Complexity of German Texts Written by Primary School Children

Jammila Laâguidi, Dana Neumann,
Ronja Laarmann-Quante, Stefanie Dipper and Mihail Chifligarov
Ruhr University Bochum, Faculty of Philology, Department of Linguistics, Germany
Dana.Neumann@edu.ruhr-uni-bochum.de
{Jammila.Laaguidi,Ronja.Laarmann-Quante,Stefanie.Dipper,Mihail.Chifligarov}
@ruhr-uni-bochum.de

Abstract

While the development of children's literacy is of large interest to researchers, few studies have yet been based on corpora of children's texts. We investigate the development of text complexity in freely-written texts of German primary school children between 2nd and 4th grade based on the longitudinal Litkey Corpus (Laarmann-Quante et al., 2019b) using NLP methods. These texts are retellings of given picture stories. Although the picture stories may constrain the vocabulary and grammar, our hypothesis is that complexity increases over time. We measure complexity using various lexical and syntactic features. The results show that our hypotheses are largely confirmed but that there are outliers that might arise because some picture stories could be more stimulating than others.

1 Introduction

An important goal of primary school education is the acquisition of written language skills. In addition to the teaching of spelling, this also includes the acquisition of a sufficiently extensive vocabulary and an arsenal of sufficiently complex syntactic constructions.

Studies of how children's (written) language abilities develop have typically been either cross-sectional or based on the development of only few children. The reason is that not many large corpora of children's text productions are available, especially longitudinal ones. One exception for German is the Litkey Corpus (Laarmann-Quante et al., 2019b), which contains texts collected from the same 251 children at 10 test points between the 2nd and 4th grade (see Section 2 for details).

The goal of this paper is to investigate whether and to what extent the complexity of vocabulary and syntax increases in the course of primary school, as reflected by the texts collected in the Litkey Corpus. Since such large corpora cannot be analyzed by hand, we apply Natural Language Processing (NLP) methods for automatic processing in this investigation.

One particularity of the corpus is that the texts are based on picture stories. This means that the vocabulary and potentially also particular syntactic constructions are to some extend bound by the picture stories. We hypothesize that with increasing written language skills over time, one can nevertheless measure an increase in linguistic complexity in the texts.

There are yet few studies that analyze children's retellings of picture stories and the ones that are available focus on oral rather than written retellings. For example, Rahayu et al. (2020) analyzed the retellings of children aged six to nine and found that their lexical diversity increases with age. Heilmann et al. (2010) analyzed the narrative macrostructure of children aged five to seven and found that narrative (macrostructure) skills are correlated with their vocabulary, grammar, and productivity skills. Bulut-Ozsezer and Canbazoglu (2018) examined the comments that seven-year-old children made on the pictures in story books and divided them into different categories (description, superficial and imaginative interpretation, and critical understanding), concluding that most of them are descriptions.

The Litkey Corpus provides the opportunity to study written retellings of picture stories. So far, the corpus has mainly been analyzed with regard to its general composition (e.g. Laarmann-Quante et al., 2019b; Laarmann-Quante et al., 2019a) and research based on the corpus has focused on spelling errors (Röhrig, 2020; Laarmann-Quante, 2021). The Litkey Corpus has not yet been used to analyze children's development concerning their lexical and syntactic complexity. This paper intends to close this gap.

To measure vocabulary complexity, we use different standardized measures for lexical diversity

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 32–39
September 10-13, 2024 ©2024 Association for Computational Linguistics

and additionally apply a new IDF-based measure. To measure syntactic complexity, we compare the distribution of part-of-speech (POS) n-grams and compute perplexity on POS n-grams based on a language model trained on a children's lexicon written by adults. We hypothesize that over time, the perplexity decreases as the children's syntax gets more similar to the one used by adults.

The main contributions of this paper are:

- A corpus-based study of the complexity of texts written by children and its development during primary school
- IDF-LDist, a new IDF-based measure of lexical distinctiveness

2 Data

This section describes the Litkey Corpus, which contains the texts produced by primary school children that we analyze, and the Klexikon Corpus, which we use as a reference corpus of texts to compare the Litkey texts with.

2.1 Litkey Corpus

The texts of the Litkey Corpus (Laarmann-Quante et al., 2019b) were collected by Frieg between 2010–2012 (Frieg, 2014). The texts were produced by 251 children in primary schools in Northrhine-Westfalia between the second half of the 2nd grade and the end of the 4th grade, i.e. the end of primary school in Germany. In total, there are 1,922 individual texts. Over the course of 10 different test points in time, children were advised to write stories retelling given picture stories.

