An Integrative Approach to Linguistic Complexity Analysis for German

Zarah Weiss

based on work done in collaboration with Detmar Meurers

Department of Linguistics & LEAD Research Network
University of Tübingen

CRC 1412 Register Colloquium, Online, June 7th, 2021

What is linguistic complexity?

- Complexity is a very productive construct in linguistics
- Application domains beyond research on language typology and second language acquisition (SLA)
  - ⇒ Increasing interest in empirical education research
- Dissertation project on broad linguistic modeling of German using concept of linguistic complexity
  - Focus on automatic assessment with NLP methods
- Two main strands of research
  1. Foundational complexity research rooted in SLA and CL
  2. Application to data from ecologically valid learning contexts

Complexity, Accuracy, and Fluency

- CAF triad describes dimensions of language performance (Housen et al. 2012)
  - Complexity = elaborateness and variation of language
  - Accuracy = native-like and error-free language use
  - Fluency = native-like production speed and smoothness
- Assessed to characterize L2 proficiency and development
  - Focus on English, but also some work on other languages (e.g., De Clercq & Housen 2019; Brezina & Pallotti 2019)
- Recent criticism on reductionist approach to complexity (Housen et al. 2019; Pallotti 2009, 2015)
  - Focus in lexical and syntactic complexity
  - Increasing interest in morphology and phraseology (De Clercq & Housen 2019; Brezina & Pallotti 2019; Paquot 2019)
- Adequacy suggested as 4th CAF dimension (e.g. Pallotti 2009)

Task Effects on Complexity, Accuracy, Fluency

- Task effects have been shown to influence CAF (Yoon & Polio 2016; Yang et al. 2015; Yoon 2017)
  - L2 corpora rarely control for comparable task conditions (Tracy-Ventura & Myles 2015; Yoon & Polio 2016)
- Debated which task factors contribute to task effects
  - Functional task factors (e.g., genre, topic)
  - Cognitive task factors (e.g., shared context, code complexity)
- Effect of cognitive task factors on CAF trade-offs debated
  - Limited Attentional Capacity Model (Skehan 1996)
  - Cognition Hypothesis (Robinson 2001)

Focus of today's talk

1. Broad modeling of linguistic complexity across linguistic domains (syntax, lexicon, morphology, and discourse)
   - Integration of psycho-linguistic measures of language use
   - Integration of measures of human processing
2. Model German L2 proficiency and readability and link linguistic characteristics of reading materials and writing
3. Study role of tasks in L2 writing across linguistic domains

Automatic Linguistic Modeling

- Extract 489 features of German linguistic complexity, language use, human processing, and discourse cohesion
  - Linguistic domains of syntax, lexicon, and morphology
- It has been successfully used for the assessment of, e.g.,
  - Text readability (Weiss & Meurers 2018; Weiss et al. 2021)
  - L2 proficiency (Weiss & Meurers 2019b, 2021)
  - Academic language development (Weiss & Meurers 2019a)
  - Teachers' grading behavior (Weiss et al. 2019)
- Features in multilingual Common Text Analysis Platform[1] (Chen & Meurers 2016; Weiss et al. 2021)
  - Intersection of 200+ features for EN, DE, FR, ES
  - Working on Dutch, Italian, Arabic, Portuguese, Turkish

[1] http://ctapweb.com

Syntactic & Lexical Complexity

Syntactic complexity

- For clausal complexity, we use established subordination and coordination ratios (Wolfe-Quintero et al. 1998)
- For phrasal complexity, we measure phrase modification and modifier coverage (focus: NP complexity, verb cluster)
- We measure grammatical constructs of academic language such as deagentivation

Lexical complexity

- We extract traditional measures of lexical diversity and lexical variation (Lu 2012; McCarthy & Jarvis 2010)
- We subsume some word-net based measures of semantic relatedness in this category

Morphological Complexity & Discourse Measures

Morphological complexity

- We assess the elaboration of compounding and derivation
- Ratios tracking the percentage of certain inflection markers
- Morphological Complexity Index for morphological variation (Brezina & Pallotti 2019)

Discourse Measures

- We assess the use of overt cohesion markers such as certain connectives (Graesser et al. 2004; Crossley et al. 2016)
- We measure implicit cohesion generated through word overlap and grammatical transitions (McNamara et al. 2014; Barzilay & Lapata 2008; Todirascu et al. 2013)

Language Use

Why language use?

- Language use measures are motivated by human behavior rather than the theoretical linguistic system
- Examples are Age of Acquisition measures but also frequency measures (traditionally lexical complexity)

Measures in our system

- Average word and lemma frequencies and word familiarity across data bases (Brysbaert et al. 2011b; Heister et al. 2011)
- Approximations of Age of Active use based on a corpus of children's texts (Lavalley et al. 2015)

Human Processing Costs

The Dependency Locality Theory

- Human processing costs are estimated by increased processing times caused by surprisal or cognitive load
- Dependency locality theory (Gibson 2000) prominent theory of sentence processing
  - Storage of incomplete discourse structures and discourse referents consumes resources
  - Distant discourse referents cause higher processing costs

Measures in our system

- We measure maximal and average DLT integration costs
- We include weighting configurations suggested by Shain et al. (2016) for verbs, coordination, and modifiers

CTAP start page (www.ctapweb.com)

CTAP dashboard (www.ctapweb.com)

CTAP

Tools

- Dashboard
- Corpus Manager
- Feature Selector
- Analysis Generator
- Result Visualizer

User

- Profile
- Feedback
- Documentation
- Sign Out

Dashboard

3 CORPORA

3 FEATURE SETS

4 ANALYSES

Corpus List

ID	Corpus Name	Description	Created On
228	Spotlight 10 (2)		10/03/21
225	Spotlight 10		09/03/21
221	OneStopEnglish		09/03/21

1-3 of 3

Feature Set List

ID	Feature Set Name	Description	Created On
185	Sentences and words (2)		10/03/21
180	Workshop features		09/03/21
179	Sentences and words		09/03/21

1-3 of 3

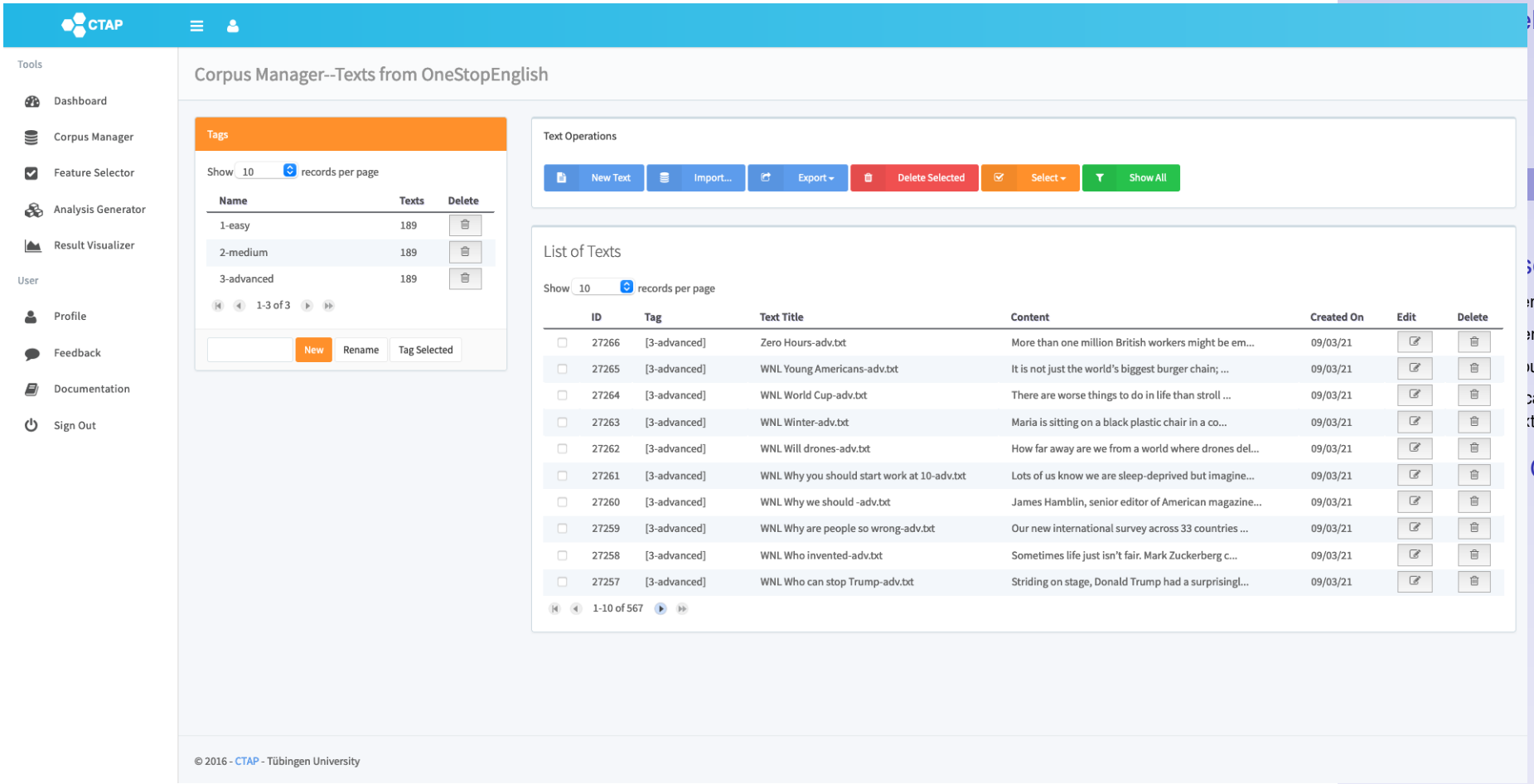
Analysis List

ID	Analysis Name	Description	Corpus	Feature Set	Language	Created On
295	Example		Spotlight 10 (2)	Sentences and words (2)	EN	10/03/21
294	Example analysis	this is just an example	Spotlight 10	Sentences and words	EN	10/03/21
291	Workshop analysis		OneStopEnglish	Workshop features	EN	09/03/21

System Provided Feature List

ID	CF Name	Description
670	Cohesive Complexity Feature: Breindl Temporal Connectives per Token	Calculates the cohesive complexity of the text. This feature calculates the temporal connectives ...
669	POS Density Feature: Singular Noun	Calculates singular noun density of the text. Singular nouns include for English: the Penn Tr...
668	Lexical Sophistication Feature: SUBTLEX Contextual Diversity (AW Type)	Calculates lexical sophistication of the text. Three sophistication measures are calculated from ...
667	Syntactic Complexity Feature: Complex T-unit Ratio	Calculates the syntactic complexity of the text. This feature calculates the complex T-unit ratio...
666	Lexical Sophistication Feature: Kucera and Francis (FW Token)	Calculates lexical sophistication of the text. Three sophistication measures are calculated from ...
665	POS Density Feature: 3rd Person Singular Verb	Calculates 3rd person singular verb density of the text. Third person singular verbs include ...
664	Lexical Sophistication Feature: Kucera and Francis (AW Type)	Calculates lexical sophistication of the text. Three sophistication measures are calculated from ...
663	POS Density Feature: Wh-Determiner	Calculates wh-determiner density of the text. Wh-determiners include for English: the Penn Tre...
662	Lexical Sophistication Feature: SUBTLEX Word Informativeness Per Million	Calculates lexical sophistication of the text. Three sophistication measures

CTAP corpus manager (www.ctapweb.com)



CTAP feature manager (www.ctapweb.com)

CTAP

Tools

- Dashboard
- Corpus Manager
- Feature Selector
- Analysis Generator
- Result Visualizer

User

- Profile
- Feedback
- Documentation
- Sign Out

Feature Set Manager: Workshop features

Selected Features

Remove Selected | Select | Show All

List of Features Selected

Show 20 records per page

	ID	Feature Name	Languages	Details	Remove
	667	Syntactic Complexity Feature: Complex T-unit Ratio	de en		
	608	Lexical Richness: Type Token Ratio (TTR)	en de		
	599	Lexical Sophistication Feature: SUBTLEX Word Frequency (AW Token)	en de		
	576	Number of Syntactic Constituents: Complex Noun Phrase	de en		
	569	Lexical Richness: Type Token Ratio (Corrected TTR)	en de		
	538	Lexical Sophistication Feature: SUBTLEX Word Frequency (LW Token)	en de		
	484	Syntactic Complexity Feature: Mean Length of Clause	en de		
	475	Number of Tokens	en de		
	470	Syntactic Complexity Feature: Dependent clause ratio	en de		
	425	Lexical Variation Feature: Corrected Verb Variation 1	en de		
	414	Mean Token Length in Syllables	en de		
	411	Mean Sentence Length in Tokens	en de		
	402	Lexical Variation Feature: Verb Variation 1	en de		

1-13 of 13

Available Features

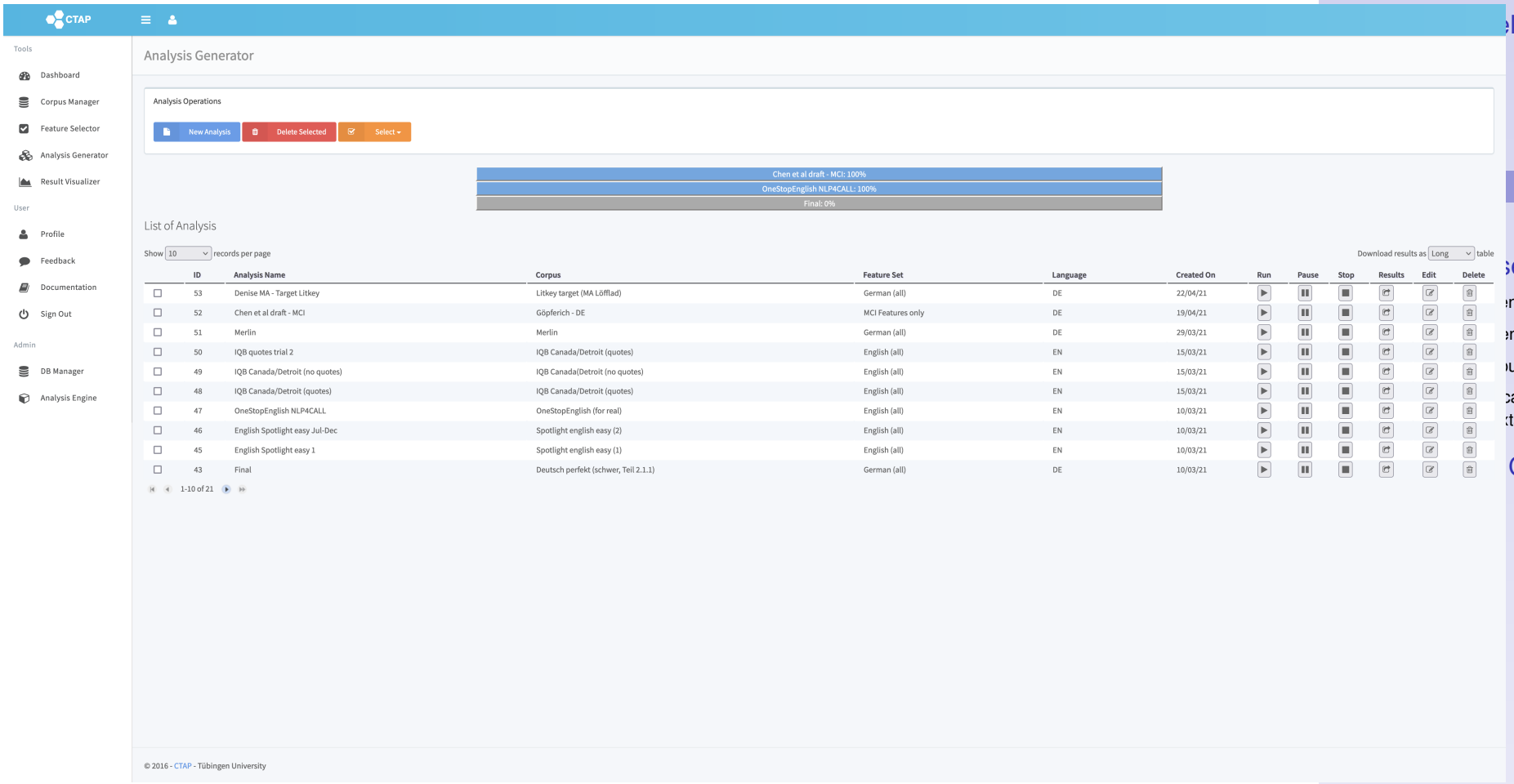
Add Selected | Select | Show All

List of Available Features

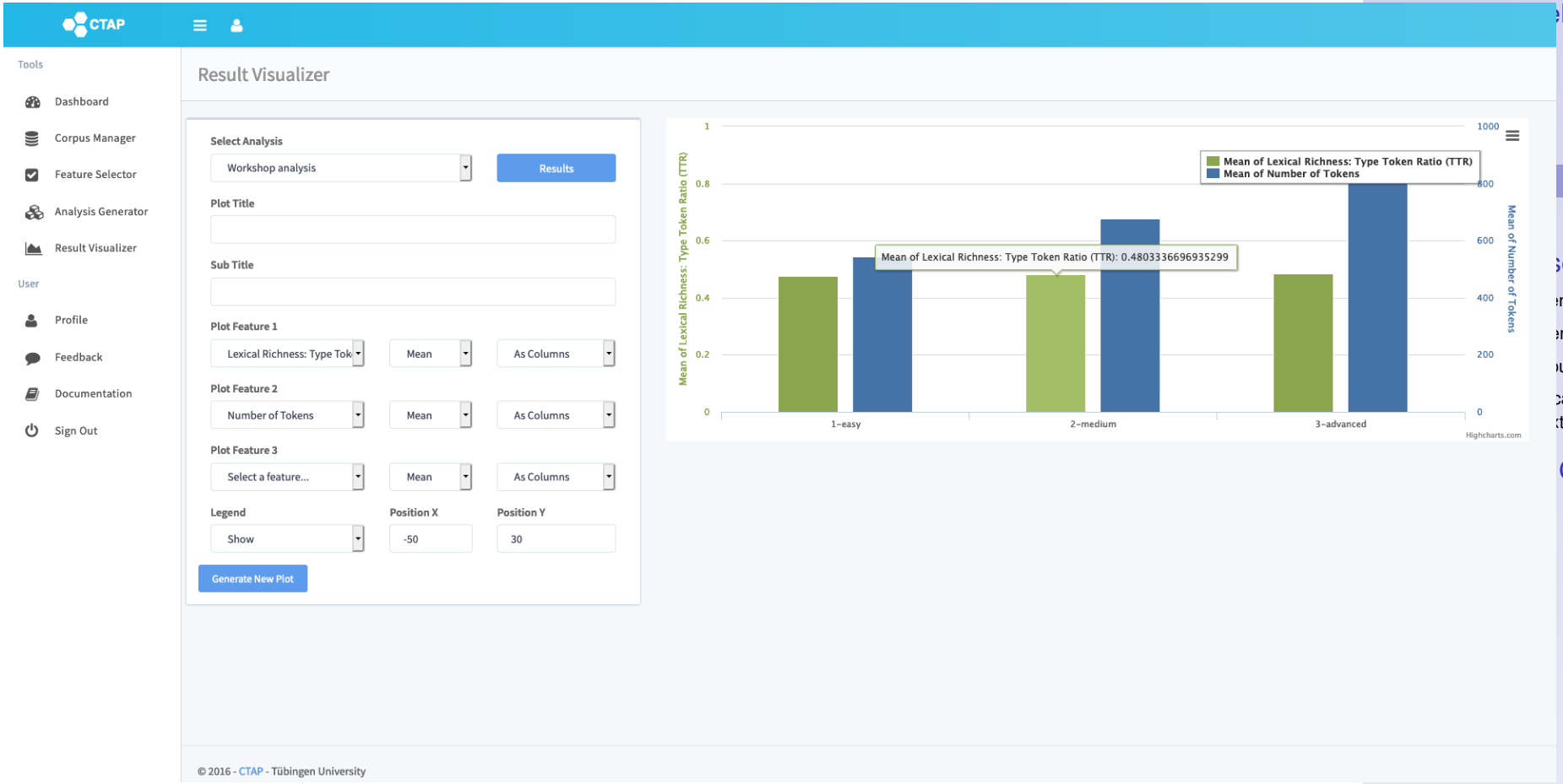
Show 20 records per page

	ID	Feature Name	Languages	Details	Add
	670	Cohesive Complexity Feature: Breindl Temporal Connectives per Token	de		
	669	POS Density Feature: Singular Noun	en		
	668	Lexical Sophistication Feature: SUBTLEX Contextual Diversity (AW Type)	en		
✓	667	Syntactic Complexity Feature: Complex T-unit Ratio	de en		
	666	Lexical Sophistication Feature: Kucera and Francis (FW Token)	en		
	665	POS Density Feature: 3rd Person Singular Verb	en		
	664	Lexical Sophistication Feature: Kucera and Francis (AW Type)	en		
	663	POS Density Feature: Wh-Determiner	en		
	662	Lexical Sophistication Feature: SUBTLEX Word Informativeness Per Million Words (AW Type)	de		
	661	Lexical Sophistication Feature: Mean Age of Active Use in KCT (LW Token)	de		
	660	Lexical Variation Feature: Squared Verb Variation 1	en de		
	659	Lexical Sophistication Feature: Google Books 2000 Word Informativeness Per Million Words (FW Type)	de		
	658	POS Density Feature: Relative Pronoun	de		
	657	Lexical Sophistication Feature: Kucera and Francis (LW Token)	en		
	655	Lexical Sophistication Feature: SUBTLEX Logarithmic Word Frequency Per Million Words (AW Type)	de		

CTAP analysis manager (www.ctapweb.com)



CTAP result visualizer (www.ctapweb.com)



Automatic proficiency assessment

- How can learners' language proficiency be characterized through their language productions?
  - Linked to tracking of learners' developmental trajectories
  - Similar to automatic essay grading (Crossley 2020)
- Common application domain of complexity research in SLA (Housen et al. 2019)
  - Typically used to track written language proficiency
- Previous studies on German L2/L1 proficiency assessment
  - Written L2 proficiency (Weiss & Meurers 2019b, 2021)
  - Oral or written early academic language (Weiss & Meurers 2019a; Weiss et al. in preparation b)
  - English/German advanced academic language (Chen et al. draft)

