Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

Assessing the Linguistic Complexity of German Abitur Texts from 1963–2013

Noemi Kapusta* and **Marco Müller†** and **Matilda Schauf*** **Isabell Siem*** and **Stefanie Dipper***
Sprachwissenschaftliches Institut
Fakultät für Philologie
Ruhr-Universität Bochum
* firstname.lastname@rub.de
† Marco.Mueller-z3b@rub.de

Abstract

This paper is about the analysis of the linguistic complexity of texts written by high school graduates as part of the final secondary-school examinations. We measure complexity on different levels (lexical diversity, perplexity of part-of-speech-based language models, and syntactic complexity) and compare the complexity of high school graduation texts from 1963–2013. It turns out that, contrary to our initial assumptions, linguistic complexity increases over time.

1 Introduction[1]

Successful literacy acquisition represents an important building block in the educational process of young people. Literacy is not only about the acquisition of correct spelling and grammar, but also about the ability to understand and produce texts with complex content, and to use appropriate registers in different situations.

Competent handling of texts with complex content is a prerequisite for successful study at university. The teaching of these skills is one of the main goals of the *Gymnasium* (secondary school). The relevant competencies are tested at the *Abitur* (the final secondary-school examinations), where school graduates must produce extensive texts as part of the German exam.

Over the past decades, the Gymnasium in Germany has changed considerably. While nationwide only a small minority of around 7% attended this type of school in the 1960s, today the figure is around 50%. This has been accompanied by a change in the composition of the student body, from a rather homogeneous, male-dominated selection of the educated population to a more diverse composition that includes children from educationally disadvantaged families and children from families with a migration background who may acquire German only as a second language.

In this paper, we investigate whether the changing composition of the school population has a measurable impact on literacy acquisition. For this purpose, we examine texts from the GraphVar corpus (Berg et al., 2021) that were written as part of the final secondary-school examinations for German in the period 1963–2013.

We focus on aspects of linguistic complexity, which we investigate at the lexical and syntactic levels. We pursue two hypotheses:

1. Because of the more homogeneous composition, the results in the 1960s are more homogeneous and have less variance.

2. Because of the more elite composition, the linguistic complexity of the texts is higher in the 1960s than nowadays.

Most work on linguistic complexity concerns data from foreign language (L2) acquisition, typically in the form of longitudinal studies over a few months in instructed settings. Such studies show that lexical and syntactic complexity typically increases over time (cf. Crossley, 2020). Besides complexity, the correctness (error rate) of texts is often investigated.

Written language acquisition in the native language is less frequently studied. A relevant corpus is the KoKo Corpus (Abel et al., 2014, 2016). It contains argumentative essays in German with about 825,000 words, written by students of graduating classes. The corpus is manually annotated for different error types (spelling, grammar). It has also been automatically enriched with part-of-speech (POS) annotations and lemmas. Additionally, it has been annotated on a textual level with

[1] All scripts, result tables and plots related to this work are available at https://github.com/rubcompling/konvens2022.

366 features related to linguistic complexity. However, we are not aware of any studies focusing on the complexity features.

The Falko corpora are a collection of different German-language corpora, mostly of L2 learners.[2] Parallel to the L2 data, there is usually a comparative corpus of L1 students. The data is richly annotated with linguistic information (lemma, POS), and errors are also annotated with corrected forms. In studies using these corpora, the L1 texts usually serve as a reference corpus, but this is not unproblematic, as Shadrova et al. (2021) show.

As a factor influencing complexity, task effects have been examined, and factors such as the task type, topic, and genre have been shown to have a significant impact on complexity (e.g., Alexopoulou et al. (2017); Weiss (2017)).

In contrast to the aforementioned corpora, the GraphVar corpus is a diachronic corpus and our focus is on the change of complexity through time. We investigate linguistic complexity using different methods: word-based measures of lexical complexity, and POS bigram probabilities and a selection of traditional syntactic features for syntactic complexity. For lexical and syntactic features, see, e.g., the overview in Crossley (2020). Further references to related literature can be found in the respective sections.

The paper is structured as follows. In Sec. 2, we present the corpora our investigations are based on. Sec. 3 introduces the different measures that we apply to assess complexity: lexical diversity, POS-based perplexity, and various syntactic features related to complexity. Sec. 4 presents the results and Sec. 5 concludes the paper.

2 Data

For our investigations, we use a subset of the GraphVar corpus (Sec. 2.1).

In addition, we use two reference corpora that we compiled in the context of this work: first, the EXPRESS corpus with a rather simple linguistic style; second, the ZEIT corpus which has a rather complex and sophisticated linguistic style (Sec. 2.2). We exploit the reference corpora in two ways:

First, for measuring POS-based perplexity we train two models on the full reference corpora. Second, for assessing lexical diversity and syntactic complexity, we compare the results from the GraphVar corpus with results from subsets of the reference corpora.

[2] https://hu-berlin.de/falko.

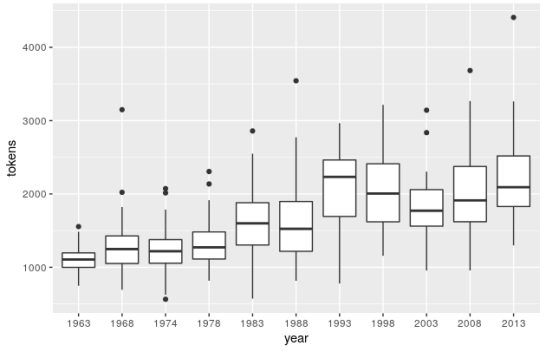


Figure 1: Boxplots of number of tokens per text, grouped by survey year.

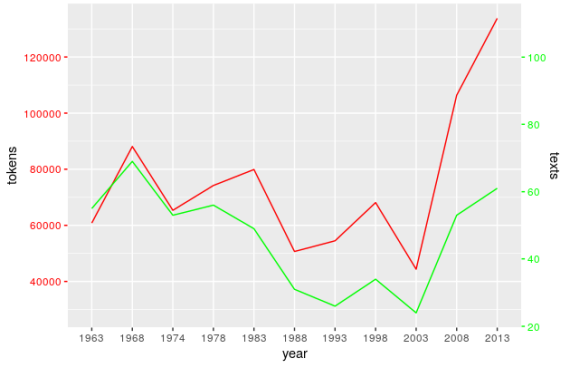


Figure 2: Plot of number of tokens (red) and total number of texts (green) per survey year (rescaled).

Text samples of each corpus can be found in Appendix A.