At each test point, a different picture story was used except for test points TP02, TP06 and TP10 (i.e., at the end of each grade), where the same story was used. At testing time, it was first made sure that the children understood the basic storyline of the pictures before they wrote a story retelling the picture story. All stories feature two children, Lea and Lars, and a dog, Dodo.

The length of the texts varies greatly (Laarmann-Quante et al., 2019b): At the first test point TP01, the texts are on average 65.9 tokens long (SD 20.3), at the last test point TP10 the average is 139.2 tokens (SD 53.5).

All texts come with an orthographic target hypothesis, i.e., a normalized version of the text where each word is corrected for spelling errors but not grammatical errors. In the present study, we use this orthographic target hypothesis. Among further annotations, the corpus comes with STTS POS tags (Schiller et al., 1999) that were created automatically using a tagger trained on children's texts, yielding an accuracy of about 93% (see Laarmann-Quante et al., 2019a, for further details).

2.2 Klexikon Corpus

Klexikon[1] is a German online lexicon similar to Wikipedia, but targeted at children. It offers simplified and summarized articles about various topics and has been written by adults. This means the texts contain standard language sentence structures without grammatical errors but at the same time the use of simplified language makes them comparable to children's writing styles. This makes the Klexikon articles a suitable dataset that children's texts can be compared with at the syntactic level.

We use the Klexikon Corpus compiled by Ortmann and Wedig (2024) as part of the KidRef Corpus, which is a collection of various German texts written by or written for children. The Klexikon subcorpus consists of 924 texts with 300,000 tokens in total. Ortmann and Wedig (2024) automatically created STTS POS tags with an accuracy of about 94%, which we use in our study.

3 Methods

In order to study the development of text complexity in the primary school children's texts, we apply different methods measuring lexical diversity (Section 3.1) and syntactic complexity (Section 3.2). Our choice of methods largely follows Kapusta et al. (2022), who assessed the development of the complexity of German Abitur texts, i.e., texts that are part of the final secondary-school examinations, between 1963 and 2013.

3.1 Lexical Diversity

A popular measure of lexical diversity is type-token ratio (TTR), which is calculated by dividing vocabulary size by text length. However, this measure is sensitive to text length since the longer a text is, the higher the probability that the following word has already occurred (see, e.g., Covington and McFall, 2010). Since the texts in the Litkey Corpus vary in length, we use variations of

[1] https://klexikon.zum.de

TTR that are independent of text length: MATTR and HD-D.[2]

Before applying these measures, we lemmatize the texts[3] and exclude tokens that contain non-alphabetic characters. We deliberately refrain from excluding function words because the acquisition of different kinds of function words constitutes important steps in the development of literacy, e.g., using anaphoric expressions like personal pronouns rather than repeating proper names.

MATTR Covington and McFall (2010) propose MATTR ("Moving Average Type-Token Ratio"). It is calculated by first choosing a window size W (e.g. 500 tokens) and then computing the TTR for each moving window: words 1 to 500, then 2 to 501, then 3 to 503, and so on until the end of the text. After that, the mean of all calculated TTRs is the MATTR of the entire text. The higher the MATTR, the higher a text's lexical diversity. Covington and McFall (2010) suggest a window size W that is smaller than the shortest text in the data, in our case 16 words. Hence, we set W to 15.

HD-D McCarthy and Jarvis (2007, 2010) propose HD-D ("Hypergeometric Distribution D"). HD-D is based on the probability of finding a type at least once in a random sample of N words, which can be estimated with the hypergeometric distribution function. The probability of occurrence is calculated for all types in a text and then summed up to make up the HD-D index of that text. McCarthy and Jarvis (2007) propose a sample size of $N = 42$, however, multiple texts in the Litkey Corpus have less than 42 words, with the shortest text having 16 words only. Therefore, we decided for a sample size of 15.

IDF-LDist In addition to the two TTR variants, we define a custom measure, IDF-LDist ("IDF-based lexical distinctiveness"), to analyze whether all children use roughly the same vocabulary to describe a picture story or to what extend a child uses distinctive words that are not used by many others.

For each child/text, we first calculate the IDF values of their word types w per test point as shown in (1):[4]

$$\text{IDF}(w) = \frac{D}{df_w} \tag{1}$$

where D is the total number of texts at that test point and df_w is the number of texts containing w.

