Becoming Linguistically Mature: Modeling English and German Children's Writing Development Across School Grades

Elma Kerz[1], Yu Qiao[1], Daniel Wiechmann [2], and Marcus Ströbel[1]
[1]RWTH Aachen University, Germany
[2]University of Amsterdam, Netherlands
{elma.kerz|marcus.stroebel}@ifaar.rwth.aachen.de
yu.qiao@rwth-aachen.de d.wiechmann@uva.nl

Abstract

In this paper we employ a novel approach to advancing our understanding of the development of writing in English and German children across school grades using classification tasks. The data used come from two recently compiled corpora: The English data come from the the GiC corpus (983 school children in second-, sixth-, ninth- and eleventh-grade) and the German data are from the FD-LEX corpus (930 school children in fifth- and ninth-grade). The key to this paper is the combined use of what we refer to as 'complexity contours', i.e. series of measurements that capture the progression of linguistic complexity within a text, and Recurrent Neural Network (RNN) classifiers that adequately capture the sequential information in those contours. Our experiments demonstrate that RNN classifiers trained on complexity contours achieve higher classification accuracy than one trained on text-average complexity scores. In a second step, we determine the relative importance of the features from four distinct categories through a Sensitivity-Based Pruning approach.

1 Introduction

There is growing recognition among researchers, educators and policymakers that literacy and the language of schooling (other terms include academic language, language of education, scientific language) are key to children's overall educational success and academic achievement (see, e.g., Commission, 2019; Lorenzo and Meyer, 2017). Children are expected to acquire the ability to comprehend and produce complex clause and sentence structures, sophisticated vocabulary and informationally dense texts characteristic of language of schooling as they progress through their school career (see, e.g., Berman, 2007; Snow, 2010, for overviews). However, this ability is acquired gradually and for many school children only with difficulty (Snow and Uccelli, 2009; Snow, 2010). Given the key role of academic language, it is somewhat surprising that relatively little empirical research has been conducted on the development of academic language skills across school grades in children's first language, in particular in the area of writing (for exceptions, see, Crossley et al., 2011; Weiss and Meurers, 2019). This paper contributes to and expands the scant literature by investigating the development of linguistic complexity in children's writing from second-, sixth-, ninth- and eleventh-grade in English schools and fifth- and ninth-grade in German schools. We employ a novel approach to the automatic assessment of text complexity. In this approach, a series of scores for a given complexity measure is obtained through a sliding window technique, tracking the progression of complexity within a text, captured in what we refer to as 'complexity contours' (Ströbel, 2014; Ströbel et al., 2020). These contours are then fed into recurrent neural network (RNN) classifiers – adequate to take into account the sequential information in the contours – to perform grade-level classification tasks. We demonstrate the utility of the approach by comparing the performance of 'contour-based' RNN models against those of 'means-based' RNN models trained on text-average performance scores. In a second step, we determine which features drive classification accuracy through a Sensitivity-Based Pruning (SBP) approach. The remainder of the paper is organized as follows: Section 2 provides a concise overview of related work on automated assessment of text complexity in combination with machine learning techniques in the language learning context. Section 3 presents the two data sets representing English and German children's school writing. Section 4 introduces our approach to assessment of text complexity based on a sliding-window technique, whereas Section 5 introduces the features

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 65–74

investigated in the paper. Sections 6 describes the model architecture and the training procedure (Section 6.1) and the SBP method used to determine the relative feature importance (Section 6.2). Sections 7 presents the results and concluding remarks follow in Section 8.