2.1 The GraphVar Corpus

The current version 1.4.2 of the GraphVar corpus (Berg et al., 2021) contains more than 1600 high school graduation essays from the years 1923–2018 from the subjects German, Biology and History. For our research, we use a subset containing only essays from the subject German from 1963–2013. The texts were collected at intervals of roughly five years.

We preprocessed the texts and excluded all tokens that were annotated as headers. Such tokens were not produced by the students but were part of the task description. Figure 1 displays information on the number of tokens per text. The boxplots show that the average text length has increased continuously since 1963. We decided to consider all data, though, because the subsets (per survey year) are rather small, with an average number of tokens of 75,000 (average per text: 1,600). In study-

ing the development of complexity over time, it is therefore important to use normalized complexity measures or measures that are not sensitive to text length.

Figure 2 shows the total number of tokens and texts per survey year. It can be seen that slightly fewer texts were included in the corpus from the 1980s and 1990s, and the total number of tokens in these years is also slightly lower. In the most recent years, 2008 and 2013, there is a clear increase in the number of texts and tokens.

The GraphVar corpus has been annotated manually and automatically with various linguistic information, including lemma, part of speech (POS) according to the STTS scheme (Schiller et al., 1999), and syntax according to the TüBa/DZ scheme (Telljohann et al., 2012). For calculating lexical diversity and syntactic complexity, we use the lemma forms and syntactic annotations provided by the corpus. Syntactic annotations are represented in GraphVar as spans spanning the dominated tokens. For further processing, we converted the GraphVar data into a column format, translating the syntactic annotation into a path notation that represents the dominating nodes (BIE tags[3]) as a path from the root to the terminal node. For instance, `I-SIMPX|B-MF|NX|PPER` is the syntactic annotation of a personal pronoun (PPER) embedded in a singleton nominal phrase (NX) which is the first node in the middle field (B-MF) inside a clause (I-SIMPX).

We randomly divided the corpus into a dev set (20%, 107 texts) and a test set (80%, 404 texts). The test set is the basis for the evaluations in Sec. 4.

2.2 Reference Corpora

For the EXPRESS corpus, we downloaded articles of the daily German newspaper "EXPRESS" from 2021/01/02 to 2022/07/03. For the ZEIT corpus, we downloaded articles of the German weekly newspaper "DIE ZEIT" from 2021/03/11 to 2022/03/02. Both data sets were downloaded from `wiso-net.de`, an online database that offers eBooks and journals as well as newspaper articles for research purposes.

We filtered out articles from categories that do not consist of plain newspaper text[4] and articles with less than 500 tokens. Both corpora contain roughly the same number of tokens, see Table 1.

Corpus	#Articles	#Tokens	#Types
EXPRESS	4,565	3.4M	180K
ZEIT	2,022	3.4M	190K

Table 1: The two reference corpora.

Subcorpus	#Fragments	#Tokens	#Sentences
EXPRESS	138	70,398	3,758
ZEIT	137	70,134	3,796

Table 2: The subsets of the two reference corpora.

We use the full corpora for training POS-based language models (Sec. 3.2).

In addition, we use randomly selected subsets of the reference corpora for assessing lexical diversity (Sec. 3.1) and syntactic complexity (Sec. 3.3) of the reference texts, see Table 2. These subsets contain about 70,000 tokens, which roughly corresponds to the median size of GraphVar texts of one survey year. The subsets consist of article fragments with at least 500 tokens each.[5]

3 Measures of Complexity

We study linguistic complexity at different levels and with different measures. First, we look at lexical diversity (Sec. 3.1); second, we use perplexity of part-of-speech (POS) based language models to estimate syntactic complexity (Sec. 3.2); third, we apply different measures to syntactic annotations (Sec. 3.3).

3.1 Lexical Diversity

Lexical complexity of learner data is measured in several ways. Lexical sophistication looks at the proportion of "complicated" words in the text. Complicated words are determined, for example, by word lists or by their general frequency: the rarer, the more complicated (Laufer and Nation, 1995).

Another aspect is lexical density, which is measured by measures such as Type-Token Ratio (TTR) or improved variants thereof. TTR is the ratio of word types to the total number of tokens in a text. However, it is well known that TTR depends on

[3] B: begin of a span/node; I: inside a span/node; E: end of a span/node. Singletons are not marked as such.

[4] E.g. "Impressum" (imprint), "Schach" (chess), "Witz der Woche" (joke of the week), "Glückszahlen" (lucky numbers), "Leserbriefe" (letters to the editor).

[5] In calculation the lexical diversity measure MATTR, we use a window of 500 tokens, so this is the minimum length for individual texts (see Sec. 3.1).

the text length, hence, it cannot be used for comparing texts of different length. Other TTR-based measures have been proposed in the past, such as Corrected TTR, Log-TTR, and Root TTR, all of which, however, have been shown to be affected by text length (e.g., Zenker and Kyle, 2021). Measures that turned out stable and are used in the current study are MTLD (McCarthy and Jarvis, 2010), MATTR (Covington and McFall, 2010), and HD-D (McCarthy and Jarvis, 2007), which we describe in the following sections. With all three measures, a higher score indicates a lexically more diverse text.

3.1.1 MTLD

McCarthy (2005) and McCarthy and Jarvis (2010) propose MTLD ("Measure of Textual Lexical Diversity") as a length-independent measure of lexical density. This measure is calculated as the mean length of segments (i.e., sequences of words) with a given TTR. The TTR is calculated for increasing bits of text, with the first round starting at the beginning of the text and going on until the given TTR threshold (default = 0.72) has been reached. At this point, the next round starts with TTR reset to 1. This process is repeated until the end of the text. Usually there are tokens left at the end of a text whose TTR does not reach the threshold. For these tokens, a partial factor is calculated, so that no data is discarded (see McCarthy and Jarvis (2010) for details). The whole process is first run forward and then reverse, hence, bidirectional, which produces consistent and accurate MTLD scores. MTLD is calculated as the total number of words in the text divided by the number of rounds.

MTLD has been proven to be a reliable measure of lexical diversity in studies such as Koizumi and In'nami (2012) and Fergadiotis et al. (2013). Only for short texts (with < 100 words), which do not even reach the given TTR score, the results are unreliable.