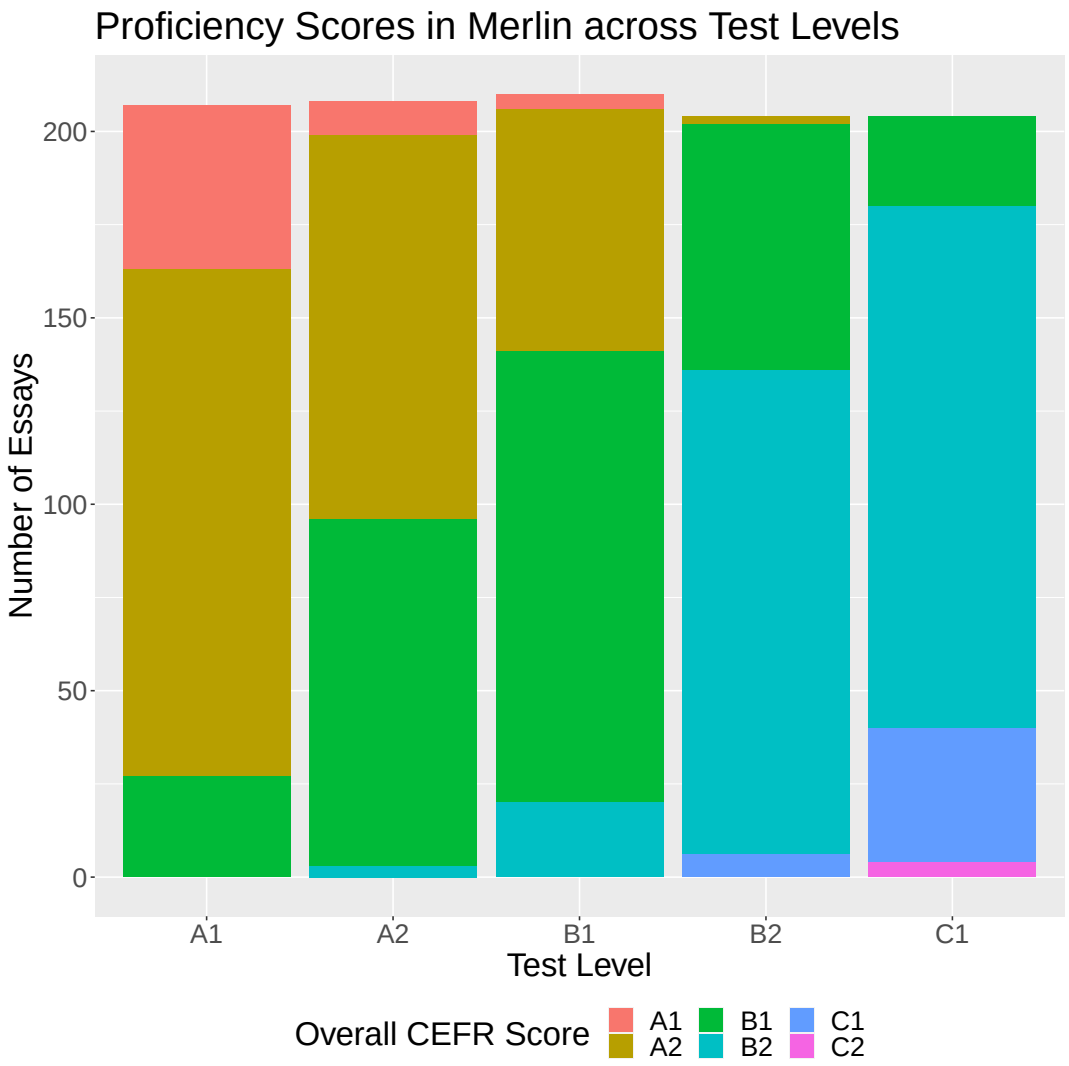
Automatic proficiency assessment

- How can learners' language proficiency be characterized through their language productions?
  - Linked to tracking of learners' developmental trajectories
  - Similar to automatic essay grading (Crossley 2020)
- Common application domain of complexity research in SLA (Housen et al. 2019)
  - Typically used to track written language proficiency
- Previous studies on German L2/L1 proficiency assessment
  - Written L2 proficiency (Weiss & Meurers 2019b, 2021)
  - Oral or written early academic language (Weiss & Meurers 2019a; Weiss et al. in preparation b)
  - English/German advanced academic language (Chen et al. draft)

The Merlin corpus

- Cross-sectional corpus of 1,033 German L2 writings (Wisniewski et al. 2013)
  - Texts rated by two experts on CEFR scale (A1 to C2)
- Elicited in official standardized language certification tests
  - Test levels A1 to C1
- Texts available as original writing and two target hypothesis
  - Form-based target hypothesis for entire corpus
  - Meaning-based target hypothesis for sub-corpus
- Three task prompts per level ⇒ 15 different task prompt

Proficiency scores and test levels



- Proficiency scores are distributed across test levels, e.g., A2 and B1 test takers' proficiency scores range from A1 to B2

Merlin corpus statistics

	N. docs	μ words $\pm SD$	M	Min	Max
A1	57	34. $\pm$ 11.5	35	6	63
A2	306	61. $\pm$ 23.3	55	24	141
B1	331	118. $\pm$ 48.3	107	28	293
B2	293	195. $\pm$ 47.4	195	66	354
C1	42	227. $\pm$ 49.9	223	150	372
C2	4	246. $\pm$ 51.5	241	192	310

- Skewed distribution with few A1 and C1/C2 texts
- Text length systematically increases with higher proficiency
- From level B1 on most texts contain more than 100 words

Merlin corpus statistics

	N. docs	μ words $\pm SD$	M	Min	Max
A1	57	34. $\pm$ 11.5	35	6	63
A2	306	61. $\pm$ 23.3	55	24	141
B1	331	118. $\pm$ 48.3	107	28	293
B2	293	195. $\pm$ 47.4	195	66	354
C1	42	227. $\pm$ 49.9	223	150	372
C2	4	246. $\pm$ 51.5	241	192	310

- Skewed distribution with few A1 and C1/C2 texts
- Text length systematically increases with higher proficiency
- From level B1 on most texts contain more than 100 words

Merlin corpus statistics

	N. docs	μ words $\pm SD$	M	Min	Max
A1	57	34. $\pm$ 11.5	35	6	63
A2	306	61. $\pm$ 23.3	55	24	141
B1	331	118. $\pm$ 48.3	107	28	293
B2	293	195. $\pm$ 47.4	195	66	354
C1	42	227. $\pm$ 49.9	223	150	372
C2	4	246. $\pm$ 51.5	241	192	310

- Skewed distribution with few A1 and C1/C2 texts
- Text length systematically increases with higher proficiency
- From level B1 on most texts contain more than 100 words

Merlin corpus statistics

	N. docs	μ words $\pm SD$	M	Min	Max
A1	57	34. $\pm$ 11.5	35	6	63
A2	306	61. $\pm$ 23.3	55	24	141
B1	331	118. $\pm$ 48.3	107	28	293
B2	293	195. $\pm$ 47.4	195	66	354
C1	42	227. $\pm$ 49.9	223	150	372
C2	4	246. $\pm$ 51.5	241	192	310

- Skewed distribution with few A1 and C1/C2 texts
- Text length systematically increases with higher proficiency
- From level B1 on most texts contain more than 100 words

Merlin corpus statistics

	N. docs	μ words $\pm SD$	M	Min	Max
A1	57	34. $\pm$ 11.5	35	6	63
A2	306	61. $\pm$ 23.3	55	24	141
B1	331	118. $\pm$ 48.3	107	28	293
B2	293	195. $\pm$ 47.4	195	66	354
C1	42	227. $\pm$ 49.9	223	150	372
C2	4	246. $\pm$ 51.5	241	192	310

- Skewed distribution with few A1 and C1/C2 texts
- Text length systematically increases with higher proficiency
- From level B1 on most texts contain more than 100 words

Machine learning experiment on Merlin

- Predict CEFR levels A1 to C1/C2 (Weiss & Meurers 2019b)
- Extract 489 features from form-based target hypothesis
- Training and testing with 10-fold cross-validation
- Compare (ordinal) random forest, SVM (polynomial/radial)
  - Ordinal random forest performs best
- Accuracy = 70.0% in 5-way classification
  - Majority baseline: 32.0%

Automatic readability assessment

- Readability assessment identifies suitable texts given
  - the language proficiency of the target audience
  - the reading purpose (e.g., learning, information retrieval)
- Complexity features successfully predict readability
  (Vajjala & Meurers 2012; Chen & Meurers 2017b; Crossley et al. 2019)
- Complexity-based readability research on German
  - Distinction of media language for adults and children
    (Weiss & Meurers 2018)
  - Readability research on schoolbooks
    (Kühberger et al. 2019; Berendes et al. 2018)
  - First multi-level readability classification for L2 materials
    (Weiss et al. 2021)

Automatic readability assessment

- Readability assessment identifies suitable texts given
  - the language proficiency of the target audience
  - the reading purpose (e.g., learning, information retrieval)
- Complexity features successfully predict readability
  (Vajjala & Meurers 2012; Chen & Meurers 2017b; Crossley et al. 2019)
- Complexity-based readability research on German
  - Distinction of media language for adults and children
    (Weiss & Meurers 2018)
  - Readability research on schoolbooks
    (Kühberger et al. 2019; Berendes et al. 2018)
  - First multi-level readability classification for L2 materials
    (Weiss et al. 2021)

Spotlight-DE data set

- 1,446 articles from monthly language learning magazine *Deutsch perfekt* (www.deutsch-perfekt.de)
  - Published in 45 issues from 01/2018 to 12/2019
  - Extract article from type-set PDFs using OCR
- Leveled reading materials aligned with CEFR levels
  - Beginning (A2), medium (B1/B2), advanced (C1)
- Part of multi-lingual data set for English and German
  - Being extended to French, Spanish, Italian
  - $\Rightarrow$ All data from magazines by Spotlight publisher

Spotlight-DE text statistics

	N. docs	μ words $\pm SD$	M	Min	Max
Easy	763	236±235	137	60	1.469
Medium	509	665±769	448	72	5.605
Advanced	174	892±537	524	91	4.161

- Contains more easy and medium than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts

Spotlight-DE text statistics

	N. docs	μ words $\pm SD$	M	Min	Max
Easy	763	236±235	137	60	1.469
Medium	509	665±769	448	72	5.605
Advanced	174	892±537	524	91	4.161

- Contains more easy and medium than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts

Spotlight-DE text statistics

	N. docs	μ words $\pm SD$	M	Min	Max
Easy	763	236±235	137	60	1.469
Medium	509	665±769	448	72	5.605
Advanced	174	892±537	524	91	4.161

- Contains more easy and medium than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts

Spotlight-DE text statistics

	N. docs	μ words $\pm SD$	M	Min	Max
Easy	763	236±235	137	60	1.469
Medium	509	665±769	448	72	5.605
Advanced	174	892±537	524	91	4.161

- Contains more easy and medium than advanced texts
- Text length increases to some degree with reading level
- All reading levels contain very long and very short texts

Machine learning experiment on Spotlight-DE

- Automatic prediction of reading levels A2, B1/B2, and C1 (Weiss et al. 2021)
- Extract 489 features from plain text articles
- Training and testing with 10-fold cross-validation
- Compare (ordinal) random forest, SVM (polynomial/radial)
  - Ordinal random forest performs best
- Accuracy = 89.0% in 3-way classification
  - Majority baseline: 52.8%

Appropriate input fosters language learning

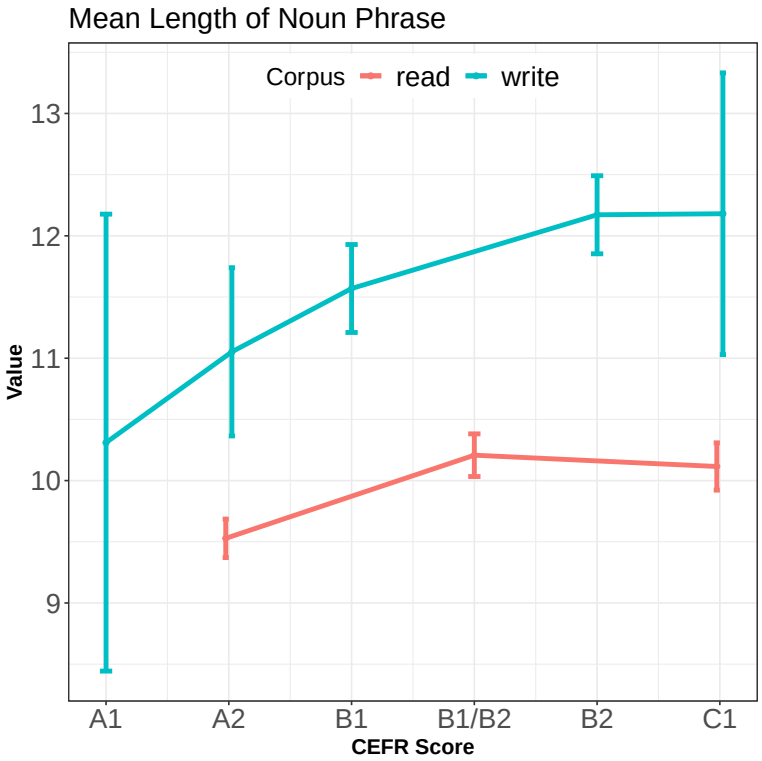
- Language input fosters (second) language learning
  - Challenging input at i+1 (Krashen 1981)
  - Zone of Proximal Development (ZPD; Lantolf et al. 2015)
- Identification of appropriate reading materials requires
  1. Assessment of proficiency of language learner
  2. Assessment of reading level of reading materials
  3. Match of proficiency and reading level

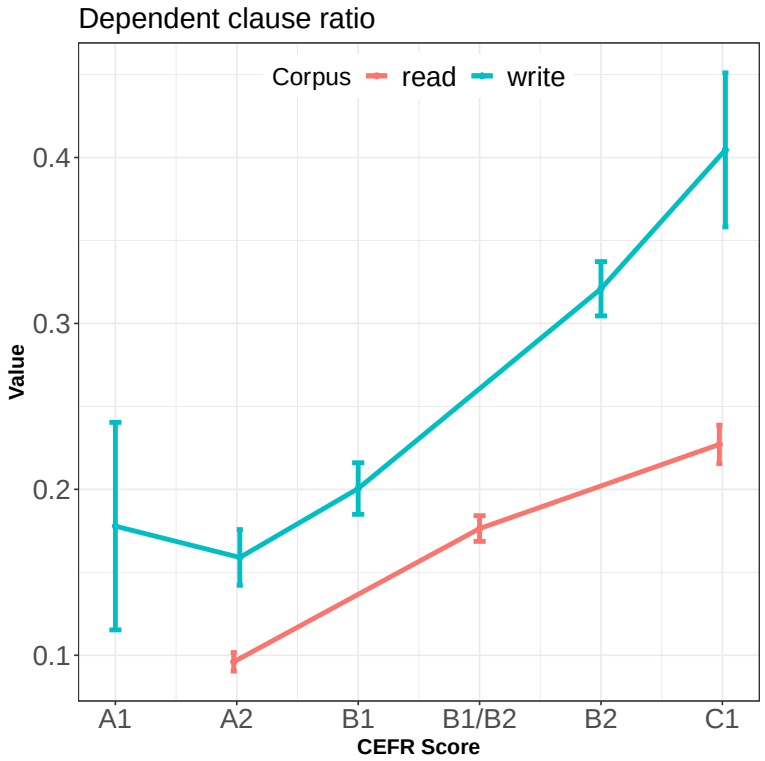
  ⇒ Reading level not necessarily at i+1 for every learner (Chen & Meurers 2017a)
- How do German L2 proficiency and reading levels match together from linguistic perspective?

Linking L2 language output and L2 reading input

- 489 features successfully discriminate CEFR levels on reading texts and written learner productions
- Compare expression of relevant features on both data sets
  - Identify informative features with information gain ranking
  - Select 422 features that are informative for both data sets
- Zoom in on features across linguistic domains to compare feature values in reading and writing data
  - Syntax, lexicon, morphology, text cohesion, language use

Zooming in on syntactic complexity

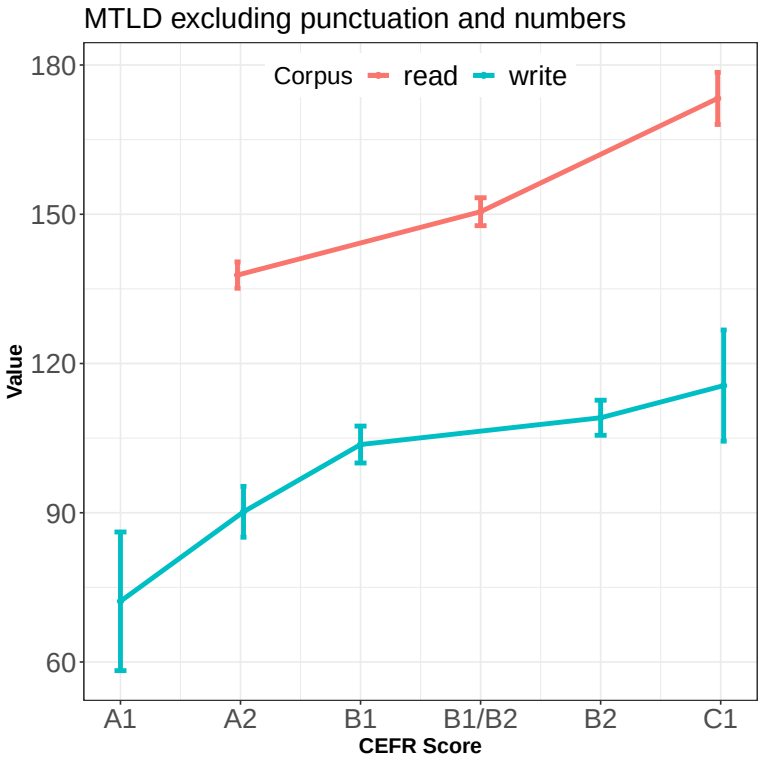


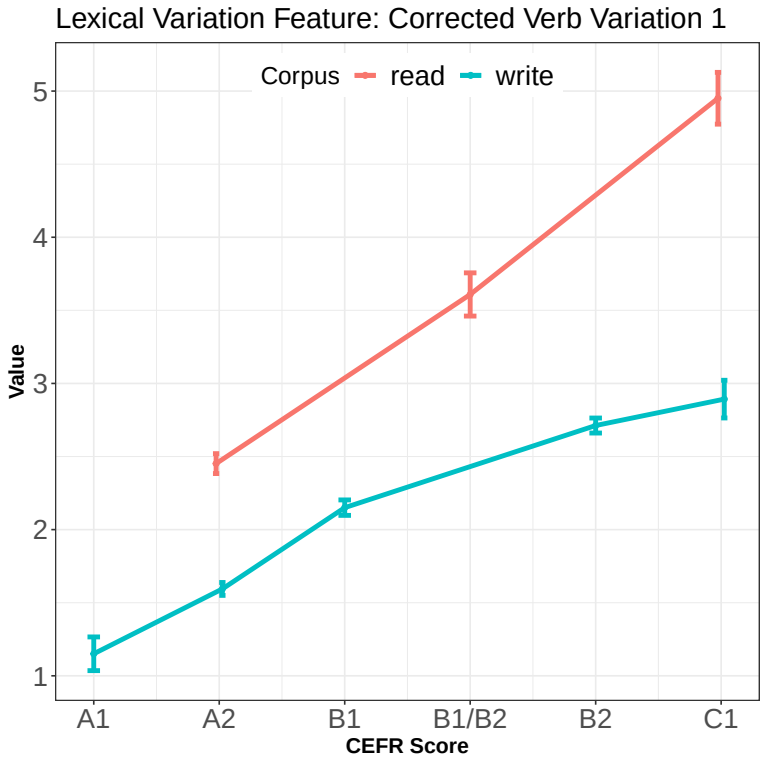


- Systematic increase syntactic elaborateness in both
- Values for reading texts are systematically lower

$\Rightarrow$ No syntactic challenge in reading texts

Zooming in on lexical complexity

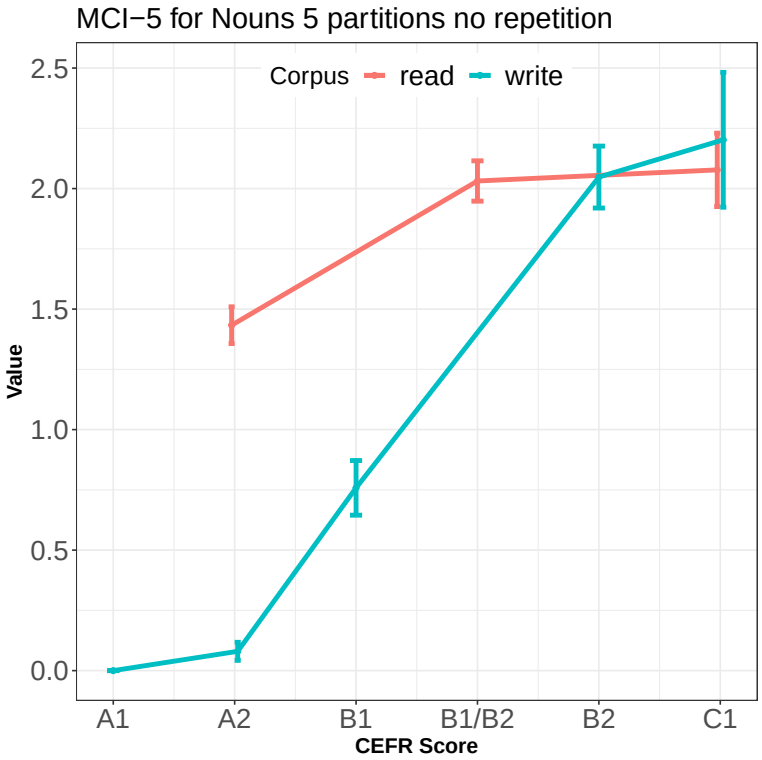


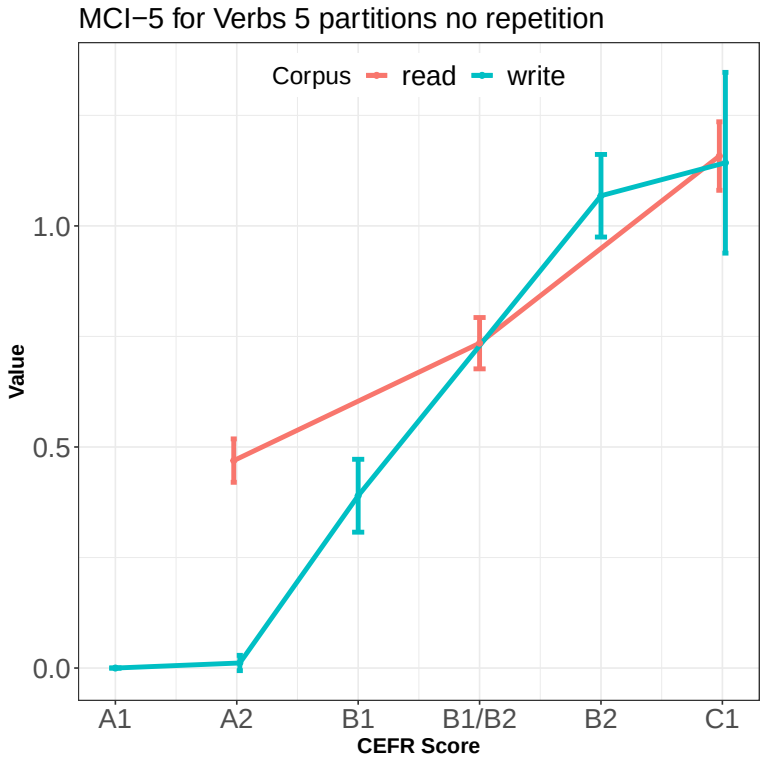


- Systematic increase in lexical variation for both
- MTLD and corrected verb variation higher for reading texts

$\Rightarrow$ Challenge for learners in lexical domain

Zooming in on morphological complexity

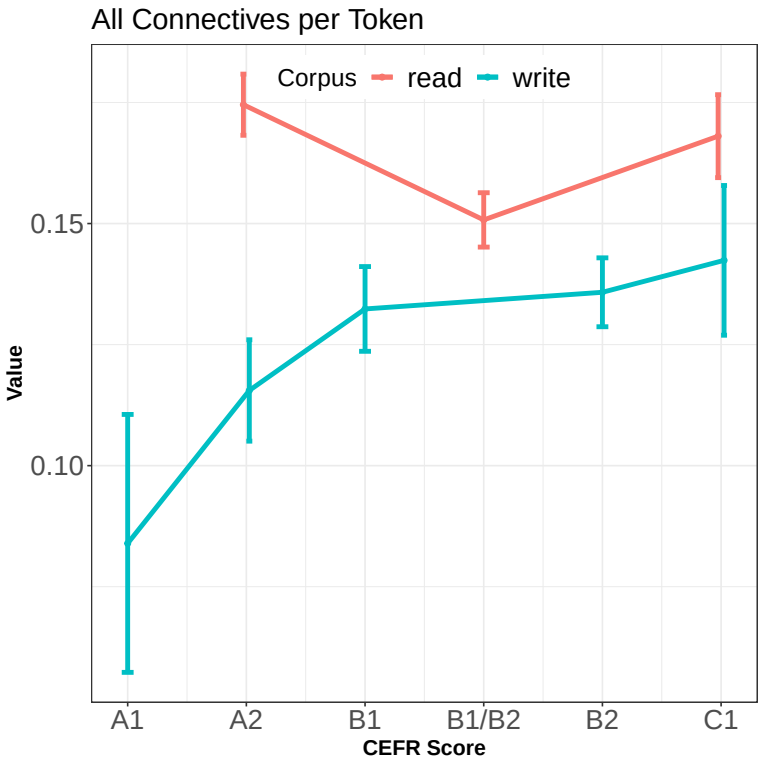


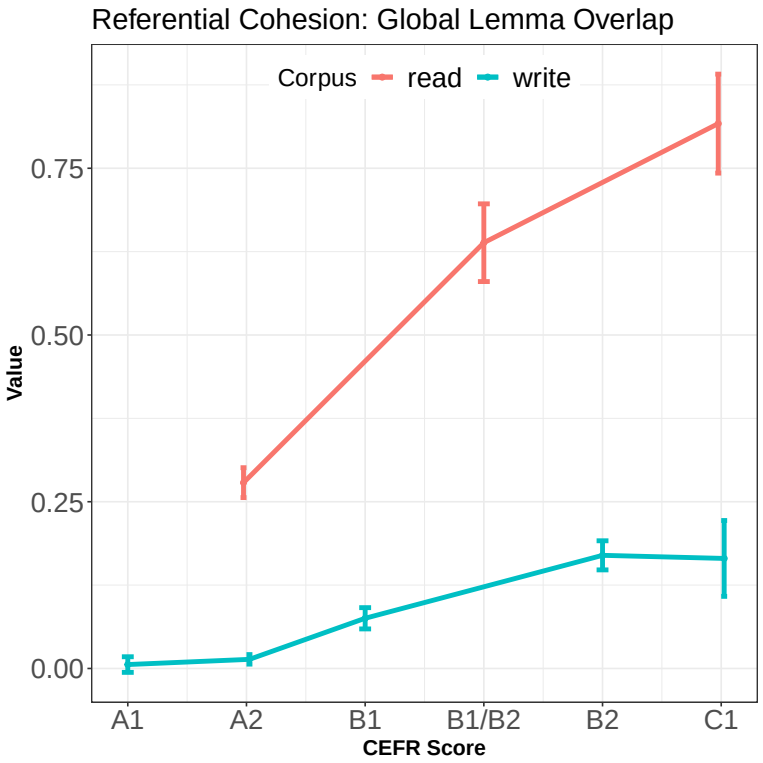


- Systematic increase in morphological variation for both
- MCI of reading texts challenging for A2 and B1 learners