We next look at all IDF values of one child and determine how many of them lie above the average value for this test point (across children), which would show to what extent the child uses more distinctive words than children use on average. We calculate the average IDF value of a test point t as in (2), where V_t is the set of words at test point t:

$$\text{IDF}_{avg}(t) = \frac{1}{|V_t|} \sum_{w \in V_t} \text{IDF}(w) \tag{2}$$

Finally, we calculate for each child the percentage of IDF values above the test point's average, as a measure of how different the vocabulary of this child is compared to the other children, as shown in (3):

$$\begin{gathered}\text{IDF-LDist}(c, t) = \\ \frac{1}{|V_{c,t}|} \sum_{w \in V_{c,t}} \mathbb{1}\{\text{IDF}(w) > \text{IDF}_{avg}(t)\}\end{gathered} \tag{3}$$

where $V_{c,t}$ is the set of words of child c at test point t. The notation $\mathbb{1}\{x\}$ means "1 if x is true, and 0 otherwise" (Jurafsky and Martin, 2024, p. 178).

The IDF-LDist measure has the following properties: If all children used the same words, the IDF-LDist score for all children would be 0. Likewise, if all children used different words, the score for all children would also be 0 but this is not realistic since at least some function words and important words in a story, e.g. the names *Lea*, *Lars*, and *Dodo* will be shared by most texts. The IDF-LDist score of a specific child is high when most other children share the same vocabulary but this child uses different words.

3.2 Syntactic Complexity

To estimate syntactic complexity, measures are typically used that measure the complexity of constituents (e.g. embedding depth) or the length of certain constituents (cf., e.g., Chen and Meurers, 2016). However, this presupposes that a syntactic

[2] Another commonly-used length-independent measure is MTLD ("Measure of Textual Lexical Diversity", McCarthy and Jarvis, 2010). However, this measure only provides reliable values for texts with at least 100 words.

[3] Most of the tokens in the Litkey Corpus come with lemma information. We added missing lemmas using simplemma (Barbaresi, 2024).

[4] Since each child contributed at most one text to each test point, the terms "child" and "text" can be used interchangeably here.

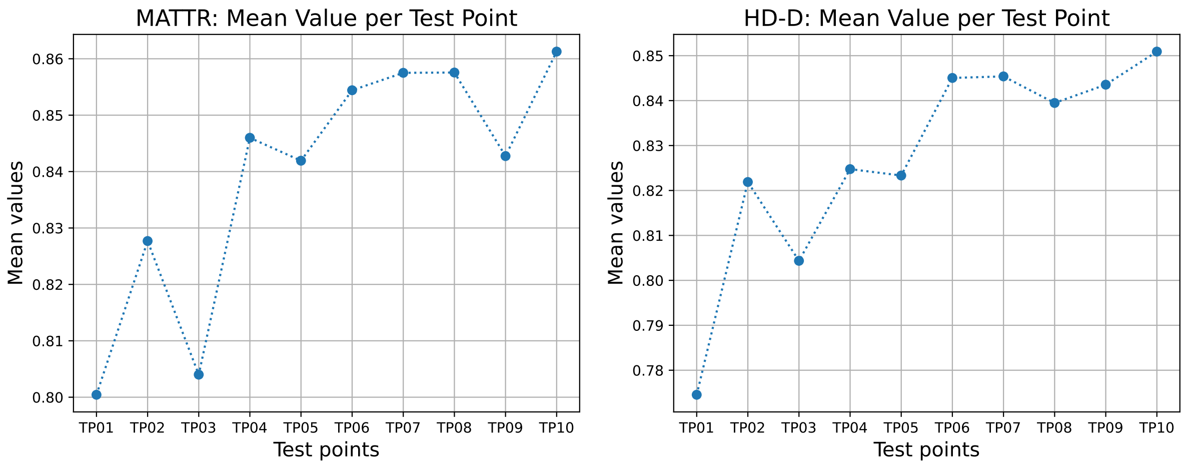

Figure 1: Development of the lexical diversity across test points measured as the mean values of MATTR (left) and HD-D (right).

annotation exists, e.g. in the form of phrase structure trees or dependency relations. However, the Litkey Corpus is not syntactically annotated except for the POS tags. Our syntactic measures are therefore based on POS tags.

