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

Zarah Weiss[a] **Anja Riemenschneider**[b] **Pauline Schröter**[b] **Detmar Meurers**[a,b]

[a] Dept. of Linguistics & LEAD University of Tübingen
zweiss@sfs.uni-tuebingen.de
dm@sfs.uni-tuebingen.de

[b] Institute for Educational Quality Improvement Humboldt-Universität zu Berlin
riemenan@iqb.hu-berlin.de
schrotpa@iqb.hu-berlin.de

Abstract

Computational linguistic research on the language complexity of student writing typically involves human ratings as a gold standard. However, educational science shows that teachers find it difficult to identify and cleanly separate accuracy, different aspects of complexity, contents, and structure. In this paper, we therefore explore the use of computational linguistic methods to investigate how task-appropriate complexity and accuracy relate to the grading of overall performance, content performance, and language performance as assigned by teachers.

Based on texts written by students for the official school-leaving state examination (*Abitur*), we show that teachers successfully assign higher language performance grades to essays with higher task-appropriate language complexity and properly separate this from content scores. Yet, accuracy impacts teacher assessment for all grading rubrics, also the content score, overemphasizing the role of accuracy.

Our analysis is based on broad computational linguistic modeling of German language complexity and an innovative theory- and data-driven feature aggregation method inferring task-appropriate language complexity.

1 Introduction

Official state education standards highlight the relevance of language complexity for the evaluation of text readability and reading skills (CCSSO, 2010) and academic writing proficiency in students first and second language (KMK, 2014a,b). The highly assessment-driven U.S. public education system has long recognized the benefits of automating the evaluation of student learning outcomes, including very substantial research, development, and commercial applications targeting automatic essay scoring (AES, Shermis and Burstein, 2013; Vajjala, 2018; Yannakoudakis et al., 2018). This situation is not transferable to other education systems, such as the German one, where so far there is hardly any discussion of automating the assessment of learning outcomes and no high-stakes testing industry. In the German *Abitur* examination, the official school-leaving state examination that qualifies students for admission to university, teachers grade language performance and content in essays without technical assistance, using grading templates that specify content and language expectations. In the language arts and literacy subject-matters (German, English, French, etc.), language performance is a crucial component of the overall grade across all states. Yet, unlike content, language requirements are only loosely specified in the education standards, mentioning complex and diverse syntax and lexis, and a coherent argumentation structure as indicators of high-quality language performance (KMK, 2014b). The exact implementation of these language requirements is left to the discretion of the teachers. Educational science has questioned to which extent teachers are biased by construct-irrelevant text characteristics while grading. There is evidence that mechanical accuracy over-proportionally influences grades and even affects the evaluation of unrelated concepts such as content (Cumming et al., 2002; Rezaei and Lovorn, 2010). Differences in lexical sophistication and diversity have been shown to impact teachers' evaluation of grammar and essay structure (Vögelin et al., 2019). This is a potentially severe issue for the German education system.

We pick up on this issue by investigating which role language complexity and accuracy play in teachers' grading of German *Abitur* essays. For this, we build upon previous work on complexity and accuracy in the context of the Complexity, Accuracy, and Fluency (CAF) framework (Wolfe-Quintero et al., 1998; Bulté and Housen, 2012) employed in Second Language Acquisition (SLA) research to model different types of language per-

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 30–45
Florence, Italy, August 2, 2019.

formance (McNamara et al., 2010; Vajjala and Meurers, 2012; Bulté and Housen, 2014). We establish an automatically obtained measure of task-appropriate overall language complexity. With this, we identify texts of more and less appropriate language complexity, which we then manually assess for their accuracy. We use this to experimentally examine teaching experts' grading behaviour and how it is influenced by accuracy and complexity. Our results show that while teachers seem to successfully identify language complexity and include it in their grading when appropriate, they are heavily biased by accuracy even when it is construct-irrelevant.

Our work innovates in exploiting computational linguistic methods to address questions of broader relevance from the domain of educational science by using sophisticated language complexity modeling. This is the first computational linguistic analysis of German *Abitur* essays and their human grading, illustrating the potential of cross-disciplinary work bringing together computational linguistics and empirical educational science. The novel approach presented for the assessment of appropriate overall language complexity also provides valuable insights into the task- or text type-dependence of complexity features. This is of direct relevance for the current discussion of task-effects in CAF research (Alexopoulou et al., 2017; Yoon, 2017).

The article is structured as follows: We briefly review related work on complexity assessment and insights from educational science into human grading behavior. We then present our data set and how we automatically extract language complexity measures. Section 5 elaborates on the construction of appropriate overall language complexity including a qualitative analysis of task-wise differences between document vectors. Section 6 reports our experiment on teacher grading behavior. We close in Section 7 with an outlook.

2 Related Work

Language complexity, commonly defined as "[t]he extent to which the language produced in performing a task is elaborate and varied" (Ellis, 2003, p. 340), has been studied extensively in the context of second language development and proficiency and text readability in particular with regard to the English language (Vajjala and Meurers, 2012; Guo et al., 2013; Bulté and Housen, 2014; Chen and Meurers, 2019). Complexity has also been investigated in relation to (academic) writing proficiency of native speakers (Crossley et al., 2011; McNamara et al., 2010). Research on languages other than English, remains rather limited, with some work on German, Russian, Swedish, Italian, and French (Weiss and Meurers, 2018; Reynolds, 2016; Pilán et al., 2015; Dell'Orletta et al., 2014; François and Fairon, 2012).

