Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School

Zarah Weiss and **Detmar Meurers**
University of Tübingen
Department for General and Computational Linguistics
{zweiss,dm}@sfs.uni-tuebingen.de

Abstract

We track the development of writing *complexity* and *accuracy* in German students' early academic language development from first to eighth grade. Combining an empirically broad approach to linguistic complexity with the high-quality error annotation included in the Karlsruhe Children's Text corpus (Lavalley et al., 2015) used, we construct models of German academic language development that successfully identify the student's grade level. We show that classifiers for the early years rely more on accuracy development, whereas development in secondary school is better characterized by increasingly complex language in all domains: linguistic system, language use, and human sentence processing characteristics. We demonstrate the generalizability and robustness of models using such a broad complexity feature set across writing topics.

1 Introduction

We model the development of linguistic complexity and accuracy in German early academic language and writing acquisition from first to eighth grade. Complexity and Accuracy are well-established notions from Second Language Acquisition (SLA) research. Together with Fluency, they form the CAF triad that has successfully be used to characterize second language development (Housen et al., 2012). Accuracy here is defined as a native-like production error rate (Wolfe-Quintero et al., 1998) and Complexity as the elaborateness and variation of the language which may be assessed across various linguistic domains (Ellis and Barkhuizen, 2005).

While there has been substantial research on the link between linguistic complexity analysis and second language proficiency and writing development for English (cf., e.g., Bulté and Housen, 2014; Kyle, 2016), much less is known about academic language development for other languages, such as the morphologically richer German. In this article, we target this gap with three contributions. We build classification models for early academic language development in German from first to eighth grade, based on a uniquely broad set of linguistically informed measures of complexity and accuracy. Our results indicate that two phases of academic language development can be distinguished: Initial academic language and writing acquisition focusing on the writing process itself, best characterized in terms of accuracy development, with little development in terms of complexity. A second stage is characterized by the increasing linguistic complexity, in particular in the domains of lexis and syntactic complexity at the phrasal level. We demonstrate the robustness and generalizability of the models informed by the broad range of linguistic characteristics – a major concern not only for obtaining practically relevant approaches for real-life use, but also for characterizing machine learning going beyond focused task to approaches capable of capturing general language characteristics.

The article is structured as follows: We first give a brief overview of research on writing development in terms of complexity and accuracy. We then present the *Karlsruhe Children's Text* corpus used as empirical basis of our work. In Section 4, we spell out our approach to assessing writing in terms of complexity and accuracy, before sections 5, 6, and 7 report on three studies designed to address the research issues introduced above.

2 Related Work

The main strand of research analyzing the complexity and accuracy constructs targets the assessment of second language development. Linguistic complexity measures have been successfully used to model the language acquisition of English

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 380–393
Florence, Italy, August 2, 2019.

as a Second Language (ESL) learners (Bulté and Housen, 2014; Crossley and McNamara, 2014). Work on first language writing development for English has also been conducted, but it is less common (Crossley et al., 2011) The same holds for the development of accuracy (Larsen-Freeman, 2006; Yoon and Polio, 2016). Most studies focus on adult ESL learners' development during periods of instruction. Vercellotti (2015) finds an increase in syntactic and lexical complexity, overall accuracy, and fluency in adult ESL speech over the course of several months. Crossley and McNamara (2014) find that advanced adult ESL learners phrasal and clausal complexity significantly increases over the course of one semester of writing instruction in particular with regard to nominal modification and number of clauses. These findings are corroborated by Bulté and Housen (2014). For uninstructed settings, however, this does not hold. Knoch et al. (2014, 2015) study university students' ESL development over 12 months and three years without instruction in an immersion context and found that only fluency but not grammatical and lexical complexity developed.

Research on languages other than English is starting to appear (Hancke et al., 2012; Velleman and van der Geest, 2014; Pilán and Volodina, 2016; Reynolds, 2016). As for English, research on German writing development has predominantly focused on German as a Second Language (GSL) in instructed settings (Byrnes, 2009; Byrnes et al., 2010; Vyatkina, 2012). Their findings suggest that as for ESL learners' writing, clausal complexity progressively increases. For lexical complexity results have been more mixed depending on the proficiency of GSL learners' proficiency level. The development of writing accuracy has also been assessed in some corpus studies using automated or manual error annotation (Lavalley et al., 2015; Göpferich and Neumann, 2016). In Weiss et al. (2019) we analyze the impact of linguistic complexity and accuracy on teacher grading behavior.

One challenge for the assessment of language performance in terms of complexity that is starting to receive attention is the influence of the task. Alexopoulou et al. (2017) demonstrate task effects, specifically task complexity and task type, on the complexity of English as a Second Language writers in the EF-Cambridge Open Language Database (EFCAMDAT) and show mixed results for accuracy. This is in line with findings by Yoon and Polio (2016), who investigate the effect of genre differences on CAF constructs. Yoon (2017) focuses on the effect of topic on the syntactic, lexical, and morphological complexity of ESL learners' writings and shows a significant influence on the complexity of writings of the same learners, similar to findings in Yang et al. (2015). Such task effects have mostly been discussed from a theoretical perspective, considering their implications for the development of CAF constructs and the two main task frameworks (Robinson, 2001; Skehan, 1996). From a more practical perspective, task, genre, and topic effects have been recognized as an important issue for machine learning for readability assessment or Automatic Essay Scoring (AES). For the real-world applicability of such approaches it is crucial for them to account for differences due to genre or topic. In their readability assessment system *READ-IT* for Italian, Dell'Orletta et al. (2014) use this issue to motivate favoring a ranking-based over a classification-based approach. A recent AES approach discussing the issue is the placement system for ESL by Yannakoudakis et al. (2018).

