Computationally Modeling the Impact of Task-Appropriate Language Complexity and Accuracy on Human Grading of German Essays

Zarah Weiss Anja Riemenschneider
Pauline Schröter Detmar Meurers

Department of Linguistics, University of Tübingen
IQB, Humboldt-Universität zu Berlin

14th Workshop on Innovative Use of NLP for Building Educational Applications
Florence, Italy, August 2nd 2019

Introduction

- Complexity and accuracy core components in national educational standards for language arts and literacy (CCSSO 2010; KMK 2012)
- Doubts about teachers' ability to evaluate complexity and accuracy of texts (CCSSO 2010; Vögelin et al. 2019)
- Assessed manually in German *Abitur*
  - → Official school-leaving state examination
  - → Determines admission to university
- Study teachers' grading behavior in authentic *Abitur* data

Research Questions and Hypotheses

How do complexity and accuracy influence teachers'

- language performance grades (partial score)?
- content grades (partial score)?
- overall grades (composite score)?

It should be the case that complexity and accuracy

- strongly affect language performance grades
- do not affect content grades
- weakly affect overall grades

Education System in the U.S. and Germany

	U.S. System	German System
Education standard	CCSSO	KMK
High-stakes testing	repeatedly	final examination
Qualitative complexity	teachers	teachers
Quantitative complexity	automatic	teachers
Automatic Testing industry	yes	no

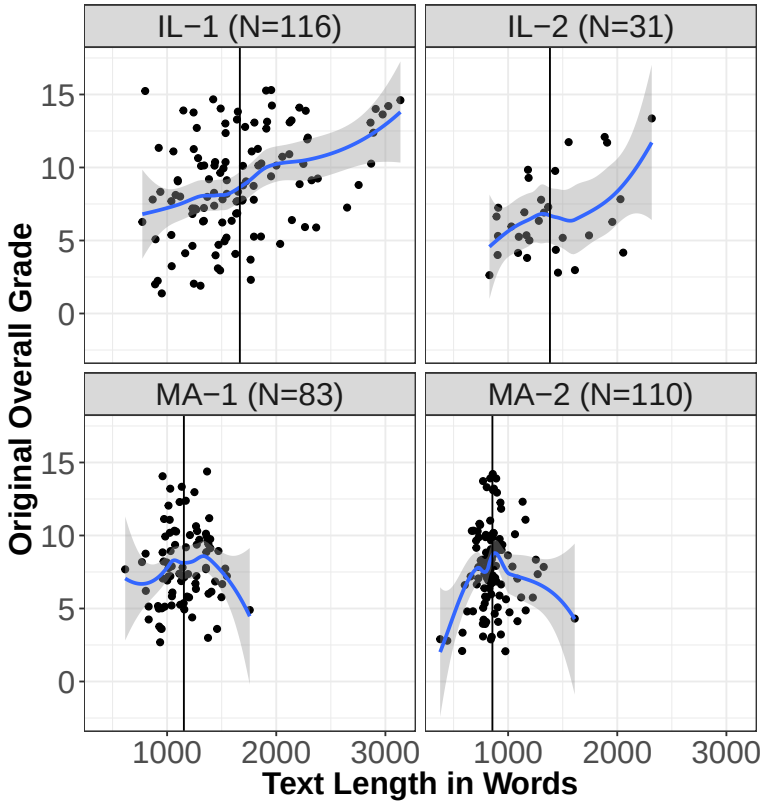
German *Abitur*, Federal States, and the IQB

- *Abitur* = official state examination required for university
- Education is a matter of the German federal states
- The Institute for Educational Quality Improvement (IQB)
  - → monitors schools' adherence to educational standards
  - → provides an official pool of tasks for the *Abitur*
  - → Includes templates for performance requirements
- States may choose and partially alter tasks from the pool

The Data

- Graded essays from German *Abitur* in 2017 ($N = 344$)
- Subject: German literature and language examination
- Collected across German states and digitized by the IQB
- Texts respond to one of four different task prompts
  - → 2 × interpretation of literature (IL-1, IL-2)
  - → 2 × material-based argumentation (MA-1, MA-2)