Recently, research has increasingly focused on the influence of task effects on language complexity in writing quality and language proficiency assessment, both in terms of their influence on CAF development in the context of the two main frameworks (Robinson, 2001; Skehan, 1996) as well as its implications for AES systems and other forms of language proficiency modeling (Yannakoudakis et al., 2018; Dell'Orletta et al., 2014). Alexopoulou et al. (2017) show that task complexity and task type strongly affect English as a Foreign Language (EFL) essay writing complexity. Topic and text type, too, have been found to impact CAF constructs in EFL writing and in particular language complexity (Yoon and Polio, 2016; Yoon, 2017; Yang et al., 2015). Vajjala (2018) demonstrates task effects across EFL corpora to the extent that text length strongly impacts essay quality negatively on one and positively on the other data set. Her results further corroborate the importance of accuracy for essay quality across data sets. Accuracy has overall received considerably less attention in SLA research than complexity (Larsen-Freeman, 2006; Yoon and Polio, 2016).

An orthogonal strand of research investigates the quality of human judgments of writing quality and how complexity and accuracy impact them. It has been demonstrated that teachers are biased by accuracy and in particular spelling even when it is irrelevant for the construct under evaluation such as content quality (Rezaei and Lovorn, 2010; Cumming et al., 2002; Scannell and Marshall, 1966). Other studies showed that characteristics such as syntactic complexity, text length, and lexical sophistication impact inter-rater agreement (Lim, 2019; Wind et al., 2017; Wolfe et al., 2016). Vögelin et al. (2019) experimentally manipulate the lexical diversity and sophistication of EFL learners' argumentative essays and let Swiss English teachers rate them for their overall quality, grammar, and essay frame. Their findings show that when the lexical diversity and sophis-

tication of an essay was manually reduced, it received lower grades not only for its overall quality but also for grammar and the essay's frame, i.e., the structured presentation of the writing objective through introduction and conclusion.

3 The *Abitur* Data

We analyzed 344 essays that were written during the German literature and language examination of the German *Abitur* in 2017. The essays were elicited across German states and collected and digitized by the Institute for Educational Quality Improvement (IQB).[1] For each essay, the final overall grade that was assigned to it in the *Abitur* serves as meta information. All essays respond to one of four task prompts.[2] Two tasks require the interpretation of literature (IL): IL-1 and IL-2. The other two elicit material-based argumentative (MA) essays based on several additional materials provided with the task: MA-1 and MA-2.[3]

Topic and task differences may substantially impact the linguistic characteristics of the resulting language (Alexopoulou et al., 2017; Yoon and Polio, 2016). For our data, this is even more the case given that MA task prompts include a recommended essay length (around 1,000 for one, around 800 words for the other), but IL task prompts do not. The effect this has on the relationship between text length and overall essay grade is shown in Figure 1. Texts elicited by MA tasks are overall shorter than answers to IL tasks and exhibit a lesser variation in length. While for IL tasks we observe a weak linear correlation between overall grade and text length, clear deviations from the expected text length seem to have a negative impact on the overall grade for MA tasks. To address this issue, we split our data for the following analyses in four data sets, one per task prompt. The data sets are henceforth referred to by the id of the respective task prompt (IL-1, IL-2, MA-1, MA-2).

4 Automatic Complexity Assessment

Our system automatically extracts 320 measures of language complexity covering a broad range of linguistic features. We include features from

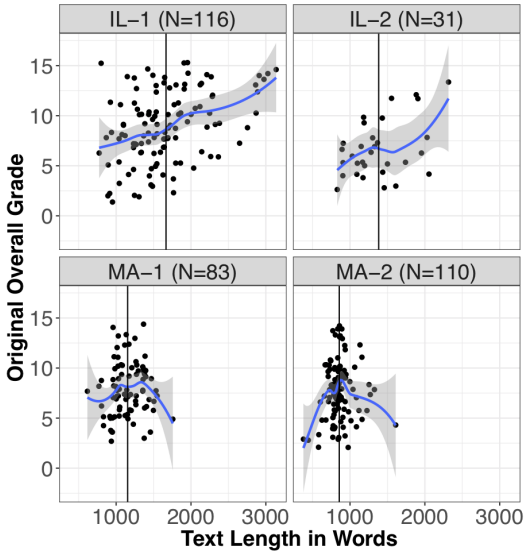

Figure 1: Text length across overall grades split by task prompts. The vertical line marks the mean length.

two main research strands on text complexity in our system: measures of the linguistic system and psycho-linguistic measures of language use and cognitive processing. An overview of all features can be found in Table 1.

Our procedure is based on our implementation of a broad range of complexity features for German which we have successfully used for the assessment of German readability of media captions for adults and children (Weiss and Meurers, 2018), German L2 proficiency (Weiss, 2017; Weiss and Meurers, in press), and German L1 writing development (Weiss and Meurers, 2019). However, for the research presented here, we altered the segmenter for sentences and tokens. Due to the specific abbreviations for line and page references systematically used in our data, we found that a rule-based segmenter combined with a customized list of abbreviations typical for German *Abitur* essays outperformed the segmentation by OpenNLP (Bohnet and Nivre, 2012).[4]

As mentioned earlier, language complexity is an important component of the German curriculum for German arts and literacy (KMK, 2014b). While it lacks a full operationalization of language complexity, it names some examples of language complexification strategies that students' writings should exhibit. Based on this, we identified a set of 75 complexity features, which implement the lan-

[1] The IQB is an academic institute that monitors if schools across Germany states adhere to the educational standards set by the Standing Conference of the Ministers of Education and Cultural Affairs of the States in Germany.

[2] Figure 4 in the Appendix shows the distribution of documents and grades across task prompts.

[3] Table 6 in the Appendix describes the task prompts.