3 Data

Our studies are based on the *Karlsruhe Children's Text* (KCT) corpus by Lavalley et al. (2015).[1] It is a cross-sectional collection of 1,701 German texts produced by students in German elementary and secondary school students from first to eighth grade. The secondary school students in the corpus attended one of two German school tracks, either a basic school track (*Hauptschule*) or an intermediate school track (*Realschule*). The texts were written on a topic chosen by the students from a set of age-appropriate options: Elementary school students were asked to continue one of two stories, one about children playing in a park, and the other about a wolf who learns how to read. Secondary school students wrote about a hypothetical day spent with their idol or their life in 20 years. All student texts in the corpus are made available in the original, including all student errors, and a normalized version, where errors and misspellings were corrected. The data is enriched with error annotations covering word splitting, incorrect word choices and repetitions, grammar, and legibility.

For our studies analyzing writing development

[1] https://catalog.ldc.upenn.edu/LDC2015T22

in terms of development across the grade levels, we made use of the normalized texts and the error annotation. Some grade levels in the corpus include only few texts, such as the 42 cases of first grade writings compared to the other grade levels with 189 to 283 writings. We thus grouped adjacent grade levels, i.e., grades 1 and 2 together, grades 3 and 4, etc., to obtain a data set with a substantial number of instances for each class.

4 Assessment of Writing Performance

To assess writing performance in terms of complexity and accuracy, we operationalized these SLA concepts in terms of several features which we automatically computed or derived from the error annotation of the KCT corpus.

4.1 Complexity

The analysis of complexity is based on our implementation of a broad range of complexity features for German (Weiss, 2017; Weiss and Meurers, 2018, in press). The features cover clausal and phrasal syntactic complexity, lexical complexity, discourse complexity, and morphological complexity. Complementing the measures of complexity of the linguistic system, we also compute two cognitively-motivated features: a characterization of language use based on word frequencies, and measures of human language processing (HLP). Table 1 summarizes the features designed to capture the elaborateness and variability in the respective domain, with more details provided in Weiss (2017) and Weiss and Meurers (in press). Overall, the studies in the current paper make use of a comprehensive set of 308 complexity features for the assessment of academic language development.[2]

4.2 Accuracy

The second dimension of language performance that we are interested in is writing accuracy. In SLA research accuracy has predominantly been assessed in terms of types of error rates or error-free T-units (Wolfe-Quintero et al., 1998; Verspoor et al., 2012). We exploited the KCT corpus' elaborate error annotation to extract a broad range of accuracy measures. Annotations on the level of individual letters and words mark (ill)legibility, word splitting errors, repetition errors, foreign words, and grammatical errors. Annotations at the sentence level mark content deletions, insertions, and incorrect word choices. In addition, we developed an approach to automatically derive additional error types by comparing the original student writings with their normalized sentence-aligned target hypotheses. This procedure allowed us to extract counts for punctuation errors, incorrect quotation marks, spelling mistakes, and word capitalization errors. The last item is a particular challenge of German orthography, given that capitalization in German is governed by a complex set of rules and conventions relating to syntactic structure.[3]

Overall, we extracted 20 accuracy counts which we aggregated and normalized by the total number or errors or the total number of words in the text as counted by the complexity analysis described in the previous subsection. The feature set measuring writing accuracy and an example feature is included as the last row in Table 1.[4]

5 Study 1: Predicting Grade-Levels across School Types

5.1 Set up

We extracted the text data from the KCT corpus, removing all texts containing less than ten words and excluding texts written by children younger than seven years and older than 15 years. This resulted in a corpus of N=1,633 texts, for which we computed the features of linguistic complexity and error rates. Table 2 shows the distribution of texts across grade levels and school tracks.

From the analyzed data set, we eliminated all complexity and error rate features that did not exhibit enough variability to be of interest for the analysis. Specifically, we excluded all features whose most common value occurred more than 90% of the time. For the remaining 262 features, we computed their z-score, centered around zero.

On this data, we performed ten iterations of 10-fold cross-validation (CV) generating different splits each time, i.e., 100 training and testing runs in total, using an SMO classifier with a linear kernel (Platt, 1998). This outperformed models using random forests or linear regression. Similarly, introducing non-linearity did not improve the clas-

[2] We are making the complexity code available as part of a multilingual version of CTAP: https://github.com/zweiss/multilingual-ctap-feature

[3] The Python script used to identify accuracy features in the KCT annotation is available at https://github.com/zweiss/KCTErrorExtractor

[4] Here and in the following, we will refer to this feature set as the *error rate* measures to avoid confusion with the term accuracy used as a classification performance metric.