⇒ No challenge for intermediate/advanced learners

Zooming in on text cohesion

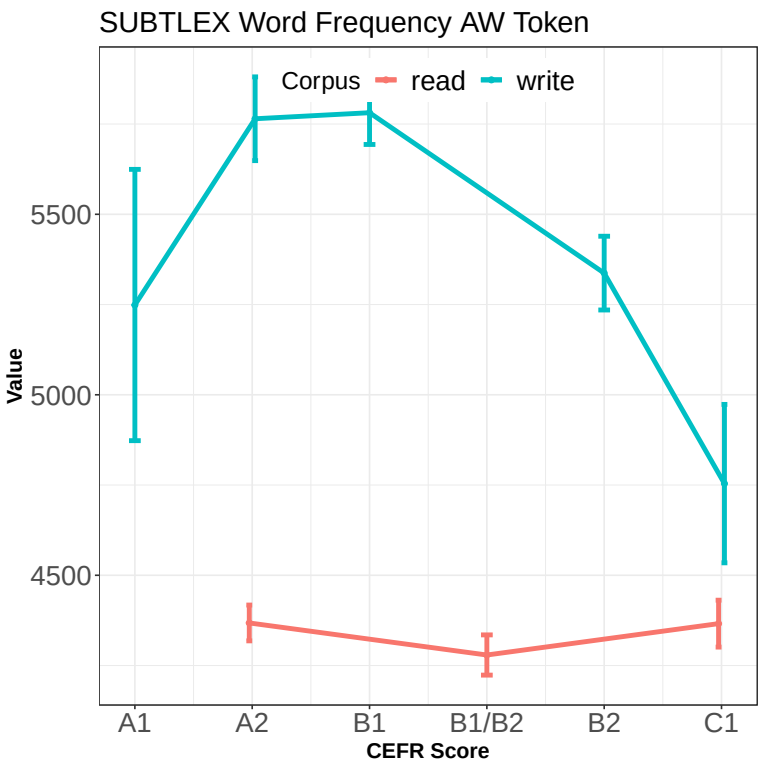


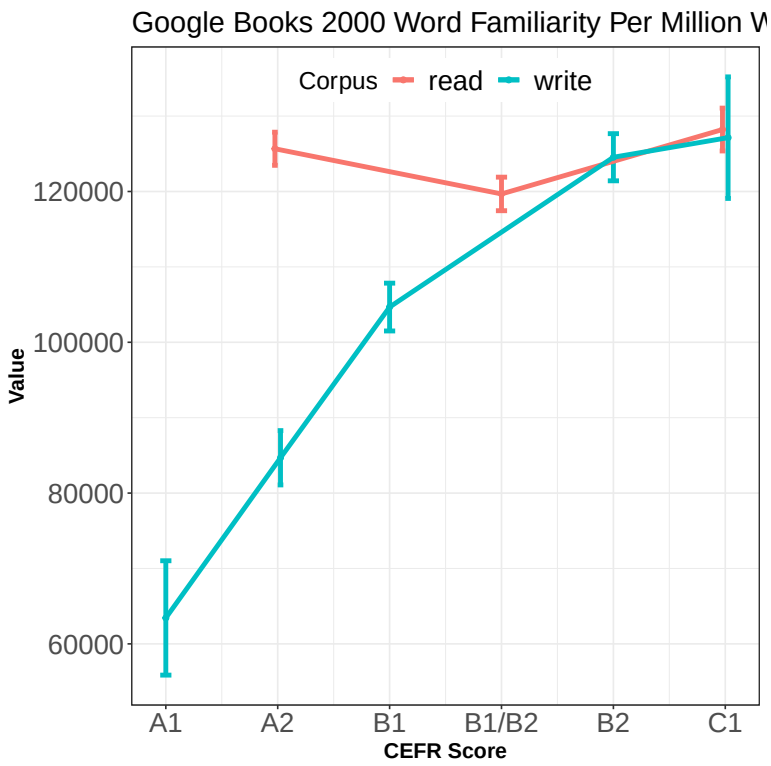


- Reading texts systematically more use of connectives and higher implicit cohesion

⇒ Challenge with respect to text cohesion

Zooming in on language use





- Reading texts use more rare words and more words with higher familiarity at constant rate
- Systematic increase in writing data with intermediate to advanced learners showing same level of familiarity

⇒ Challenge with respect to language use

Discussion

- Reading input lies systematically above language production in all domains but syntax
- Is this challenge indeed in the ZPD or above?
  - Study by Chen & Meurers (2017a) for individual features
- Is there a trade-off of acceptable challenges between different linguistic domains?

$\Rightarrow$ Research program needed on interaction between proficiency and input to foster learning

Moving beyond foundational research

- Linguistic complexity as valuable concept for understanding of learning processes and outcomes
- More research needed to understand implications for real-life teaching and learning
- Studies on data from ecologically valid contexts needed

$\Rightarrow$ Interdisciplinary work to connect insights from complexity research with empirical education research

CA(F) in authentic education contexts

- Study linguistic complexity in oral classroom interactions
  - Student utterances mirror written language proficiency (Weiss et al. in preparation b)
  - Complexity of teacher utterances differs across school types (Weiss et al. in preparation a)
  - Teachers' adaptation skills in L2 classrooms (COLD project)[2]
- Complexity & genre impact text item difficulty in reading comprehension tests in grade 3 & 4 (Ludewig et al. in preparation)
- It links to complexity of historical thinking processes
  - ...displayed student writing (Bertram et al. in preparation)
  - ...required by tasks history books (Kühberger et al. 2019)
- Task-appropriate complexity and accuracy influence teachers' grading behavior in Abitur data (Weiss et al. 2019)

[2] www.die-bonn.de/cold

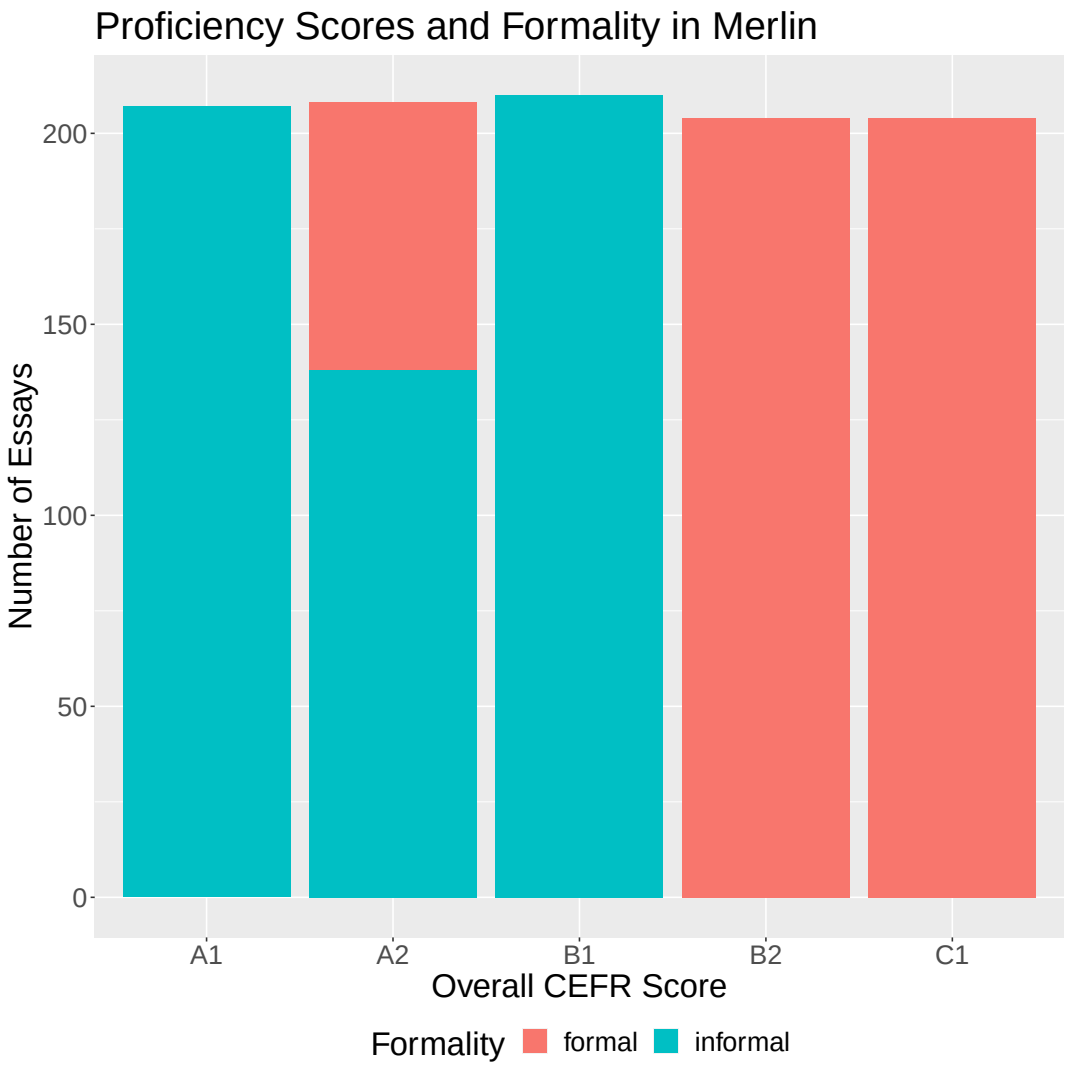
Narrowing in on task effects

- Describe interaction of task and complexity
  - How do task effects differ across linguistic domains?
- Use Merlin corpus for investigation of task effects
  1. Predict 15 tasks to gauge strength of overall effect
  2. Zoom in on task effects across linguistic domains
- Focus on task rather than individual task factors
  - Annotated 15 task factors for task prompts[3] (Weiss 2017)
  - Used approach by Alexopoulou et al. (2017)

  ⇒ Nearly all factors fully confounded with test level

[3] Cognitive factors: code and cognitive complexity, shared context, reasoning demands, referenced elements, perspective; functional factors: genre, audience, task theme and type

Writing formality confounds with test level

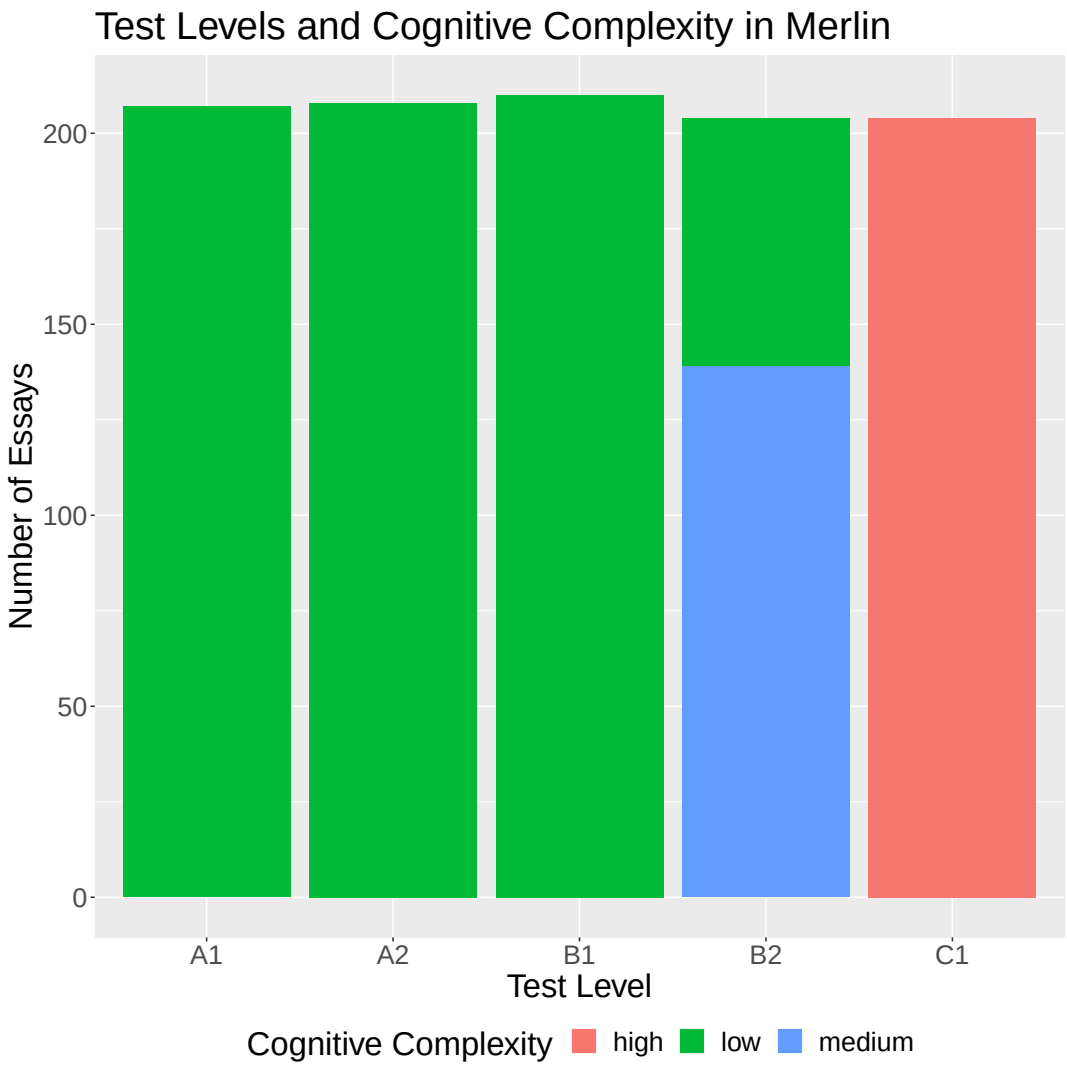


Text type confounds with test level

Code complexity confounds with test level

- Code complexity depends on provision of no, few, or detailed language material in instructions

Cognitive complexity confounds with test level

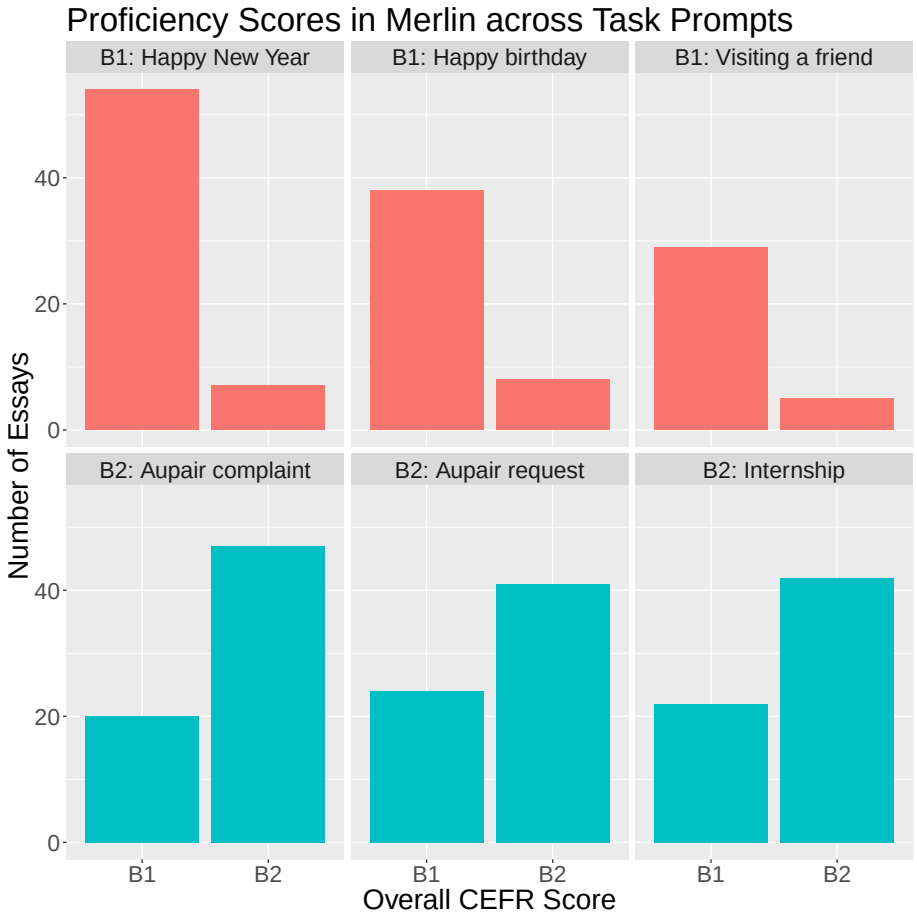


- Did the task clearly outline the intended text content and structure or did learners need to structure texts while writing?

Task classification on Merlin corpus

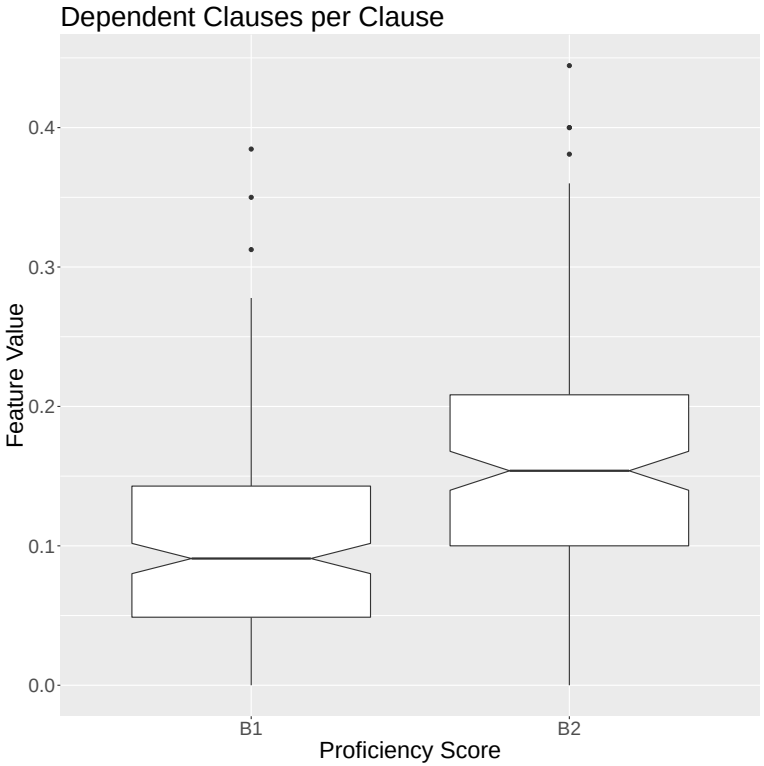
- Automatic prediction of 15 tasks (3 per test level)
- Extract 489 features from form-based target hypothesis
- Training and testing with 10-fold cross-validation
- Compare random forest, SVM (polynomial/radial)
  - SVM with polynomial kernel performs best
- Accuracy = 89.8% in 15-way classification
  - Majority baseline: 8.0%
  - Exceeds 70.0% accuracy for 5-way proficiency classification

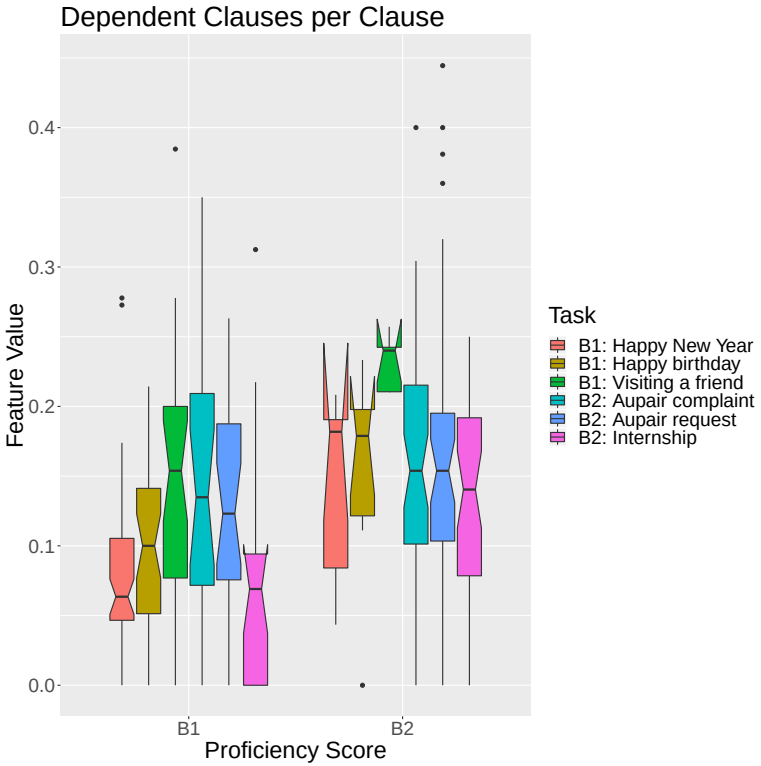
Zooming in on the Merlin corpus



- Subset of the Merlin corpus used for classification studies
- Focus on B1/B2 learners taking B1/B2 tests (N=337)
  - ⇒ Overall six different elicitation tasks
  - ⇒ This also includes, e.g., B1 learners who failed a B2 test

Clausal Complexity





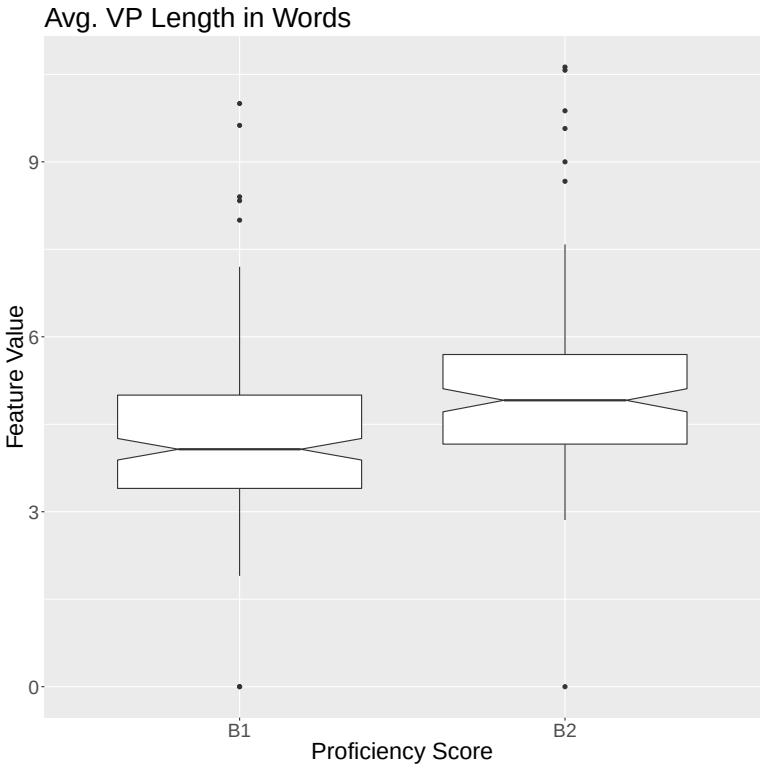
⇒ B2 texts use more elaborate subordination than B1 texts