[4] We used the segmenter by Stefanie Dipper available at https://www.linguistics.ruhr-uni-bochum.de/~dipper/resources/tokenizer.html

Feature Set	Description
Lexical complexity	measures lexical density, variation, sophistication, and relatedness; e.g., type token ratio
Discourse complexity	measures use of cohesive devices; e.g., connectives per sentence
Phrasal complexity	measures phrase modification; e.g., NP modifiers per NP
Clausal complexity	measures clausal elaboration; e.g., subordinate clauses per sentence
Morphological complexity	measures inflection, derivation, and composition; e.g., average compound depth per compound noun
Language Use	measures word frequencies based on frequency data bases; e.g., mean word frequency in SUBTLEX-DE (Brysbaert et al., 2011)
Language Processing	measures cognitive load during human sentence processing, mostly based on Dependency Locality Theory (Gibson, 2000) e.g., average total integration cost at the finite verb

Table 1: Overview over the feature sets used to capture language complexity

guage requirements that were pre-defined for our data. These may be grouped into three categories:

Argumentation Structure Texts should be structured coherently, clearly, be compelling and provide clear guidance for the reader. The author's reasoning should be made explicit. Both, the text's general structure as well as the language used should facilitate this (KMK, 2014b, p. 17). We operationalized these aspects by measuring various uses of connectives and the local and global co-occurrence of arguments, nouns, and word stems.

Lexical Complexity Texts should be lexically elaborate and varied. Stylistically, vocabulary choice should adhere to a task-appropriate written register (KMK, 2014b, e.g., pp. 42, 52). We cover this by including a range of measures of lexical diversity and density.

Syntactic Complexity Texts should be syntactically elaborate and varied and include connected and subordinated clauses to reflect a coherent structure. Stylistically, they should adhere to a task-appropriate written register. Students should also make appropriate use of tenses (KMK, 2014b, e.g., pp. 42, 52). To measure syntactic complexity, we include sentence length and several clause to sentence ratios, e.g., complex t-units per sentence and relative clauses per sentence.

Due to the repeatedly named focus on stylistically and norm-appropriate writing (KMK, 2014b, p. 16f), we also include prominent measures of German academic language which constitutes the appropriate written register for all four tasks represented in our data. There is a broad consensus that in particular complex noun phrases are a prominent feature of academic language (Hennig and Niemann, 2013; Morek and Heller, 2012; Schleppegrell, 2001), thus we include a series of measures of noun phrase elaboration and the variability of noun complexity. Another prominent aspect of academic language is deagentivization (Hennig and Niemann, 2013; Snow and Uccelli, 2009; Bailey, 2007), which entails passivization, verb modification and verb cluster. Hence, we specifically include measures of verb complexity and the variation of verb clusters as well as the coverage of deagentivization patterns in general. Finally, we include measures of tense usage to cover the specific request for appropriate tense usage across text types. Note that while across tasks the notions of what constitutes appropriate tense use may differ, within tasks these are fixed, e.g., favoring the use of past tense over present tense or vice versa.[5]

5 Complexity-Based Essay Selection

In order to evaluate how language complexity impacts grading behavior, we first needed to identify texts of high and low language complexity for our experiment (Section 6). For this, we followed a two-step approach: First, we transformed each student essay into a vector representation of relevant features of language complexity (Section 5.1). Then, we ranked them with regard to

[5]The complete list of theoretically motivated features may be found in Table 7 in the Appendix.

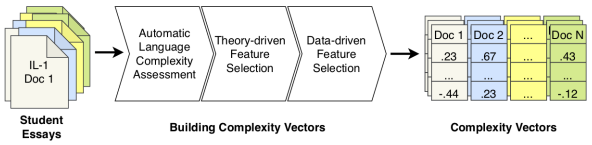

Figure 2: Task-wise transformation of essays to language complexity vector representations.

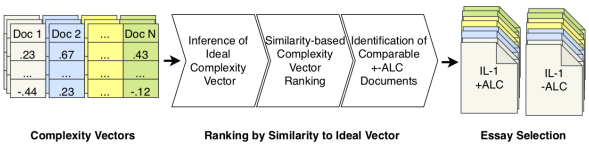

Figure 3: Selection of essays with more and less task-appropriate overall language complexity.

their similarity to an artificial ideal vector and selected for each task two essays of high and two of low language complexity (Section 5.2).

5.1 Building Complexity Vectors

Figure 2 outlines the procedure used to build language complexity vectors tailored towards the individual task prompts. We extracted the 320 measures of language complexity from the *Abitur* data as discussed in Section 4. We then removed all outliers that deviated more than two standard deviations from the mean and calculated the z-score of each feature. Based on this, we identified which of the dimensions of linguistic complexity that we measured are relevant for a given task.

We defined relevance in terms of correlation with the overall grade an essay received. These grades represent teachers' judgments of essay quality under consideration of language performance in a high stakes testing situation. We used a hybrid approach combining theory-driven and data-driven feature selection. First, we calculated the Pearson correlation between the z-scores of 75 theoretically relevant features and the overall grade each essay had received in the *Abitur* examination. We did so separately for each data set. Features with a significant ($p < .05$) absolute correlation of $r \geq .2$ were included in the complexity vector if they did not correlate more than $r = .8$ with another feature in the vector. For highly correlated features, we only kept the feature most highly correlated with the overall grade.