Task-Effects

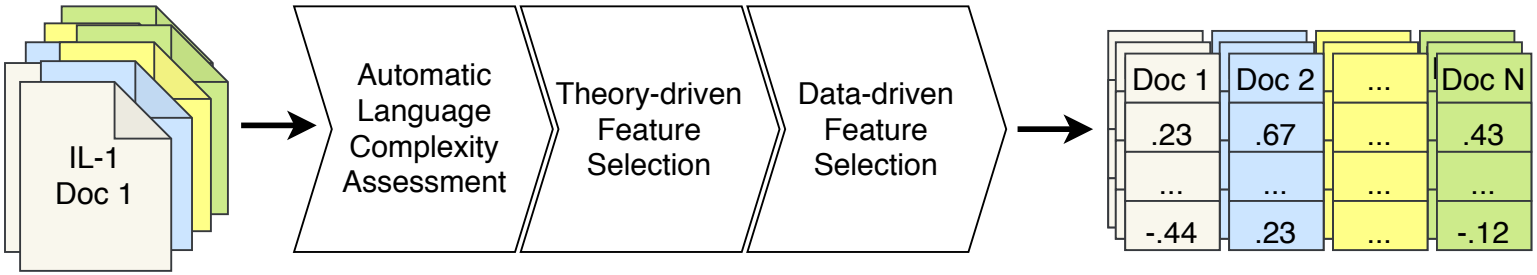


- Task prompts request and elicit texts of different length
- Influences correlation of text length and overall grade
- Task-effects are known to influence linguistic complexity (Alexopoulou et al. 2017; Yoon & Polio 2016)

Selecting and Representing Writing Complexity

- Select authentic texts of more and less task-appropriate overall linguistic complexity for the experiment (±ALC)
- Two-fold strategy:
  1. Build document vector representations capturing relevant dimensions of complexity
  2. Create a ranking of these vector representations to identify more and less complex documents
- Separately for each task to account for task-differences

Step 1: Creating Complexity Vectors



Student Essays

Building Complexity Vectors

Complexity Vectors

Automatic Complexity Assessment

- Automatically extract 320 complexity features (Weiss 2017)
- Successfully used to assess German readability and L1/L2 development (Weiss & Meurers 2018, 2019, in press)
- Measures of human processing, language use, and lexical, morphological, syntactic, and discourse complexity
- Based on SLA research where Complexity, Accuracy, and Fluency are dimensions of language performance (Bulté & Housen 2012; Wolfe-Quintero et al. 1998)

Theoretically-Motivated Complexity Features

- Education standards name examples of welcome writing strategies to make language more complex (KMK 2012)
- Includes argumentation structure, lexical complexity, and syntactic complexity (as well as accuracy)
- Register and norm-appropriateness → academic language (Hennig & Niemann 2013; Snow & Uccelli 2009)
- We identify 75 theoretically-motivated complexity features that are extracted by the system

Theory- and Data-Driven Feature Selection

1. Automatic extraction of 320 complexity features
2. Outlier removal and z-score calculation
3. Calculate the Pearson correlation (r) of each complexity feature with essays' original overall grade (r_g)
4. Add theoretically-motivated feature f ranked by decreasing r_g, if f correlates
   a. $abs(r_g) \geq 0.2$; $p < 0.05$ with the overall grade, and
   b. $abs(r_f) \leq 0.8$ with an already added feature
5. Repeat Step 4 for all other features with $abs(r_g) \geq 0.3$

Theory- vs. Data-Driven Feature Contribution

Task	Theory-Driven	Data-Driven	Total
IL-1	20	13	33
IL-2	32	13	45
MA-1	13	0	13
MA-2	9	4	13

- Resulting complexity vectors differ in length
- Most pronounced differences between task objectives (interpretation of literature, material-based argumentation)
- Overall mostly theoretically-motivated features selected

