)
jetzt (now)	4, 2 (40)
aber (however)	4,0 (38)

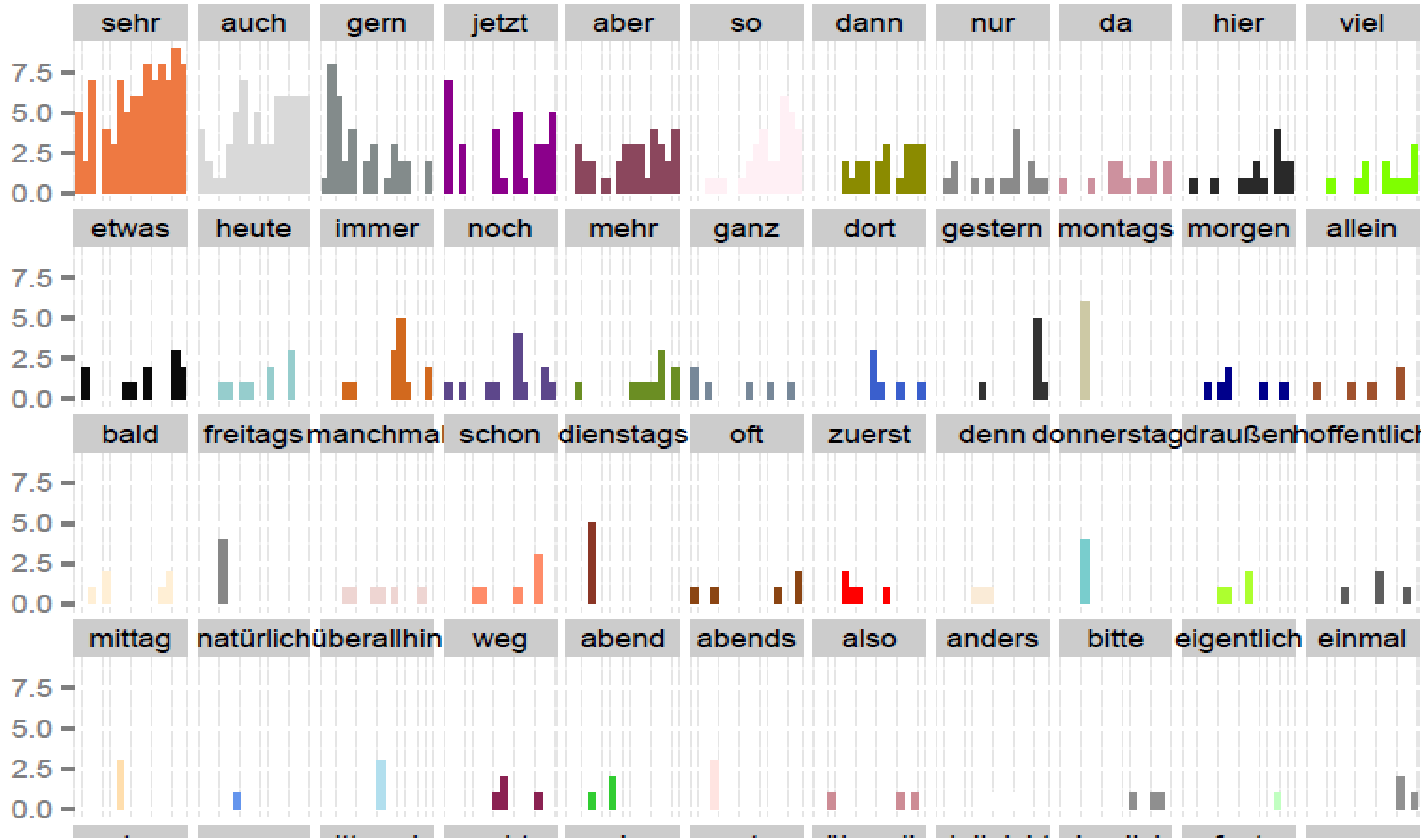
Results for ADV



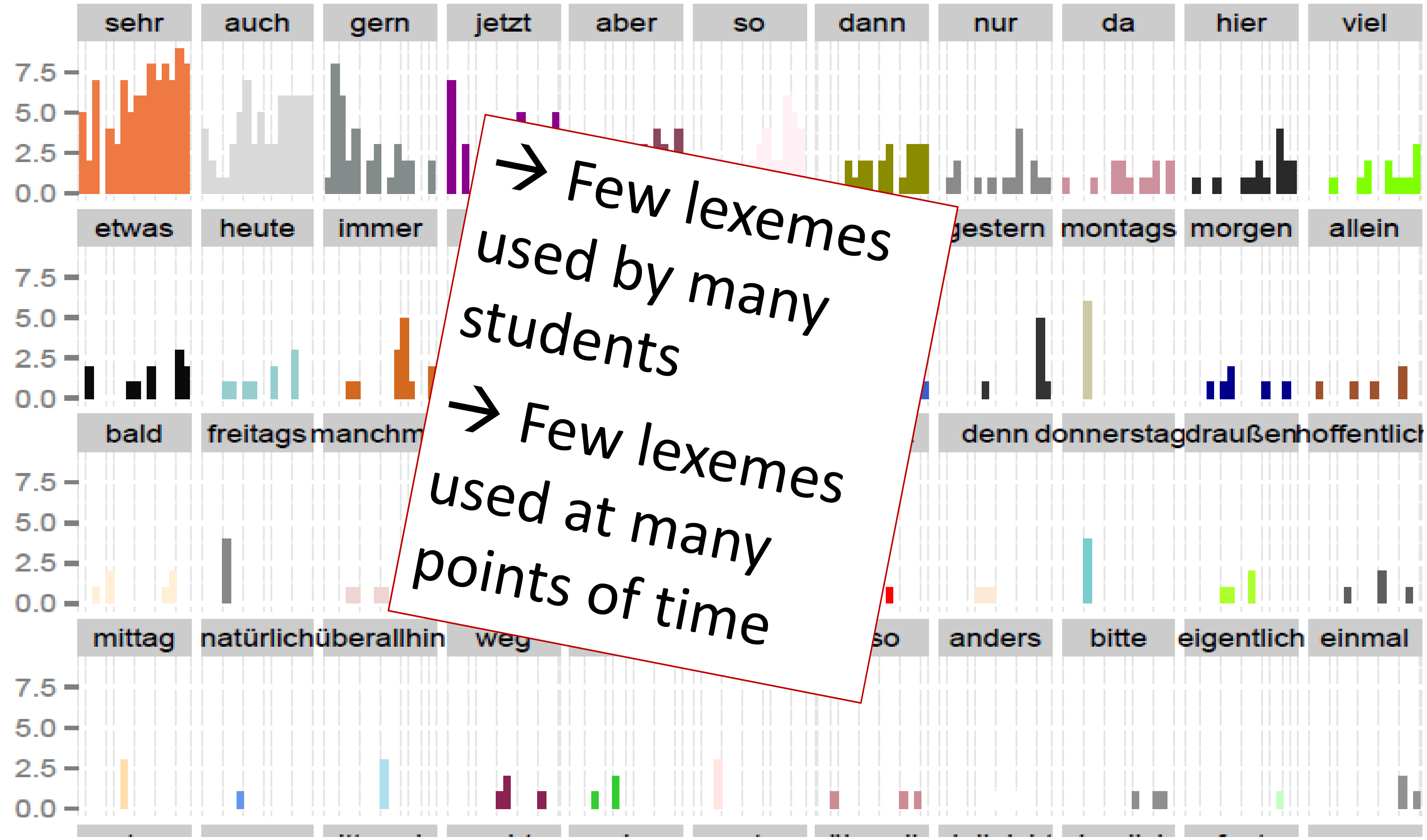
Use of category adverb (ADV)
according to Vyatkina&Hirschmann&Golcher 2015

- Now taking a look at heterogeneous category ADV:
  - – Variation of lexeme use within group
  - – Frequencies of adverb subcategories