Feature Set	Size	Description
Lexical complexity	31	measures vocabulary range (lexical density and variation) and sophistication, measures of lexical relatedness; e.g., type token ratio
Discourse complexity	64	measures the use of cohesive devices such as connectives; e.g., connectives per sentence
Phrasal complexity	47	measures of phrase modification; e.g., NP modifiers per NP
Clausal complexity	27	measures of subordination or clause constituents; e.g., subordinate clauses per sentence
Morphological complexity	41	measures inflection, derivation, and composition; e.g., average compound depth per compound noun
Language Use	33	measures word frequencies based on frequency data bases; e.g., mean word frequency in Subtlex-DE (Brysbaert et al., 2011)
Human Language Processing	24	measures of cognitive load during human sentence processing, mostly based on Dependency Locality Theory (Gibson, 2000) e.g., average total integration cost at the finite verb
Error Rate	41	measures ratios of error types per error or word; e.g., spelling mistakes per word

Table 1: Overview over the feature sets used to capture linguistic complexity and accuracy

	1/2	3/4	5/6	7/8	all
Elementary	203	524	0	0	727
Realschule	0	0	297	236	533
Hauptschule	0	0	165	208	373
all	203	524	462	444	1633

Table 2: Text distribution across grades & school tracks

sification. For each feature set introduced in Section 4, we trained a separate classifier to support a comparison of the different complexity and error feature sets. In addition, we built one classifier based on the combination of all complexity feature sets and one combining all feature sets including error rate. Finally, we built a classifier also including the meta information about the school track and topic chosen, to investigate their influence on the complexity features and the comparability of grade-levels across school types.

As reference for evaluating classifier performance, we use a majority baseline assigning always the most common grade level, and a second baseline inspired by traditional readability formulas, for which we trained a classifier using text length and average word length features.

5.2 Results & Discussion

Table 3 shows the performance of the classifiers in terms of mean accuracy and standard deviation across iterations and folds, and the feature set size. The majority baseline and the traditional readability feature baseline displayed above the dashed line are both around 32%. All linguistically informed classifiers clearly outperform these two baselines. The best performing model with an accuracy of 72.68% combines linguistic complexity features and error rate with information on topic and school track.[5] Adding this meta-information, which in most real-life application contexts is readily available, accounts for an 1.72% increase in accuracy. But also without this meta-information, the combination of linguistic complexity features and error rate is highly successful with an accuracy of 70.96%.

Let us take a look at the individual contributions of the different feature sets. The overall linguistic complexity classifier clearly outperforms the one informed by the error rate features. This comparison may be biased towards the linguistic complexity classifier because it is informed by six times more features. However, the impression that complexity features are more indicative for writing development as a function of grade level is supported by the classifiers based on individual domains of linguistic complexity, which are more comparable in size to the error rate based classifier. The lexical complexity, discourse complexity, and phrasal complexity classifiers all clearly outperform the classifier informed by error rate with accuracies between 60.10% and 61.29% compared to 54.47%. The same holds for morphological

[5] The confusion matrix for all ten iterations of the 10-CV may be found in Table 10 in the Appendix.

	Size	μ-Acc.	SD-Acc.
Majority baseline	1	32.08	0.14
Traditional baseline	2	32.56	0.80
All Features + Meta	264	**72.68**	1.94
All Features	262	**70.96**	2.01
Complexity	225	**68.35**	2.25
Error Rate	37	54.47	2.11
Lexical	31	60.10	1.69
Discourse	48	60.10	1.66
Phrasal	41	61.29	1.73
Clausal	26	52.95	1.56
Morphological	27	56.45	1.47
Language Use	30	45.45	1.28
Human processing	20	42.18	1.55

Table 3: Grade-level classification of elementary & secondary school texts, ten iterations of 10-fold CV, distinguishing levels 1st/2nd, 3rd/4th, 5th/6th, 7th/8th

complexity (56.45%), although the difference is less pronounced. However, not all dimensions of linguistic complexity outperform error rate. This holds only for features measuring the linguistic system. While psycho-linguistic measures of language use and human language processing clearly outperform the baselines, they are performing significantly worse than the error rate features. Language experience and cognitive measures of the complexity in processing language does not seem to be the factor limiting academic writing performance, which is intuitively plausible considering that, especially in the early school years, the language experience and language processing will be mostly shaped by spoken language interaction.

6 Study 2: Writing Development in Elementary vs. Secondary School

6.1 Set-Up

Having established that linguistic complexity and error rate successfully predict writing performance across academic writing development, let us compare the development in early writing with that in secondary school. For this, we split the KCT data into two subsets: one containing only elementary school writing ($N = 727$), the other the secondary school writing from the different school tracks ($N = 906$). We applied the same pre-processing steps described in Section 5.1 including feature reduction and scaling of all predictor variables, obtaining 256 features for the elementary school and 255 for the secondary school data set (with numbers differing slightly since the feature reduction is performed separately on each data set).

We then followed a two-fold approach: First, we again tested and trained the same SMO classifiers as in Study 1 with linear kernels and 10 iterations of 10-fold CV (Section 6.2). Although the classifiers were informed by the same feature sets, due to the reduction of the sample size some sets were reduced more in the aforementioned preprocessing step which may result in slightly deviating feature set sizes across tables. For the elementary school data set, only topic was added as meta information, because there are no different elementary school tracks in Germany.