2 Related work

In recent years, there has been an increased interest in automated assessment of text complexity in authentic contextualized language samples in combination with machine learning techniques (Meurers, 2020, for a recent overview). As a valuable complement to experimental research, this research has the potential to advance our current understanding of (both first and second) language learning and development (Rebuschat et al., 2017; Ellis, 2019). Important steps have been made in this direction through both language input and language output perspectives: Regarding the former, a number of studies have examined whether and to what extent learning materials show an adequate level of linguistic complexity considered to be of crucial importance for successful learning outcomes (see, e.g., François and Fairon, 2012; Pilán et al., 2016; Xia et al., 2019; Chen and Meurers, 2018; Berendes et al., 2018). For example, Berendes et al. (2018) employ a text classification approach to examine to whether and to what extent reading complexity of school textbooks differ systematically across grade levels in line with the so-called 'systematic complexification assumption'. They build text classification models using a Sequential Minimal Optimization (SMO) algorithm trained on a wide range of lexical, syntactic, morphological, and cohesion-related features to predict the grade level (fifth to tenth grade) and school track (high vs. low). The best performing model reached a grade-level classification accuracy of 53.7%, corresponding to a 20.7% over the random baseline, providing only partial support for thee systematic complexification assumption. In addition, they report significant differences across grade levels and tracks for some of the ten linguistic features. Regarding the latter, a rapidly growing body of research has focused on language output aiming to determine to what extent L2 writing and speaking differs from that of their L1 peers and expert writers/speakers, to differentiate levels of language proficiency, to predict human ratings of the quality of learner productions, and to examine to the relationship between L1 and L2 writing complexity and speaking fluency (see, e.g., Crossley et al., 2014; Lu, 2017; Duran-Karaoz and Tavakoli, 2020; Ströbel et al., 2020). Much research in this area has focused on English and on populations of upper intermediate to advanced L2 learners (but see Crossley et al., 2011; Durrant and Brenchley, 2019; Weiss and Meurers, 2019, for L1 English and German, respectively). Two recent studies are particularly relevant for the purposes of the present study. Durrant and Brenchley (2019) zoom-in on the development of vocabulary sophistication in English children's writing across second-, sixth-, ninth- and eleventh grade. Their corpus (also used in the present paper) consists of 2,898 texts of children's writing produced by 983 children. Through a mixed-effects regression modeling approach, they assess the effects of grade level and genre on lexical sophistication - measured through children's use of low-frequency words and register appropriate words. Their analysis reveals no significant differences with regard to the average frequency of the lexical words used by younger and older children. However, with increasing age children's writing display a shift from a more fiction-like vocabulary to a more academic-like vocabulary, reflecting a development towards more register appropriate word use. Weiss and Meurers (2019) focus on German children's writing development through a text classification approach based on a broad range of complexity and accuracy measures. Their dataset includes 1,633 texts of writing from 727 German elementary school children from first to fourth grade and 906 secondary school students from fifth to eighth grade, who attended either a basic or an intermediate school track. Using SMO classifiers with a linear kernel, their best performing classification model employed a combination of linguistic complexity features, error rate and meta information on topic and school track to reach an accuracy of 72.68% in classifying four grade level categories. Their analysis further revealed a shift in the primary locus of development from accuracy to complexity within elementary school and an increasing linguistic complexity in secondary school, in particular in the lexical complexity domain.

3 Data

The data used in this study come from two recently compiled corpora representing school writing: The English data come from the the Growth in Grammar corpus (GIG, https://gigcorpus.com/) that

comprises 2,898 texts produced by 983 children in 24 different schools from 14 cities in Great Britain. The texts in the GiG corpus were sampled at four points that mark 'key stages' of the English school system: the ends of Key Stage (KS) 1 (Year 2, when children are 6-7 years old) and KS2 (Year 9, when children are 10-11 years old), encompassing the primary phase of the school system, and the ends of KS3 (Year 9, when children are 13-14 years old) and KS4 (Year 11, when children are 15-16 years old), encompassing the secondary stage. The texts were classified into two text types (literary and non-literary texts) on the basis of their overall purpose. Approximately 13% of the texts were written by children categorized as speaking English as an additional language. The German data come from the Forschungsdatenbank Lernertexte (FD-LEX; https://fd-lex.Uni-koeln.de/), a research database of learner texts compiled in joint project of the Mercator Institute for Language Promotion and German as a Second Language. It contains a total of 5,628 texts from two text types (report and argumentation) collected from a total of 930 school children in grades five (when children are 10-11 years old) and nine (when children are 14-15 years old) at comprehensive and grammar schools in two German cities. These texts were elicited using a narrative and an argumentative writing prompt. The database contains information on a number of learner background variables, including the learners language background distinguishing monolingual German students from students who have German as a their first language (L1) and know at least one additional language and students for whom German is not their first language. Table 1 shows the distribution of texts along with descriptive statistics of text sizes across grade levels and registers for each language.

Grade	Register	N Texts	M	SD
English data: GIG				
2	lit	263	83.56	58.36
2	non-lit	376	71.18	43.09
4	lit	23	169.7	111.6
4	non-lit	26	151.58	96.65
6	lit	293	371.58	200.61
6	non-lit	575	208.68	104.53
9	lit	220	422.22	186.69
9	non-lit	584	277.25	187.95
11	lit	63	422.22	186.69
11	non-lit	475	415.3	264.46
German data: FD-LEX				
5	arg	1462	49.26	27.94
5	nar	1460	67.65	32.68
9	arg	1282	70.67	32.24
9	nar	1305	80.69	32.54

Table 1: Composition of two corpora of children's school writing. 'lit' = literary, 'non-lit' = non-literary, 'arg' = argumentative, 'nar' = narrative; M = mean number of words, SD = standard deviation