Zooming in on Complexity Vectors

Feature	IL-1	IL-2	MA-1	MA-2
MTLD	.2014	.4358	.2876	.3361
Dependent clauses per sentence	*.3040*	.2528	.2046	-.0380
Derived nouns per noun phrase	*.2394*	.4751	.1604	.3301
Average total integration cost at finite verb	.4093	*.4909*	.0708	.0308
Complex noun phrases per noun phrase	.4177	*.3186*	.1316	-.0353
Relative clauses per sentence	.3027	.1814	.1381	-.0077
Dep. clauses w/o conjunction per sentence	.1414	.2460	.0744	.0058
Conjunctional clauses per sentence	.1632	.2433	.0744	-.0285

- The four vectors include overall 75 unique features
- 18 features generalize across at least three vectors
- Mostly lexical and clausal complexity and nominal style

→ Known features of German academic language

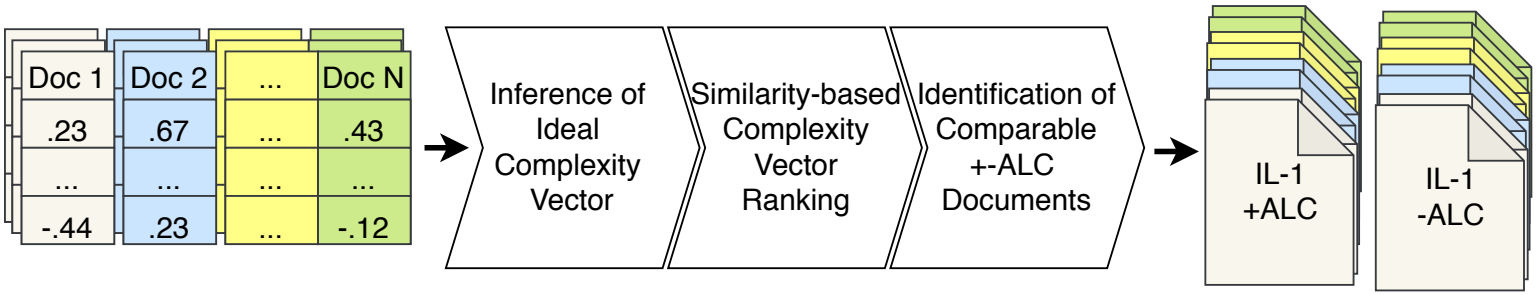
Differences between Task Objectives and Types

Feature	IL-1	IL-2	MA-1	MA-2
MTLD	.2014	.4358	.2876	.3361
Dependent clauses per sentence	*.3040*	.2528	.2046	-.0380
Derived nouns per noun phrase	*.2394*	.4751	.1604	.3301
Average total integration cost at finite verb	.4093	*.4909*	.0708	.0308
Complex noun phrases per noun phrase	.4177	*.3186*	.1316	-.0353
Relative clauses per sentence	.3027	.1814	.1381	-.0077
Dep. clauses w/o conjunction per sentence	.1414	.2460	.0744	.0058
Conjunctional clauses per sentence	.1632	.2433	.0744	-.0285

- The interpretation of literature task vectors are similar
- 21/26 features occurring twice are shared by IL-1 and IL-2
- Mostly (noun) phrase complexity and human processing

→ Generalizable characteristics of task objective (interpretation) and type (essay)?

Step 2: Ranking Complexity Vectors



Inferring Task-Wise Ideal Complexity Vectors

- Reference vector identifying the polarity of the correlation
- Assign maximal and minimal feature values to feature dimensions of appropriate and inappropriate complexity
- Positive correlations with original overall grade → 1
- Negative correlations with original overall grade → 0

Document Ranking

- Force each feature in the complexity vectors to range from 0 to 1 using min-max scaling
- Calculate Manhattan distance between each document vector and its corresponding ideal vector
- Rank documents task-wise by increasing distance
- Rank from more to less appropriate language complexity ($\pm ALC$)

Document Selection and Further Processing

- Consider only documents with a medium overall grade → often more difficult to rate and avoids ceiling/floor effects
- Select texts of comparable length from top and bottom rank
  → 16 documents selected (2 +ALC and 2 -ALC per task)
- Manual extraction of punctuation, spelling, and grammar errors by the IQB to assess text accuracy