We augmented this feature selection with the remaining features of linguistic complexity in our document vector that had a significant ($p < .05$) absolute Pearson correlation with the overall grade of $r \geq .3$. Features were required to correlate less than $r = .8$ with other features selected for the complexity vector. For highly inter-correlated features, the feature with the highest correlation with the overall grade or the theoretically motivated feature was favored. This lead to complexity vectors of size 33 for IL-1, 45 for IL-2, and 13 for MA-1 and 13 for MA-2.[6, 7]

5.2 Ranking by Similarity to Ideal Vector

We selected essays for our experiment using the similarity of complexity vectors to a reference vector representing the artificial ideal use of each complexity feature as illustrated in Figure 3. We assigned the values 1 for feature dimensions with a positive correlation with the original overall grade and 0 for those with a negative correlation with the original overall grade. Conceptually, this represents the ideal language complexity for a given task: Features that are associated with low performance are not present and features associated with high performance are maximally represented.

For each feature in the complexity vector, we replaced the previously introduced z-scores with a min-max normalization to enforce a scale from 0 to 1. We calculated the similarity between each essay and the reference vector using Manhattan distance and ranked all essays based on their distance to the artificial ideal document vector.

Based on this ranking, we chose four essays per task which were comparable with each other in terms of their text length: two from the top of our ranking, i.e. closer to the ideal vector, and two from the bottom of our ranking, i.e. more distant to the ideal vector. We limited our choice to essays that had received a medium overall grade between 7 and 9 points in the German grading system for the final three years of German high school. This corresponds to essays with a point percentage between 55% and 69% (KMK, 2018, p. 22).[8] This restriction ensures on the one hand that essays are comparable in terms of their overall and content performance. On the other hand, it prevents ceiling and floor effects in teachers' grades.

[6] The final feature selection for all four vector representations and the correlation of all features with the original overall grade may be found in Table 8 in the Appendix.

[7] Table 9 in the Appendix shows for each task how many features were selected using the theory-driven and the data-driven selection step.

[8] An overview relating this system to percentage points may be found in Table 10 in the Appendix.

We labeled the resulting eight texts close to the ideal vector as essays with more appropriate language complexity (+ALC) and the eight texts relatively distant from the ideal vector as essays with less appropriate language complexity (-ALC).

5.3 Task-Wise Vector Differences

Comparing the features that were selected for the vector representations across tasks reveals some interesting structures which are relevant for the ongoing discussion of task effects on language performance. Overall, 75 unique features are included across all vectors. Table 2 shows a selection of 10 features chosen to illustrate patterns across vectors.[9]

Nearly a quarter of features (18 of 75) re-occurs in at least three of the four vectors. We take this as an indication of generalizable characteristics of language performance. This group is predominantly comprised of features of lexical sophistication in form of lexical diversity and verb variation (6/18), clausal elaboration in form of words, clauses, dependent clauses, and dependent clauses with conjunctions per sentence as well as the overall use of connectives (6/18), and nominal writing style in form of post-nominal modifiers, genitives, and nominalization strategies (4/18), all of which are positively correlated with the overall grade. These groups are represented in Table 2 by MTLD, dependent clauses per sentence, and the percentage of derived nouns. Taken together, they represent important markers of German academic language (Hennig and Niemann, 2013; Morek and Heller, 2012). Lexical sophistication has also repeatedly been observed as an important indicator of English first and second language writing performance (Guo et al., 2013; Crossley et al., 2011). Evidence that the relevance of these features for writing performance persists across task contexts is highly relevant as it provides empirical underpinning to the mostly theoretical concept of German academic language.

Aside from this general overlap across task prompts, we observe considerable similarities between both IL task prompts indicating that the features represent a coherent subgroup of appropriate linguistic complexity for interpretative writing rather than idiosyncratic properties of the specific task prompts. Of 26 features that are relevant across two tasks, 21 are shared between the IL tasks. This is a remarkable overlap given the respective vector sizes. Characteristic for IL tasks are especially features of phrasal modification (9/21), predominantly but not exclusively with regard to noun phrase modification, and clausal elaboration resulting in higher cognitive load in form of integration cost and dependency lengths (5/21). All of these are positively correlated with the overall grade. The two groups are represented in Table 2 by the percentage of complex noun phrases and the average total integration cost. Several of the features not shared across both IL tasks relate to different realizations of clausal elaboration: while for IL-2 several subtypes of subordination are relevant, such as interrogative clauses, conjunctional clauses, clauses without conjunction, various types of connectives, for IL-1 only relative clauses occur as specific type of clausal elaboration. Table 2 displays this contrast for relative clauses, dependent clauses without conjunction, and conjunctional clauses per sentence. Material-based argumentation does not exhibit such a pattern which may be due to the fact that both MA prompts request different text types, once a commentary (MA-2) and once an essay (MA-1), while both IL tasks share not only a task objective (interpretation) but also the same text type (essay).

6 Experiment

6.1 Set-Up

We recruited 33 teachers (14 female, 19 male) from different schools across German states.[10] Their teaching experience ranges from 5 to 38 years ($\mu = 19.9; SD = 9.1$). All of them have participated in grading German subject-matter *Abitur* tasks at least twice, most of them more than eight times. We asked them to grade essays for their language, content, and overall performance using the grading scale used for the German *Abitur* ranging from 0 to 15 points. Teachers were provided with a grading template for each task prompt, which is a standard feature in the German *Abitur*. The template states the expectations of students' answers with regard to content and language. Each teacher received 8 texts from over-

[9] The selection was taken from the aforementioned full table displaying all 75 features relevant for the vector representations in Table 8 in the Appendix.

[10] We recruited 32 teachers plus one replacement teacher to cover an anticipated drop-out. Since all teachers completed the study, eight texts were graded by an additional teacher (i.e. 17 instead of 16 teachers).