Then, for both data sets we selected the most informative features of each feature set in order to zoom in on how they differ across grade-levels (Section 6.3). This more fine grained analysis allows us to complement the broader perspective gained form the classification experiments with a more concrete sense of which features matter and how they change. For the selection, we ranked all features by their information gain for the distinction of grade-levels in the respective data set and selected the most informative feature of each feature set resulting in overall 16 features chosen for closer inspection. We then conducted two-tailed t-tests to test for significant differences across grade-levels in both data sets. To avoid redundancy in our comparison, if the most informative feature for a given feature set in both data subsets assessed the same concept, we chose the next-most informative feature.[6]

6.2 Results & Discussion

Table 4 shows the classifiers performance on the elementary school data subset.

Unlike in the previous study, the majority baseline for this binary classification task is relatively high with 71.72% given that there is less data for the first and second grade. As in the first study, the second baseline using the traditional readability formula features text length and average word length performs only at the level of the majority baseline. The classifier combining evi-

[6] For example, the most informative feature of lexical complexity is in both subsets a measure of lexical diversity (Yule's k and root type-token ratio). Due to its higher ranking (overall most informative for secondary school) and its reduced sensitivity to text length, we chose to keep Yule's k and included the second most informative lexical complexity feature for elementary school: corrected verb variation (measuring lexical variation).

	Size	μ-Acc.	SD-Acc.
Majority baseline	1	71.72	0.35
Traditional baseline	2	71.72	0.35
All Features + Meta	256	**82.81**	2.11
All Features	255	**82.60**	1.97
Complexity	218	77.93	2.42
Error Rate	37	**81.56**	1.27
Lexical	31	77.32	1.92
Discourse	46	75.18	1.71
Phrasal	39	76.77	2.18
Clausal	26	72.44	0.49
Morphological	27	71.72	0.35
Language Use	30	71.72	0.35
Human processing	19	71.72	0.35

Table 4: Grade-level classification of elementary school texts, ten iterations of 10-fold CV, distinguishing levels *1st/2nd* and *3rd/4th*

dence from linguistic complexity features and error rate clearly outperforms the baselines with an accuracy of 82.60%.[7] Adding meta-information, which here means adding the writing topic, does not make a significant contribution.

Looking at the classifiers for the subsets of features, we see that error rate features make a significant contribution. While the difference in performance still is significant,[8] the classifier informed only by error rate features with an accuracy of 81.56% performs close to the combined model with an accuracy of 82.60%. The classifier using only complexity features performs worse, with an accuracy of 77.93%, even though this classifier is informed by considerably more features. When looking at the individual domains of linguistic complexity, again lexical complexity, discourse complexity, and phrasal complexity are the most informative features, but they perform significantly lower than the error rate features. The other domains of linguistic complexity seem to be uninformative for the grade level distinction in elementary school student writings – clausal and morphological complexity, language use, and human language processing all perform at baseline level.

Our findings show that early writing and academic language development predominantly focuses on establishing writing correctness rather than language complexification. However, in certain domains writing performance also advances in terms of complexity, namely the lexicon, discourse, and phrase complexity. Systematic improvements in the domains of clausal and morphological complexity or language use and human language processing, however, do not take place.

Turning to the secondary school data set, Table 5 shows the classification results for that subset.

	Size	μ-Acc.	SD-Acc.
Majority baseline	1	51.15	0.27
Traditional baseline	2	51.56	1.75
All Features + Meta	258	**65.66**	2.13
All Features	255	**63.71**	1.82
Complexity	220	**64.16**	1.63
Error Rate	35	54.34	2.48
Lexical	30	62.74	1.58
Discourse	45	57.13	1.75
Phrasal	41	57.64	2.10
Clausal	25	58.70	2.37
Morphological	27	54.31	2.39
Language Use	30	55.73	2.34
Human processing	18	52.67	1.90

Table 5: Grade-level classification on secondary school texts, ten iterations of 10-fold CV, distinguishing levels: *5th/6th* and *7th/8th*

The data set is more balanced across grouped grade levels, with a majority baseline of 51.15%. Traditional readability features again perform at the same level as the majority baseline. The best performing classifier again combines the features encoding linguistic complexity and error rate with information on topic and school track. It reaches an accuracy of 65.66%, performing nearly 2% better than the model without the meta-information.[9] Different from the elementary school data classifier, we here also distinguish the two secondary school tracks, which apparently differ in the complexity of the texts written in a given grade level.

A comparison of the classifiers based on error rate features versus the complexity features shows that for secondary school grade levels linguistic complexity is more indicative for differentiating grade levels. The classifiers differ in terms of their accuracy by nearly 10%. When comparing the performance of error rate features with the individual domains of linguistic complexity, we see that this difference cannot merely be explained by

[7] The confusion matrix for all ten iterations of the 10-CV may be found in Table 11 in the Appendix.

[8] One-sided t-test: t = -4.3978, df = 169.34, p = 9.63e-06

[9] The confusion matrix for all ten iterations of the 10-CV may be found in Table 12 in the Appendix.

the difference in feature set size. Lexical complexity, in particular, but also discourse complexity, phrasal complexity, and clausal complexity significantly outperform error rate features. This clear development of clausal complexity in secondary school writing is another difference to the development of writing of elementary school students. Language use and morphological complexity also show more development and significantly outperform the baselines. Human language processing features do not show a significant development.