Teachers Participating in Essay (Re-)Grading

- 33 subjects (14 female, 9 male, 0 diverse)
- Age $\mu = 46.4 \pm 8.7$ years; range = [34; 65]
- Teaching experience $\mu = 19.9 \pm 9.1$ years; range = [5; 38]
- Graded *Abitur* at least twice, mostly more than 8 times

Provided Materials and Grading Set-Up

- Each text was graded by 16 teachers
- Mail with 8 texts without original grades (50:50 $\pm ALC$)
- Grading at home with *Abitur* scale: 0 (worst) to 15 (best)
- Grading template with content and language requirements
- Best approximation of real-life *Abitur* grading

Evaluation

- Linear mixed regression model for each grade
- Response variable: language, content, or overall grade (re-)assigned by teachers in the experiment
- Predictor variables: $\pm ALC$ and z-scores of $\frac{\sum errors}{word}$
- Random intercept for task (IL-1, IL-2, MA-1, MA-2)

Results: Influence on Language Performance Grades

	Estimate	SE	t-value	p-value
(Inter.)	6.989	0.561	12.468	< 0.001
+ALC	1.374	0.368	3.732	< 0.001
Error	-1.992	0.211	-9.459	< 0.001

- +ALC texts get higher language performance grades
- More errors lead to lower language performance grades
- This confirms our expectations as complexity and accuracy are components of language performance

Results: Influence on Content Grades

	Estimate	SE	t-value	p-value
(Inter.)	6.138	0.772	7.948	0.003
+ALC	0.614	0.393	1.562	0.120
Error	-1.265	0.227	-5.586	< 0.001

- No evidence that complexity influences content grading
- More errors lead to lower content grades
- Punctuation, spelling, and grammar errors individually show the same kind of influence
- This partially violates our expectations as complexity and error rate are conceptually unrelated to content quality

Results: Influence on Re-Assigned Overall Grades

	Estimate	SE	t-value	p-value
(Inter.)	6.460	0.696	9.278	0.002
+ALC	0.703	0.359	1.962	0.051
Error	-1.518	0.208	-7.316	< 0.001

- Marginally significant impact of +ALC on overall grades
- More errors lead to lower overall grades
- Corresponding to the results for the partial grades the impact of error rate is over-proportionally strong

Discussion

Complexity

- Language performance grades successfully reflect differences in quantitative complexity
- Grades experienced teachers assign to ecologically valid texts are not unduly influenced by complexity differences
- Earlier findings for teachers in training do not carry over (Vögelin et al. 2019)

Accuracy

- Accuracy influences all grades – even when it is irrelevant
- This is a problematic issue for German *Abitur*
- Confirms Rezaei & Lovorn (2010); Cumming et al. (2002)

Conclusion & Outlook

- First results from collaboration of computational linguistic and education science research
- Novel methodology to identify task-appropriate language complexity for document selection
- Teachers identify and modularize language complexity but are clearly biased by accuracy across all grades
- Future work will investigate further the link between automatic and human complexity assessment and grading

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.

Birchenough, J., R. Davies & V. Connelly (2017). Rated age-of-acquisition norms for over 3,200 German words. Behavior research methods 49(2), 484–501.

Bulté, B. & A. Housen (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (eds.), Dimensions of L2 Performance and Proficiency, John Benjamins, pp. 21–46. URL https://doi.org/10.1075/lllt.32.02bul.

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.

CCSSO (2010). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects. Tech. rep., National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C. URL http://dx.doi.org/10.2139/ssrn.1965026.

Chen, X. & D. Meurers (2016). Characterizing Text Difficulty with Word Frequencies. In Proceedings of the 11th Workshop on Innovative Use of NLP

for Building Educational Applications. San Diego, CA: Association for Computational Linguistics, pp. 84–94.

Cumming, A., R. Kantor & D. E. Powers (2002). Decision Making while Rating ESL/EFL Writing Tasks: A Descriptive Framework. The Modern Language Journal 86(1), 67–96. URL https://doi.org/10.1111/1540-4781.00137.