3.1.2 MATTR

Covington and McFall (2010) introduce MATTR ("Moving Average Type-Token Ratio"). Similar to MTLD, MATTR is based on TTR. Yet, while MTLD uses segments that can be of different length, MATTR uses a window of a fixed size that moves forward by one token at a time and whose TTR is calculated in each case. Covington and McFall (2010) suggest a large window for lexical diversity. Since the shortest GraphVar texts contain roughly 550 tokens, we chose a window size of 500. The MATTR score of the text is the mean of all these TTR scores.[6]

3.1.3 HD-D

McCarthy and Jarvis (2007) propose HD-D ("Hypergeometric Distribution D"), which is a simplified version of vocd-D (Malvern et al., 2004). vocd-D calculates TTR scores for random samples of different size. In contrast, HD-D is based on probabilities: For every type in a text, the probability of occurring in a sample of n tokens is calculated. As recommended by McCarthy and Jarvis (2007), we use $n = 42$. HD-D is the sum of all probabilities.

3.2 Perplexity of POS-based Language Models

Perplexity is a common measure to evaluate language models, by comparing perplexity of two models on a test set. The model with the lower perplexity score fits the test data better.

We assume that the ZEIT corpus has a more complex language style than the EXPRESS corpus. A language model trained on the ZEIT corpus should therefore have a lower perplexity on a linguistically complex test text than a language model trained on the EXPRESS corpus. However, the perplexity of two models can only be compared if they use identical vocabularies. Therefore, it is not possible to compare language models based on word ngrams here. Instead, we compare POS ngrams (more precisely: POS bigrams), since here the vocabulary of both training corpora is identical. So essentially we compare syntactic properties.

We calculated the perplexity as described in Jurafsky and Martin (2022) with the log probabilities of the bigrams. For the test set, we randomly extracted the same number of bigrams from each text of the same year such that a total of 5000 bigrams per survey year are included in the test set.

3.3 Syntactic Complexity

For measuring syntactic complexity, we use the syntactic annotation provided by the GraphVar corpus, which we converted into path representations (Sec. 2.1). We implemented a range of measures that have been listed in Chen and Meurers (2016) for measuring syntactic complexity, in particular measures that relate to complex constituents (like

[6] MATTR is an improved version of MSTTR ("Mean Segmental Type-Token Ratio"). MSTTR uses non-overlapping segments and has to discard remaining words at the end of the text (for details, see the description in Covington and McFall (2010)).

No	Feature	Definition
1	Mean Sentence Length	#tokens / #sentences
2	Clauses per Sentence	#(SIMPX + R-SIMPX + P-SIMPX) tokens / #sentences
3	Subordinate Clauses per Sentence	#C / #sentences
4	Mean Clause Length	#(SIMPX + R-SIMPX + P-SIMPX) tokens / #(SIMPX + R-SIMPX + P-SIMPX)
5–6	Mean {Simplex | Relative} Clause Length	#{SIMPX | R-SIMPX} tokens / #{SIMPX | R-SIMPX}
7–9	{Simplex | Relative | Paratactic} Clauses Ratio	#{SIMPX | R-SIMPX | P-SIMPX} / #(SIMPX + R-SIMPX + P-SIMPX)
10–12	Mean {Prefield | Middle Field | Postfield} Length	#{VF | MF | NF} tokens / #{VF | MF | NF}
13–14	Mean {NP | PP} Length	#{NX | PX} tokens / #{NX | PX}
15–16	{Verbs | NPs} per Sentence	#{VXFIN + VXINF | NX} tokens / #sentences
17	Verb/Noun Ratio	#VV.* / #NN
18	Mean Token Embedding Depth	#nodes / #tokens
19	Mean Maximum Embedding Depth per Sentence	sum of maximum embedding depth per sentence / #sentences

Table 3: Syntactic complexity features and their definitions.

embedded clauses) within sentences, or length of specific constituents. In addition, we included measures that relate to topological fields, in particular the prefield ("Vorfeld", VF), the middle field ("Mittelfeld", MF), and postfield ("Nachfeld", NF) (cf. Telljohann et al., 2012). Similar features have been used in other studies for automatically evaluating syntactic complexity (Chen and Zechner, 2011; Meyer et al., 2020).

Table 3 shows all of our features along with their definitions.[7] Mean length of constituents is calculated as follows: First, all tokens within the relevant constituents are counted by counting all nodes pertaining to the constituent (i.e., singletons and BIE nodes). Next, this sum is normalized by the total number of relevant constituents, which is calculated by counting the number of nodes marking the beginning of the constituent (singletons and B nodes). For instance, mean length of SIMPX is calculated as shown in (1). In Table 3, we use the simplified notation "#SIMPX tokens / #SIMPX" for the formula in (1).

(1) Mean length of SIMPX

$$= \frac{\#\text{SIMPX}+\#\text{B-SIMPX}+\#\text{I-SIMPX}+\#\text{E-SIMPX}}{\#\text{SIMPX}+\#\text{B-SIMPX}}$$

Features 1–3 concern the complexity of sentences, measured in number of tokens, clauses, and subordinate clauses.[8]

Features 4–9 concern the complexity of clauses in general and specific clause types. Features 7–9 record the proportions of different clause types. Unfortunately, the annotation scheme only distinguishes between relative clauses, paratactic (i.e., coordinated) clauses, and the rest, called simplex clauses. Simplex clauses cover a huge and heterogeneous class with verb-second main clauses as well as verb-final subordinate clauses.[9]

Features 10–12 and 13–14 measure the length of the topological fields and of NPs and PPs, respectively.

Features 15–17 concerns the number and ratio of verbs and nouns, which can indicate a more verbal (i.e., oral) style vs. a more nominal (i.e., written) style.

Features 18 and 19 concern the depth of embedding in general. Feature 18 calculates an overall mean embedding depth, considering all tokens in the text. The embedding depth is measured by the number of nodes which form the path from the root node to a token's terminal node. Topological field nodes do not contribute to the path length. Feature 19 considers only the maximum embedding depth per sentence, and calculates the mean over all sentences in a text.

Appendix B illustrates the syntactic annotation and the resulting complexity scores with an exam-

[7] "X" as part of a syntactic label stands roughly for "phrase"; e.g., "NX" corresponds to "NP". Syntactic nodes labeled "VXFIN" and "VXINF" dominate a finite or infinite verb (infinitives and participles), respectively (Feature 16). For the exact definitions of the syntactic labels, see Telljohann et al. (2012). "VV.*" and "NN" refer to POS tags (Feature 17).

[8] Virtually all subordinate clauses contain a node labeled "C", which hosts the subordinating conjunction in complementizer and adverbial clauses, the relative pronoun in relative clauses, and the interrogative pronoun in (embedded) interrogative clauses. An exception are embedded verb-second clauses, which do not contain a node C and are therefore not covered here.

[9] We do not include mean length of paratactic clauses because they connect two or more simplex clauses, whose length we include. Moreover paratactic clauses are very rare, as shown by Feature 9.