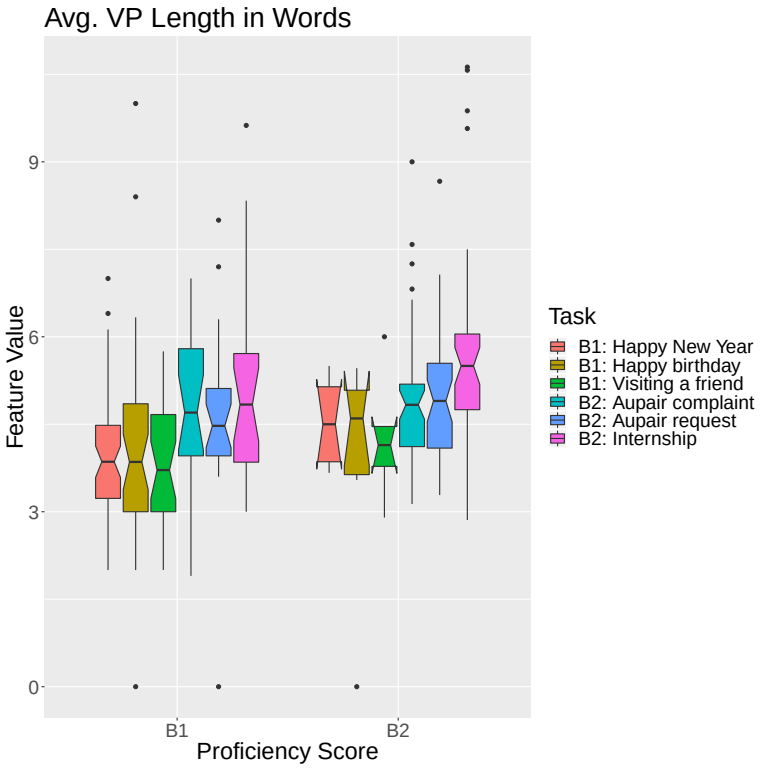
- This holds across measures of clausal complexity

⇒ We see some some task-specific variation at B1 and B2

- Specific tasks elicit more or less clausal elaboration
- This holds across measures of clausal complexity

Phrasal Complexity

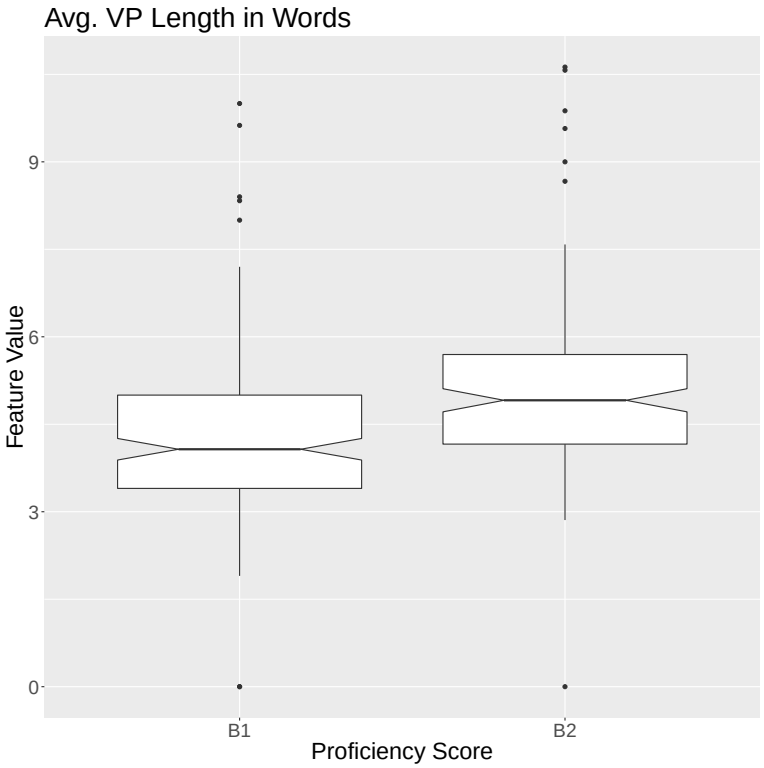


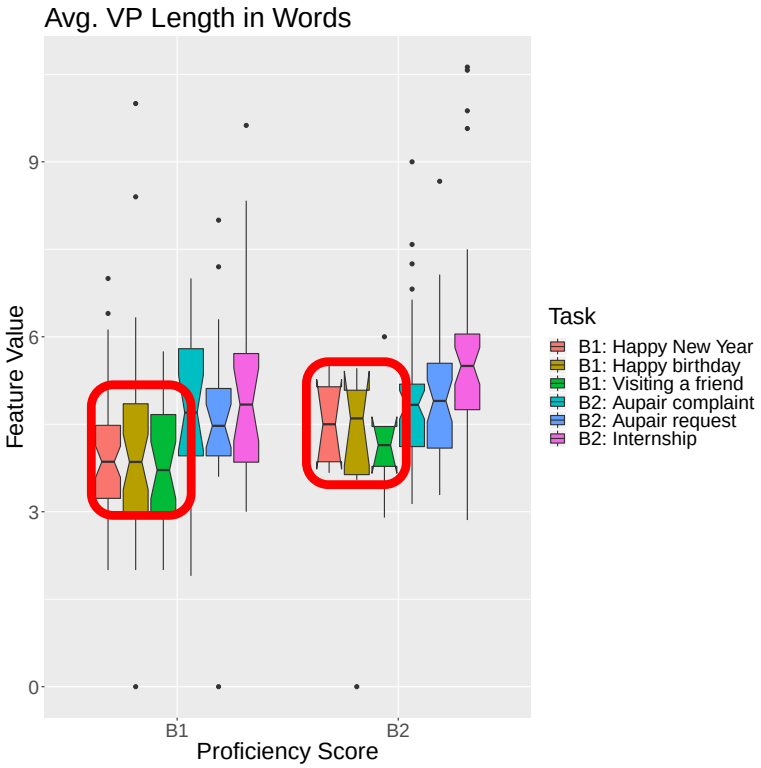


⇒ Verb phrase complexity is influenced by test level but also shows differences between B1 and B2 scores

- No clear proficiency differences for NP or PP complexity

Phrasal Complexity

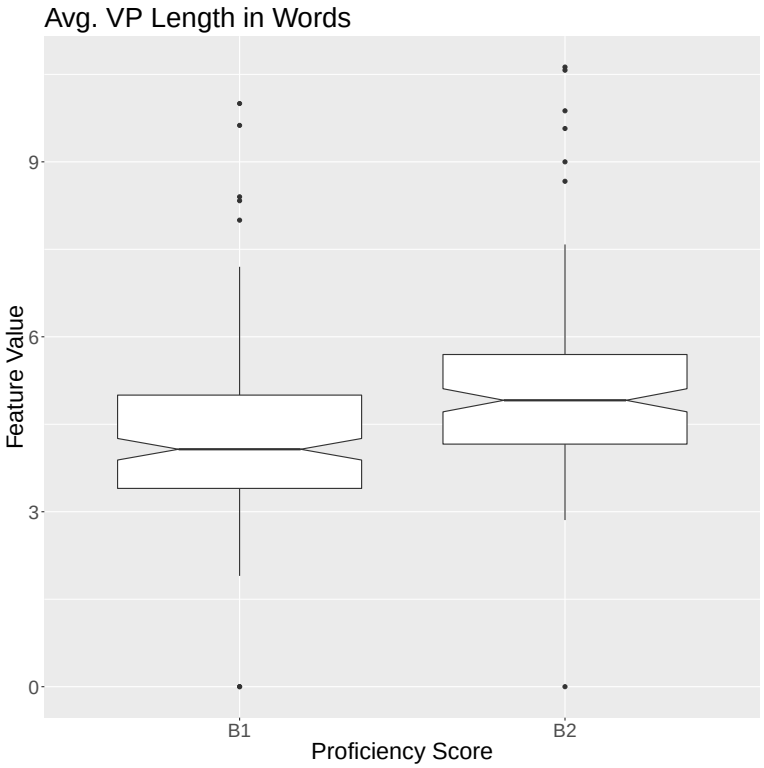


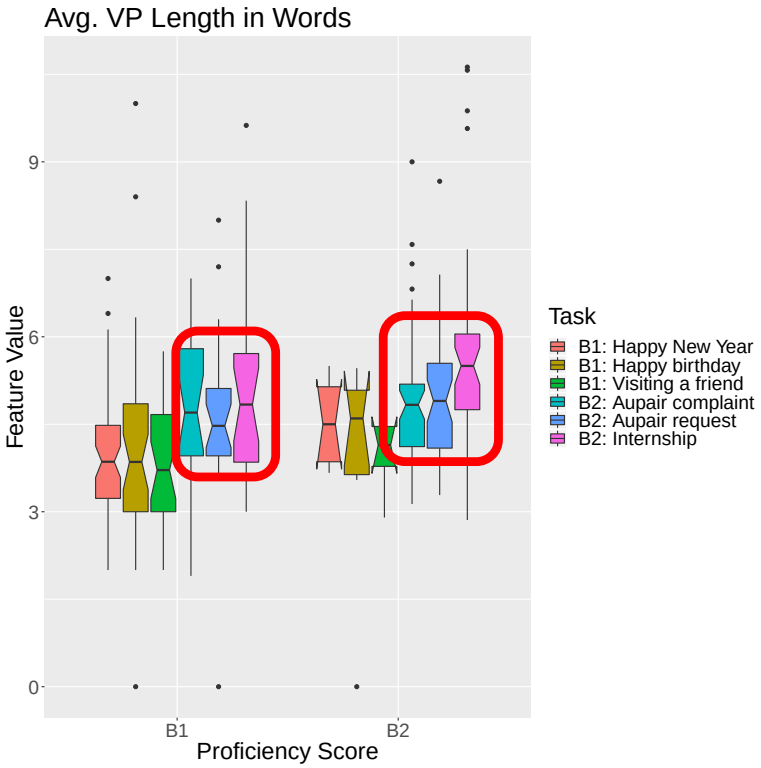


⇒ Verb phrase complexity is influenced by test level but also shows differences between B1 and B2 scores

- No clear proficiency differences for NP or PP complexity

Phrasal Complexity

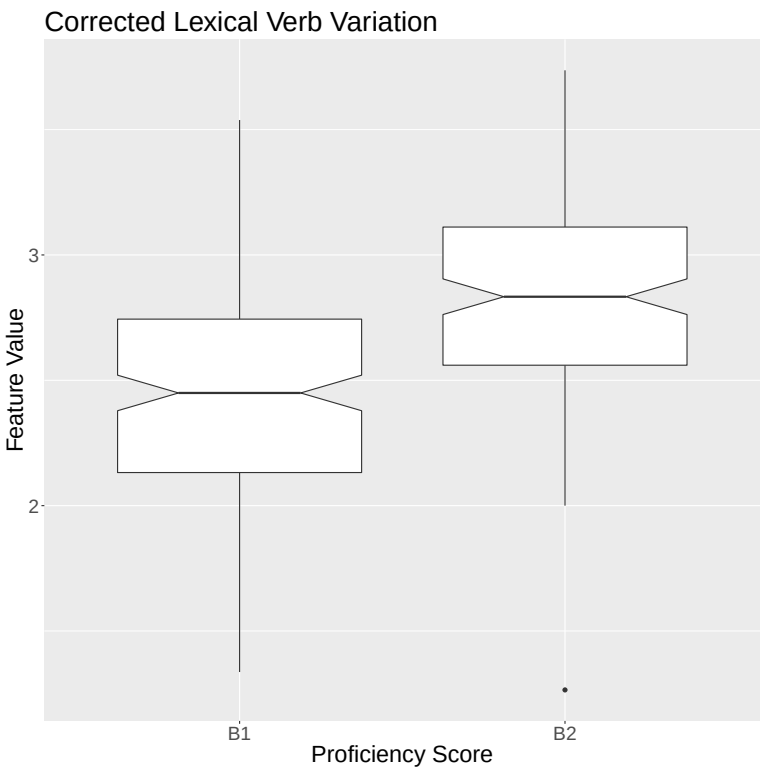


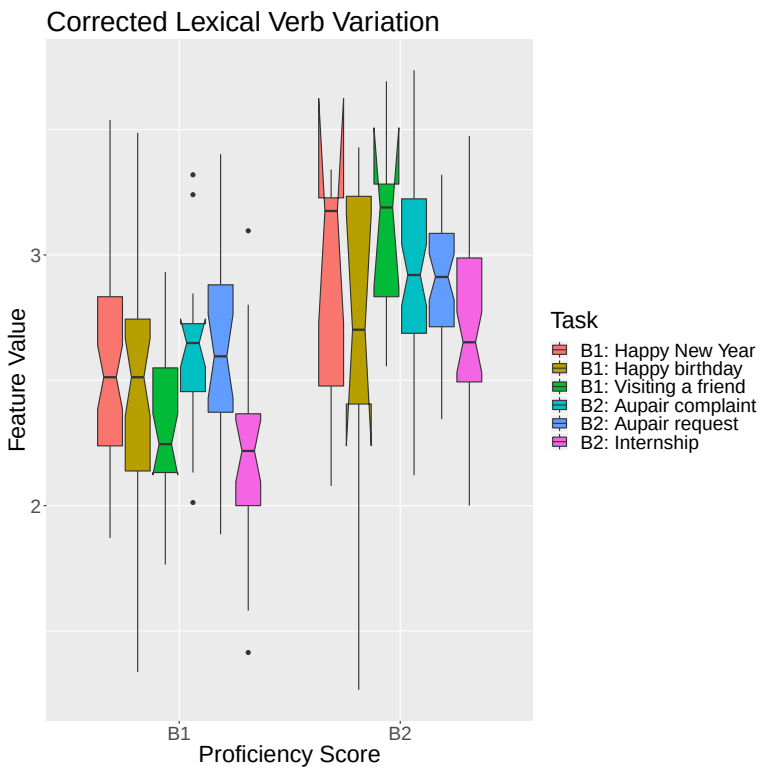


⇒ Verb phrase complexity is influenced by test level but also shows differences between B1 and B2 scores

- No clear proficiency differences for NP or PP complexity

Lexical Complexity

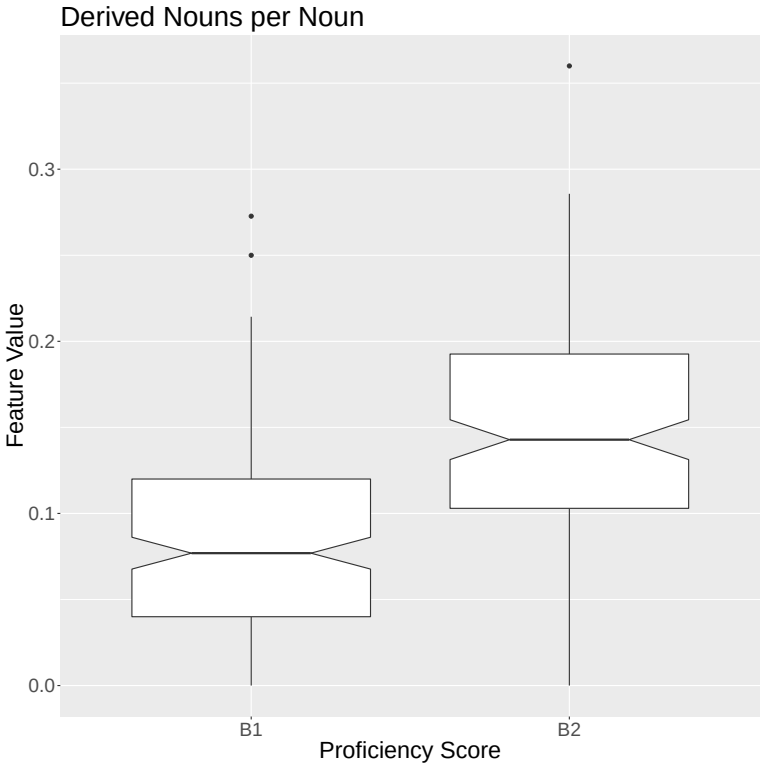


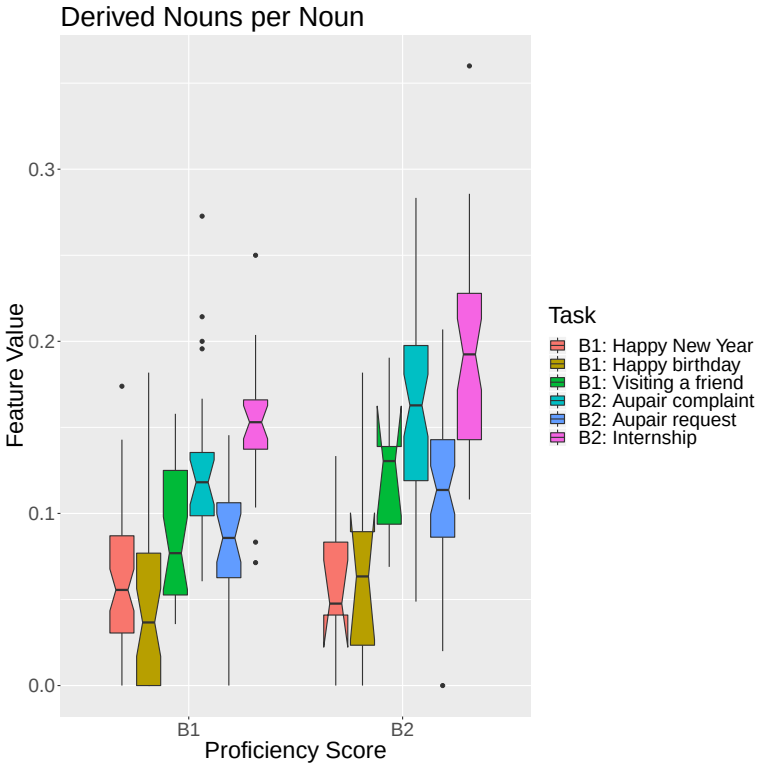


$\Rightarrow$ B2 texts have a higher lexical variation than B1 texts

- Some tasks have a very strong effect on lexical variation
  - Lexical diversity (MTLD) shows no significant differences
  - Semantic relatedness is highly sensitive to task differences

Morphological Complexity



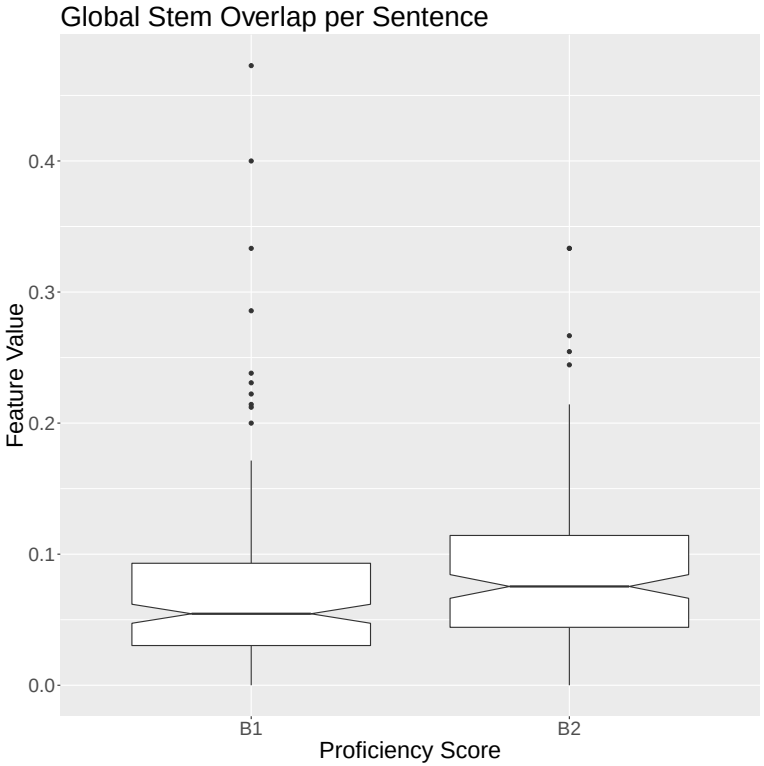


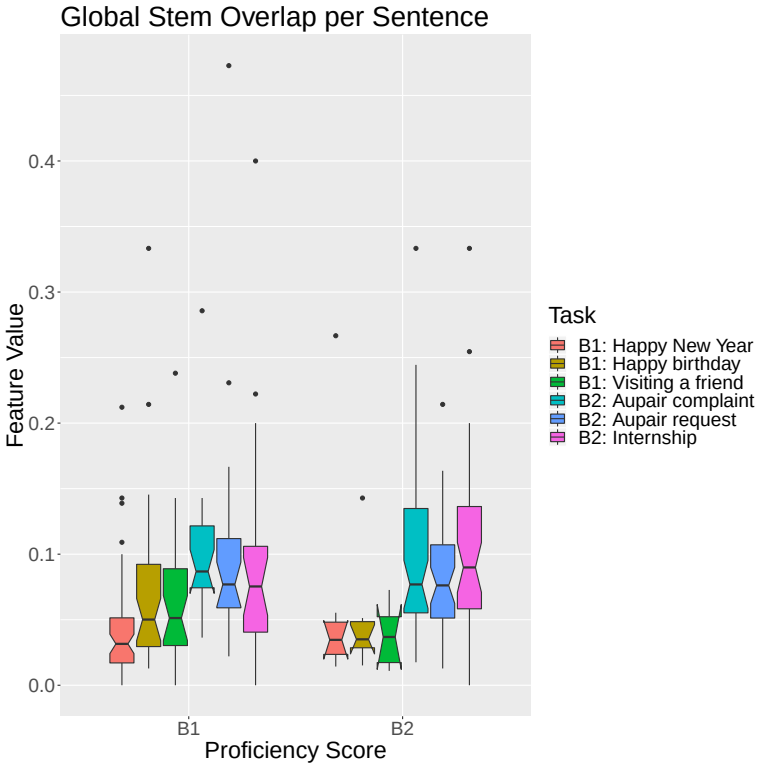
⇒ The use of derived forms is highly task dependent but for most tasks, B2 texts contain more derivation than B1 texts

⇒ Most stable morphological complexity feature we found

- Use of composition more task than proficiency dependent
- Inflection features show even stronger task dependence

Discourse Complexity





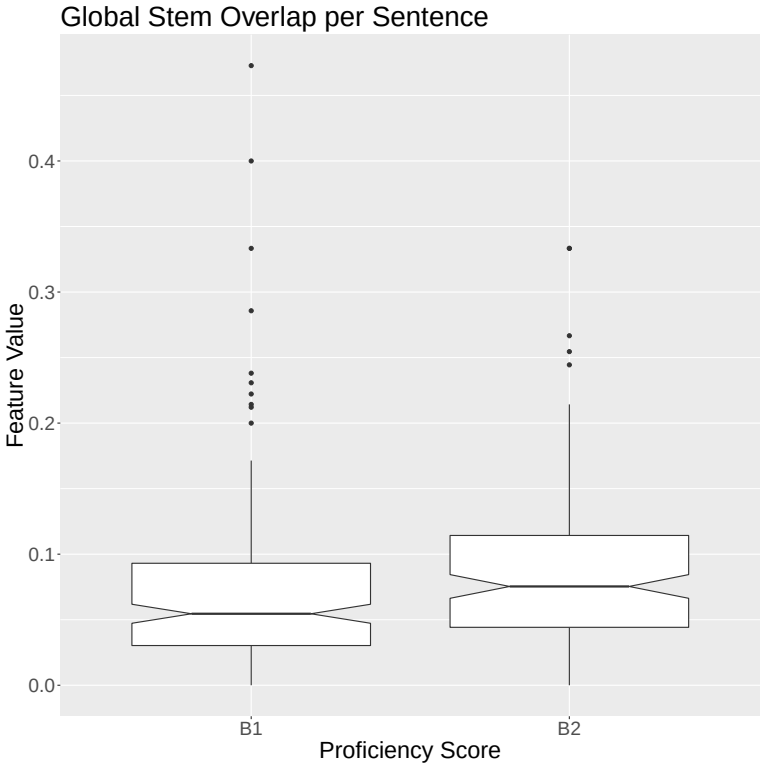
⇒ B2 test takers write more implicitly cohesive texts