Ellis, R. (2003). Task-based Language Learning and Teaching. Oxford, UK: Oxford University Press.

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.

Hancke, J., S. Vajjala & D. Meurers (2012). Readability Classification for German using lexical, syntactic, and morphological features. In Proceedings of the 24th International Conference on Computational Linguistics (COLING). Mumbay, India, pp. 1063–1080. http://aclweb.org/anthology-new/C/C12/C12-1065.pdf.

Hennig, M. & R. Niemann (2013). Unpersönliches Schreiben in der Wissenschaft. Informationen Deutsch als Fremdsprache 4, 439–455. URL https://doi.org/10.1515/infodaf-2013-0407.

KMK (2012). Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife. Beschluss der Kultusministerkonferenz vom 18.10.2012, https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Fortgef-FS-Abi.pdf.

Kyle, K. (2016). Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Ph.D. thesis, Georgia State University. URL http://scholarworks.gsu.edu/alesl_diss/35.

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.

Rezaei, A. R. & M. Lovorn (2010). Reliability and validity of rubrics for assessment through writing. Assessing Writing 15, 18–39. URL https://doi.org/10.1016/j.asw.2010.01.003.

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.

Snow, C. E. & P. Uccelli (2009). The Challenge of Academic Language. In D. R. Olson & N. Torrance (eds.), The Camebridge Handbook of Literacy, Camebridge: Cambridge Univerisity Press, pp. 112–133.

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.

Vögelin, C., T. Jansen, S. D. Keller, N. Machts & J. Möller (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. Assessing Writing 39, 50–63. URL https://doi.org/10.1016/j.asw.2018.12.003.

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. & 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 (2019). 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 (in press). 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.

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.

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.

Task Prompts

Interpretation of literature

- IL-1: Interpretation and comparison of poems
- IL-2: Interpretation of novel ending with given focus

Material-based argumentation

- MA-1: Essay on social media and communication
- MA-2: Comment on dialect use in modern societies

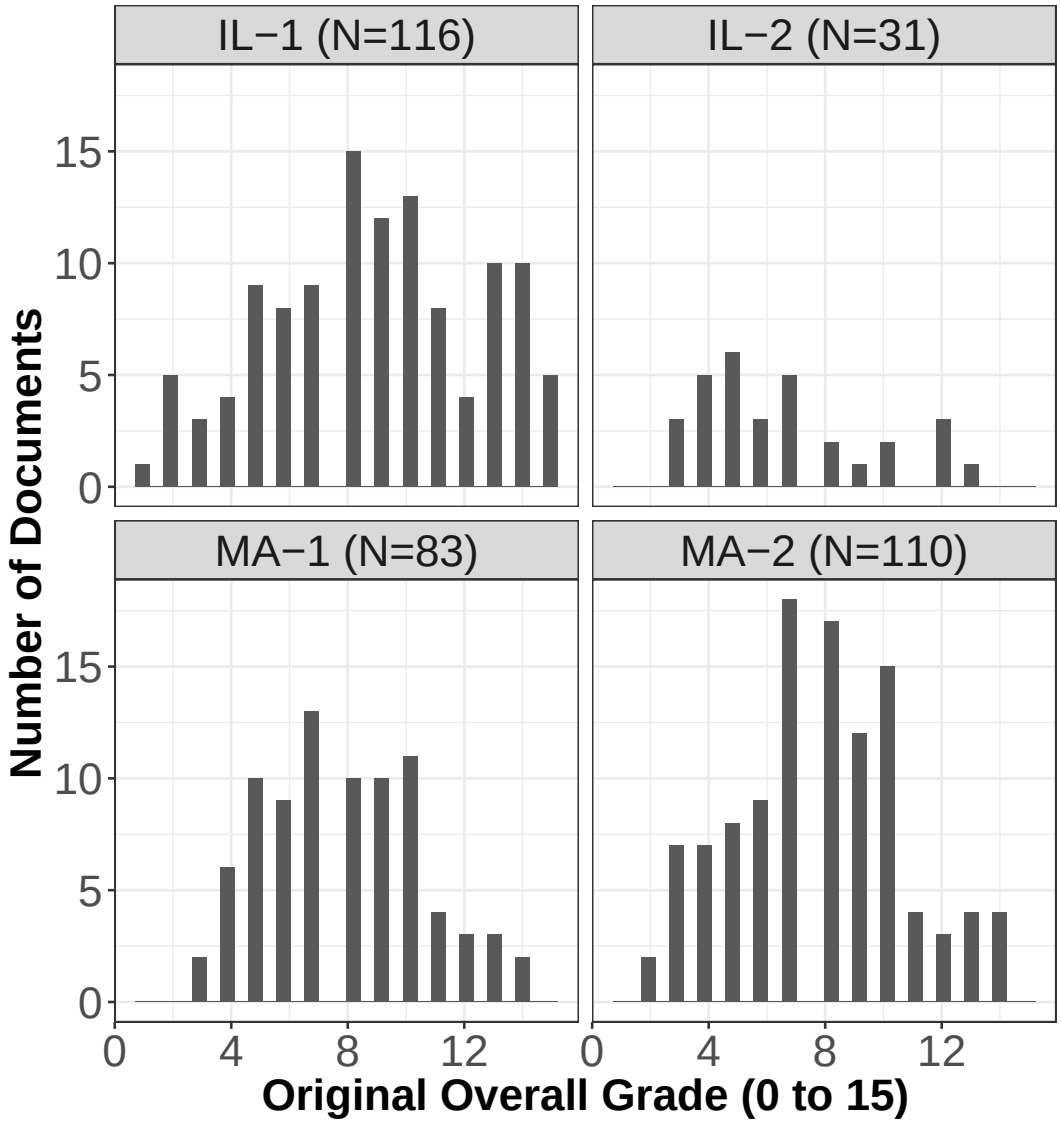
→ Based on 7 to 8 materials (essays, statistics, graphics, ...)