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

Table 2: Selection of features in the complexity vectors and their correlation with the original overall grade. Gray font marks uncorrelated features. Italics mark correlated but redundant features.

all 2 tasks: 4 +ALC and 4 -ALC texts. Each text was graded by 16 teachers independently. Teachers did not know the original grades that their texts had received, neither were they aware of the ranking-based selection. This grading situation was maximally familiar to our subjects, because it mimics teachers' real-life experience for essay grading in the context of German *Abitur*.

For each of the three grades (overall, content, and language performance), we built a linear mixed regression model fitted by REML. The respective grade served as response variable and we included task prompt as random effect. Each model had two predictor variables: ±ALC and error rate. We included error rate (in form of z-scores) as a predictor, because accuracy is an important criterion for the evaluation of students' language performance and thus overall performance in the German *Abitur* and to investigate its influence on teachers' grading. We manually extracted spelling mistakes, punctuation errors, and grammatical errors from each essay and aggregated them into one overall error score by dividing the total number of errors by the number of words.

6.2 Results

Tables 3, 4, and 5 show the respective model fits for each grade. For all three models, the residuals were homoscedasticly distributed around a zero mean. Table 3 shows that +ALC affects language performance grades by raising it about 1.37 points (± 0.37 SE) for essays with more appropriate linguistic complexity. Error rate, too, clearly affects the grade, lowering it about -1.99 points (± 0.21 SE). The model overall explains 37.5% of the variance, 29.3% of which are attributed to both error rate and ±ALC. Although error rate is the stronger of the two predictors, ±ALC does significantly improve the model fit ($\chi^2 = 1277.7, p < 0.001$). The random intercept for the four tasks accounts for 1.0% of the variance (±1.0 SD). The residuals account for 7.6% of the variance (±2.8 SD).

Table 4 shows the fit for the content grades the teachers assigned. We do not see evidence that the content grade is affected by +ALC in our ratings. Error rate, however, influences the grade negatively, lowering it about -1.265 points (± 0.227 SE). The model overall explains 29.1% of the variance. 11.9% are attributed to error rate and ±ALC but complexity does not make a significant contribution to the overall model fit. The random intercept for the four tasks accounts for 2.1% of the variance (±1.4 SD). The residuals account for 8.8% of the variance (±2.9 SD). In order to rule out that this influence of error rate on the content grade is caused by certain errors obstructing understanding, we refitted the content grade model with each of the individual error types instead of overall error rate. We find that all three error types impact content grade. Spelling significantly lowers it ($t = -4.651, p = 0.000$) about -1.197 points (± 0.257 SE). Punctuation significantly lowers it ($t = -3.078, p = 0.002$) about -0.597 points (± 0.194 SE). Grammar significantly lowers it ($t = -7.836, p = 0.000$) about -1.560 points (± 0.199 SE).

Table 5 shows the fit for the overall grades assigned by the teachers. The overall grade is marginally affected by +ALC . The overall grade is about 0.703 points higher (± 0.359 SE) for text with more appropriate linguistic complexity. As for the other grades, error rate strongly influences the overall rating lowering it about -1.518 points (± 0.208 SE). The model overall explains 31.1% of the variance. Of this, 17.3% are attributed to

	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

Table 3: Estimates for language performance grade.

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

Table 4: Estimates for content grade.

	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

Table 5: Estimates for re-assigned overall grade.

+ALC and error rate. Again, error rate is the stronger predictor and ±ALC does not make a significant contribution to the overall model fit. The random intercept for the four task accounts for 1.7% of the variance (±1.3 SD). The residuals account for 7.3% of the variance (±2.7 SD).

6.3 Discussion

Our results show that the language performance grades based on criteria stated in the grading template reflect differences between essays exhibiting more and less appropriate language complexity (±ALC). This result is not trivial, because previous research suggests that the assessment of quantitative aspects of text complexity is not a key competence of teachers (CCSSO, 2010). We do not find evidence that teachers are unduly influenced by differences in language complexity when assigning content grades. This is an encouraging finding in light of Vögelin et al. (2019)'s study on the effect of differences in lexical complexity on construct-unrelated grades. Our study differs in several aspects from their set-up: We asked experienced teachers rather than pre-service teachers, and we used the set-up of the *Abitur* they are familiar with. We provided them with texts that differed not only in terms of their lexical complexity (although these dimensions are represented in each of the document vector representations) but rather across various linguistic domains. While they altered texts experimentally, we used essays that are ecologically valid. We find that teachers include language complexity to a limited extent in the overall grades they assign. This is in line with the grading template stating that language performance should account for 30% of the overall performance.

As for accuracy, our results clearly show that all three grades are heavily influenced by error rate. For the language performance grade, this is motivated insofar as correctness is one of the criteria named in the corresponding grading template. Similarly, accuracy may be reflected in the overall grade as it is part of the overall evaluation. However, its weighting in both models is disproportionate. For content grading, accuracy is conceptually irrelevant, which is also stated in the grading template. Yet, teachers are clearly biased against essays with higher error rates, which is in line with previous research findings (Rezaei and Lovorn, 2010; Cumming et al., 2002). All three individual error types (punctuation, spelling, and grammar) show the same kind of influence on the content grade as the overall error rate. This demonstrates that the effect is not restricted to error types that may impede understanding, such as grammar errors. All error types affect content grading. Essays with a lower overall error rate receive higher content grades. This strong bias for a construct-irrelevant characteristic that is already included in another grading component, namely language performance, is highly problematic. Note, however, that we cannot rule out the possibility that students with better spelling in fact coincidentally also produce texts with better content. This is one of the limitations of our research design, which focuses on ecological validity. We will address this issue in a follow-up study, in which we will include corrected versions of the texts studied here. This way, we can keep essay content fixed while varying error rate. Overall our results indicate that although teachers can successfully capture different dimensions of language performance, such as complexity, accuracy, and content, they fail to modularize them clearly into separate grades.