Summarizing the findings from Table 4 and Table 5, we saw that the early writing and academic language development seemed to predominantly focus on establishing writing correctness rather than complexification. However, despite this focus on correctness, writing performance exhibits also in early stages of writing acquisition advances in terms of linguistic complexity in the domains of lexicon, discourse, and phrasal complexity. Systematic improvements in the other domains of linguistic complexity only take place at later stages of writing development. The beginning of this trend may be seen in the evidence from secondary school writings, for which clausal complexity and to some extent also morphological complexity and language use become increasingly relevant. Lexical complexity, phrasal complexity, and discourse complexity develop throughout all stages of writing acquisition.

6.3 Zooming in on Individual Features

Table 6 shows the most informative features from each feature set, their group means across grade-levels in elementary and secondary school, and the results of the t-tests.[10] In the first step (Section 6.2), we found that error rate as well as lexical, phrasal, and discourse complexity develop in both, elementary and secondary school writing. Zooming in on these domains, we see that some features systematically develop throughout grade-levels. Overall error rate and capitalization errors are highly informative in both data sets and decrease significantly across all grade-levels. Similarly, for lexical complexity, lexical diversity measured by Yule's k significantly decreases with progressing grade-levels (from 217 in grade-level 1/2 to 128 in grade-level 7/8). However, not in all cases the results are as clear. Lexical variation measured as corrected verb ratio significantly increases from grade-levels 1/2 to 3/4 and 5/6 to 7/8. Yet, the lexical variation of grade-level 7/9 writing is closer to that of grade-level 3/4 than 5/6, leaving unclear to which extent we see systematic development in this subdomain of lexical complexity.

For discourse complexity, the transition probability of dropping the subject in a following sentence, i.e., not repeating it as, e.g., the subject or object, significantly decreases with increasing grade-level in elementary school, i.e., the discourse becomes more coherent. The probability remains stable at a low level in secondary school. There, discourse complexity seems to develop rather in terms of use of connectives such as temporal connectives which significantly increase with progressing grade-level, while showing inconclusive results for elementary school. The two most informative features from the domain of phrasal complexity behave similarly: The coverage of noun phrase modifiers for elementary school which significantly increases from grades 1/2 to grades 3/4 from 0.31 to 0.42 but stagnates around 0.52 in secondary school. For secondary school, it is represented by the ratio of verb modifiers per verb, which significantly increases across all grade-levels from 0.29 to 0.65.

In contrast to phrasal complexity, clausal complexity represented by conjunction clauses per sentence and verbs per t-unit does not significantly change throughout elementary school. However, it significantly increases in secondary school from 0.13 conjunction clauses per sentence to 0.18 and from 1.69 verbs per t-unit to 1.8. This is in line with our previous observation that elementary school writing rather develops in terms of phrasal but not clausal complexity, while clausal complexity gains importance in secondary school.

The same holds for morphological complexity and language use, which we found to only play a role in the development of secondary school writing. Accordingly, we do not see a significant difference in either across elementary school grade-levels for the most informative features of these domains. For secondary school writing, however, the number of derived nouns per noun significantly increases, indicating a stronger nominal style in students writing and we see a significant increase in vocabulary overlap with dlexDB, which consists of frequencies from news

[10] The appendix contains the information gain ranking for the 16 most informative features for both data sets, see Tables 15 and 16 as well as boxplots visualizing of all features across grade-levels, see Figures 2 to 1.

Feature name	Set	Elementary school				Secondary school			
		1/2	3/4	t	p	5/6	7/8	t	p
Overall errors / W	Error Rate	0.68	0.37	11.53	.000	0.28	0.22	5.60	.000
Corrected verb variation	Lexical	1.62	2.13	-11.55	.000	1.88	2.01	-3.03	.003
P(Subject → Nothing)	Discourse	0.15	0.10	3.40	.001	0.05	0.06	-1.35	.177
Avg. NP modifier types	Phrasal	0.31	0.42	-8.93	.000	0.52	0.52	-0.21	.831
Conjunction clauses / S	Clausal	0.11	0.13	-0.96	.339	0.13	0.18	-3.47	.001
Finite verbs / verb	Morph.	0.82	0.81	1.63	.105	0.71	0.70	0.88	.381
Pct. LW in Subtlex	Language Use	0.04	0.05	-1.71	.089	.085	.077	1.82	.069
DLT-IC (M) / finite verb	Human Processing	1.09	1.11	-1.96	.051	1.22	1.25	-1.65	.099
Capitalization errors / W	Error Rate	0.15	0.07	9.87	.000	0.05	0.04	5.61	.000
Yule's K	Lexical	217.	153.	7.21	.000	152.	128.	5.60	.000
Temp. connectives / S	Discourse	0.73	0.63	1.85	.066	0.47	0.62	-4.10	.000
Verb modifiers / VP	Phrasal	0.29	0.49	-4.85	.000	0.55	0.65	-2.86	.004
Verbs / t-unit	Clausal	1.67	1.57	-0.97	.333	1.69	1.81	-3.18	.002
Derived nouns / noun	Morph.	0.02	0.02	-0.38	.708	0.04	0.05	-2.66	.008
Pct. LW in dlexDB	Language Use	0.62	0.60	1.60	.111	0.60	0.63	-3.27	.001
($\sum$ max. dep.) / S	Human Processing	5.12	5.60	-2.64	.009	6.30	6.97	-4.59	.000

Table 6: Across-grade level group means of the most informative features of each feature set for distinguishing grade-levels in elementary school (above dashed line) and secondary school (below dashed line).

texts. This might indicate that language use becomes more similar to news language in secondary school, as dlexDB is based on news paper data.