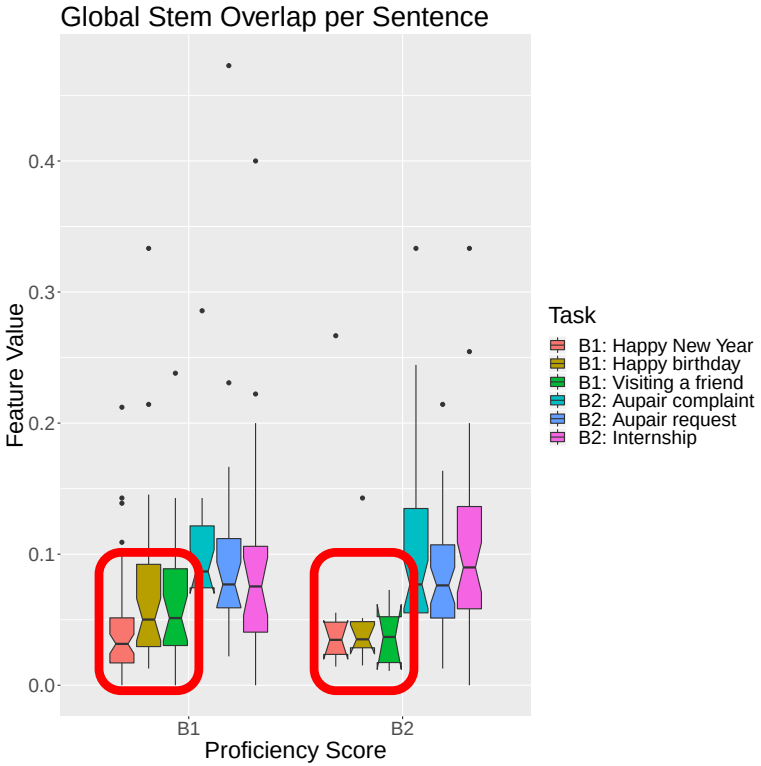
- Limited within task difference between B1 and B2 texts

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

- B2 tests were more often taken by B2 than B1 learners

Discourse Complexity





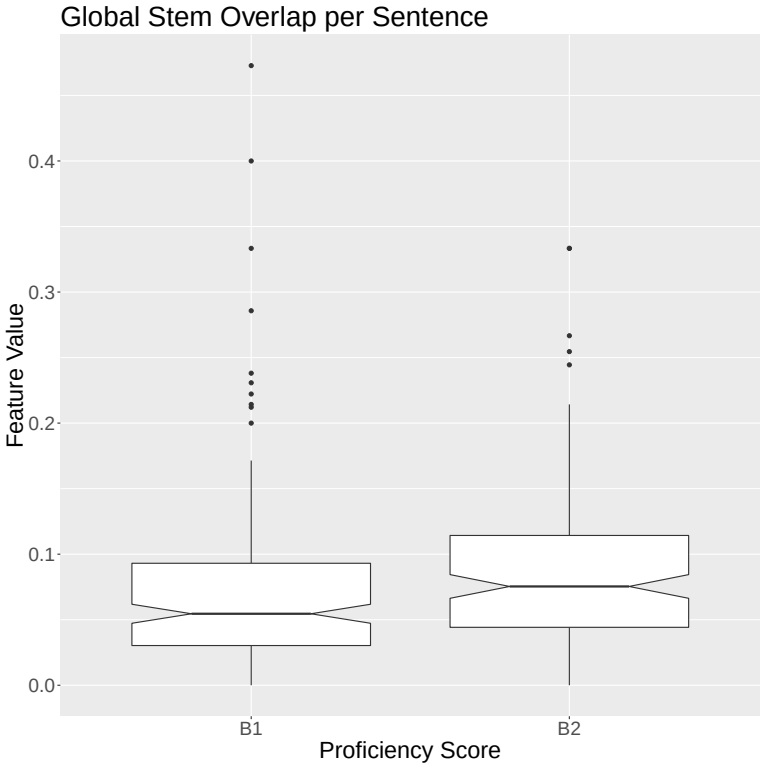
⇒ B2 test takers write more implicitly cohesive texts

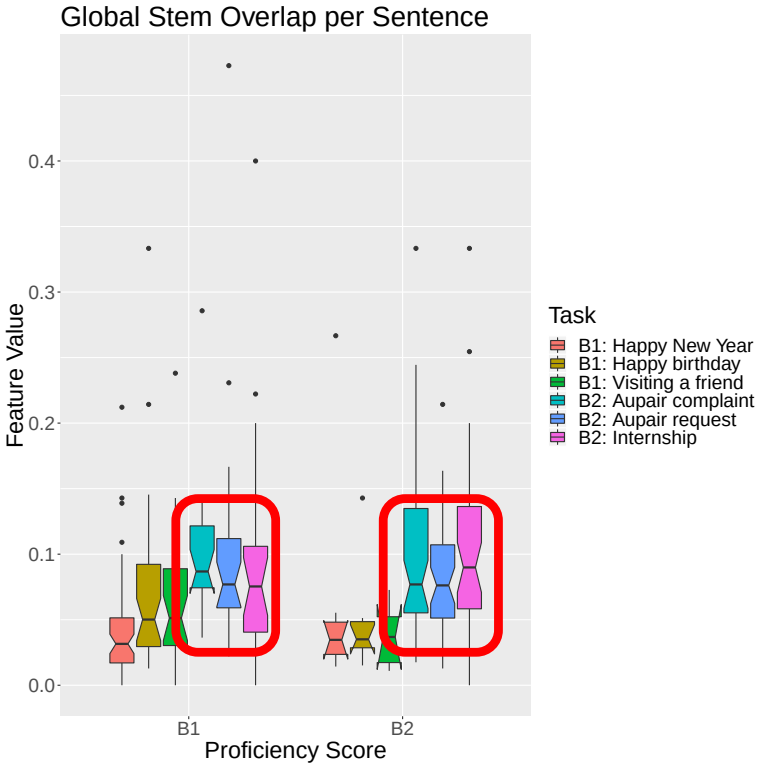
- Limited within task difference between B1 and B2 texts

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

- B2 tests were more often taken by B2 than B1 learners

Discourse Complexity





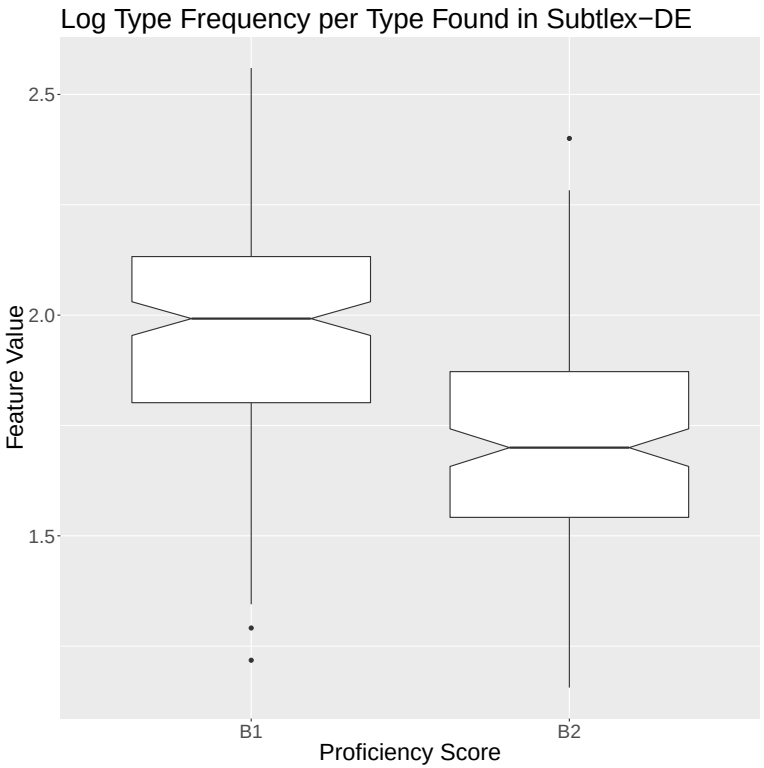
⇒ B2 test takers write more implicitly cohesive texts

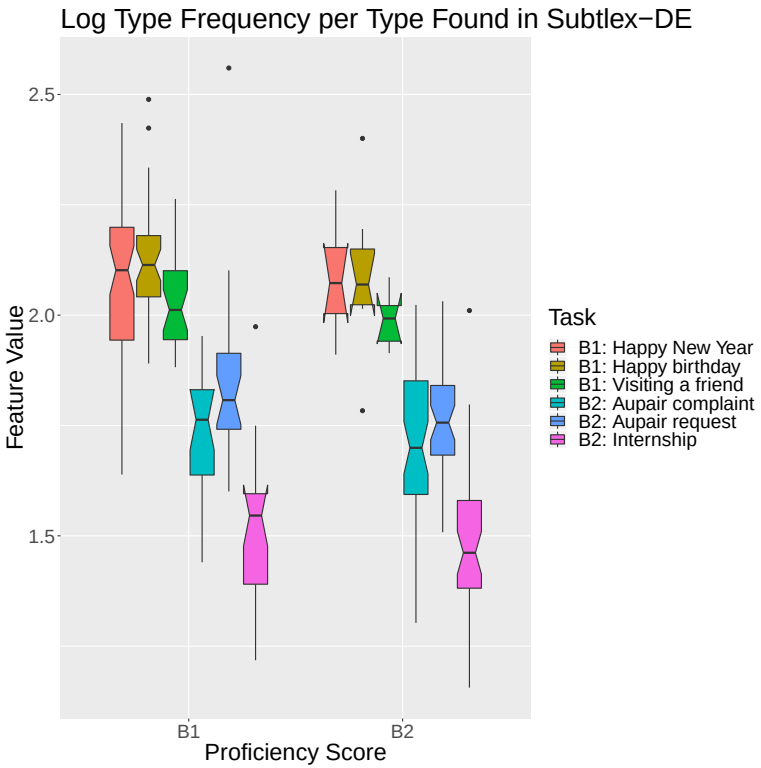
- Limited within task difference between B1 and B2 texts

⇒ Overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

- B2 tests were more often taken by B2 than B1 learners

Language Use



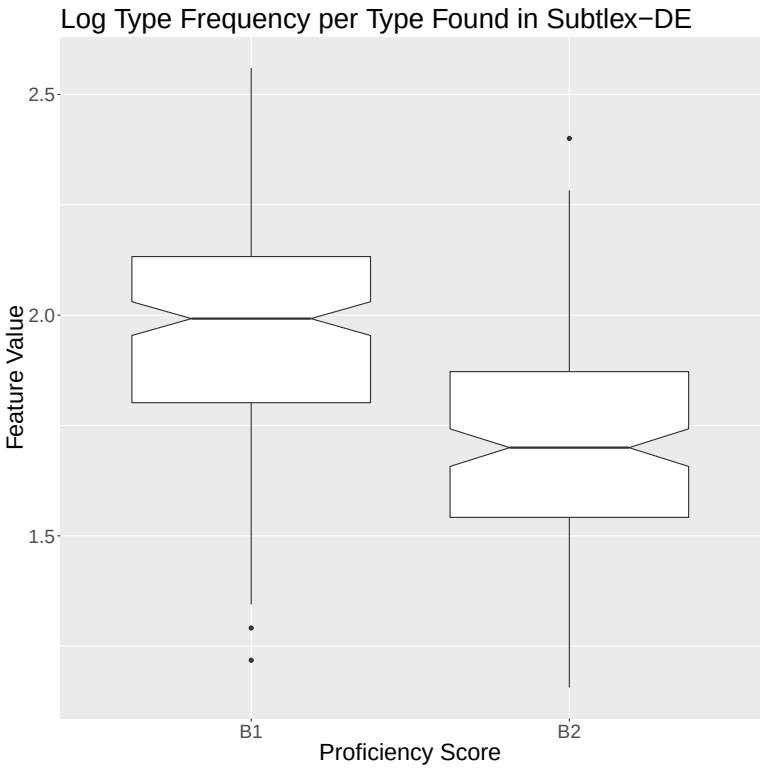


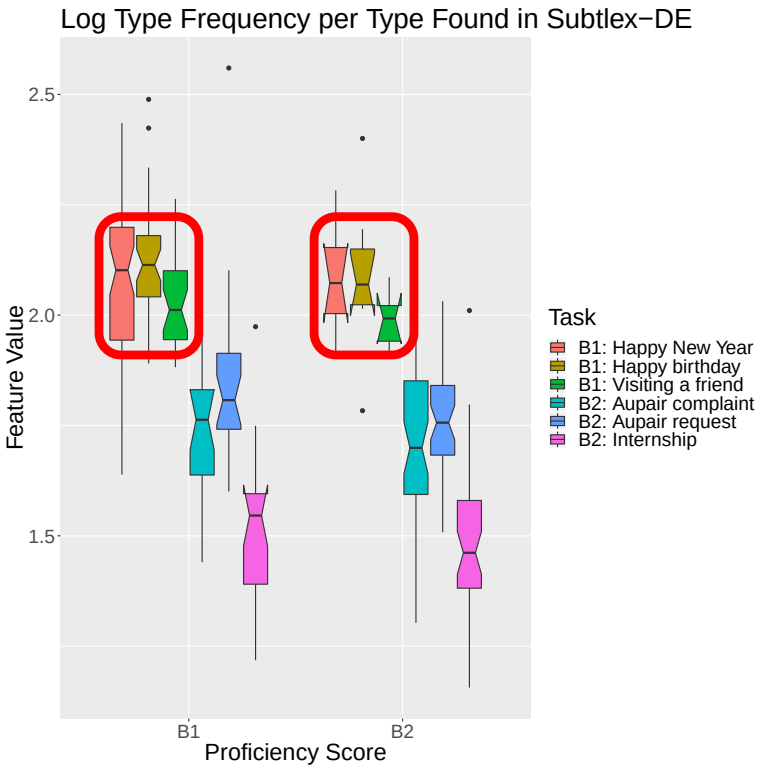
⇒ Tasks for B2 testing elicit more rare vocabulary

- Within tasks, texts by B1 and B2 learners do not differ
- This pattern holds for all frequency features we measured

⇒ Again, overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

Language Use



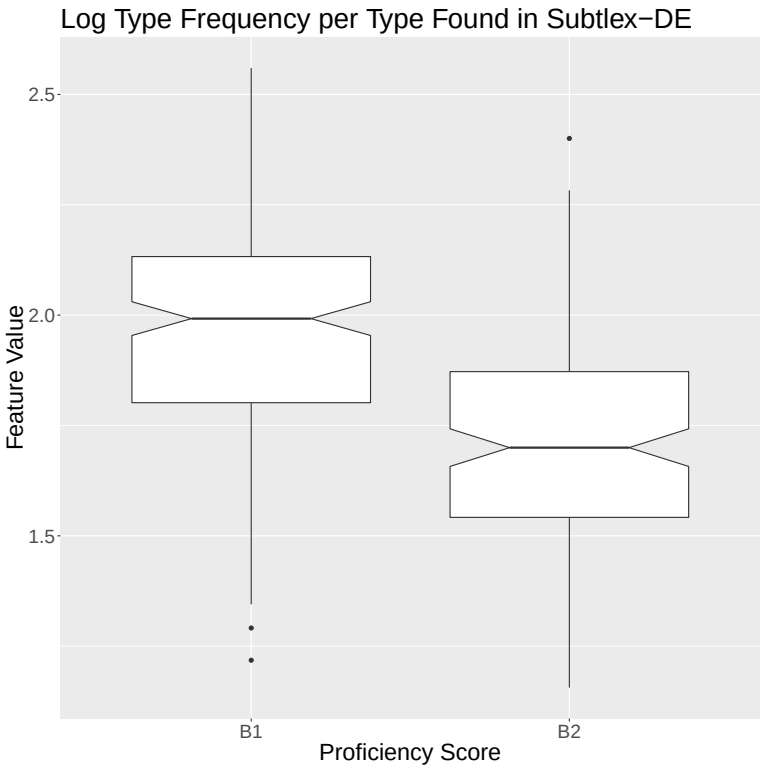


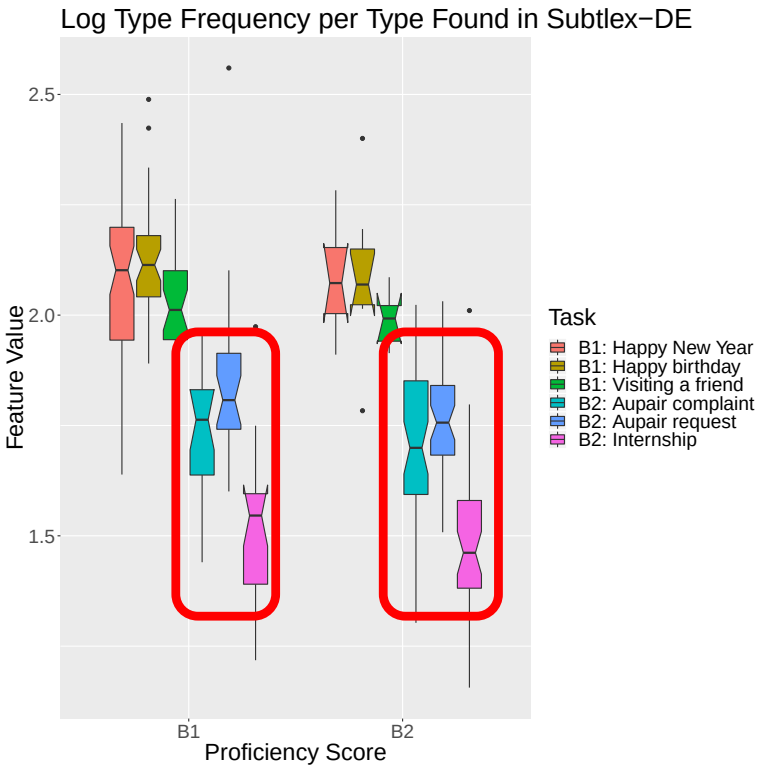
⇒ Tasks for B2 testing elicit more rare vocabulary

- Within tasks, texts by B1 and B2 learners do not differ
- This pattern holds for all frequency features we measured

⇒ Again, overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

Language Use



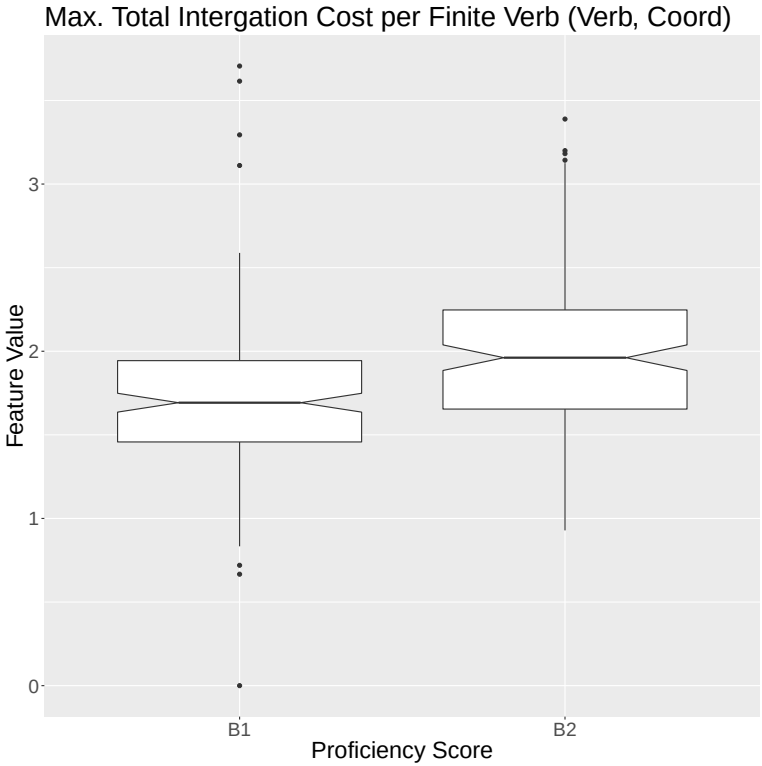


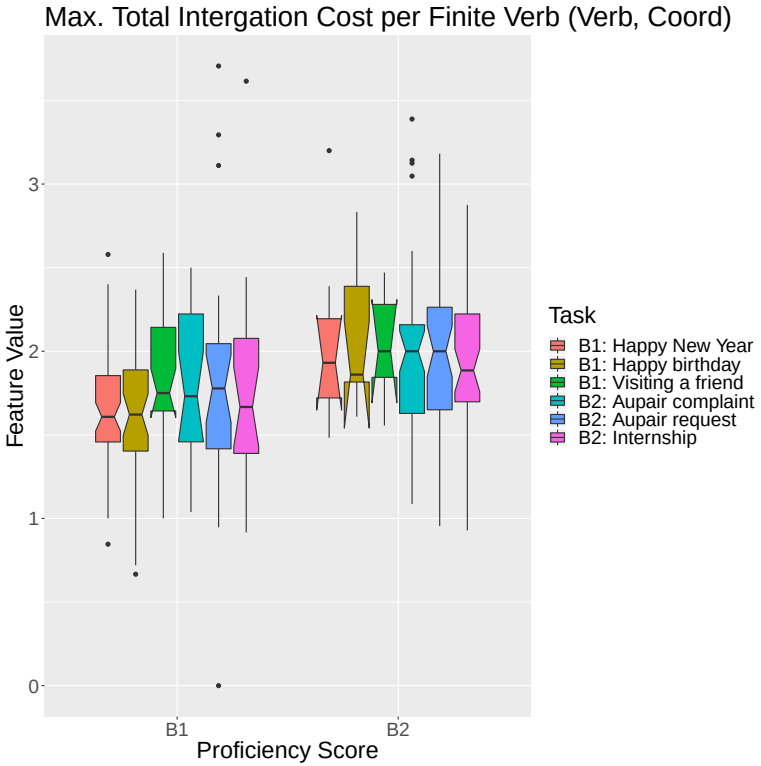
⇒ Tasks for B2 testing elicit more rare vocabulary

- Within tasks, texts by B1 and B2 learners do not differ
- This pattern holds for all frequency features we measured

⇒ Again, overall differences between B1 and B2 texts can be explained by the elicitation bias of CEFR testing

Human Processing Costs





⇒ Integration costs are higher for texts with B2 scores

⇒ Withing proficiency levels, they are stable across tasks

Discussion

- Comparison showed that some presumed proficiency differences can be explained by task differences
  - Morphological complexity
  - Discourse complexity
  - Language use
- Some measures of lexical and phrasal complexity characterize proficiency differences despite task sensitivity
- Two complexity domains were systematically stable
  - Human processing features
  - Clausal complexity features
- Which (combination of) task factors causes these effects?
  - More research and data needed

Conclusion

- System for broad linguistic modeling of linguistic complexity, language use, processing, and discourse
  - Applicable to German, English, French, Spanish
  - Being extended to Italian, Turkish, Arabic, Portuguese
- Characterized differences in proficiency and reading differences to compare L2 input and output
  - Reading input lies systematically above language production in all domains but syntax
  - Is this challenge indeed in the ZPD or above?
- Showed influence of tasks on complexity in L2 writing
  - Demonstrated remarkable role of tasks on complexity
  - Morphology, language use & discourse task dependent
  - Processing and clausal features seem more stable
  - Role of individual task factors remains unclear

Outlook

- Specify interaction between proficiency & appropriate input
  - Is there a trade-off of acceptable challenges between different linguistic domains?
  - Study link in spoken language, e.g., classroom interactions (see COLD project, www.die-bonn.de/cold)
- Interaction of task factors, proficiency, and linguistic domains
  - How does proficiency mediate task effects on CAF?
  - Extension across languages to test generalizability of findings
  - Separating task factors to understand cause of task effects

Outlook

- Specify interaction between proficiency & appropriate input
  - Is there a trade-off of acceptable challenges between different linguistic domains?
  - Study link in spoken language, e.g., classroom interactions (see COLD project, www.die-bonn.de/cold)
- Interaction of task factors, proficiency, and linguistic domains
  - How does proficiency mediate task effects on CAF?
  - Extension across languages to test generalizability of findings
  - Separating task factors to understand cause of task effects

References

Alexopoulou, T., M. Michel, A. Murakami & D. Meurers (2017). Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. *Language Learning* 67, 181–209. URL https://doi.org/10.1111/lang.12232.

Barzilay, R. & M. Lapata (2008). Modeling Local Coherence: An entity based approach. *Computational Linguistics* 34(1), 1–34.