Top POS n-grams We first look at the most frequent POS n-grams for each test point. This allows us to see whether the children use different constructions in different acquisition phases and which type of construction becomes more frequent with increasing literacy.

Perplexity In addition, we apply perplexity of POS-based language models. Perplexity is a standard metric in natural language processing (Jurafsky and Martin, 2024). It is usually used to assess the performance of language models, by comparing perplexity of two models on a test set. The model with the lower perplexity score fits the test data better.

In our study, we train a language model on the Klexikon corpus and investigate how the perplexity of this model changes over time when applied to texts from different test points, reflecting the evolving writing skills and practice of the children.

We hypothesize that the texts from the Litkey Corpus show decreasing perplexity over time as children's linguistic abilities improve with age and experience. This assumption is based on the premise that a language model trained on the Klexikon corpus, which shows no grammatical errors and contains more complex sentence structures, would yield higher perplexity scores when applied to texts written by elementary school children at the beginning of learning how to write, compared to the same children at the end of elementary school. To measure the syntactic complexity of the texts, we use a POS trigram language model with Kneser-Ney smoothing.[5]

4 Results

4.1 Lexical Diversity

We calculated both measures of lexical diversity, i.e. MATTR and HD-D, per text. Figure 1 shows the mean value at each test point (TP).

Both measures show that overall the lexical diversity increases over time, proving the initial hypothesis right that we can see an increase in spite of different picture stories used. However, the increase is rather small and not homogeneous. Both measures show similar patterns: There is a drop in each measure at TP03 and TP05 and another drop for HD-D at TP08 and for MATTR at TP09. It is likely that these drops are indeed caused by the different picture stories used in that some of them elicited a more diverse vocabulary than others. This assumption is supported by the observation that we see a clear upward trend between TP02, TP06 and TP10 where the same picture stories were used. The results emphasize the importance of taking into account the stimulus material with which texts are elicited when interpreting the results in a longitudinal study.

IDF-LDist The results for our new measure IDF-LDist are shown in Figure 3. For each test point, we see the distribution of the percentage of

[5] We used the NLTK module `nltk.lm` with default settings for calculating the model and perplexity.

Example 1 (IDF-LDist = 0.50):
Dodo ist verschwunden
An einem schönen warmen Sommertag ging Lea unten auf dem Bürgersteig hektisch umher. Sie sah ziemlich traurig aus. Sie klebte an jedem Baum, Haus, oder am einer Mauer Zettel auf.
'Dodo has disappeared. On a beautiful warm summer's day, Lea was walking frantically along the sidewalk below. She looked quite sad. She stuck notes on every tree, house, or on a wall.'

Example 2 (IDF-LDist = 0.07):
Lea sucht Dodo. sie klebt Bilder von Dodo.
'Lea is looking for Dodo. She sticks pictures of Dodo.'

Figure 2: Two (normalized) example texts from TP02 describing the same situation of a picture story: the dog Dodo has disappeared and the girl Lea hangs up 'missing dog' posters. Example 1 is the text with the top IDF-LDist score of TP02, Example 2 has a very low score.

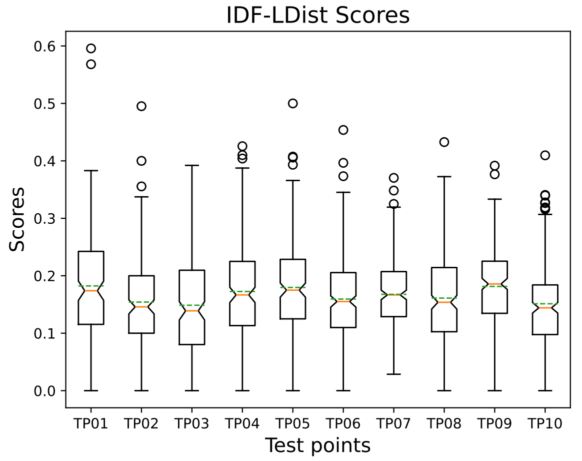

Figure 3: Distribution of the children's IDF-LDist scores per test point.

Rank	TP01		TP10	
1	NE	17.54	VVFIN	11.24
2	NN	12.55	NN	10.91
3	VVFIN	10.98	NE	10.71

Table 1: Top-frequent POS unigrams (percentages) at TP01 and TP10 (NE: proper nouns; NN: common nouns; VVFIN: finite verbs).

words above the test point's average IDF value.