ADV lexemes used by number x of learner per point in time



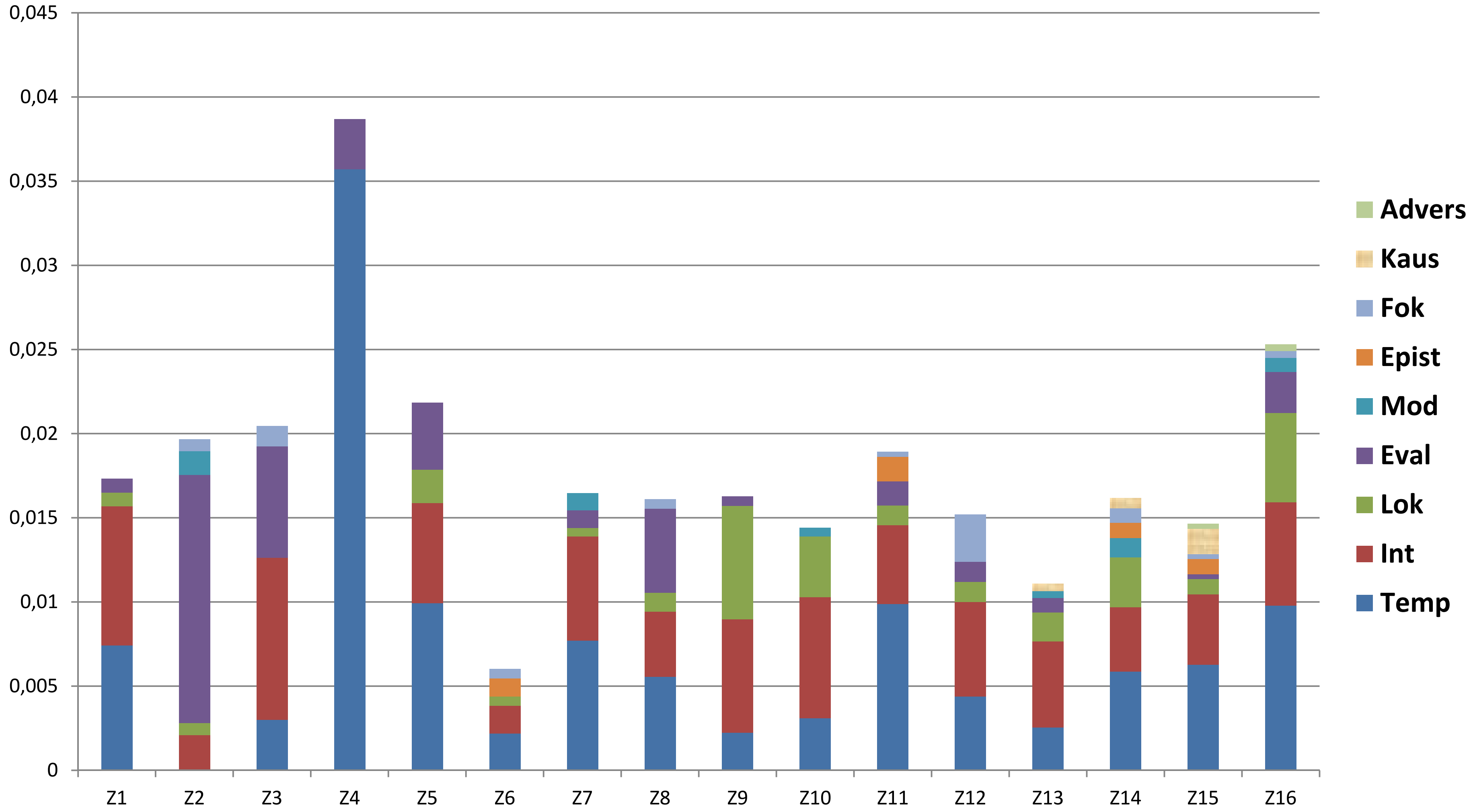
ADV lexemes used by number x of learner per point in time