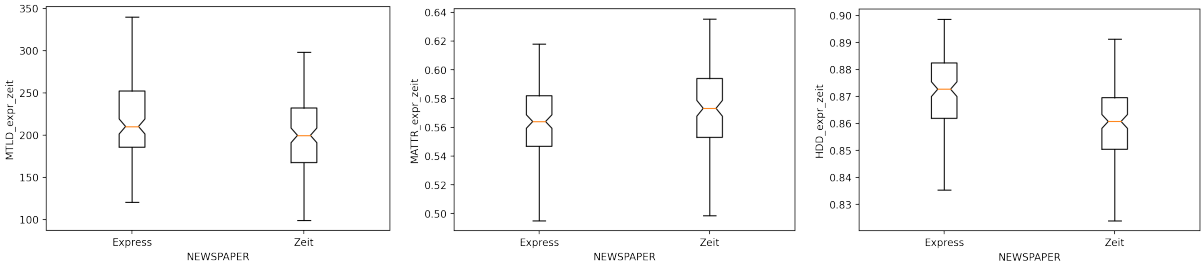


Figure 3: Boxplots of the scores according to *MTLD* (left), *MATTR* (center), and *HD-D* (right) for the EXPRESS and ZEIT corpora (left vs. right box, respectively).

ple sentence from the GraphVar corpus.

Our basic assumption is that a higher number of clauses and a greater length of clauses is an indicator of a higher syntactic complexity.[10] Expectations concerning the topological fields are less straightforward. A complex middle field is often considered a feature of the written register. In contrast, a complex postfield typically results from postponing complex constituents from the middle field and, hence, can possibly be considered a characteristics of the oral register and less complex. Regarding length and embedding depth of constituents, higher scores also imply higher complexity.

4 Results

4.1 Lexical Diversity

4.1.1 Reference Corpora

For the two reference corpora, we assumed that the ZEIT corpus should result in higher scores of lexical diversity than EXPRESS corpus. To validate this assumption, we lemmatized the reference subsets with the TreeTagger (Schmid, 1994) and determined MTLD, MATTR, and HD-D scores for both subsets.

The results vary, as shown in Fig. 3: Contrary to our assumption, the EXPRESS corpus achieves slightly higher MTLD and HD-D scores than the ZEIT corpus, i.e., it is lexically more diverse than the ZEIT corpus according to these scores (the difference is not significant with MTLD, though). Only with MATTR the ZEIT corpus achieves the higher scores (no significant difference, though).

Perhaps this unexpected result can be attributed to the way the subcorpora were sampled, see our considerations in Sec. 5.

[10] However, as mentioned above, a nominal style (i.e., using nominalizations instead of clauses) is also an indication of high complexity (see Features 15–17).

4.1.2 GraphVar Corpus

With regard to the GraphVar corpus, we assumed that due to the changing composition of the students (i) the results from the early years would be more homogeneous and have less variance, and (ii) the lexical diversity of texts written in the 1960s and 1970s would be rather high and would gradually decrease when progressing in time.

However, the results from the lexical diversity study do not confirm our hypothesis. We calculated the measures for each text separately, and computed mean and standard deviation per year.

We start with the second hypothesis. All three measures show an increasing trend over time, see Fig. 4. This is especially clear with MTLD and HD-D, so our hypothesis is clearly refuted. With regard to the first hypothesis, the boxplots in Fig. 4 show that variance is smallest in 2003–2013, again contrary to our expectations.

The texts from 1998 seem to be an interesting outlier: The mean is very clearly below the trend line, and there is also an unusually high variance this year.

Compared to the EXPRESS and ZEIT corpora, the GraphVar texts turn out lexically less diverse than both the EXPRESS and ZEIT texts, with all measures.[11] Presumably, this can be attributed to the different tasks: Essays written as part of the German exam deal with one predefined topic, e.g. a question on a novel that has to be answered and discussed, and therefore tend to use recurring vocabulary rather than newspaper texts, aimed at a broad public.

Regarding the first hypothesis, there seems to

[11] Means per subcorpus:

Measure	EXPRESS	ZEIT	GraphVar
MTLD	215.60	203.11	74.74
MATTR	0.56	0.57	0.41
HD-D	0.87	0.86	0.77

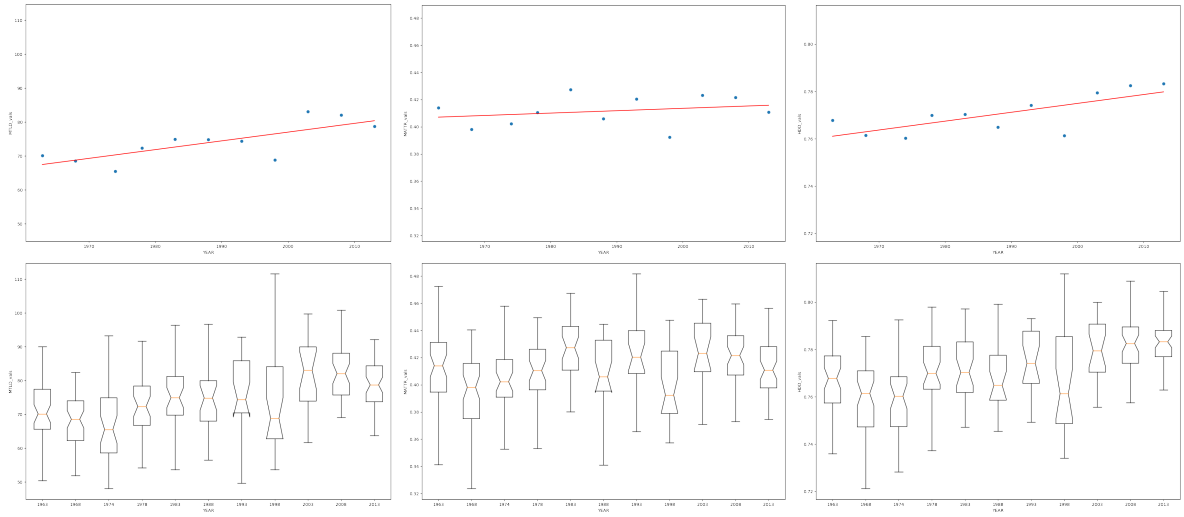


Figure 4: MTLD (left), MATTR (center), and HD-D (right) scores for the GraphVar corpus: means (top) and boxplots (bottom) per year.

be a trend toward less variance, i.e., toward more homogeneous texts, which again contradicts the hypothesis.