Figure 2 shows two example texts, one with a very high IDF-LDist score and one with a very low score. The IDF-LDist score of a specific text becomes high when most other texts share the same vocabulary but this text uses different words. We see such outliers at almost each test point, most notably at TP01. At later test points, the variance and the outliers tend to decrease. This means that the distinctiveness of the children's vocabulary tends to become more homogeneous in that either the children all tend to use more similar words or – more likely given the increase in lexical variation reported above – all children tend to write in a more distinctive manner so that individual texts do not stick out anymore. One explanation could be that at early test points, some children start off with a broader or more different vocabulary than others, depending on their personal backgrounds. Then, the older the children become and the longer they have attended school, the more they reach a similar level of vocabulary. Hence, previous advantages some children might have had at the first test point are equalized to some extend. Nevertheless, this is only a rather subtle trend. Overall, we see that across all test points some individual differences remain.

Again, we must not forget a potential influence of the picture story. But when we compare TP02, TP06 and TP10 where the same story was used, we see a similar decrease in variance, especially between TP06 and TP10, as described above for all test points.

4.2 Syntactic Complexity

Top POS n-grams We start by comparing the two extremes, TP01 and TP10, see Table 1. The three most frequent POS tags are the same in both cases but appear in different order. It is noticeable that in TP01 NE, i.e. proper names, are by far the most frequent POS, with 17.54% of all tokens. It is obvious that the names of the two children and the dog occur disproportionately in the early texts.

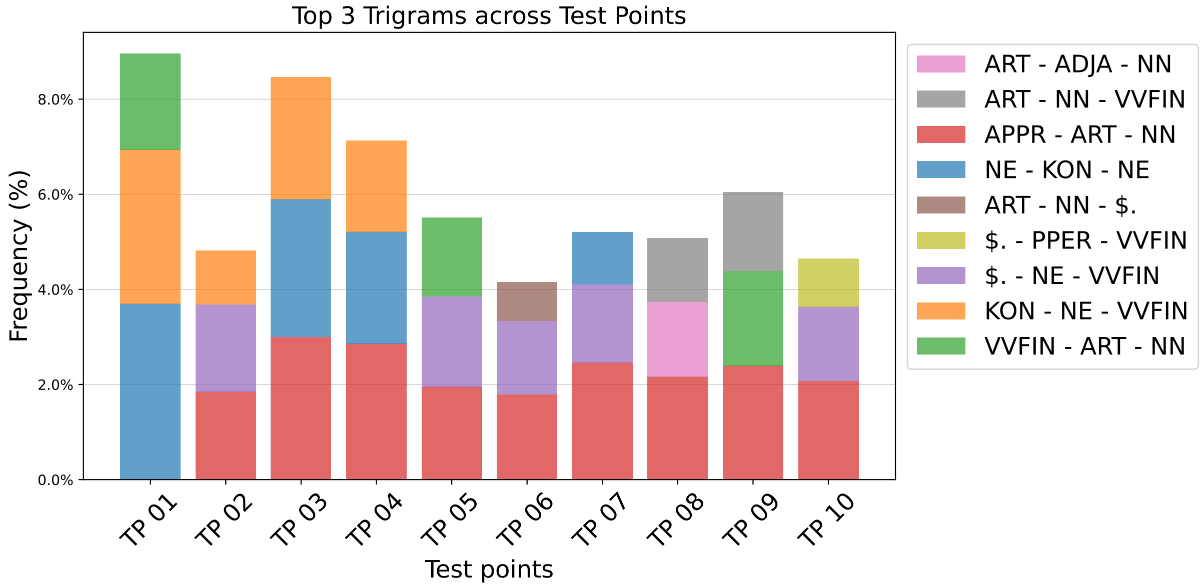

Figure 4: Top-frequent POS trigrams across all test points.

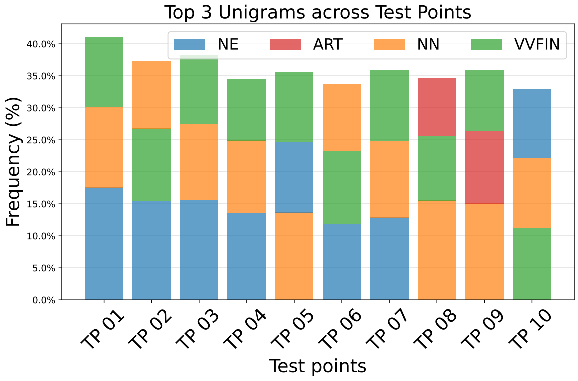

Figure 5: Top-frequent POS unigrams across all test points, stacked according to frequency.