Interestingly, for human language processing, there seems to be a marginally significant increase in DLT processing costs at the finite verb (with decreased modifier weight as defined in Shain et al. 2016) and a significant increase in the mean maximal dependency length per sentence across all grade-levels in elementary and secondary school.

	1/2	3/4	5/6	7/8	all
Wolf	133	353	0	0	466
Park	90	171	0	0	261
Future	0	0	332	333	665
Idol	0	0	130	111	241
all	203	524	462	444	1,663

Table 7: Distribution of grade levels across topics

7 Study 3: Cross-Topic Testing of Academic Language Development Across Topics

7.1 Set Up

In our final study, we want to test whether the results we obtained generalize across topics. Elementary school and secondary school students were both allowed to freely choose from two different topics for their writing as spelled out in Section 3. We used the two data subsets from Study 2, but additionally split them by topics, obtaining four data sets: i) elementary school: *Wolf* topic, ii) elementary school *Park* topic, iii) secondary school: *Future* topic, and iv) secondary school *Idol* topic. Table 7 shows the distribution of texts across grade levels and topics.

We used the data sets of *Wolf* topic writings and *Future* topic writings as training data sets and tested the resulting model on *Park* topic and *Idol* topic texts, respectively. We chose this set-up since the two test data sets are too small to allow for training and testing with reversed data sets. We do not use cross-validation here, because we specifically want to study transfer across different topics rather than just different folds. In the new set-up, we cross-topic trained and tested the SMO classifiers based on the combination of complexity and error rate features and separately for the error rate and for the complexity features. We compared the results against the majority baseline and the traditional readability baseline containing measures of text and word length. For the secondary school data, we trained one model with and one without meta information on school tracks.

7.2 Results & Discussion

Table 8 shows the cross-topic classification performance on elementary school students' writings.

Feature Set	Train	Test	Acc.
Majority baseline	*n.a.*	Park	65.52
Traditional baseline	Wolf	Park	65.52
All Features	Wolf	Park	76.63
Complexity	Wolf	Park	68.58
Error Rate	Wolf	Park	81.61

Table 8: Cross-topic results for elementary school data

The majority baseline for elementary school writings' on the *Park* topic is more balanced than the one for the *Wolf* topic. For both topics, 3rd/4th grade was the most common grade-level. Training on *Wolf* texts and testing on *Park* texts with the SMO classifier yields an accuracy of 76.63%. While this does constitute a drop in accuracy as compared to Study 2, which may at least partially be explained by the reduced size of the training data set, the model clearly generalizes across topics. When taking a closer look at the difference between the purely error rate-based informed classifier and the complexity feature based classifier, we see that both generalize across topics. However, error rate clearly outperforms the complexity features and in fact hardly drops in performance when compared to the results obtained in Study 2.[11] The better performance of the classifier informed by error rate compared to both complexity-based classifiers indicates that error rate is more robust across topics than complexity. It also further corroborates the particular importance of writing correctness for early writing and academic language development.

Table 9 shows the results of the classifiers for the secondary school writing.

Feature Set	Train	Test	Acc.
Majority baseline	*n.a.*	Idol	50.01
Traditional baseline	Future	Idol	43.15
All Features + Meta	Future	Idol	62.66
All Features	Future	Idol	59.33
Complexity	Future	Idol	59.34
Error Rate	Future	Idol	55.19

Table 9: Cross-topic results for secondary school data

Unlike for the elementary school data, grade-levels are more or less balanced across topics for this data set, leading to a majority baseline around 50%. As before, we see that all SMO classifier generalize across topics when training on the larger data set (*Future*) and testing on the smaller one (*Idol*). In line with their relative importance for this school level established in the second study, the complexity features play more of a role and interestingly generalize well, while the error rate measures known to play less of a role at this level of development are also less robust.[12]

8 Conclusion and Outlook

We presented the first approach modeling the linguistic complexity and accuracy in German academic language development across grades one to eight in elementary and secondary school. Our models are informed by a conceptually broad feature set of linguistic complexity measures and accuracy features extracted from error annotations. The computational linguistic analysis made it possible to empirically identify a shift in the developmental focus from accuracy as the primary locus of development in elementary school to the increasing complexity of the linguistic system in secondary school. Our results also show where both domains advance in parallel, in particular in the lexical complexity domain, which plays an important role throughout. Despite the emerging focus on complexity throughout secondary school, accuracy also continues to play a role. Investigating the generalizability of our results and the approach to complexity and accuracy development, we demonstrated the cross-topic robustness of our classifiers. The use of cross-topic testing to establish the robustness of machine learning models thus supports the applicability of language development modeling in real life.

These first results provide insights into the complexity and accuracy development of academic writing across the first eight years in German. Yet, they are based on the quasi-longitudinal operationalization of writing development as a function of grade level. Tracking genuine longitudinal develop of individual students across extended periods of time is a natural next step, which will make it possible to study individual differences and learning trajectories rather than overall group characteristics. We plan to follow up on this in future work.

[11] The confusion matrix for all ten iterations of the 10-CV may be found in Table 13 in the Appendix.

[12] The confusion matrix for all ten iterations of the 10-CV may be found in Table 14 in the Appendix.

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.

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. 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.