→ Word limits of 1,000 and 800 words

Task Prompts (cont.)

Task	Text Type	Description
IL-1	Interpretation of literature	Interpret poem *A* written in the 1950s and compare it with poem *B* written in the 1980s.
IL-2	Interpretation of literature	Interpret the given excerpt from novel *A*. Focus on the conflicts with which the protagonist struggles.
MA-1	Material-based argumentation	Write a newspaper essay on the influence social media has on our communication. Use around 1,000 words. Include the following materials in your argumentation: 6 essays, 1 poem, 1 statistic.
MA-2	Material-based argumentation	Write a newspaper commentary on the influence of dialects and sociolects on success in society. Use around 800 words. Include the following materials in your argumentation: 4 essays, 1 interview, 2 graphics.

Original Overall Grades Split By Task Prompt



German *Abitur* Grading System

Grade	Points	Percentage
excellent +	15	100–95
excellent	14	94–90
excellent -	13	89–85
good +	12	84–80
good	11	79–75
good -	10	74–70
satisfying +	9	69–65
satisfying	8	64–60
satisfying -	7	59–55
sufficient +	6	54–50
sufficient	5	49–45
sufficient -	4	44–40
insufficient +	3	39–33
insufficient	2	32–27
insufficient -	1	26–20
failed	0	19–0

Content Grades and Spelling Error Rate

	Estimate	SE	t-value	p-value
(Inter.)	5.976	0.802	3.335	0.003
+ALC	0.934	0.444	2.101	0.037
Spelling	-1.197	0.257	-4.651	< 0.001

Content Grades and Grammar Error Rate

	Estimate	SE	t-value	p-value
(Inter.)	5.954	0.392	15.172	< 0.001
+ALC	0.943	0.379	2.489	0.013
Grammar	-1.197	0.257	-4.651	< 0.001