→ Few lexemes used by many students

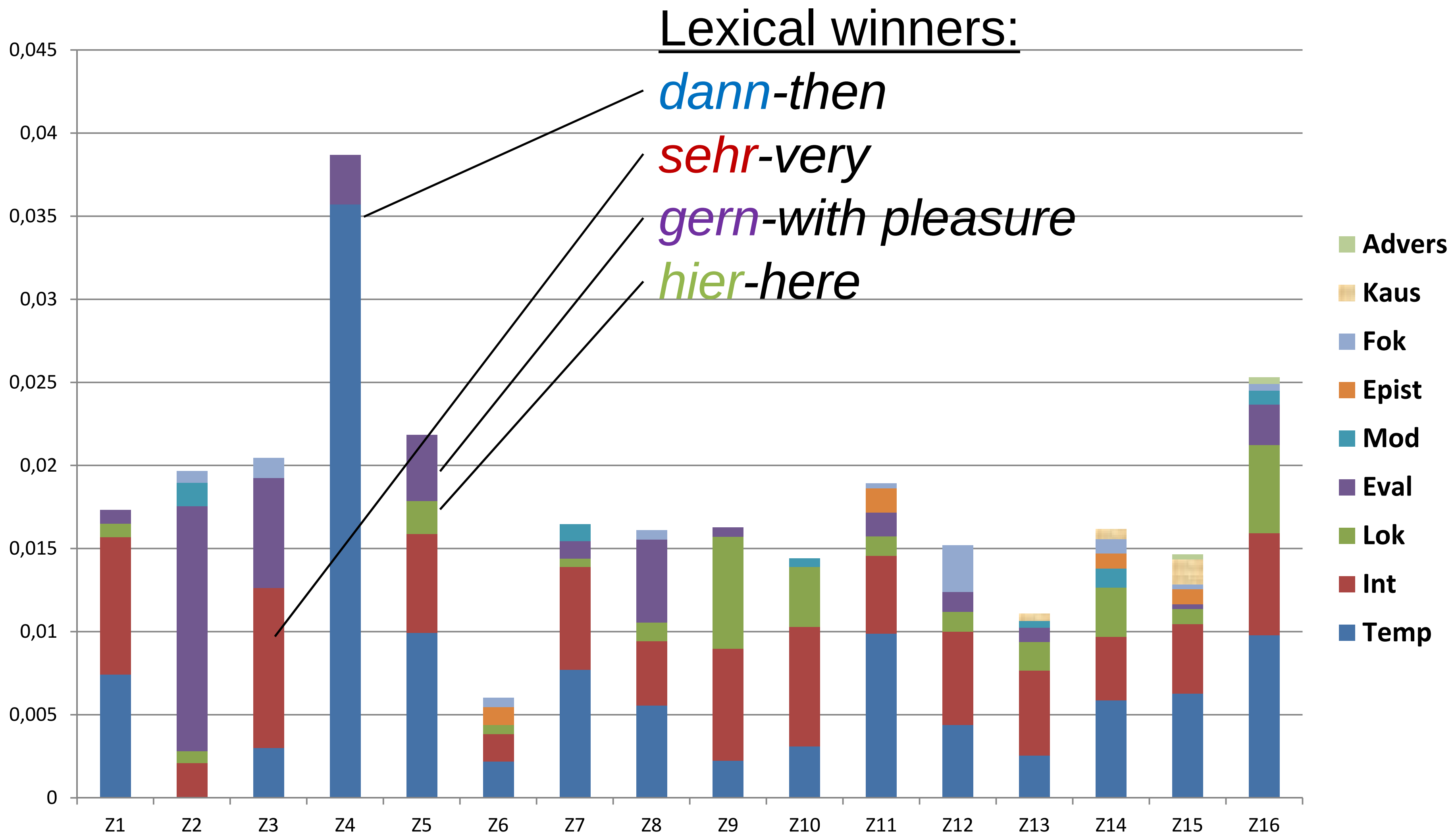
→ Few lexemes used at many points of time

Results: ADV – semantic categories



- Semantic categories: temporal, intensifying, locative, evaluative, modal, epistemic, focus modification, causal, adversative

Results: ADV – semantic categories



Lexical winners:

dann-then

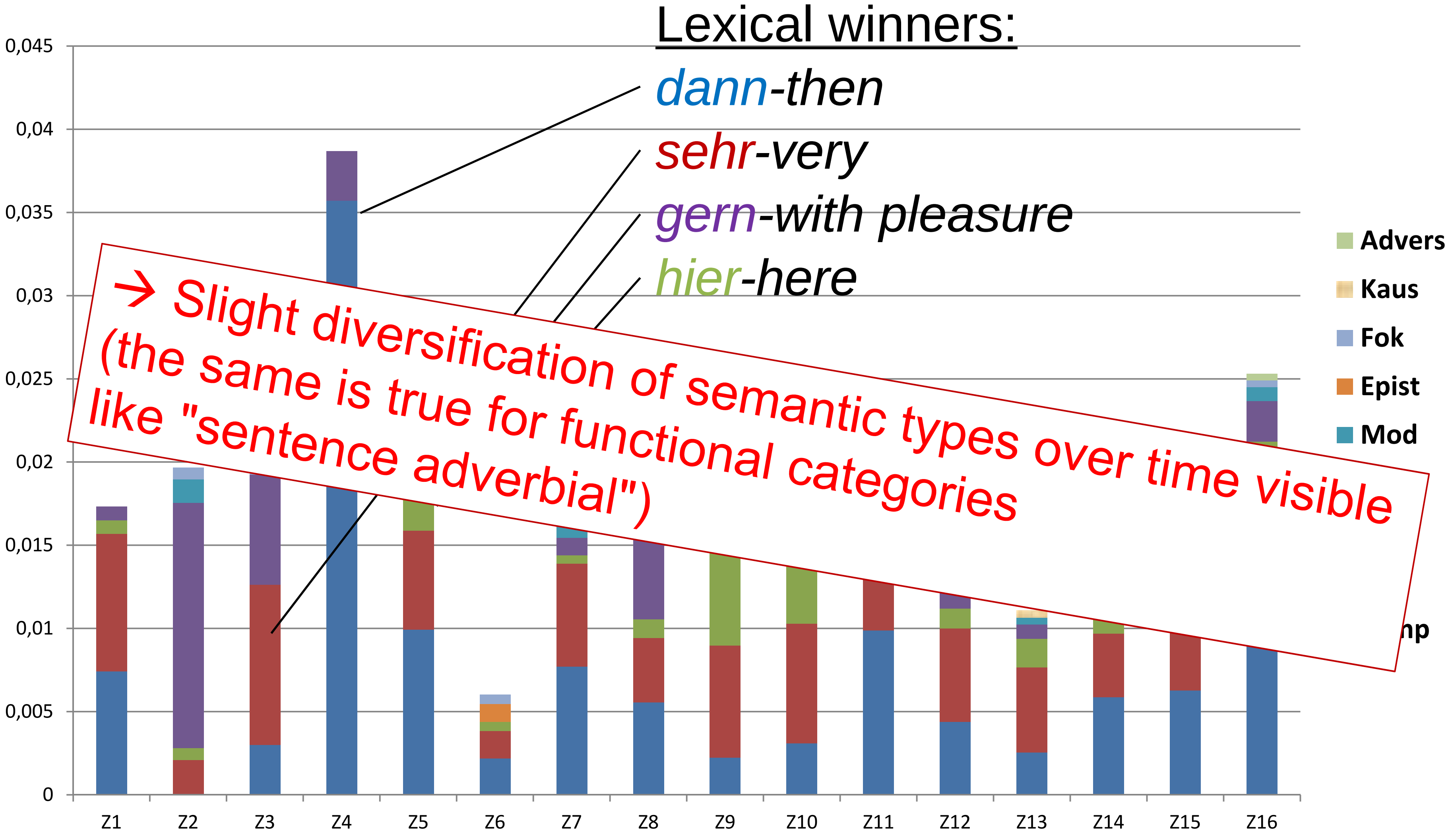
sehr-very

gern-with pleasure

hier-here

- Semantic categories: temporal, intensifying, locative, evaluative, modal, epistemic, focus modification, causal, adversative

Results: ADV – semantic categories



Lexical winners:

dann-then

sehr-very

gern-with pleasure

hier-here

→ Slight diversification of semantic types over time visible (the same is true for functional categories like "sentence adverbial")

- Semantic categories: temporal, intensifying, locative, evaluative, modal, epistemic, focus modification, causal, adversative

Conclusions: correlations between lexical and syntactic measures

- Longitudinal KANDEL data allows for qualitative and quantitative descriptions of learner development
- Correlation of lexical "concepts" with categorial use in KANDEL data:
  new types per point in time > types per point of time > TTR per point of time
- Generally, less systematic growth of lexical diversity than expected
  →Hypothesis "RH" not confirmed
- Huge individual differences (despite homogeneous learner group)
- But systematic developments on different grammatical levels:
  - semantic categories: adversative and causal adverbs
  - functional categories: sentence adverbs and modal particles
- ‘lexical teddybears’ in many subclasses
  (e.g. *sehr*–‘very’ – an absolute winner for intensifiers)
- Task and topic effects observed especially on semantic level

Future research directions

- Correlations between complexity and accuracy
- Analysis of lexico-grammatical constructions
- Analysis of pseudo-longitudinal (cohort) data
  – much larger KANDEL subcorpora

Thanks for your comments!

Acknowledgments:

- Marc Reznicek (for contributing to earlier stages of this study)
- Emily Hackmann and Michael Grünbaum (University of Kansas) for data annotation
- The German and American Fulbright Commission
- KU Institute for Digital Research in the Humanities, *Digital Humanities* Seed Grant
- *Language Learning* Small Grants Research Program

Nina Vyatkina	vyatkina@ku.edu
Hagen Hirschmann	hirschhx@hu-berlin.de
Felix Golcher	felix.golcher@hu-berlin.de