Heidi Byrnes. 2009. Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education*, 20(1):50 – 66.

Heidi Byrnes, Hiram H. Maxim, and John M. Norris. 2010. Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment. *The Modern Language Journal*, 94.

Scott A Crossley and Danielle S McNamara. 2014. Does writing development equal writing quality? a computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26:66–79.

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.

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 and Gary Barkhuizen. 2005. *Analysing learner language*. Oxford University Press.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain: papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT.

Susanne Göpferich and Imke Neumann. 2016. Writing competence profiles as an assessment grid? – students' L1 and L2 writing competences and their development after one semester of instruction. In *Developing and Assessing Academic and Professional Writing Skills*, pages 103–140. Peter Lang, Bern, Switzerland.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India.

Alexis Housen, Folkert Kuiken, and Ineke Vedder. 2012. Complexity, accuracy and fluency: Definitions, measurement and research. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency*, Language Learning & Language Teaching, pages 1–20. John Benjamins.

Ute Knoch, Amir Roushad, Su Ping oon, and Neomy Storch. 2015. What happens to ESL students' writing after three years of study at an English medium university? *Journal of Second Language Writing*, 28:39–52.

Ute Knoch, Amir Roushad, and Neomy Storch. 2014. Does the writing of undergraduate ESL students develop after one year of study in an english-medium university? *Assessing Writing*, 21:1–17.

Kristopher Kyle. 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.

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.

Rémi Lavalley, Kay Berkling, and Sebastian Stüker. 2015. Preparing children's writing database for automated processing. In *LTLT@ SLaTE*, pages 9–15.

Ildikó Pilán and Elena Volodina. 2016. Classification of language proficiency levels in swedish learners' texts. In *Proceedings of Swedish language technology conference*.

John C. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research.

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.

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

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka.

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

Eric Velleman and Thea van der Geest. 2014. Online test tool to determine the cefr reading comprehension level of text. *Procedia computer science*, 27:350–358.

Mary Lou Vercellotti. 2015. The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 38(1):90–111.

Marjolijn Verspoor, Monika S. Schmid, and Xiaoyan Xu. 2012. A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, 21(3):239–263.

Nina Vyatkina. 2012. The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*, 96(4):576–598.

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.

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.

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar 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.

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 Appendices

↓Obs/Pred→	1/2	3/4	5/6	7/8	$\sum$
1/2	**1217**	813	0	0	2030
3/4	430	**4810**	0	0	5240
5/6	0	0	**3029**	1591	4620
7/8	0	0	1590	**2850**	4440
$\sum$	1647	5623	4619	4441	16330

Table 10: Confusion matrix for the best model in study 1 (all feat. + meta) summed across iterations

↓Obs/Pred→	1/2	3/4	$\sum$
1/2	**1232**	798	2030
3/4	449	**4791**	5240
$\sum$	1681	5589	7270

Table 11: Confusion matrix for best elementary school model in study 2 (all feat. + meta) summed across iterations

↓Obs/Pred→	5/6	7/8	$\sum$
5/6	**3049**	1571	4620
7/8	1497	**2943**	4440
$\sum$	4546	4514	9060

Table 12: Confusion matrix for best secondary school model in study 2 (all feat. + meta) summed across iterations

↓Obs/Pred→	1/2	3/4	$\sum$
1/2	**51**	39	90
3/4	9	**162**	171
$\sum$	60	201	261

Table 13: Confusion matrix for the best model for elementary school in study 3 (Error rate)

↓Obs/Pred→	5/6	7/8	$\sum$
5/6	**91**	39	130
7/8	51	**60**	111
$\sum$	142	99	241

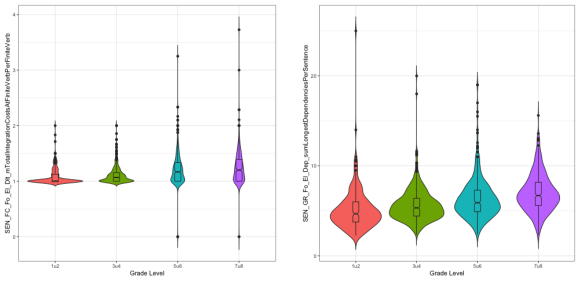
Table 14: Confusion matrix for the best model for secondary school in study 3 (all feat. + meta)

Feature name	Set	Merit
Overall errors / W	Error rate	.166
Root type-token ratio	Lexical	.150
Corrected type-token ratio	Lexical	.150
Number of words	Clausal	.137
Capitalization errors / W	Error rate	.128
HDD	Lexical	.124
Corrected verb variation	Lexical	.110
Squared verb variation	Lexical	.110
Word splitting + hyphenation errors / W	Error rate	.108
P(Subject→Nothing)	Discourse	.106
P(Nothing→Nothing)	Discourse	.104
P(Nothing→Subject)	Discourse	.099
Number of sentences	Clausal	.094
P(Nothing→Object)	Discourse	.093
Yule's K	Lexical	.091
MTLD	Lexical	.088

Table 15: Top features in information gain ranking for grade-level distinction in elementary school