4.2 Perplexity of POS-based Language Models

As argued in Sec. 3.2, we assume that a POS-based language model trained on the ZEIT corpus should have a lower perplexity on a linguistically complex text than a POS-based language model trained on the EXPRESS corpus.

We tagged both reference corpora with the SoMeWeTa POS tagger (Proisl, 2018) with the model "german_newspaper_2020-05-28"[12] and trained two models on the POS tags of the ZEIT corpus and the EXPRESS corpus, respectively. We used the same tagger to re-tag the GraphVar corpus such that the annotation can be compared to the reference corpora.[13]

[12]The SoMeWeTa tagger comes with two pre-trained models: "german_newspaper_2020-05-28", which was trained on German newspaper texts, and "german_web_social_media_2020-05-28", which was trained on German web and social media data. In an informal evaluation, we compared these models and evaluated 50 randomly selected tokens from each of the three corpora (EXPRESS, ZEIT, GraphVar) where the models yielded different results. It turned out that the model "german_newspaper_2020-05-28" performed slightly better than the model "german_web_social_media_2020-05-28". In addition, we evaluated the model "german_newspaper_-2020-05-28" on 100 randomly selected tokens from each of the three corpora. The tagger achieved very good accuracies of 97% (ZEIT and GraphVar) and 96% (EXPRESS).

[13]We used the NORMAL forms of the GraphVar texts for tagging. These are normalized word forms with (corrected) modernized spellings.

Fig. 5 displays the result from the POS-based models trained on the ZEIT and EXPRESS corpora when applied to the GraphVar corpus. For each year, first the perplexity of the EXPRESS model is shown, followed by the one of the ZEIT model.

Overall, later years tend to yield higher perplexities, i.e., the syntactic distance between the GraphVar texts and the two newspapers models increases over time. This is remarkable because the newspaper models have been trained on data from 2021 and 2022, but perplexity is very low with the GraphVar data from the 1960s. Interestingly, however, the upward trend breaks off abruptly in 2008 (assuming that 1998 is again an outlier and that the upward trend continues to 2003).

Concerning the reference corpora, it is interesting to note that most of the time, the ZEIT-based perplexity is lower than the EXPRESS-based one, even though the differences are not significant (as indicated by the overlapping regions of the notches).

With regard to our first hypothesis, the boxplots show a relatively high variance for the entire period.

4.3 Syntactic Complexity

4.3.1 Reference Corpora

For the two reference corpora, we assumed that the ZEIT corpus should have a higher syntactic complexity than the EXPRESS corpus. For the comparison, we parsed the subsets of the reference corpora with the Berkeley Parser (Petrov and Klein, 2007), using a model for German that provides

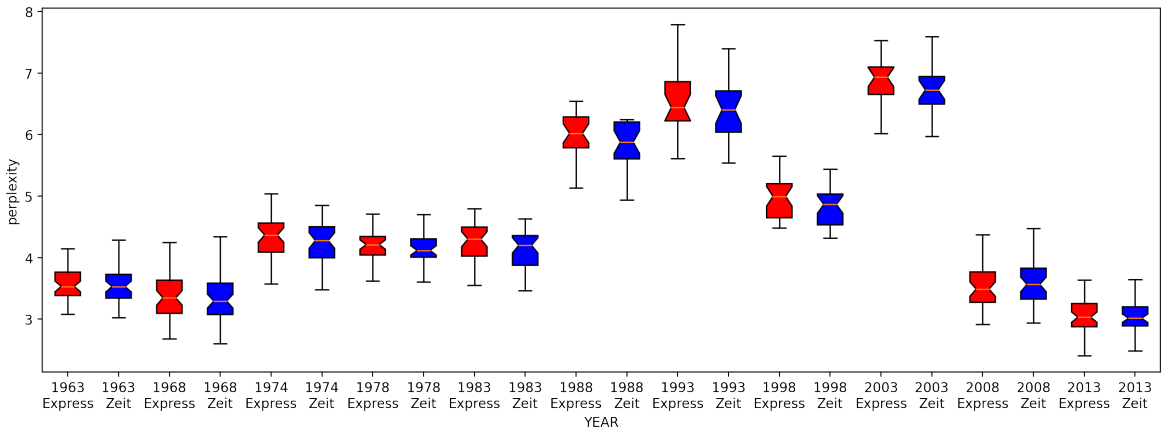


Figure 5: Mean perplexity per year using the EXPRESS and ZEIT models

syntactic as well as topological field annotations (Cheung and Penn, 2009).[14]

Table 4 lists the different measures and scores of the subsets (columns "EXPRESS" and "ZEIT"). As the table shows, ZEIT texts tend to have higher scores (with 12 out of 19 measures, column "E/Z"), although the scores are often close to each other. With Features 3, 6 and 12, the differences are more pronounced. At least for Features 3 and 6, a higher score clearly indicates higher complexity.

We conclude that the ZEIT texts are generally more syntactically complex than the EXPRESS texts, so that our assumption is confirmed here.

4.3.2 GraphVar Corpus

Table 4 shows that the GraphVar corpus achieves higher scores than the reference corpora with most of the measures. In fact, there is often a very clear gap to the scores of the reference corpora, in particular for Features 1–5 and 18–19, which are all clearly related to syntactic complexity.

The final column "Trend" shows that the vast majority of the features tend to have lower scores in early years (1963–1978) and higher scores in late years (1983–2013), clearly contradicting our second hypothesis. These features are marked by "+" in Table 4.[15]

Texts written in 1998 represent a remarkable exception, again, showing low average scores for most of these features, see the plots in Fig. 7 in Appendix C.

Contrary to our initial hypothesis, these results suggest that the syntactic complexity of the GraphVar texts is higher in late years.

With regard to our first hypothesis, the tendencies are less clear and there is a relatively high variance for the entire period, as in the case of perplexity.

5 Conclusion

In this paper, we examined high school graduation texts over five decades (1963–2013). Our initial hypotheses were: (i) variance increases; (ii) complexity decreases. However, these hypotheses were not confirmed by our tests.

Lexical diversity does not distinguish clearly between the two reference corpora EXPRESS and ZEIT. For the GraphVar corpus, diversity increases over time according to all three measures, but variance seems to decrease. The results by perplexity show a growing distance to both reference corpora, with an abrupt break in the year 2008. Variance is rather high for the entire period. There is no real difference in perplexity between the two reference models. According to the syntactic measures, the GraphVar texts are clearly more complex than both of the reference corpora, and the ZEIT texts are slightly more complex than the EXPRESS texts. The GraphVar corpus shows an increase in syntactic complexity over time with most features. Again, variance is rather high for the entire period. In summary, GraphVar texts are becoming more complex over time.