7 Outlook

We addressed the question to which extent German teachers are able to identify differences in

appropriate language complexity across tasks and how complexity and accuracy bias grading when they are construct-relevant or -irrelevant. For this, we proposed a novel similarity-based approach for the identification of task-appropriate language complexity in student essays. This also yielded some interesting insights in task differences between writing objectives and task prompts confirming common but so far empirically not sufficiently validated assumptions about German academic language. While our results indicate that teachers successfully identify and modularize the concept of language complexity, we show a clear bias for higher language accuracy across all grades. Teachers not only consider accuracy over-proportionally for the grading of language performance, it also influences their assessment of construct-irrelevant aspects such as content. This is in line with previous research findings (Rezaei and Lovorn, 2010; Cumming et al., 2002).

We see our work as a first step towards the analysis of the grading behaviour in the German education system using computational linguistic methods. In future work, we plan to build on this by exploring the grading behavior of teachers in greater depth, clustering teachers in terms of their characteristics and grading behavior. In particular, there is evidence that teachers' personal evaluation of the complexity of a text impacts their perception and, consequently, their grading of its language quality. We will explore this in a follow-up study. We will also follow-up on the question to which extent better accuracy and content quality coincide in ecologically valid texts by studying the link between content grades and writing accuracy in a more controlled setting with experimentally manipulated texts with corrected errors.

References

Theodora Alexopoulou, Marije Michel, Akira Murakami, and Detmar 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.

Allison L. Bailey, editor. 2007. *The Language Demands of School. Putting Academic English to the Test*. Yale University Press, New Haven and London.

Bernd Bohnet and Joakim 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*, pages 1455–1465, Jeju Island, Korea. Association for Computational Linguistics.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58:412–424.

Bram Bulté and Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency*, pages 21–46. John Benjamins.

Bram Bulté and Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26(0):42 – 65. 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. Technical report, National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.

Xiaobin Chen and Detmar Meurers. 2019. Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer-Assisted Language Learning*.

Scott A. Crossley, Jennifer L. Weston, Susan T. McLain Sullivan, and Danielle S. McNamara. 2011. The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, 28(3):282–311.

Alister Cumming, Robert Kantor, and Donald E. Powers. 2002. Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(1):67–96.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of the International Journal of Applied Linguistics*, 165(2):163–193.

Rod Ellis. 2003. *Task-based Language Learning and Teaching*. Oxford University Press, Oxford, UK.

Thomas François and Cedrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126.

Liang Guo, Scott A Crossley, and Danielle S. McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18:218–238.

Mathilde Hennig and Robert Niemann. 2013. Unpersönliches Schreiben in der Wissenschaft. *Informationen Deutsch als Fremdsprache*, 4:439–455.

KMK. 2014a. *Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulre*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK), Wolters Kluwer Deutschland GmbH, Köln.

KMK. 2014b. *Bildungsstandards im Fach Deutsch für die Allgemeine Hochschulreife*. Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK), Wolters Kluwer Deutschland GmbH, Köln.

KMK. 2018. *Vereinbarung zur Gestaltung der gymnasialen Oberstufe und der Abiturprüfung. Beschluss der Kultusministerkonferenz vom 07.07.1972 i.d.F. vom 15.02.2018*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK), Wolters Kluwer Deutschland GmbH.

Diane Larsen-Freeman. 2006. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27(4):590–619.

Jungmin Lim. 2019. An investigation of the text features of discrepantly-scored ESL essays: A mixed methods study. *Assessing Writing*, 39:1–13.

Danielle S. McNamara, Scott A. Crossley, and Philip M. McCarthy. 2010. Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

Miriam Morek and Vivien Heller. 2012. Bildungssprache – kommunikative, epistemische, soziale und interaktive Aspekte ihres Gebrauchs. *Zeitschrift für angewandte Linguistik*, pages 67–101.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2015. A readable read: Automatic assessment of language learning materials based on linguistic complexity. In *Proceedings of CICLING 2015- Research in Computing Science Journal Issue (to appear)*.

Robert Reynolds. 2016. *Russian natural language processing for computer-assisted language learning: capturing the benefits of deep morphological analysis in real-life applications*. Ph.D. thesis, UiT - The Arctic University of Norway.

Ali Reza Rezaei and Michael Lovorn. 2010. Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15:18–39.

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

Dale P. Scannell and Jon C. Marshall. 1966. The effect of selected composition errors on grades assigned to essay examinations. *American Educational Research Journal*, 3(2):125–130.

Mary J. Schleppengrell. 2001. Linguistic features of the language of schooling. *Linguistics and Education*, 12(4):431–459.

Mark D. Shermis and Jill Burstein, editors. 2013. *Handbook on Automated Essay Evaluation: Current Applications and New Directions*. Routledge, Taylor & Francis Group, London and New York.

Peter Skehan. 1996. A framework for the implementation of task-based instruction. *Applied Linguistics*, 17(1):38.

Catherine E. Snow and Paola Uccelli. 2009. The challenge of academic language. In David R. Olson and Nancy Torrance, editors, *The Camebridge Handbook of Literacy*, pages 112–133. Cambridge Univerisity Press, Camebridge.

Sowmya Vajjala. 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1):79–105.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 163–173, Montréal, Canada. ACL.

Cristina Vögelin, Thorben Jansen, Stefan D. Keller, Nils Machts, and Jens Möller. 2019. The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, 39:50–63.