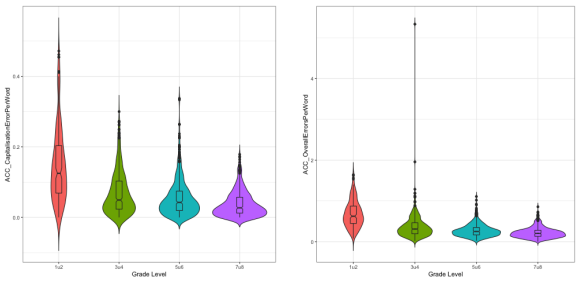
Feature name	Set	Merit
Yule's K	Lexical	.030
Capitalization errors / W	Error rate	.029
($\sum$ max. dep.) / S	Human processing	.026
MTLD	Lexical	.023
Verbs / t-unit	Clausal	.023
Verbs / S	Clausal	.023
HDD	Lexical	.022
Overall errors / W	Error rate	.022
Nouns / W	Lexical	.021
$\sum$ Non-terminal nodes / tree	Clausal	.021
W / S	Clausal	.021
to infinitives / S	Lexical	.020
Uber index	Lexical	.020
Temporal connectives / S	Discourse	.019
$\sum$ Non-terminal nodes / W	Clausal	.019
Clauses / S	Clausal	.017

Table 16: Top features in information gain ranking for grade-level distinction in secondary school



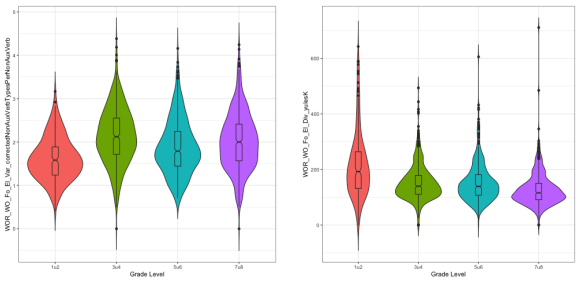
(a) DLT integration cost (m) (b) Max. dependency / S

Figure 1: Most informative human processing features



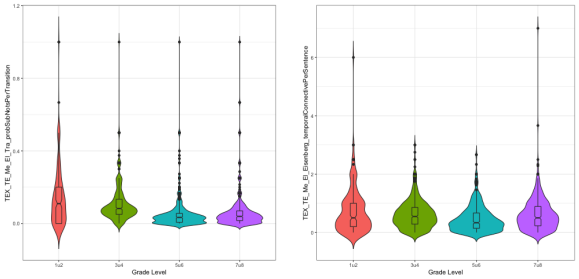
(a) Capitalization errors (b) Overall errors

Figure 2: Most informative error rate features



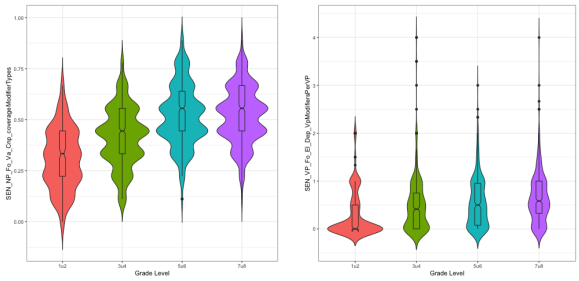
(a) Corrected verb variation (b) Yule's K

Figure 3: Most informative lexical features



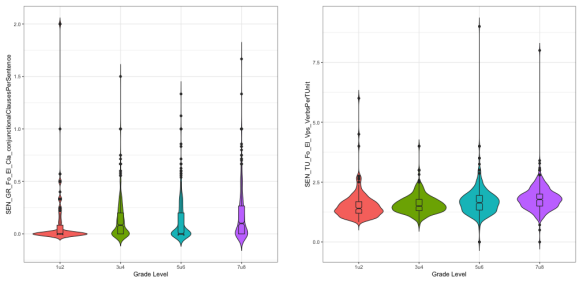
(a) Subject transitions (b) Temporal connectives

Figure 4: Most informative discourse features.



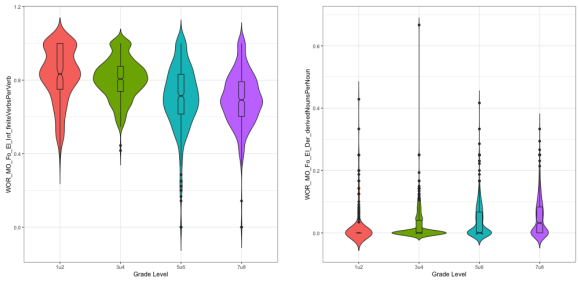
(a) NP modifier coverage (b) Verb modifiers / VP

Figure 5: Most informative phrasal features.



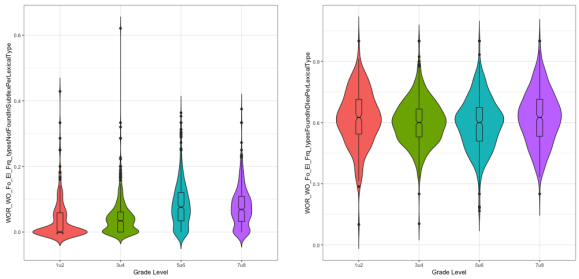
(a) Conjunction clauses / S (b) Verbs / t-unit

Figure 6: Most informative clausal features.



(a) Finite verbs / verb (b) Derived nouns / noun

Figure 7: Most informative morphology features.



(a) Words in Subtlex-DE (b) Words in dlexDB

Figure 8: Most informative language use features