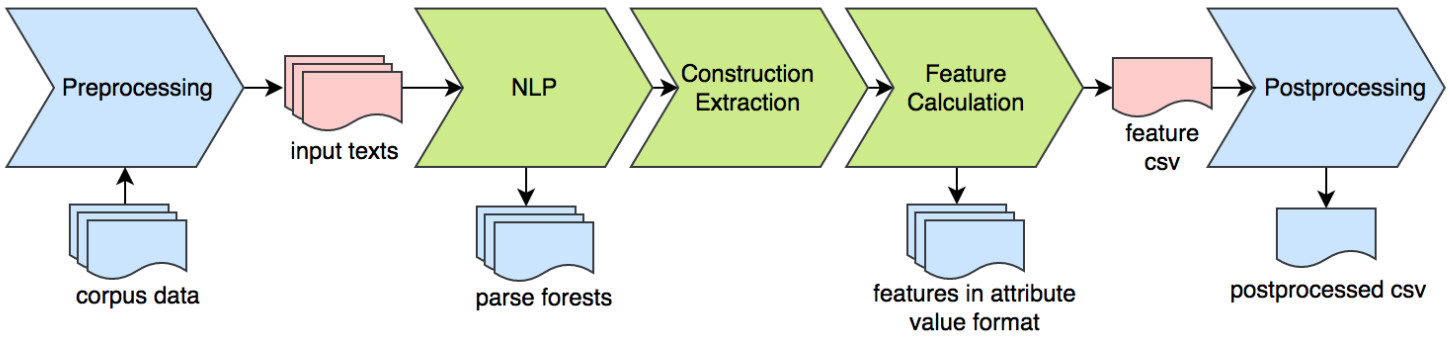
Content Grades and Punctuation Error Rate

	Estimate	SE	t-value	p-value
(Inter.)	6.484	0.534	12.136	< 0.001
+ALC	-0.1016	0.382	-0.266	0.790
Punctuation	-0.5968	0.1939	-3.078	0.002

Complexity in Second Language Acquisition

- Complexity is an important construct in SLA research
- Language performance = Complexity, Accuracy, Fluency (Bulté & Housen 2012; Wolfe-Quintero et al. 1998)
- Complexity = language elaboration and variety (Ellis 2003)
- Accuracy = native speaker-like error rate (Pallotti 2009)
- Fluency = native speaker-like production rate (Pallotti 2009)

NLP Pipeline



Lexical Complexity

- Describes the elaboration, inter-relatedness, and variation of the lexical system
- Measures vocabulary range and size as well as semantic relatedness
- E.g., type token ratio, lexical density, hyponyms per word
- Bulté & Housen (2014); Wolfe-Quintero et al. (1998)

Syntactic Complexity

- Describes the elaboration and variation of the syntactic domain (often split in clausal and phrasal complexity)
- Measures clausal and phrasal modification and variation
- E.g., % dependent clauses types, NP modifiers per NP
- Kyle (2016); Bulté & Housen (2014); Wolfe-Quintero et al. (1998)

Morphological Complexity

- Describes the elaboration and variation of the morphological system
- Measures derivation, composition, and inflection
- E.g., periphrastic tenses per verb, avg. compound depth
- Pallotti (2015); Bulté & Housen (2014); Hancke et al. (2012)

Discourse Complexity

- More elaborate, inter-related, and varied discourse relations are more complex
- Includes measures of cohesive markers, transition probabilities, co-reference chains
- E.g., connectives per sentence, probability subject of drops
- Origin from theoretical and psycho-linguistic research
- Todirascu et al. (2013); Barzilay & Lapata (2008); Graesser et al. (2004)

Language Use

- Assume that less frequently used or later acquired constructions are more complex
- Includes word or phrase frequency measures from large corpora or age of acquisition (AoA) measures
- E.g., mean AoA per word, mean frq. in dlexDB per word
- Origin from corpus- and psycho-linguistic research
- Paquot (2019); Chen & Meurers (2016); Birchenough et al. (2017)

Human Language Processing

- Measures cognitive complexity through processing times as measures by eye-tracking and reading time
- Includes measures of cognitive load and surprisal
- E.g., maximal DLT integration cost per verb
- Origin from cognitive science, psycho-linguistics, and information theory
- Shain et al. (2016); Gibson (2000)