Zarah Weiss. 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.

Zarah Weiss and Detmar 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. International Committee on Computational Linguistic.

Zarah Weiss and Detmar 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.

Zarah Weiss and Detmar Meurers. in press. Broad linguistic modeling is beneficial for German L2 proficiency assessment. In *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, Louvain-La-Neuve. Presses Universitaires de Louvain.

Stefanie A. Wind, Catanaya Stager, and Yogendra J. Patil. 2017. Exploring the relationship between textual characteristics and rating quality in rater-mediated writing assessments: An illustration with l1 and l2 writing assessments. *Assessing Writing*, 34:1–15.

Edward W. Wolfe, Tian Song, and Hong Jiao. 2016. Features of difficult-to-score essays. *Assessing Writing*, 27:1–10.

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, Honolulu.

Weiwei Yang, Xiaofei Lu, and Sara Cushing 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.

Helen Yannakoudakis, Øistein E. Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

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

Hyung-Jo Yoon and Charlene Polio. 2016. The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, pages 275–301.

A Appendix

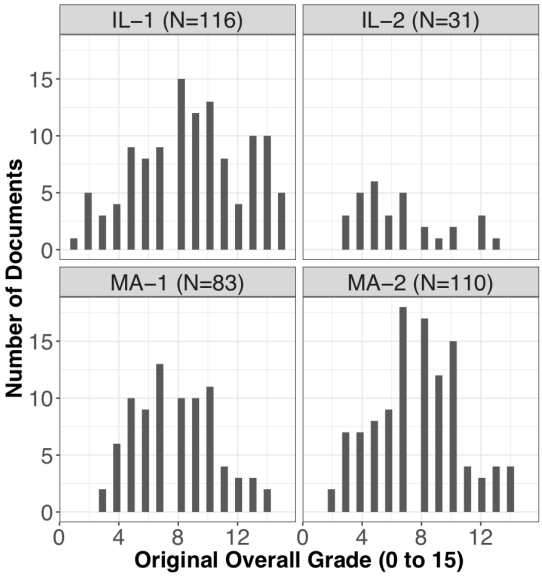

Figure 4: Original overall grades split by task prompt.

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.

Table 6: Overview of the four task prompts used to elicit the *Abitur* data.

Domain	Feature
Argumentation structure	Number of Paragraphs
	Adversative and concessive connectives (Breindl) per sentence
	Additive connectives (Breindl) per sentence
	Adversative connectives (Breindl) per sentence
	All connectives (Breindl) per sentence
	All multi word connectives (Breindl) per sentence
	All single word connectives (Breindl) per sentence
	Causal connectives (Breindl) per sentence
	Concessive connectives (Breindl) per sentence
	Other connectives (Breindl) per sentence
	Temporal connectives (Breindl) per sentence
	Adversative and concessive connectives (Eisenberg) per sentence
	Additive connectives (Eisenberg) per sentence,

	Adversative connectives (Eisenberg) per sentence
	All connectives (Eisenberg) per sentence
	All multi word connectives (Eisenberg) per sentence
	All single word connectives (Eisenberg) per sentence
	Causal connectives (Eisenberg) per sentence
	Concessive connectives (Eisenberg) per sentence
	Other connectives (Eisenberg) per sentence
	Temporal connectives (Eisenberg) per sentence
	Global argument overlap per sentence
	Global content overlap per sentence
	Global noun overlap per sentence
	Global stem overlap per sentence
	Local argument overlap per sentence
	Local content overlap per sentence
	Local noun overlap per sentence
	Local stem overlap per sentence
Lexical complexity	HDD
	MTLD
	TTR
	Bilogarithmic TTR
	Corrected TTR
	Root TTR
	Uber index
	Yule's K
	Adjectives and adverbs per lexical word
	Adjectives per lexical word
	Adverbs per lexical word
	Corrected lexical verb type per lexical per token
	haben instanced per verb
	Lexical types per lexical token
	Lexical types per token
	Lexical verb type per lexical token
	Lexical verb type per lexical verb token
	Lexical verb per token
	Nouns per lexical verb
	Lexical verbs per word
	Nouns per lexical word
	Nouns per word
	sein instances per verb
	Squared lexical verb types per lexical verb
	Verbs per noun
Syntactic complexity	Clauses per sentence
	Conjunctional clauses per sentence
	Dependent clauses per sentence
	Relative clauses per sentence
	Dependent clauses with conjunction per sentence
	Dependent clauses without conjunction per sentence
	Interrogative clauses per sentence
	Words per sentence
	Complex t-units per sentence
	Complex nominals per sentence

Postnominal modifiers per noun phrase
Prenominal modifiers per noun phrase
Noun phrase modifiers per noun phrase
Coverage of noun phrase modifier types
Verb modifiers per verb phrase
Coverage of verb modifier types
Coverage of verb cluster sizes
Coverage of verb cluster types
Standard deviation of verb cluster sizes
Mean verb cluster size
Coverage of Periphrastic tenses
Coverage of tenses
Coverage of deagentivization patterns

Table 7: List of all complexity features that are theoretically motivated by the German curriculum (KMK, 2014b).