With regard to the reference corpora, we could hypothesize that the unexpected results could be

[14] We downloaded the parser and the model "tuebadz_topf_no_edge.gr" from https://www.cs.mcgill.ca/~jcheung/topoparsing/topoparsing.html.

[15] We fit linear models for each of the features, with the year as the predictor and the score as the dependent variable (in R: `lm(formula = score ~ year)`). If the year has a highly significant effect ($p < 0.001$), the feature is marked as "+" in Table 4. A (weak) significant effect ($p < 0.05$) is recorded as "(+)" in the table.

No	Feature	E/Z	EXPRESS	ZEIT	GraphVar	Trend
1	Mean Sentence Length	E	*17.59*	17.57	**21.30**	+
2	Clauses per Sentence	Z	1.90	*1.96*	**2.21**	ns
3	Subordinate Clauses per Sentence	Z	0.40	*0.51*	**0.73**	(+)
4	Mean Clause Length	Z	13.03	*13.30*	**14.63**	+
5	Mean Simplex Clause Length	Z	13.34	*13.61*	**15.19**	+
6	Mean Relative Clause Length	Z	9.04	**10.05**	*9.42*	+
7	Simplex Clauses Ratio	E	**0.92**	*0.90*	0.88	ns
8	Relative Clauses Ratio	Z	0.07	*0.09*	**0.11**	ns
9	Paratactic Clauses Ratio	Z	0.00	0.00	**0.01**	ns
10	Mean Prefield Length	E	**3.64**	3.35	*3.46*	+
11	Mean Middle Field Length	E	*5.14*	5.02	**5.30**	+
12	Mean Postfield Length	Z	9.48	*10.36*	**10.98**	+
13	Mean NP Length	Z	2.46	*2.55*	**2.57**	+
14	Mean PP Length	Z	3.57	*3.72*	**3.82**	+
15	Verbs per Sentence	E	*2.55*	2.53	**2.97**	ns
16	NPs per Sentence	E	*6.96*	6.84	**7.62**	+
17	Verb/Noun Ratio	Z	0.49	*0.51*	**0.52**	ns
18	Mean Token Embedding Depth	E	3.18	*3.27*	**4.13**	+
19	Mean Maximum Embedding Depth per Sentence	Z	*4.62*	4.59	**5.95**	+

Table 4: Results of syntactic complexity measures. Column "E/Z" marks which of the reference corpora achieves the higher score for the respective feature. Columns "EXPRESS", "ZEIT" and "GraphVar" list the average scores of each subcorpus. For each feature, the highest score is in bold, the second highest in italics. The column "Trend" shows the GraphVar trend over the survey years: "+" means that late years show significantly higher scores than early years. The feature marked by "(+)" still shows similar tendencies but the difference is less pronounced. "ns" marks features that do not show clear trends between the scores of the different years of the GraphVar corpus.

Corpus		#Articles	Avg. #Tokens
EXPRESS	complete	30K	295
	filtered	4.6K	740
ZEIT	complete	7.5K	1,094
	filtered	2K	1,670

Table 5: The two reference corpora, complete and filtered.

due to the way the text fragments were sampled. Only articles that were at least 500 tokens long were considered. This excludes a large number of articles, especially in the EXPRESS corpus: out of almost 30,000 articles, only 4,565 remain. The average length of an EXPRESS article before this filtering is 295 tokens, after the filtering 740 (see Table 5). That is, it could be that the filtering sorts out the "typical", linguistically simple EXPRESS articles and the more unusual, more complex articles remain. In contrast, the filter effect with the ZEIT corpus is much smaller.

This could explain why the EXPRESS corpus is lexically more diverse than ZEIT according to MTLD and HD-D, and could also be a reason why the EXPRESS corpus gets quite similar scores as the ZEIT corpus with many syntactic features.

Concerning the GraphVar corpus, we have observed two striking anomalies. First, texts from 1998 stood out as outliers in all studies. Second, perplexity results indicate a major break in 2008. Maybe these anomalies can be explained by some external factor such as an important change in the task.[16]

In general, increasing complexity of GraphVar texts could be traced back to different reasons, all of which require further investigation: Teaching methods could have improved and students are achieving better results in later years. The type of task might have changed more than expected over the years and therefore the results differ. We leave this question open for future research.

Acknowledgments

We would like to thank the reviewers for their constructive and valuable feedback. Many thanks also to Kristian Berg (Bonn), who provided us with the GraphVar corpus and answered numerous questions about it.

[16]The anomalies cannot be due to the spelling reform from 1996: The lexical measures are based on normalized lemma forms, which are not affected by the reform. The perplexity and syntactic measures refer to abstract syntactic categories.

References

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2014. KoKo: an L1 learner corpus for German. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2414–2421, Reykjavik, Iceland.

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2016. An extended version of the KoKo German L1 learner corpus. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLIC-it 2016)*, pages 13–18, Naples, Italy.

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(S1):180–208.

Kristian Berg, Jonas Romstadt, and Cedrek Neitzert. 2021. GraphVar – Korpusaufbau und Annotation. Version 1.0. Friedrich-Wilhelms-Universität Bonn, https://graphvar.uni-bonn.de/dokumentation.html.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 722–731, Portland, Oregon, USA. Association for Computational Linguistics.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 113–119, Osaka, Japan. The COLING 2016 Organizing Committee.

Jackie Chi Kit Cheung and Gerald Penn. 2009. Topological field parsing of German. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 64–72, Suntec, Singapore. Association for Computational Linguistics.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

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

Gerasimos Fergadiotis, Heather Wright, and Thomas West. 2013. Measuring lexical diversity in narrative discourse of people with aphasia. *American journal of speech-language pathology / American Speech-Language-Hearing Association*, 22:397–408.

Daniel Jurafsky and James H. Martin. 2022. *Speech and Language Processing*. Draft from Jan 12, 2022.

Rie Koizumi and Yo In'nami. 2012. Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40(4):554–564.

Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3):307–322.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Duran. 2004. *Lexical diversity and language development: Quantification and assessment. Basingstoke, Hampshire: Palgrave Macmillan*. Cited in McCarthy and Jarvis (2010).

Philip M. McCarthy. 2005. *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M. McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

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

Jennifer Meyer, Torben Jansen, Johanna Fleckenstein, Stefan Keller, and Olaf Köller. 2020. Machine Learning im Bildungskonstext: Evidenz für die Genauigkeit der automatisierten Beurteilung von Essays im Fach Englisch. *Zeitschrift für Pädagogische Psychologie*, 0:1–12.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York. Association for Computational Linguistics.