Berendes, K., S. Vajjala, D. Meurers, D. Bryant, W. Wagner, M. Chinkina & U. Trautwein (2018). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology* 110(4), 518–543. URL https://doi.org/10.1037/edu0000225.

Bertram, C., Z. Weiss, L. Zachrich & R. Ziai (in preparation). Artificial Intelligence in the Research of History Teaching: Automating Assessments of Open-Ended Histoy Tasks Through Computational Linguistic Content and Complexity.

Bohnet, B. & J. Nivre (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1455–1465. URL https://www.aclweb.org/anthology/D12-1133.

Brezina, V. & G. Pallotti (2019). Morphological complexity in written L2 texts. *Second language research* 35(1), 99–119.

Brysbaert, M., M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte & A. Böhl (2011a). The word frequency effect: A review of recent developments and implications

for the choice of frequency estimates in German. *Experimental Psychology* 58, 412–424.

Brysbaert, M., E. Keuleers & B. New (2011b). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology* 2(27).

Bulté, B. & A. Housen (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26(0), 42 – 65. URL http://www.sciencedirect.com/science/article/pii/S1060374314000666. Comparing perspectives on L2 writing: Multiple analyses of a common corpus.

Chen, X. & D. Meurers (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*. Osaka, Japan: COLING, pp. 113–119.

Chen, X. & D. Meurers (2017a). Challenging Learners in Their Individual Zone of Proximal Development Using Pedagogic Developmental Benchmarks of Syntactic Complexity. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017*. Gothenburg, Sweden: ACL, Linköping Electronic Conference Proceedings 134, pp. 8–17. http://aclweb.org/anthology/W17-0302.pdf.

Chen, X. & D. Meurers (2017b). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading* 41(3), 486–510.

Chen, X., Z. Weiß& D. Meurers (draft). Developmental trade-off of complexity and accuracy in L1 and L2 acqusition. Ms.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research* 11(3), 415–443.

Crossley, S. A., K. Kyle & D. S. McNamara (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 48(4), 1227–1237.

Crossley, S. A., S. Skalicky & M. Dascalu (2019). Moving beyond classic readability formulas: new methods and new models. *Journal of Research in Reading* 42(3-4), 541–561.

De Clercq, B. & A. Housen (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research Special Issue on Linguistic Complexity* 35(1), 71–97.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita & W. O'Neil (eds.), *Image, language, brain: papers from the First Mind Articulation Project Symposium*, MIT, pp. 95–126.

Graesser, A. C., D. S. McNamara, M. M. Louwerse & Z. Cai (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods* 36(2), 193–202.

Heister, J., K.-M. Würzner, J. Bubenzer, E. Pohl, T. Hanneforth, A. Geyken & R. Kliegl (2011). dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau* 62, 10–20.

Housen, A., B. De Clercq, F. Kuiken & I. Vedder (2019). Multiple approaches to complexity in second language research. *Second Language Research. Special Issue on Linguistic Complexity* 35(1), 2–31.

Housen, A., F. Kuiken & I. Vedder (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken & I. Vedder (eds.), *Dimensions of L2 Performance and Proficiency*, John Benjamins, Language Learning & Language Teaching, pp. 1–20. URL https://doi.org/10.1075/lllt.32.01hou.

Krashen, S. D. (1981). The Fundamental Pedagogical Principle in Second Language Teaching. *Studia Linguistica* 35(1–2), 50–70. URL http://www3.interscience.wiley.com/journal/121500538/abstract?CRETRY=1&SRETRY=0.

Kühberger, C., C. Bramann, Z. Weiss & D. Meurers (2019). Task Complexity in History Textbooks. A Multidisciplinary Case Study on Triangulation in History Education Research. *History Education International Research Journal (HEIRJ)* 16(1). URL https://doi.org/10.18546/HERJ.16.1.12. Special Issue on Mixed Methods and Triangulation in History Education Research.

Lantolf, J. P., S. L. Thorne & M. E. Poehner (2015). Sociocultural theory and second language development. In *Theories in second language acquisition: An introduction*, Routledge New York, NY.

Lavalley, R., K. Berkling & S. Stüker (2015). Preparing Children's Writing Database for Automated Processing. In *LTLT@ SLaTE*. pp. 9–15.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Languages Journal* pp. 190–208.

Ludewig, U., Z. Weiss & D. Meurers (in preparation). Text comprehension and reading demands: What Features Make Text Comprehension Difficult Across Elementary School? .

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard & D. McClosky (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pp. 55–60. http://aclweb.org/anthology/P/P14/P14-5010.

McCarthy, P. M. & S. Jarvis (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2), 381–392.

McNamara, D. A., A. C. Graesser, P. M. McCarthy & Z. Cai (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, M.A.: Cambridge University Press.

Pallotti, G. (2009). CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics* 30(4), 590–601. URL http://applij.oxfordjournals.org/content/30/4/590.full.pdf.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research* 31(1), 117–134.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research* 35(1), 121–145.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics* 22(1), 27–57.

Shain, C., M. van Schijndel, R. Futrell, E. Gibson & W. Schuler (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Osaka, pp. 49–58. URL https://aclweb.org/anthology/W16-4106.

Siemen, P., A. Lüdeling & F. H. Müller (2006). FALKO - ein fehlerannotiertes Lernerkorpus des Deutschen. In *Proceedings of Konvens*. Konstanz. URL http://www2.informatik.hu-berlin.de/~siemen/daten/falko.pdf.

Skehan, P. (1996). A Framework for the Implementation of Task-based Instruction. *Applied Linguistics* 17(1), 38.

Todirascu, A., T. François, N. Gala, C. Fairon, A.-L. Ligozat & D. Bernhard (2013). Coherence and Cohesion for the Assessment of Text Readability. In *Proceedings of the 10th International Workshop on Natural Language Processing and Cognitive Science*. URL http://www.kicorangel.com/wp-content/uploads/2013/10/NLPCS2013-proceedings.pdf.

Tracy-Ventura, N. & F. Myles (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research* 1(1), 58–95.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. pp. 163–173. http://aclweb.org/anthology/W12-2019.pdf.

Weiss, Z. (2017). Using Measures of Linguistic Complexity to Assess German L2 Proficiency in Learner Corpora under Consideration of Task-Effects. Master's thesis, University of Tübingen, Germany. URL http://www.sfs.uni-tuebingen.de/~zweiss/ma-thesis/weiss2017-distr.pdf.

Weiss, Z., X. Chen & D. Meurers (2021). Using Broad Linguistic Complexity Modeling for Cross-Lingual Readability Assessment pp. 38–54.

Weiss, Z., K. Lange-Schubert, B. Geist & D. Meurers (in preparation a). Sprachliche Komplexitä"t im Unterricht: Eine computerlinguistische Analyse der gesprochenen Sprache von Lehrenden und Lernenden im naturwissenschaftlichen Unterricht in der Primar- und Sekundarstufe. *Zeitschrift fü"r Germanistische Linguistik* .

Weiss, Z. & D. Meurers (2018). Modeling the Readability of German Targeting Adults and Children: An Empirically Broad Analysis and its Cross-Corpus Validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*. Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1026.

Weiss, Z. & D. Meurers (2019a). Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School. In *Proceedings of the 14th Workshop on Innovative Use of*

NLP for Building Educational Applications (BEA). Florence, Italy: Association for Computational Linguistics.

Weiss, Z. & D. Meurers (2019b). Broad Linguistic Modeling is Beneficial for German L2 Proficiency Assessment. In A. Abel, A. Glaznieks, V. Lyding & L. Nicolas (eds.), *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Louvain-La-Neuve: Presses Universitaires de Louvain.

Weiss, Z. & D. Meurers (2021). Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. *International Journal of Learner Corpus Research* 7(1), 84–131.

Weiss, Z., S. Prediger, K. Erath & D. Meurers (in preparation b). Assessing academic language proficiency by monitoring language complexity in authentic oral language use in school.

Weiss, Z., A. Riemenschneider, P. Schröter & D. Meurers (2019). Computationally Modeling the Impact of Task-Appropriate Language Complexity and Accuracy on Human Grading of German Essays. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. Florence, Italy.

Wisniewski, K., K. Schöne, L. Nicolas, C. Vettori, A. Boyd, D. Meurers, A. Abel & J. Hana (2013). MERLIN: An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In *ICT for Language Learning 2013, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni*.

Wolfe-Quintero, K., S. Inagaki & H.-Y. Kim (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity*. Tech. rep., Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, Manoa, Hawaii. URL https://doi.org/10.2307/3587656.

Yang, W., X. Lu & S. C. Weigle (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing* 28, 53–67. URL https://doi.org/10.1016/j.jslw.2015.02.002.

Yoon, H.-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System* 66, 130–141.

Yoon, H.-J. & C. Polio (2016). The Linguistic Development of Students of English as a Second Language in Two Written Genres. *TESOL Quarterly* pp. 275–301. URL https://doi.org/10.1002/tesq.296.

How do we define complexity?

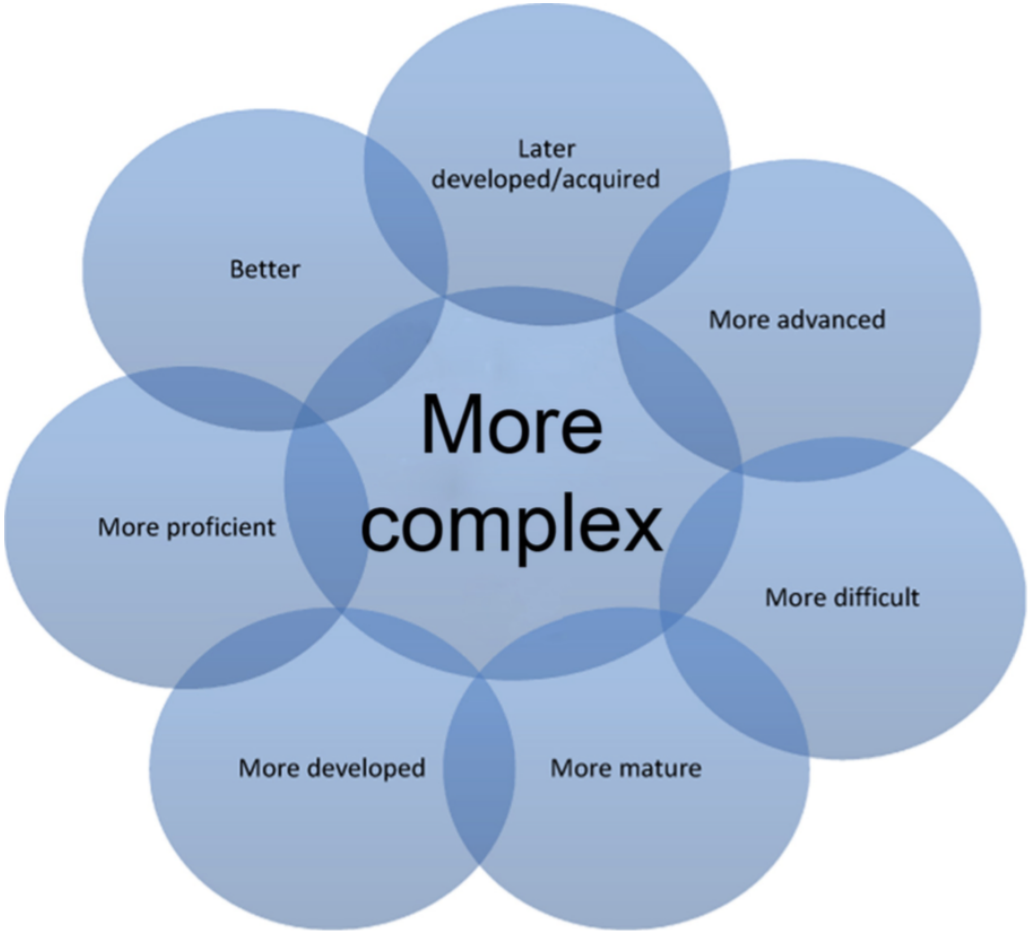


Figure from Bulté & Housen (2014)

Taxonomy of Complexity

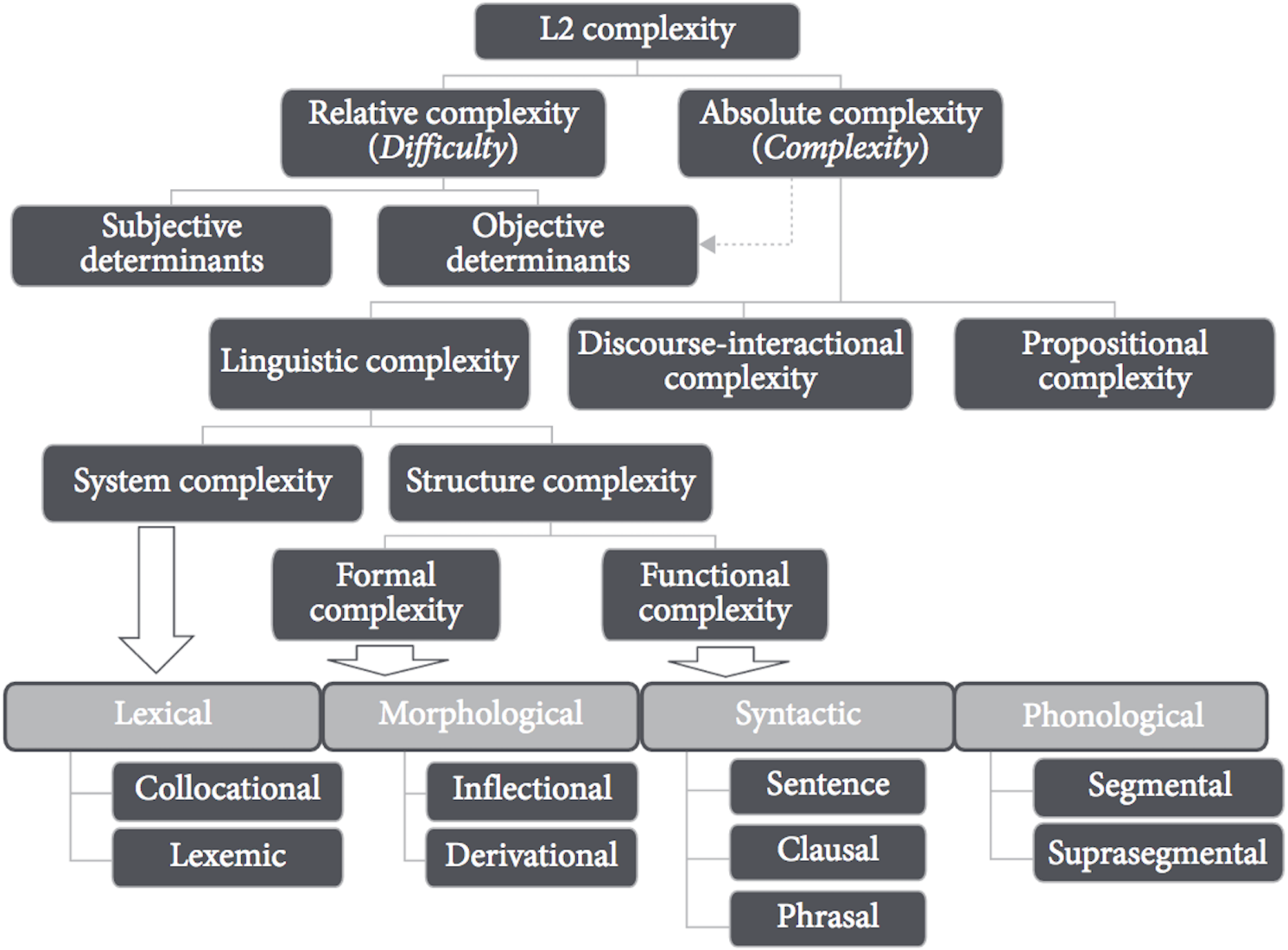


Figure from Housen et al. (2012)

Taxonomy of Complexity

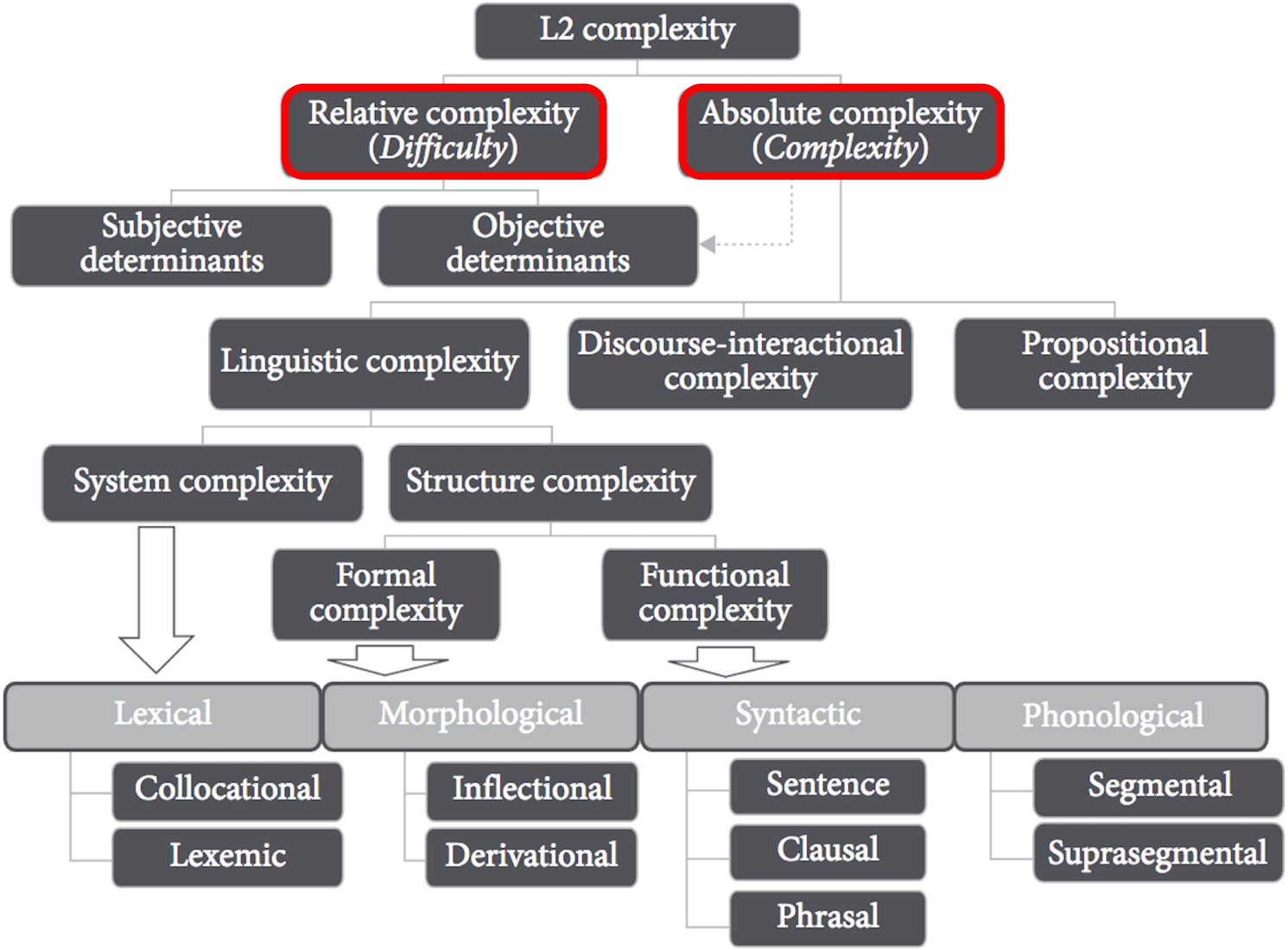


Figure from Housen et al. (2012)

Taxonomy of Complexity

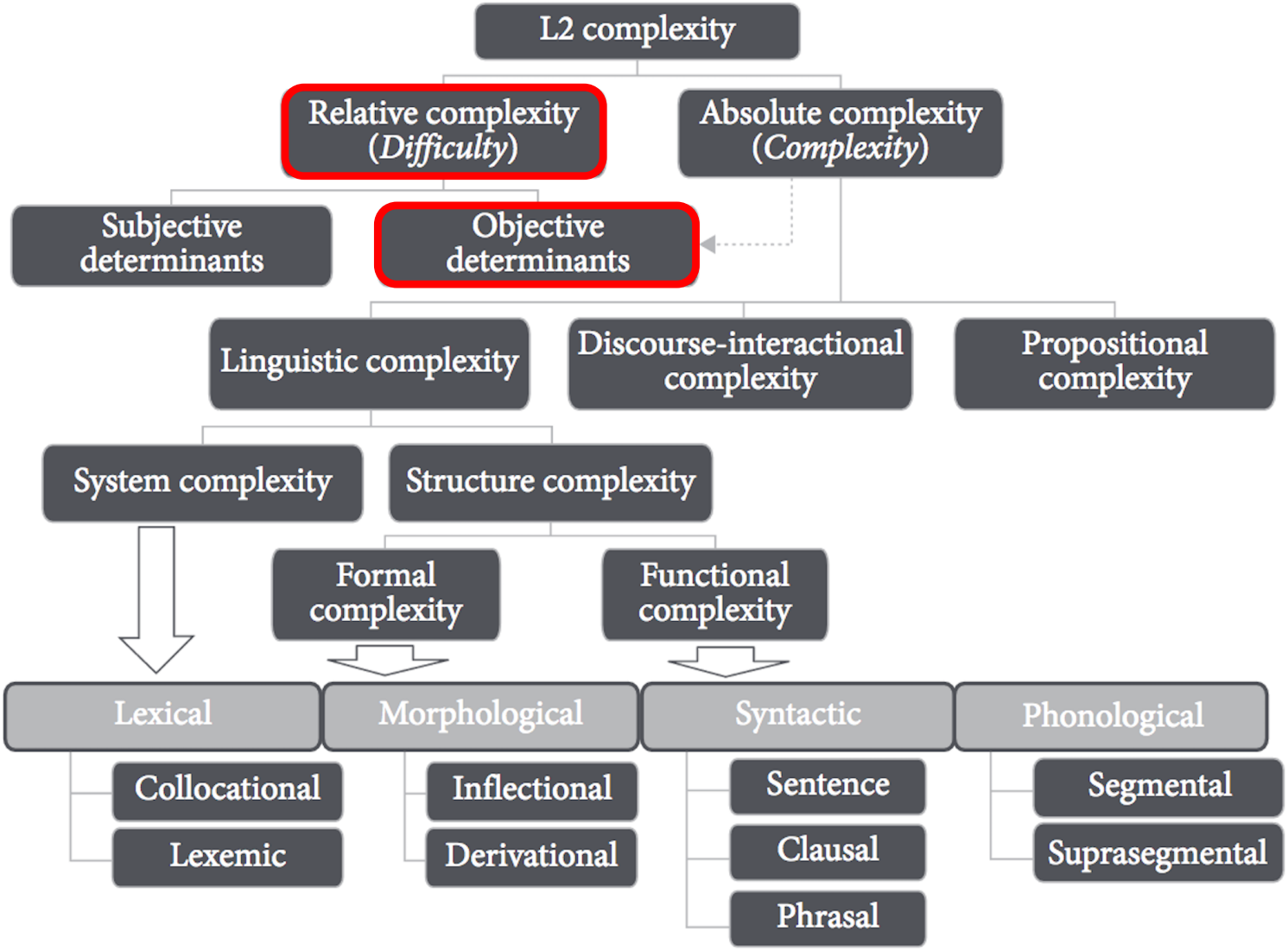


Figure from Housen et al. (2012)