Figure 5 plots the distribution of the top-frequent POS unigrams across the ten test points. The blue part of the bar plots corresponds to the proper nouns (NE). Proper names are almost always the most common POS up to TP07. In TP08, the article (ART, red part) appears as the first function word among the top three POS, and proper nouns become less important.

Figure 4 shows the top-frequent trigrams across all test points. TOP01 shows a special pattern here: the combination NE–KON–NE (blue part; KON for conjunction) is the most common, followed by KON–NE–VVFIN (orange) and VVFIN–ART–NN (green). These three patterns are typical for sentences in which the two children and/or the dog appear as the subject, as in Example (4). Parts of these patterns also show up in TP02–TP04.

(4) *Lars* *und* *Lea* *kaufen* *ein* *Eis.*
NE KON NE VVFIN ART NN
'Lars and Lea buy an ice cream.'

A similar pattern is the trigram $.–NE–VVFIN (purple): These are sentence beginnings (after $., the period) in which only one proper noun occurs as the subject.

From TP02 on, however, the most common construction are prepositional phrases (APPR–ART–NN, red) and it remains so until TP10. It can be assumed that such prepositional phrases are frequently used to indicate place and time.

TOP10 shows an interesting distribution: In addition to the prepositional phrases (red) and the sentences starting with proper names (purple), a new pattern appears here among the top trigrams: $.–PPER–VVFIN (ligth green, PPER for personal pronouns). Instead of mostly repeating the proper nouns, the children now begin to start sentences with a personal pronoun more regularly, as in Example (5), so that this pattern shows up among the top trigrams.

(5) *Sie* *sah* *Lars* *mit* *Dodo*
PPER VVFIN NE APPR NE
'She saw Lars with Dodo'

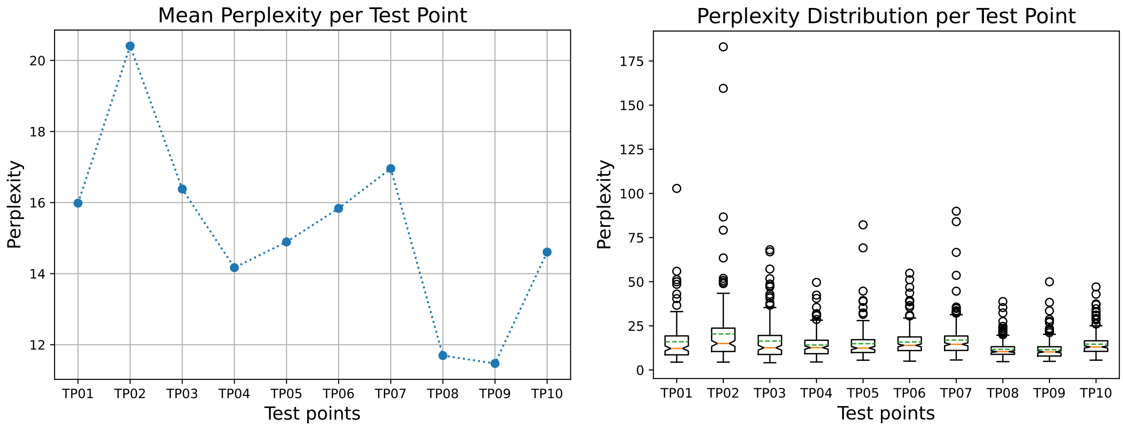

Figure 6: Mean perplexity (left) and perplexity distribution (right) over test points.

POS-based perplexity We calculated perplexity separately for each text. Figure 6 plots the mean values and distribution of perplexity of the texts written at the ten different test points. Looking first at the mean values (left), we observe an overall downward trend in the perplexity scores over time. The three test points with the same story, TP02, TP06 and TP10, also show a clear downward trend. As perplexity values indicate how well the trained language model fits the sample, the overall downward trend shows that in general the children's texts become more similar to the Klexikon in terms of POS trigrams. There are, however, peaks and troughs indicating exceptions to the overall trend. These need to be examined further to see if, e.g., there is a story-related reason for the outliers.

The boxplots (right) show that there are more outliers at earlier test points, i.e. texts that deviate clearly from the style of the Klexikon-based language model. The later the test points, the more homogeneously the children write. We could already observe such a development in Fig. 3 for the IDF-LDist scores.