Thomas Proisl. 2018. SoMeWeTa: A part-of-speech tagger for German social media and web texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 665–670, Miyazaki. European Language Resources Association ELRA.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technical report, Universitäten Stuttgart und Tübingen, http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Anna Shadrova, Pia Linscheid, Julia Lukassek, Anke Lüdeling, and Sarah Schneider. 2021. A challenge for contrastive L1/L2 corpus studies: Large inter- and intra-individual variation across morphological, but not global syntactic categories in task-based corpus data of a homogeneous L1 German group. *Frontiers in Psychology, Section Language Sciences*, 12.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Katrhin Beck. 2012. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen.

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.

Fred Zenker and Kristopher Kyle. 2021. Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47:100505.

A Text Samples

GraphVar corpus (1963)
Franz Werfel setzt über das Gedicht einen lateinischen Spruch, der übersetzt heißt: Komm Schöpfer Geist. So gemahnt dies Gedicht an einen liturgischen Hymnus. In den Versen und mit Endreimen erhält das Gedicht eine andere Form als ein mittelalterlicher Hymnus. Der Dichter hat wohl diese Überschrift gewählt, um den Menschen heute, die auf der Suche nach einem Weltbild sind, die Geschlossenheit des mittelalterlichen Weltbildes zu zeigen, damit sie aus diesem lernen. Rainer Maria Rilke setzt keine Überschrift über das Gedicht. Er gibt keinen Fingerzeig, sondern stellt uns so vor das Gedicht, das kein Versmaß hat, sondern unregelmäßige Langzeilen mit Endreimen. B I Gott kommt zu den Menschen nur durch schöpferische Tätigkeit. Der Mensch muss sich Gott wie ein großes Kunstwerk erst erarbeiten. Er muss um Gott kreisen, "um den alten Turm". Hat der Mensch ihn gefunden, dann kommt er "mit ihm" - mit Gott - "aus der Nacht." Gott führt ihn aus dem Chaos zum Licht.

GraphVar corpus (2013)
In der damaligen Ständegesellschaft waren ständeübergreifende Beziehungen sehr problematisch. Mit einer solchen Beziehung zwischen einer Bürgerlichen und einem Adligen beschäftigt sich auch Theodor Fontane in dem Auszug aus seinem Roman "Irrungen und Wirrungen", erschienen im Jahre 1887. In dem Textauszug aus dem fünften Kapitel findet ein Dialog zwischen der Bürgerlichen Lene und ihrem adeligen Geliebten Botho statt, in welchem die Aussichtslosigkeit der Liebe der beiden aufgrund der Ständegesellschaft thematisiert wird. Das Paar trifft sich bei Nacht in einem Garten zum Spaziergang. Sie unterhalten sich zunächst über die Mutter von Botho, wobei Lene ihre Furcht vor dieser Person äußert. Botho ist der Ansicht, dass sie seine Mutter falsch einschätzt, woraufhin Lene ihre Bedenken bezüglich ihrer Liebe und ihrer Beziehung anspricht.

EXPRESS
Heftiger Regen. Und das fast den ganzen Tag. Zig Straßen sind überflutet, Hunderte Keller sind vollgelaufen, Menschen müssen raus aus ihren Wohnungen, es gibt Vermisste. Tief “Bernd” setzt fast ganz Deutschland mächtig zu. Besonders hart hat der Starkregen Nordrhein-Westfalen getroffen. In Hagen musste ein Altenheim evakuiert werden, weil Wassermassen einströmten. Es ist unbewohnbar geworden. Eltern wurden gebeten, ihre Kinder nicht in die Kita zu schicken. Eine verschüttete Person wurde leicht verletzt gerettet worden. Mehrere Fahrer mussten aus ihren von Wassermassen eingeschlossenen Autos befreit werden. Es gab mindestens 200 Einsatzorte. Einige Ortsteile waren zum Teil nicht mehr zu erreichen. “Die Leute sind verzweifelt”, sagte ein Sprecher des Polizeipräsidiums Hagen. Bundeswehrpanzer sollen helfen, die Straßen wieder frei zu machen.

ZEIT
Der zerbrochene Krug, der chaotische Schreibtisch oder die Fahrt nach Rimini mit einem Diesel verbrennenden alten Opel - das alles sind Anwendungsfälle des Zweiten Hauptsatzes der Thermodynamik. Der besagt in aller Kürze, dass jedes System den Zustand höchster Unordnung anstrebt - solange niemand Extraenergie reinsteckt. Dieses »Extraenergiereinstecken« aber ist die vornehmste Aufgabe der Politik. Ein hervorragendes Beispiel dafür ist die Mülltrennung. Früher (bis in die Sechzigerjahre) gab es für den gesamten Müll eine einzige große Tonne : für Zeitungen und faule Äpfel, für leere Flaschen und Konservendosen, für alte Batterien, löchrige Socken und Asche aus dem Kohleofen. Manchmal war die noch heiß, dann fing der Mülleimer an zu qualmen. In dieser (guten) alten Zeit - in Teilen der USA ist das heute noch so - war die einzige ernst zu nehmende Frage: Wer bringt den Müll runter?

B Syntactic Complexity: An Example

We illustrate the Syntactic Complexity measures with an example sentence from the GraphVar corpus, shown in (i).

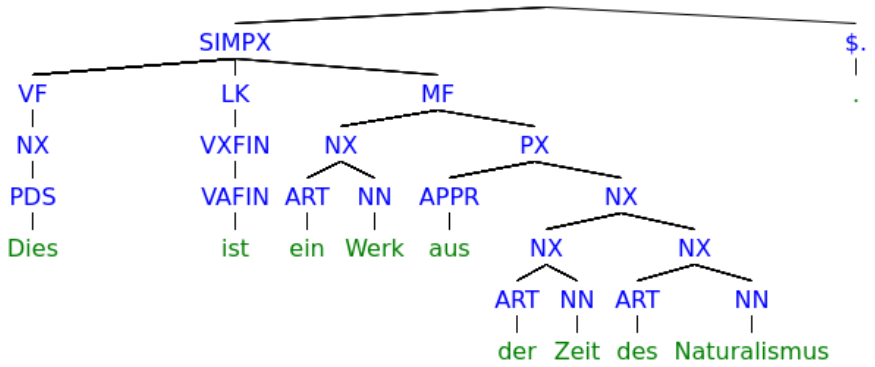
(i) *Dies ist ein Werk aus der Zeit des Naturalismus.*
'This is a work from the period of naturalism.'