Taxonomy of Complexity

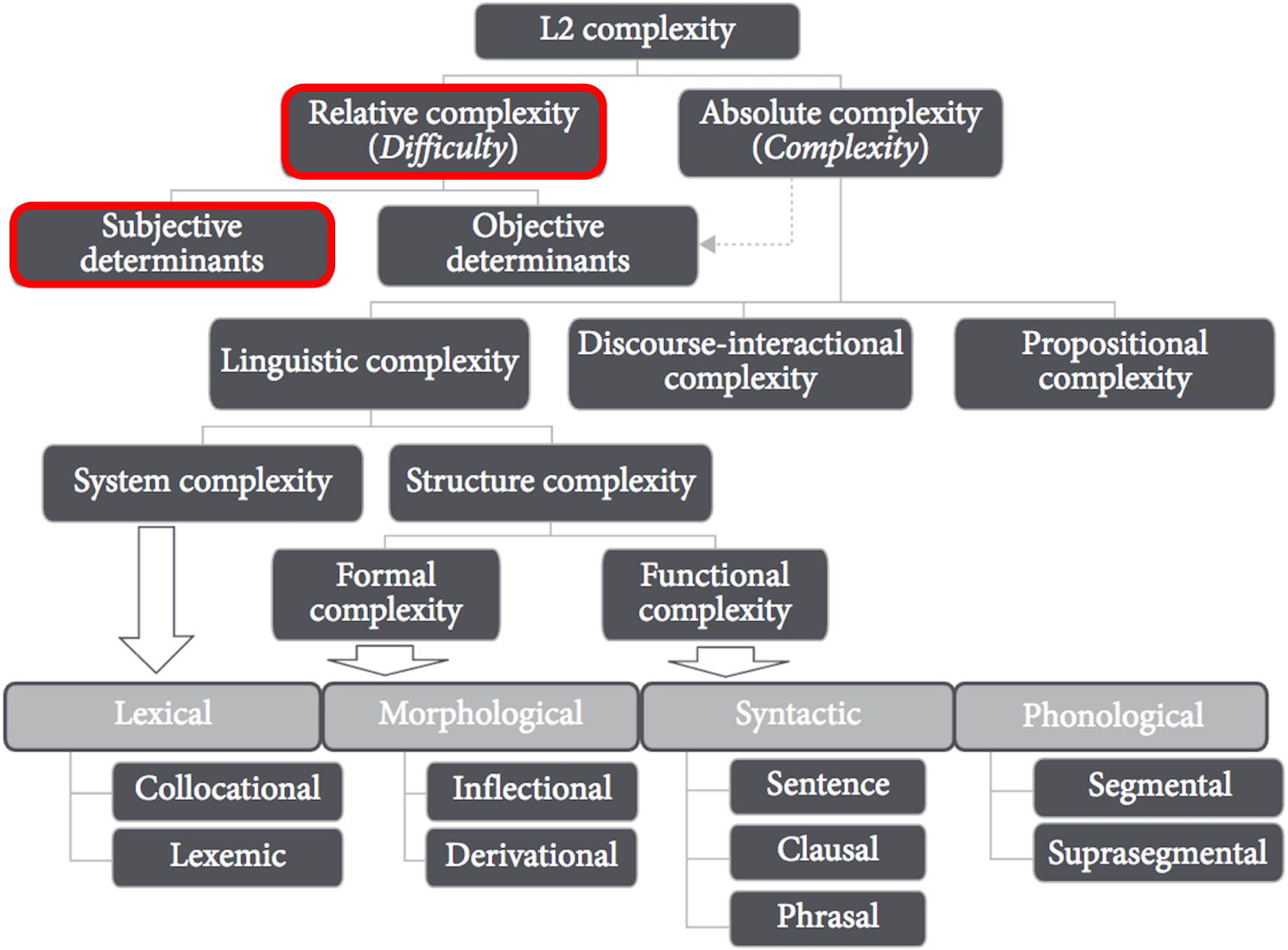


Figure from Housen et al. (2012)

Taxonomy of Complexity

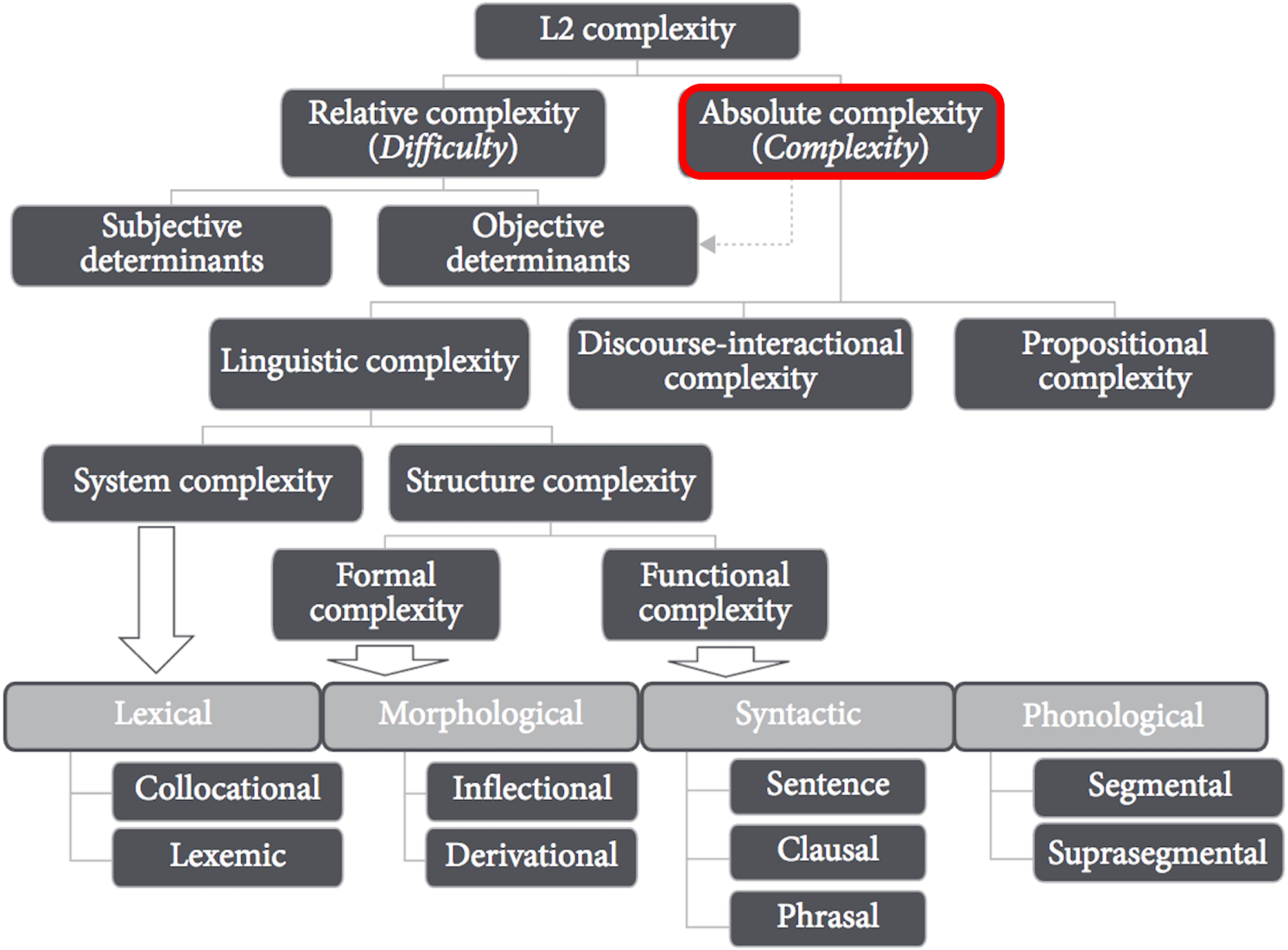


Figure from Housen et al. (2012)

Taxonomy of Complexity

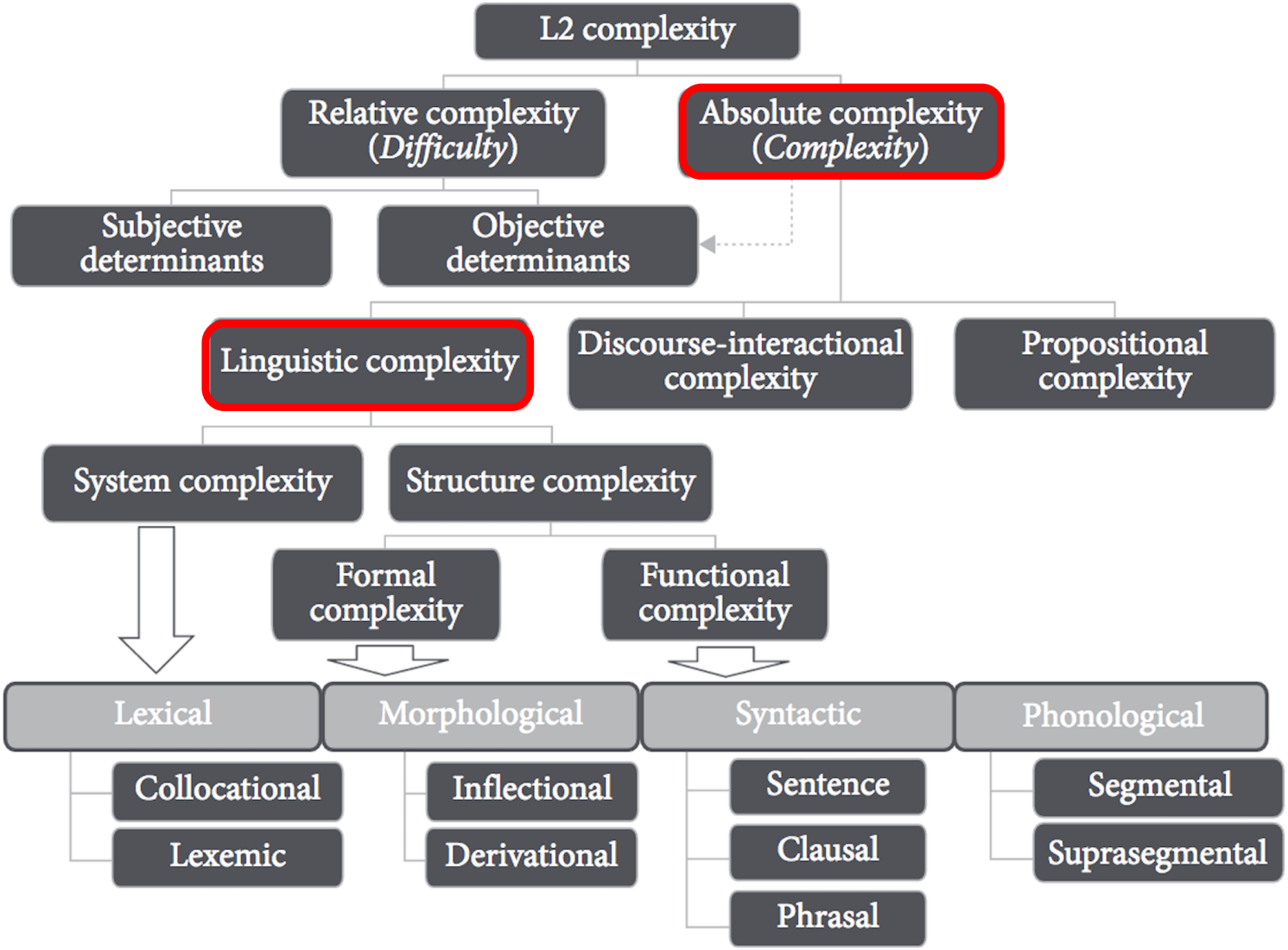


Figure from Housen et al. (2012)

Taxonomy of Complexity

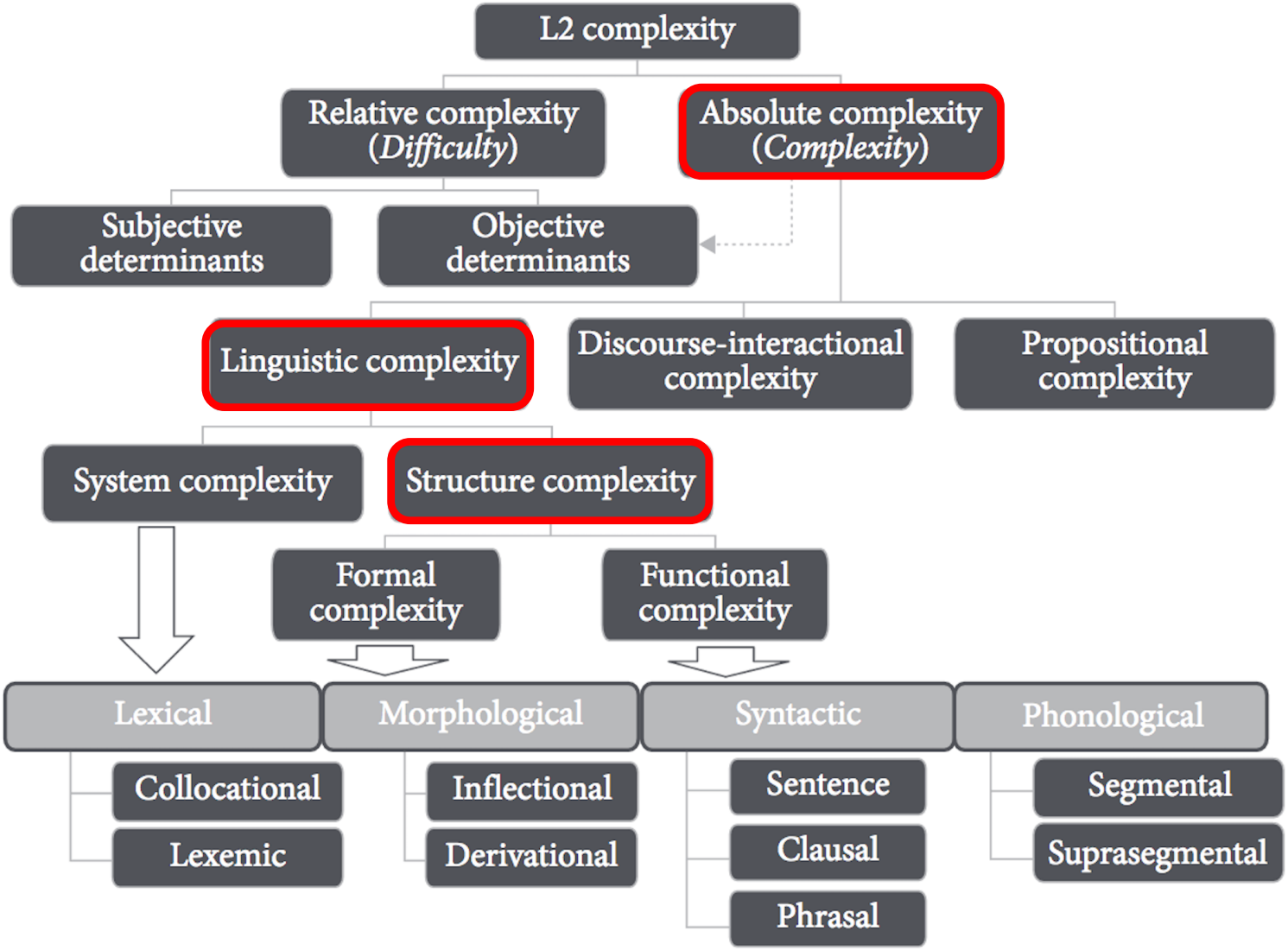
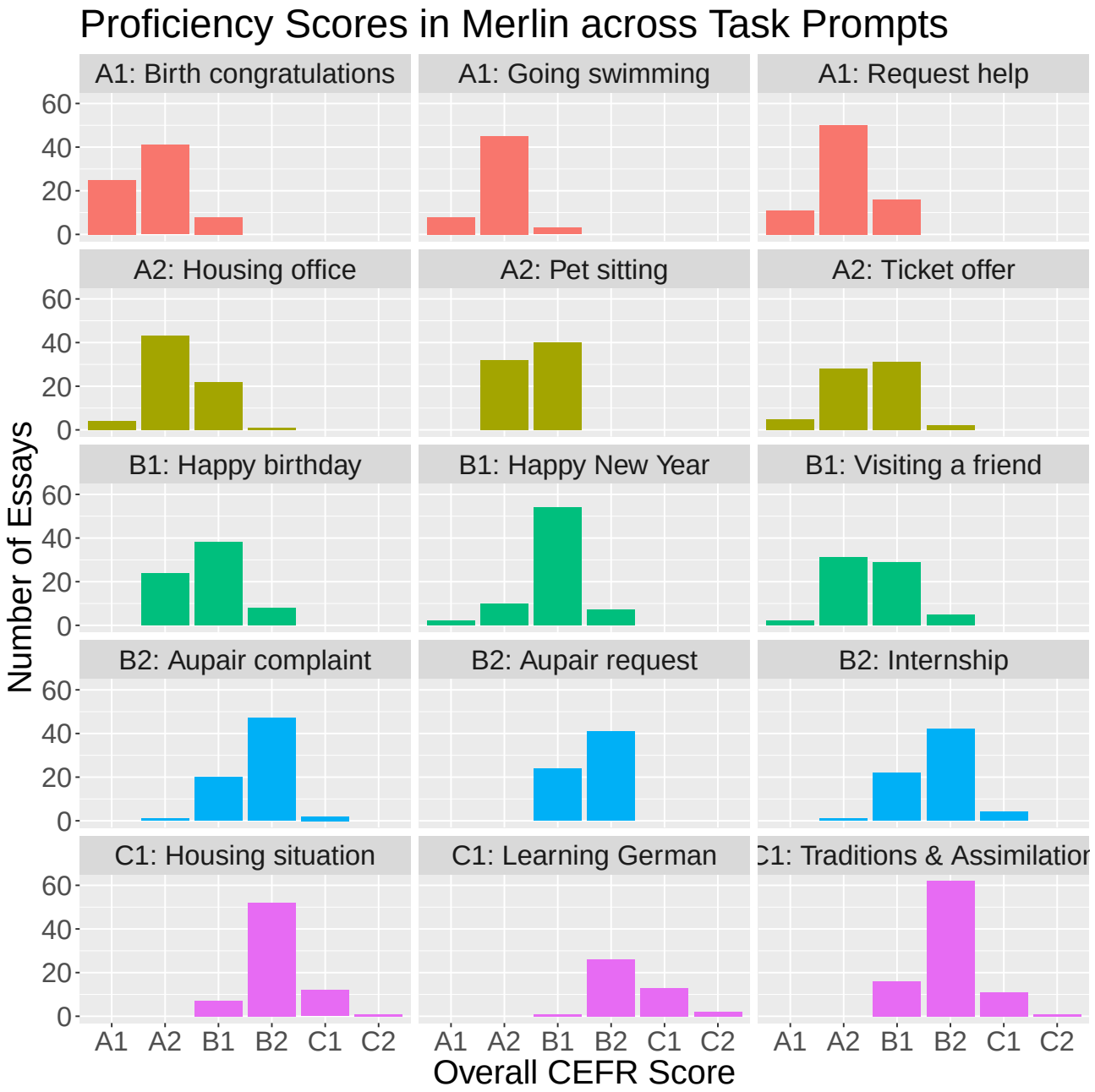


Figure from Housen et al. (2012)

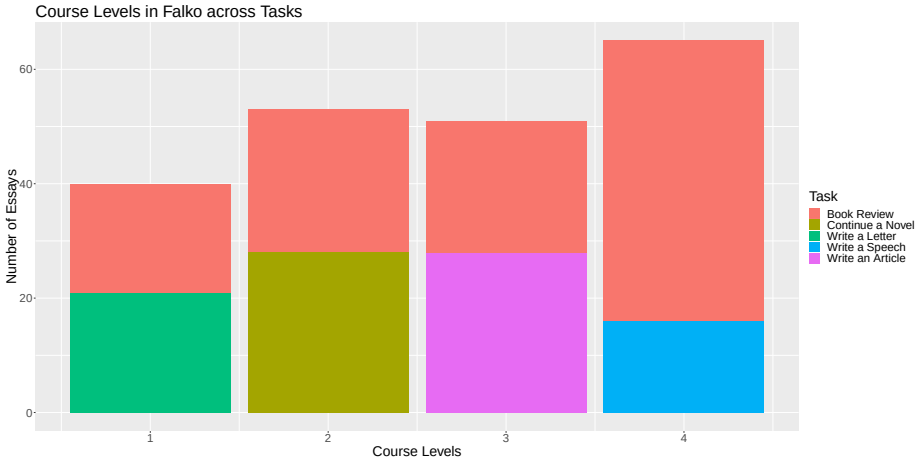
Broad multi-lingual complexity modeling

1. Shared NLP pipeline to obtain
   - Sentences, words, POS tags, constituency/dependency parses with Stanford CoreNLP (Manning et al. 2014)
   - Lemmas with Mate tools (Bohnet & Nivre 2012)
   - Stems with OpenNLP Snowball stemmer
2. Identify linguistic constructions with extraction rules
   - Language-specific rules for syllables, POS, constituents
   - Language-independent rules for all other
   - Frequency features based on Subtlex-US and Subtlex-DE (Brysbaert et al. 2011a,b)
3. Fully language-independent calculation of feature ratios

Proficiency split by tasks in Merlin

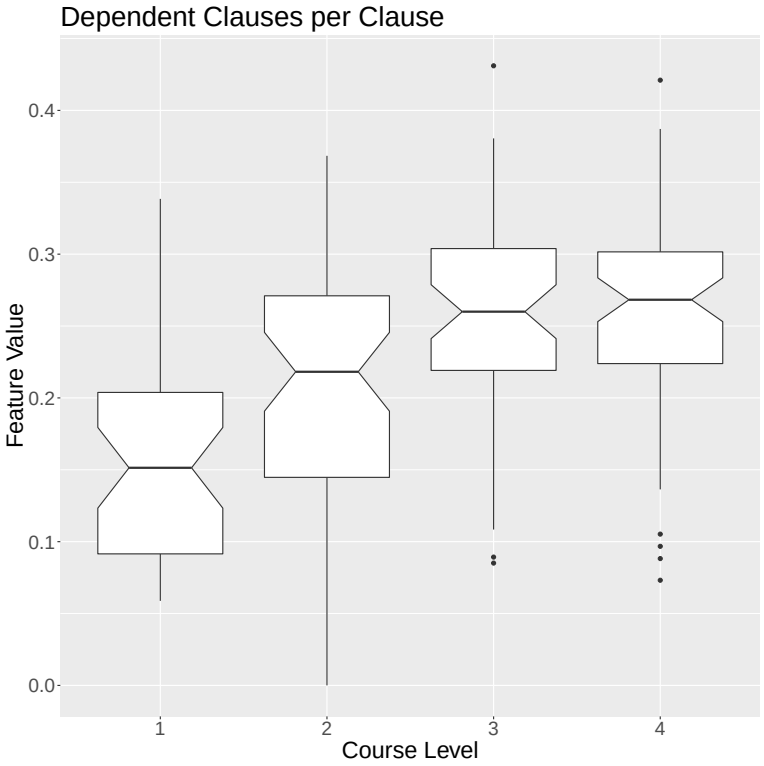


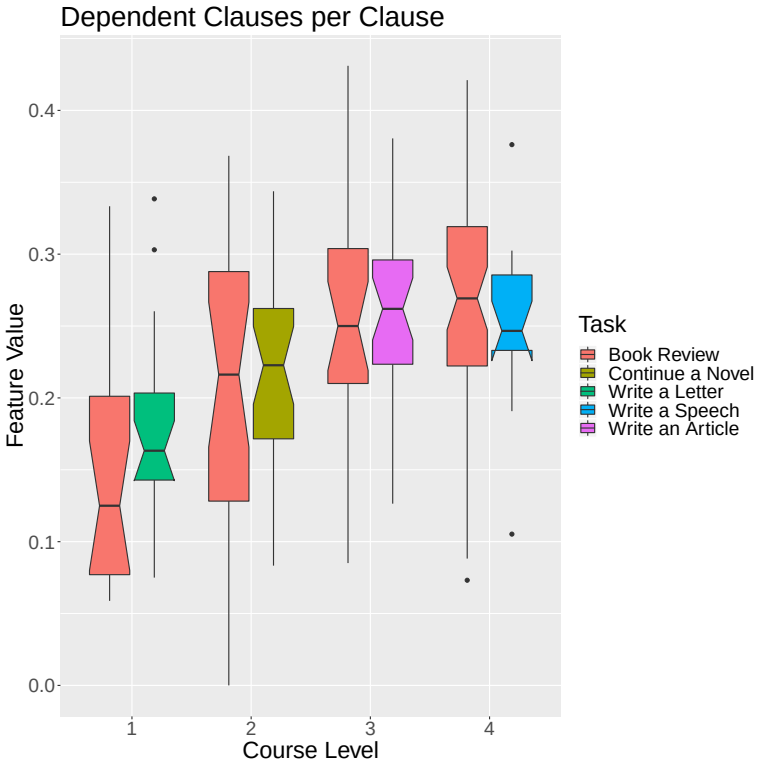
Falko Georgetown corpus



- Corpus of 209 texts written by 123 German L2 learners (Siemen et al. 2006)
- Elicited in curricular writing courses at the Georgetown University ranging from medium (1) to advanced (4)
  - Take-home final exams written by students
  - Task prompt either curriculum dependent or reference task
  - Several students attended consecutive courses
- No additional external validation of proficiency available