4 Automatic Assessment of Text Complexity through a Sliding Window Technique

Text complexity of the writing samples in the two corpora is automatically assessed using the Complexity Contour Generator (CoCoGen), a computational tool that implements a sliding-window technique to generate a series of measurements for a given complexity measure (CM) (Ströbel, 2014; Ströbel et al., 2018; Ströbel et al., 2020). This approach enables a 'local assessment' of complexity within a text, in contrast to the standard approach that represents text complexity as a single score, providing a 'global assessment' of the complexity of a text. A sliding window can be conceived of as a window with a certain size (ws) defined by the number of sentences it contains. The window is moved across a text sentence-by-sentence, computing one complexity score per window for a given CM. The series of measurements generated by CoCoGen track the progression of linguistic complexity within a text captured in what we refer to as 'complexity contours'. These contours faithfully represent that complexity is typically not uniformly distributed within a text but rather by characterized peaks and troughs and that complexity contours of individual measures may exhibit different trajectories (see Figure 1. For a text comprising n sentences, there are $w = n - ws + 1$ windows.[1] To compute the complexity score of a given window, a measurement function is called for each sentence in the window and returns a fraction wn_S/wd_S, where wn_S is the numerator of the complexity score for a sentence and wd_S is the denominator of the complexity score for that sentence. If the window size is specified to be greater than one sentences, the denominators and numerators of the fractions from the first to the last sentence

[1] Given the constraint that there has to be at least one window, a text has to comprise at least as many sentences at the ws is wide $n \geq w$.

in the window are added up to form the denominator and numerator of the resulting complexity score of a given window (see Figure 3 in the Appendix). The size of the window is user-defined parameter whose value depends on the goals of the analysis: When windows is set to the minimum, i.e. complexity is measured at each sentence of a text, the resulting complexity contour will typically exhibit many sharp turns. By increasing the window size, i.e. the number of sentences in a window, the complexity contour can be smoothened akin to a moving average technique (see Figure 4 in the Appendix). To compute the complexity scores, CoCoGen uses the Stanford CoreNLP suite (Manning et al., 2014) for performing tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic parsing using the probabilistic context free grammar parsers for English (Klein and Manning, 2003) and German (Rafferty and Manning, 2008).

5 Features

In its current version CoCoGen features 57 complexity measures (CMs) for English of which 13 are also available for German.[2] These features cover (1) surface measures, (2) measures of syntactic complexity, (3) measures of lexical richness, (4) information theoretic measures, and (5) register-based n-gram frequency measures. The operationalizations of the syntactic and lexical CMs follow those given in Lu (2011) and Lu (2012). For details on the operationalization of the information theoretic CMs, see Ströbel (2014). The operationalization of the register-based n-gram frequency measures is provided below. Surface measures concern the length of production units and include Mean Length of Words in characters (MLWc), Mean Length of Words in syllable (MLWs), Mean length of clause (MLC), Mean length of sentence (MLS), and Mean length of T-Unit (MLT). Syntactic complexity is typically quantified in terms of measures of the type and incidence of embeddings (Sentence complexity ratio (C/S), T-Unit complexity ratio (C/T), Complex T-Unit ratio (CT/T), Dependent clause ratio (DC/C), Dependent clauses per T-Unit (DC/T), T-Units per Sentence (T/S), and Verb Phrases per T-Unit (VP/T)), the types and number of coordinations between clauses and phrasal units (Coordinate phrases per clause (CP/C), Coordinate phrases per T-Unit (CP/T)), and the type of particular structures (Complex nominals per T-Unit(CN/T), Complex nominals per Clause (CN/C), Noun Phrase Premodification in words (NPpreW), Noun Phrase Postmodification in words (NPpostW)) (see Lu, 2017, for a recent overview). The lexical richness measures fall into three distinct sub-types: (1) Lexical density, i.e. the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text (Lexical Density (LD)), (2) Lexical variation, i.e. the range of a learner's vocabulary as displayed in his or her language use (Number of Different Words (NDW), Type-Token Ratio (TTR), Log Type-Token Ratio (logTTR), Root Type-Token Ratio (rTTR), Corrected Type-Token Ratio (cTTR)) and (3) Lexical sophistication, i.e. the proportion of relatively unusual or advanced words in the learner's text (words from the New Academic Word List (NAWL), words from the New Academic Formula List (NAFL), words that are not part of the New General Service List (NGSL), Lexical Sophistication BNC (LS.BNC), Lexical Sophistication ANC (LS.ANC)). The three information-theoretic measures are Kolmogorov Deflate (KolDef), Kolmogorov Deflate Syntactic (KolDefSyn), Kolmogorov Deflate Morphological (KolDefMor) (see Ehret and Szmrecsanyi, 2019, for the benefits of using these measures in the assessment of text complexity in the context of language learning). These measures, use the Deflate algorithm (Deutsch and Gailly, 1996) to compress a given text and obtain performance scores by relating the size of the compressed file to the size of the original file (Ströbel, 2014). The fifth group of register-based n-gram frequency measures was based on list of the top 100,000 most frequent ngrams (for $n \in [1, 5]$) from the five register subcomponents of the COCA corpus[3] (spoken, magazine, fiction, news, academic language). The general definition of these CMs is given in (1) and

[2] CoCoGen was designed with extensibility in mind, so that additional CMs can easily be added. It uses an abstract measure class for the implementation of CMs. Currently, additional CMs from the cognitive science (psycholinguistic) literature are being implemented for both English and German.