Feature	IL-1	IL-2	MA-1	MA-2
MTLD	.2014	.4358	.2876	.3361
Root type token ratio	.3140	.3361	.3355	.2179
Corrected lexical verb types per lexical verb	*.2338*	.3103	.2105	*.2294*
Squared lexical verb types per lexical verb	.2588	*.3022*	.1998	.2458
Lexical verb types per lexical verb	.0587	.2257	.2291	.2446
Uber Index	.1153	.2412	.3131	*.2281*
Lexical word types found in dlexDB	-.3367	*-.4004*	-.1795	*-.2597*
Lexical word types not found in KCT	*.3901*	.4959	*.2770*	.1495
Clauses per sentence	*.2198*	.4681	.2304	-.0623
Dep. clauses per sentence	*.3040*	.2528	.2046	-.0380
Dep. clauses with conjunction per sentence	.3055	*.2013*	*.2029*	-.0484
Words per sentence	.3546	.4698	.2197	-.0403
Additive conn. per sentence (Breindl)	.2974	.2319	.2073	.1500
1-word conn. per sentence (Breindl)	*.2131*	.2855	*.2044*	.0745
Genitive case per noun phrase	*.2853*	.4689	.1869	*.2044*
-ung nominalizations per word	*.2080*	.4286	.1122	*.2339*
Derived nouns per noun phrase	*.2394*	.4751	.1604	.3301
Postnominal modifiers per noun phrase	.3064	.4510	.2031	.1113
Probability(other→other) per sentence	.1194	*.2077*	.1152	.3054
Probability(object→object) per sentence	-.1419	-.4929	.0545	*-.2068*
Global noun overlap per sentence	.2686	*.3072*	.1066	-.1590
Local content overlap per sentence	-.1359	-.2527	-.1725	-.3631
Global stem overlap per sentence	*.2587*	.4042	-.1162	-.0647
Temporal conn. per sentence (Breindl)	.2769	-.0185	.2206	.0408
Causal conn. per sentence (Eisenberg)	.3096	.3876	.0485	.0761
1-word conn. per sentence (Eisenberg)	*.2733*	.5241	.1068	.0275
Maximal total integration cost at finite verb (C)	*.2739*	.5062	-.0398	.0514
Average total integration cost at finite verb	.4093	*.4909*	.0708	.0308
Syll. between non-adjacent 1. argument & VFIN	.3158	*.2757*	.0210	.0815
Syllables in middle field per MF	.4244	.4286	.0351	.1092
Longest dependency in words	.3929	*.3207*	.0146	.1740
Prenominal modifiers per noun phrase	.2442	.5263	.0229	.1039
Possessive noun modifiers per NP	*.2378*	.4167	.1802	-.0308
Complex noun phrases per noun phrase	.4177	*.3186*	.1316	-.0353

Noun modifiers per noun phrase	.3357	*.2045*	.0689	.0648
NP deps. per NP with dependents	*.2855*	.4180	.1321	-.0798
Complex noun phrases per sentence	*.4177*	.3186	.1316	-.0353
Verb modifiers per verb phrase	.3565	.4219	.1761	.0375
Prepositional verb modifier per sentence	*.2184*	.4347	.0658	-.1204
Coordinated phrases per sentence	.3413	*.3299*	.0465	.1603
Average log type frequency in Google Books '00	-.4396	*-.4289*	-.1903	-.0994
Accusative case per noun phrase	-.3169	*-.2909*	-.0131	.0996
Lexical types per token	.2413	.1043	.0050	*.2446*
Verbs per noun	-.2213	-.3284	-.1294	-.1475
Nouns per lexical word	*-.2667*	.1709	.1916	.2415
Temporal conn. per sentence (Eisenberg)	.2225	.2244	.1665	-.2012
Determiners per noun phrase	-.3139	*.3066*	-.0006	.0023
Lexical verb types per lexical word	-.3142	-.0391	.1019	.0736
Yule's K	-.1144	-.2352	-.1663	-.0534
Lexical verbs per token	-.2667	-.1414	-.1022	-.0588
Adverbs per lexical word	-.0281	-.2781	-.0311	-.0401
Adjectives per lexical word	.1259	.3089	.1534	.0970
Dative case per noun phrase	-.1291	.1071	-.0440	-.3914
Third person markings per VFIN	-.0097	-.4361	-.1556	-.0727
-ist nominalizations per word	.0128	.4197	-.1266	.0122
Local argument overlap per sentence	.0547	-.1601	-.0256	-.2787
Local noun overlap per sentence	-.0007	-.0650	-.1356	-.2188
Causal conn. per sentence (Breindl)	.1512	.0658	.2936	-.0194
Concessive conn. per sentence (Eisenberg)	.0984	.2497	.0855	.0136
Other conn. per sentence (Breindl)	.1757	.2458	-.0343	.0181
Connectives per sentence (Eisenberg)	.1989	.3342	-.0400	.0386
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
Interrogative clauses per sentence	.0982	.4078	.0506	-.0574
Auxiliary verb cluster per verb cluster	.0460	.0569	-.0375	-.3221
haben instances per word	-.1818	-.2031	.-0251	-.1989
Coverage of verb cluster sizes	.1617	-.2824	-.1325	-.0088
Non-modal VP deps. per verb with dependents	.3219	.1250	.1804	.1116
Coverage of verb modifier types	.0758	.2216	.1706	.0119
Coverage of deagentivization patterns	.0763	.0227	.2020	-.0097
Passives per sentence	.1879	.4329	-.1692	-.0660
Average lemma frequency in dlexDB	-.4126	-.1037	-.1461	.0255
Average log lemma frequency in dlexDB	-.3890	-.1767	.0589	.0042
Hyponyms per type in GermaNet	-.3018	-.0741	-.1354	-.0926

Table 8: Features used in at least one of the four complexity document vectors and their correlation with the original overall grade across tasks. Gray font marks uncorrelated features. Italics mark relevant features that were excluded from the respective vector due to redundancy.

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

Table 9: Contribution of theory- and data-driven feature selection to each language complexity vector.

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

Table 10: German *Abitur* Grading System (KMK, 2018, p. 22).