Clausal Complexity





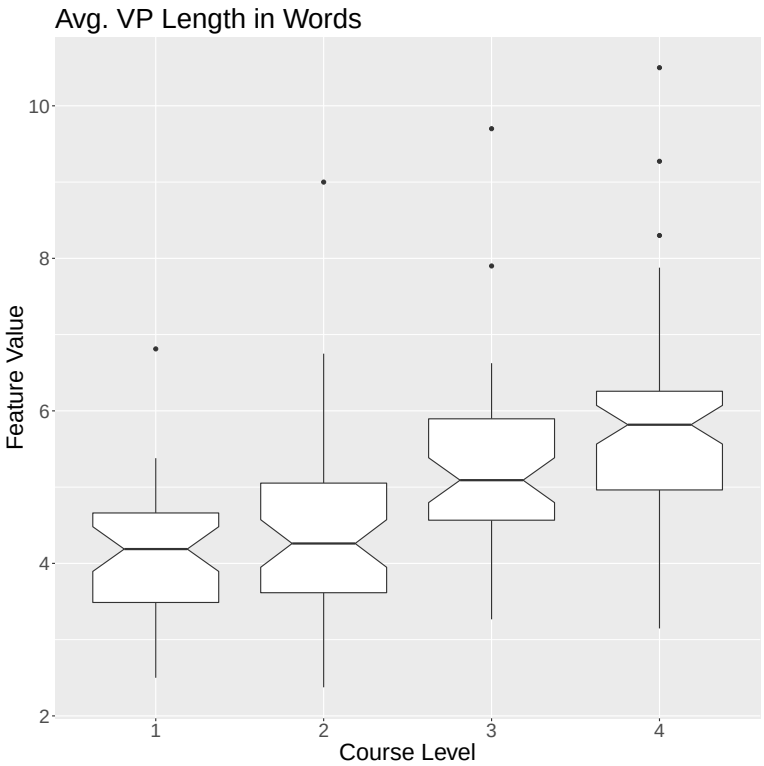
⇒ Students systematically use more dependent clauses per clause with increasing course levels

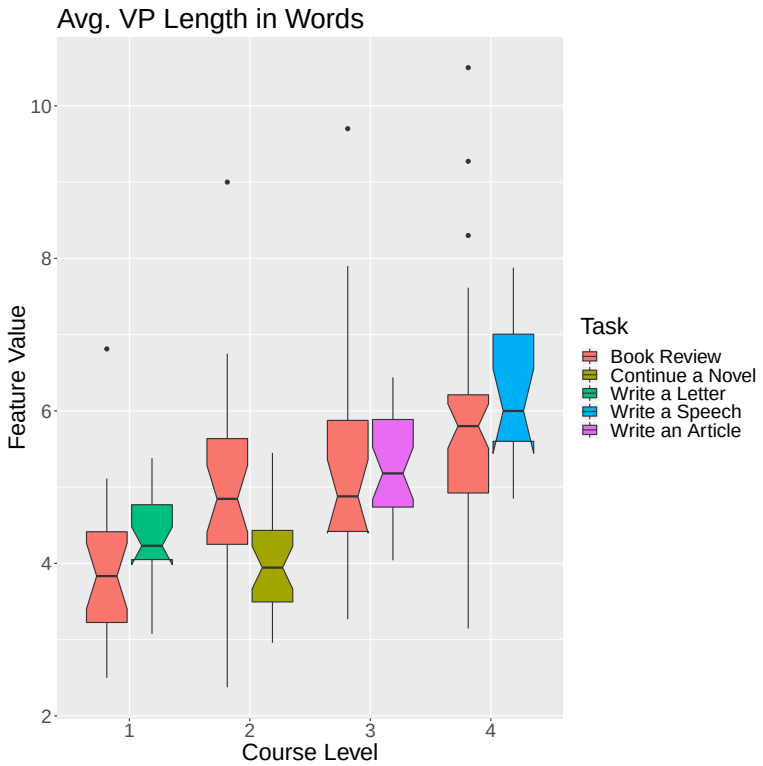
- This holds across measures of clausal complexity

⇒ Relative stability within a course level

- More pronounced differences for some features

Phrasal Complexity

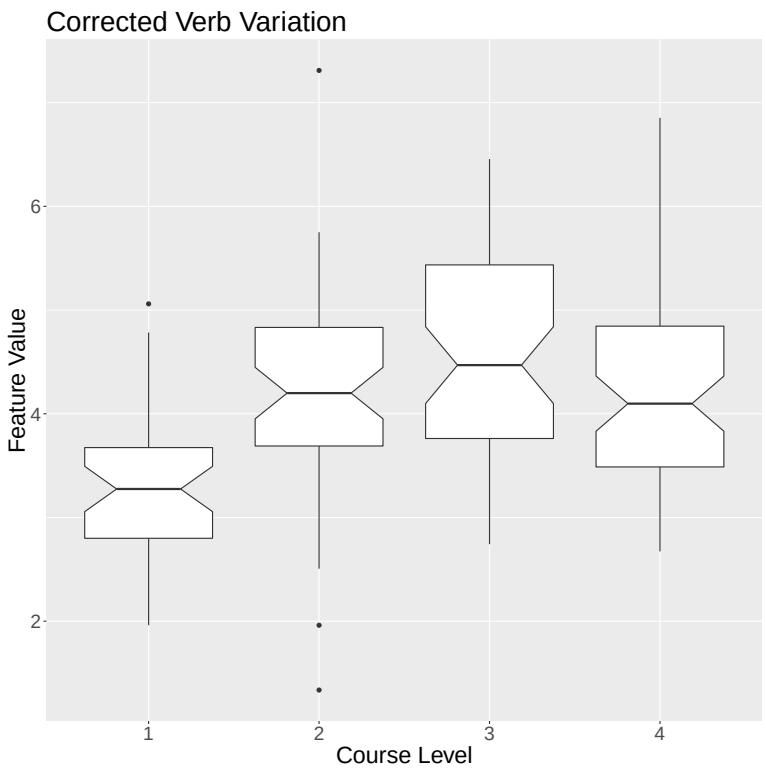


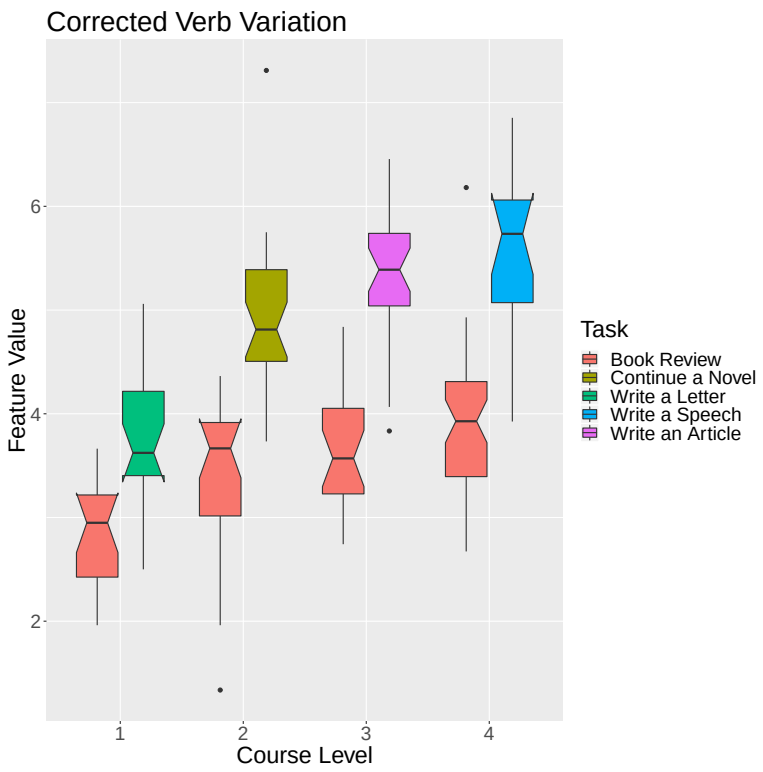


⇒ VP length increases across course levels relatively stable

- PP and NP complexity features are more task dependent and show a less clear development on the reference task

Lexical Complexity



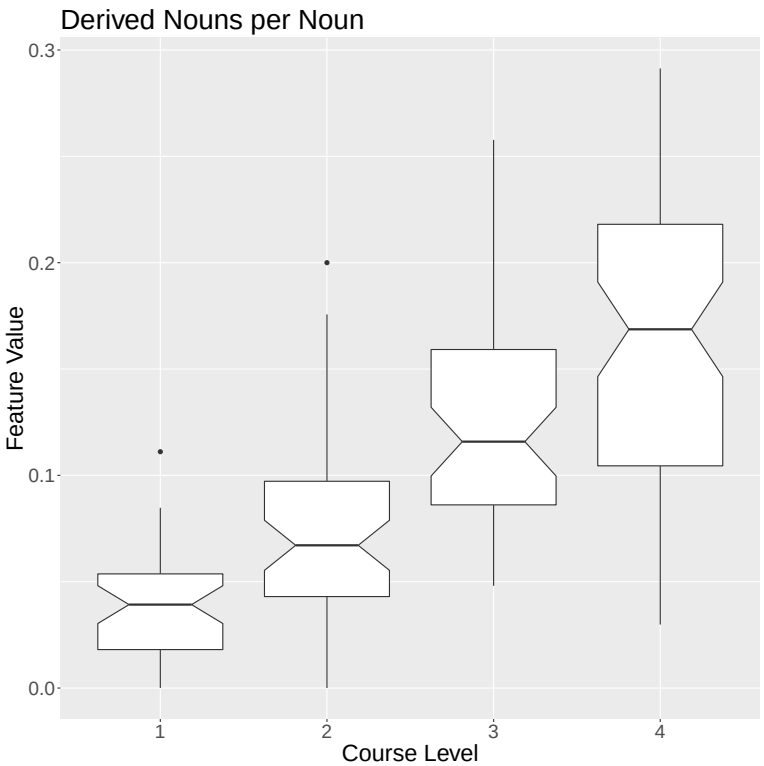


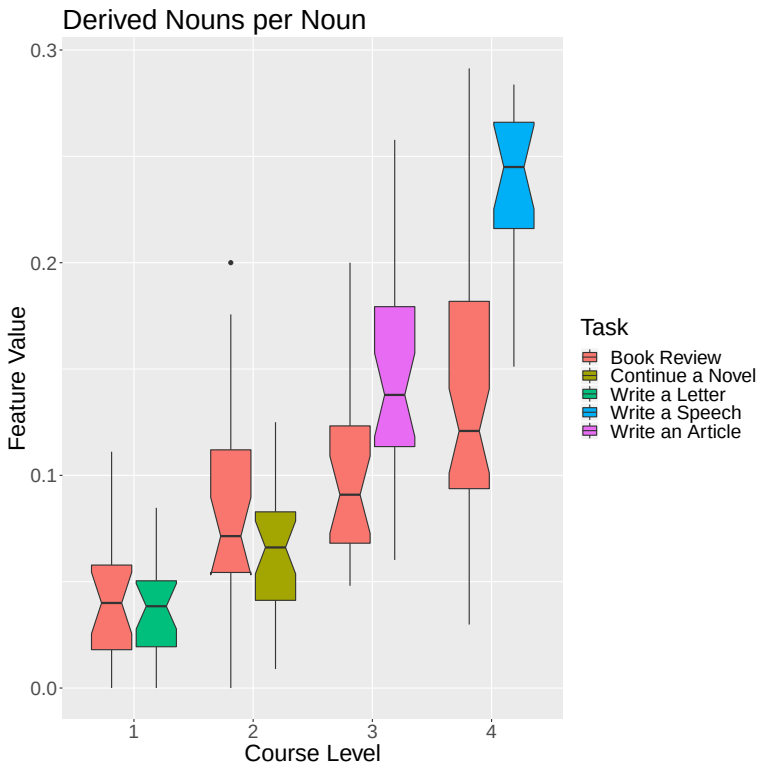
⇒ More lexically varied book reviews at higher course levels

⇒ Curriculum tasks elicit more lexically varied writing

- Lexical diversity shows a comparable pattern
- Semantic relatedness is only sensitive to task differences

Morphological Complexity

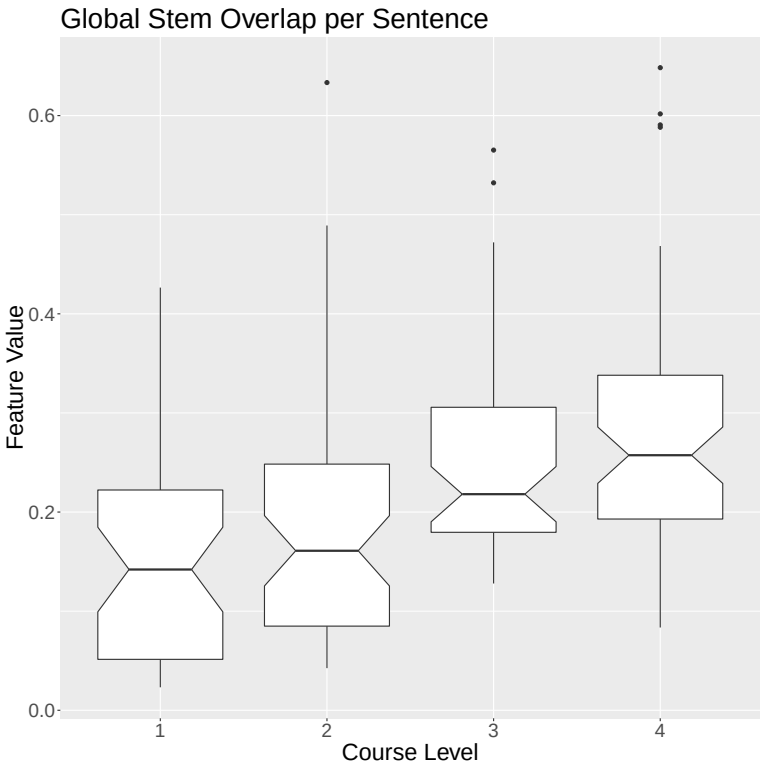


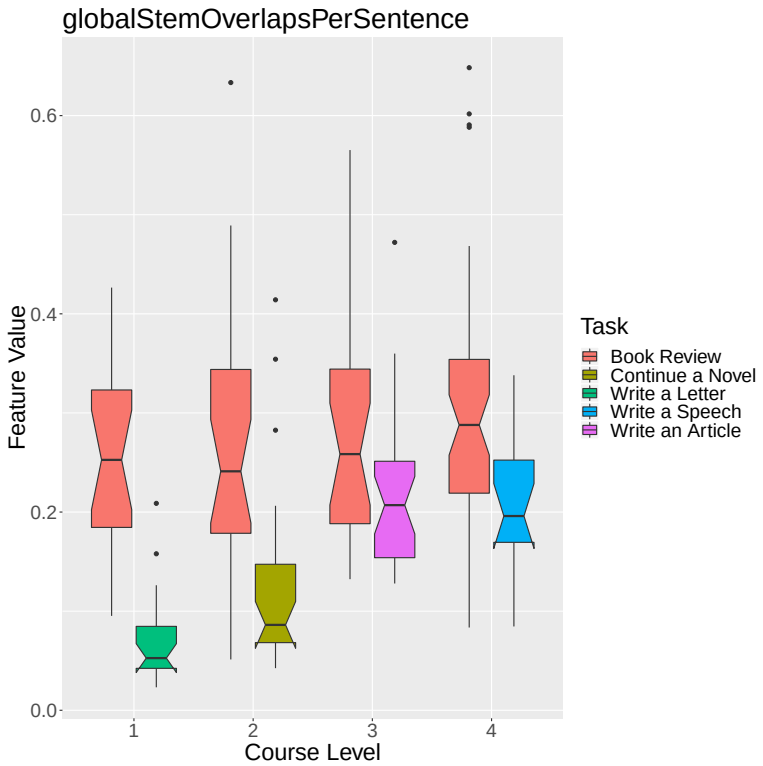


⇒ At higher course levels, learners use more derivation but task differences are also pronounced

⇒ Differences for inflection and composition features can all be attributed to task differences

Discourse Complexity

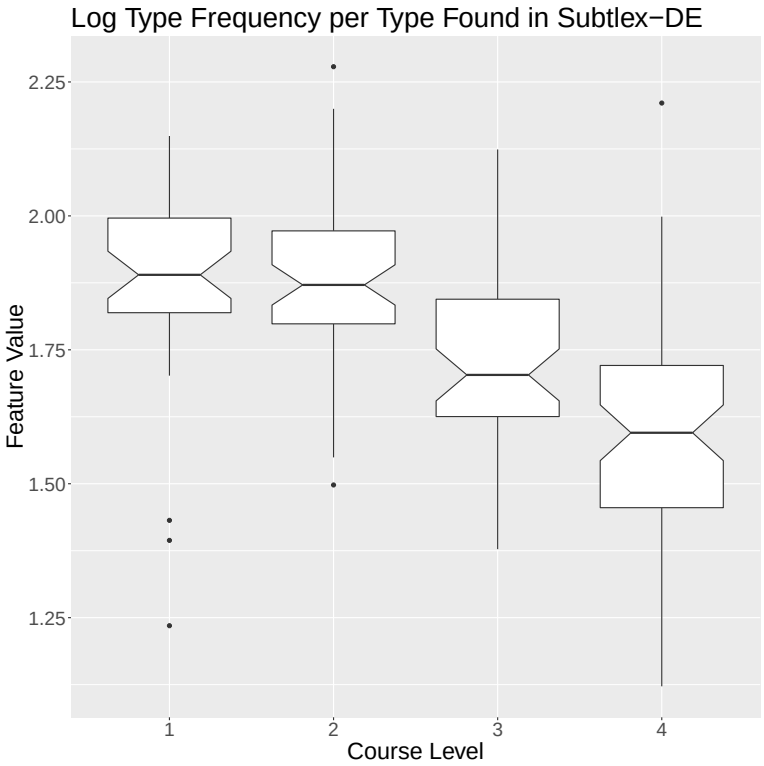


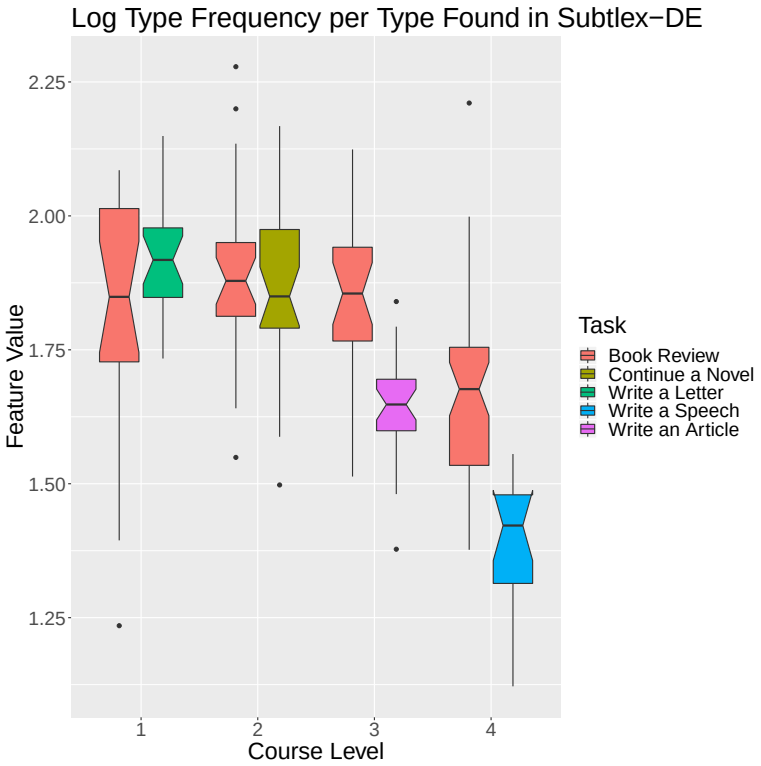


⇒ Only curriculum tasks show a systematic increase in implicit cohesion while book reviews remain stable

- Similar pattern across measures of implicit cohesion

Language Use



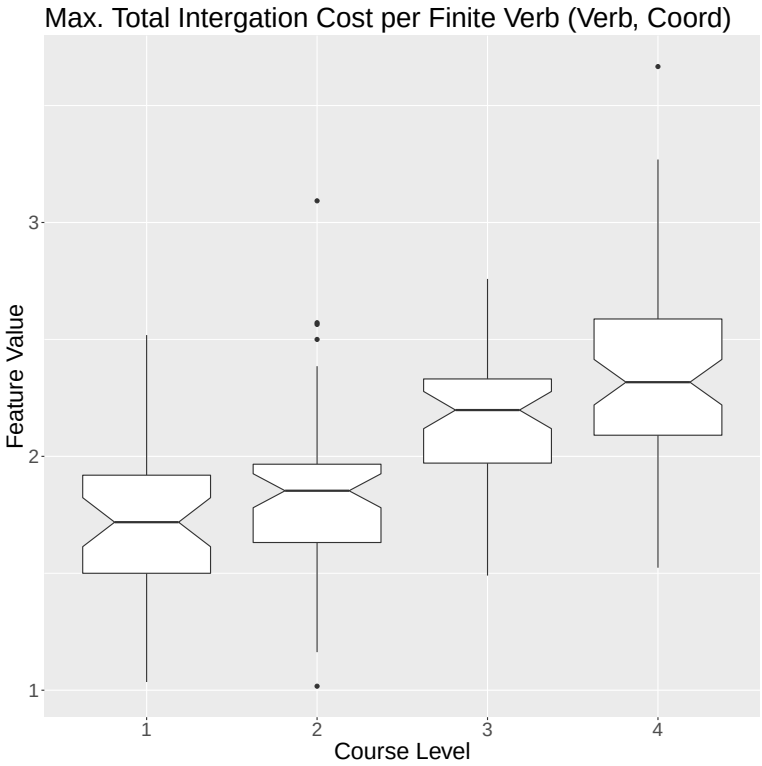


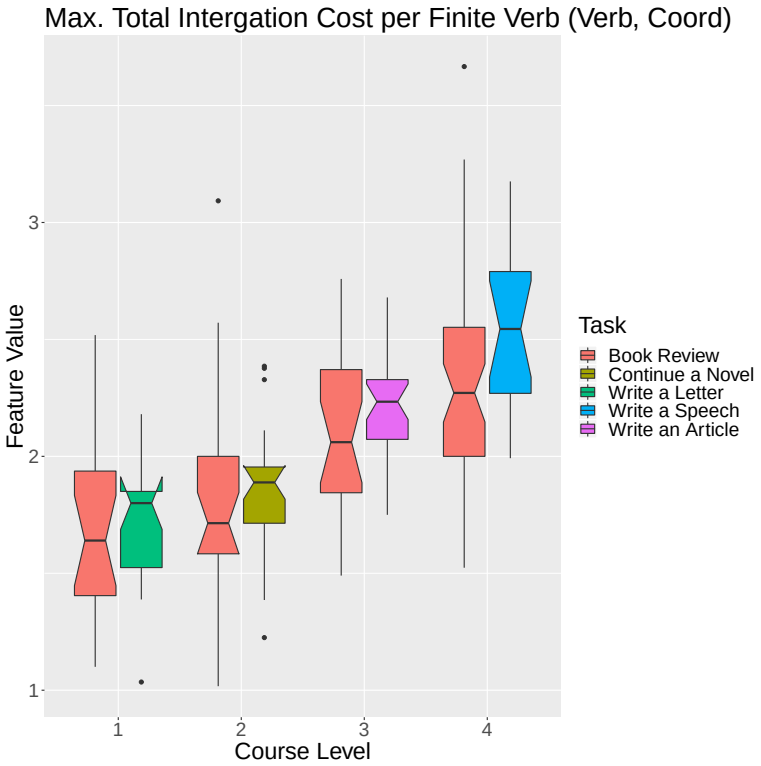
⇒ The vocabulary frequency of book reviews is relatively stable across course levels (except level 4)

- At level 4 significant frequency decrease for some features

⇒ Curriculum-tasks strongly differ vocabulary frequency from each other and the reference task

Human Processing Costs





⇒ Integration costs increase with course levels

⇒ Curriculum and reference tasks exhibit comparable values

- But they are systematically higher in curriculum-tasks
- Less pronounced difference between course levels 1 and 2