Fig. 6 displays the syntactic analysis produced by the Berkeley parser (Petrov and Klein, 2007), using the model "tuebadz_topf_no_edge.gr" (Cheung and Penn, 2009).[17] It further shows the corresponding BIE path notation and presents the results for the individual syntactic complexity measures.

C Syntactic Complexity: Results

Fig. 7 shows the means and boxplots per survey year for all syntactic features. The numbers refer to the numbered features listed in Table 4 in Sec. 4.3.

[17] The tree view has been produced by the Syntax Tree Generator, http://mshang.ca/syntree/.



Word	Lemma	POS	Syntax
Dies	dies	PDS	B-SIMPX | VF | NX | PDS
ist	sein	VAFIN	I-SIMPX | LF | VXFIN | VAFIN
ein	eine	ART	I-SIMPX | B-MF | B-NX | ART
Werk	Werk	NN	I-SIMPX | I-MF | E-NX | NN
aus	aus	APPR	I-SIMPX | I-MF | B-PX | APPR
der	die	ART	I-SIMPX | I-MF | I-PX | B-NX | B-NX | ART
Zeit	Zeit	NN	I-SIMPX | I-MF | I-PX | I-NX | E-NX | NN
des	die	ART	I-SIMPX | I-MF | I-PX | I-NX | B-NX | ART
Naturalismus	Naturalismus	NN	E-SIMPX | E-MF | E-PX | E-NX | E-NX | NN
.	.	$.	$.

No	Feature	Score
1	Mean Sentence Length	10
2	Clauses per Sentence	1
3	Subordinate Clauses per Sentence	0
4	Mean Clause Length	9.0
5	Mean Simplex Clause Length	9.0
6	Mean Relative Clause Length	–
7	Simplex Clauses Ratio	1
8	Relative Clauses Ratio	0
9	Paratactic Clauses Ratio	0
10	Mean Prefield Length	1.0
11	Mean Middle Field Length	7.0
12	Mean Postfield Length	–
13	Mean NP Length	2.2
14	Mean PP Length	5.0
15	Verbs per Sentence	1
16	NPs per Sentence	5
17	Verb/Noun Ratio	0
18	Mean Token Embedding Depth	3.6
19	Mean Maximum Embedding Depth	5

Figure 6: Syntactic analysis of the example sentence. The tree (top) shows the output of the parser. The first table (center) shows the corresponding path notation using BIE tags in the column “Syntax”; the last node of each path consists of the POS tag. The second table (bottom) lists the scores of the syntactic complexity measures that result for the example sentence; note that Features 18 and 19 do not consider the topological nodes (VF, LK, MF in the example)

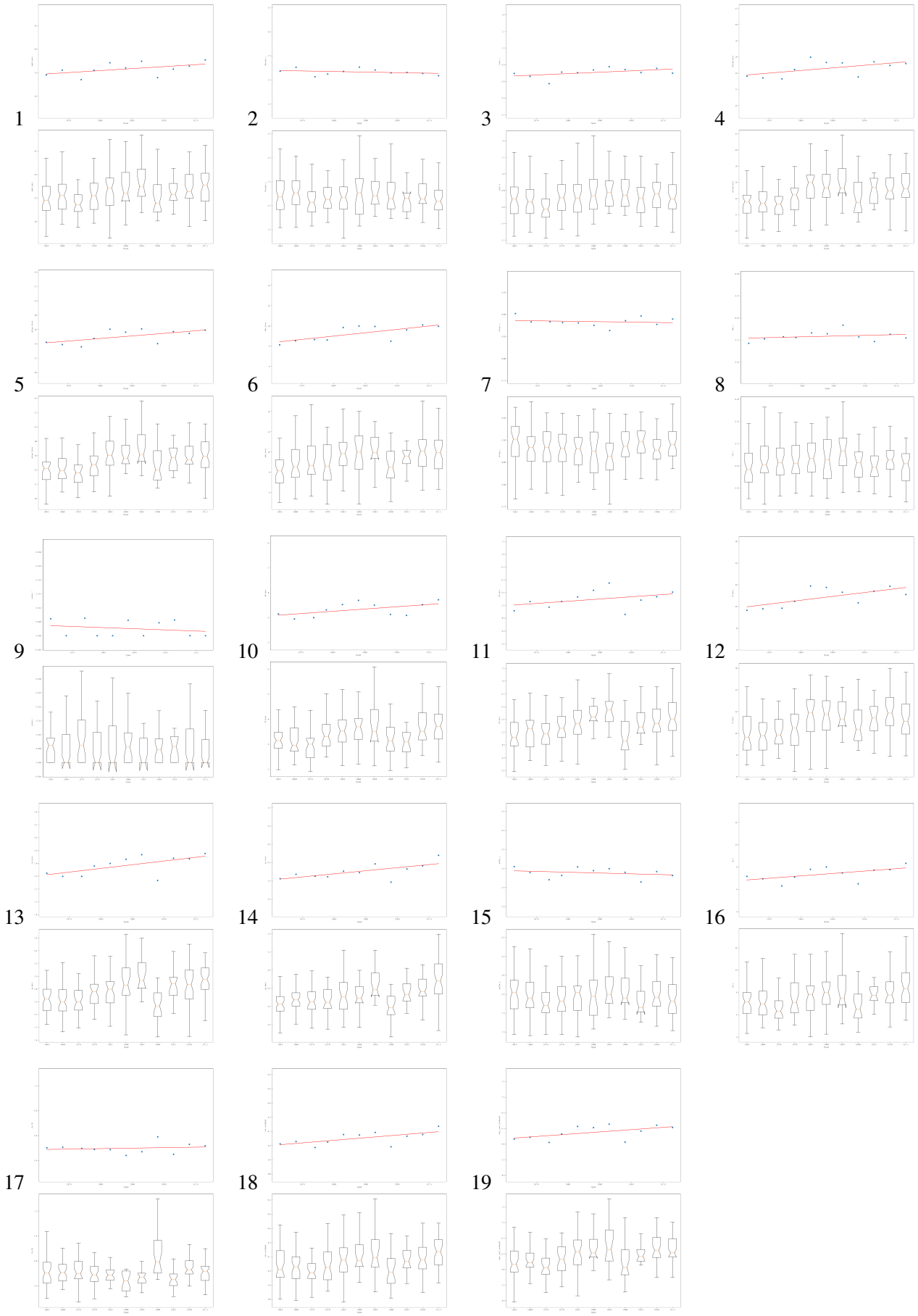


Figure 7: Syntactic features: mean and boxplot per survey year.