5 Conclusion

The aim of this paper was to investigate the development of complexity in texts produced by primary school children. We measure complexity on a lexical and syntactic level with different measures based on the Litkey Corpus.

The different measures of lexical diversity confirm our expectations: the children's vocabulary in describing the picture stories becomes increasingly diverse over time, despite the fact that the children were limited in their text production by the given picture stories.

The new measure of lexical distinctiveness, IDF-LDist, shows that the texts become more homogeneous overall, i.e., the older children tend to write similarly diverse texts. We hypothesized that personal background may play a greater role at the beginning of elementary school, which would explain the greater variance and the extreme outliers. At later test points, the children's competencies become more and more similar.

At the syntactic level, the distribution of the POS n-grams shows that the syntactic structures used by the children when writing are developing further and that, for example, function words such as articles and personal pronouns are being added.

Perplexity on POS trigrams shows an overall downward trend, as expected. However, there are also outliers, which require further investigations. Similar to lexical distinctiveness, perplexity becomes more homogeneous over time.

Ethical Considerations

We do not see a direct harm that could follow from the research reported in this study. However, the analyses could inherit potential biases present in the Litkey Corpus and not reflect all populations of primary school children in Germany equally well.

Limitations

A limitation of the present study is that we measure linguistic complexity using only a small subset of potential measures, focusing on lexical diversity and syntactic complexity based on POS sequences. Incorporating further measures, e.g. based on syntactic dependencies, would be necessary in order to draw a more complete picture of the development of linguistic complexity in primary school children's texts. However, this is yet infeasible because the Litkey Corpus lacks gold-standard annotations of structures above the word level.

Acknowledgments

We would like to thank the anonymous reviewers for their very helpful comments.

References

Adrien Barbaresi. 2024. Simplemma: A simple multilingual lemmatizer for Python [computer software] (version 0.9.1). Berlin, Germany: Berlin-Brandenburg Academy of Sciences. Available from https://github.com/adbar/simplemma.

Muzaffer Sencer Bulut-Ozsezer and Hatice Beyza Canbazoglu. 2018. Picture in children's story books: Children's perspective. *International Journal of Eductional Methodology*, 4(4):205–217.

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.

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

Hendrike Frieg. 2014. *Sprachförderung im Regelunterricht der Grundschule: Eine Evaluation der Generativen Textproduktion*. Ph.D. thesis, Ruhr-Universität Bochum.

John Heilmann, Jon F. Miller, Ann Nockerts, and Claudia Dunaway. 2010. Properties of the narrative scoring scheme using narrative retells in young school-age children. *American Journal of Speech-Language Pathology*, 19(2):154–166.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing*. (3rd edition, draft of Feb 3, 2024).

Noemi Kapusta, Marco Müller, Matilda Schauf, Isabell Siem, and Stefanie Dipper. 2022. Assessing the linguistic complexity of German abitur texts from 1963–2013. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 48–62, Potsdam, Germany.

Ronja Laarmann-Quante. 2021. *Prediction of spelling errors in freely-written texts of German primary school children*. Ph.D. thesis, Ruhr-Universität Bochum.

Ronja Laarmann-Quante, Stefanie Dipper, and Eva Belke. 2019a. The making of the Litkey Corpus, a richly annotated longitudinal corpus of German texts written by primary school children. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 43–55, Florence, Italy. Association for Computational Linguistics.

Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert, Simon Masloch, Doreen Scholz, Eva Belke, and Stefanie Dipper. 2019b. The Litkey Corpus: A richly annotated longitudinal corpus of German texts written by primary school children. *Behavior Research Methods*, 51(4):1889–1918. (Shared senior authorship).

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

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

Katrin Ortmann and Helena Wedig. 2024. KidRef: Ein Kinderreferenzkorpus. *Bochumer Linguistische Arbeiten*, 26.

Famala Eka Sanhadi Rahayu, Aries Utomo, and Ririn Setyowati. 2020. Investigating lexical diversity of children's oral narratives: A case study of L1 speaking. *Register Journal*, 13(2):371–388.

Jan Thomas Röhrig. 2020. Empirisch ermittelte Muster in Rechtschreibfehlern für die Automatisierung qualitativer Rechtschreibdiagnostik. Poster at the 23. Symposion Deutschdidaktik.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).