[3] The Contemporary Corpus of American English (Davies, 2008) is the largest genre-balanced corpus of American English, which at the time the measures were derived comprised 560 million words.

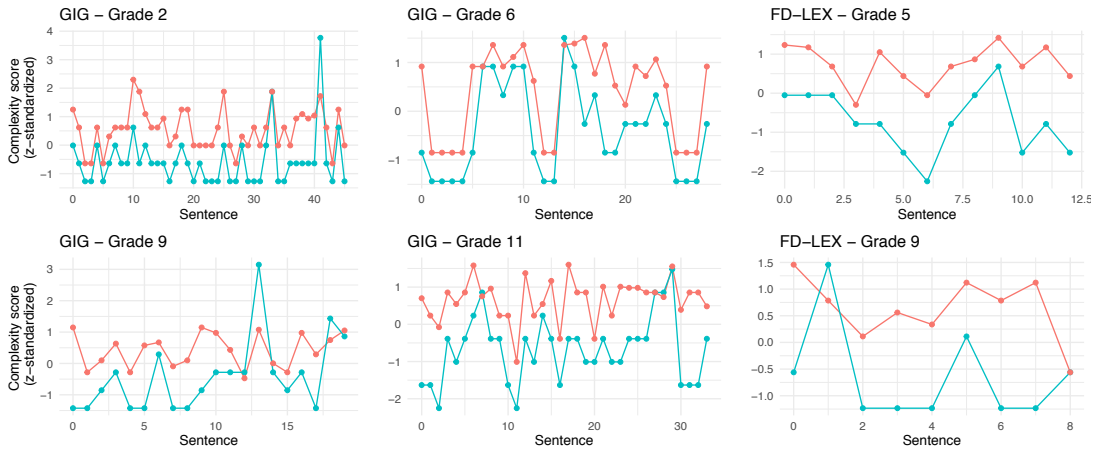

Figure 1: Complexity contours for two different measures (red: Type Token Ratio, blue: Clauses per Sentence) for six randomly selected texts from different grade levels for each language

(2):

$$\text{Score}_{n,s,r} = \frac{|C_{n,s,r}| \cdot \log\left[\prod_{c \in |C_{n,s,r}|} freq_{n,r}(c)\right]}{|U_{n,s}|} \quad (1)$$

where

$$C_{n,s,r} = A_{n,s} \cap B_{n,r} \quad (2)$$

Let $A_{n,s}$ be the list of n-grams ($n \in [0,5]$) contained within a sentence s, $B_{n,r}$ the list of n-grams on the n-gram frequency list of a register r ($r \in \{\text{acad}, \text{acad}, \text{fic}, \text{mag}, \text{news}, \text{spok}\}$) and $C_{n,s,r} = A_{n,s} \cap B_{n,r}$ the intersection list. Furthermore, $U_{n,s}$ denotes the list of unique n-grams in s, and $freq_{n,r}(a)$ the frequency of n-gram a in the n-gram frequency list of register r. The score of a given n-gram-based CMs is thus obtained by multiplying the number of n-grams in a sentence that are on the n-gram list with the log of the product of the corresponding frequencies of those n-grams divided by the number of distinct n-grams in the sentence.

6 Classification Models

6.1 Model Architecture

We used a Recurrent Neural Network (RNN) classifier, specifically a dynamic RNN model with Gated Recurrent Unit (GRU) cells (Cho et al., 2014). A dynamic RNN was chosen as it can handle sequences of variable length[4]. As shown in Figure 2, the input of the contour-based model is a sequence $X = (x_1, x_2, \ldots, x_l, x_{l+1}, \ldots, x_n)$, where x_i, the output of CoCoGen for the ith window of a document, is a 13 dimensional vector (for German) or a 57 dimensional vector (for English), l is the length of the sequence, $n \in \mathbb{Z}$ is a number, which is greater or equal to the length of the longest sequence in the dataset and $x_{l+1}, \cdots, x_n$ are padded **0**-vectors. The input of the contour-based model is be fed into a RNN which consists of two layers of GRU cells with 20 hidden units each. To predict the class of a sequence, the last output of the RNN, i.e. the output of RNN right after the feeding of x_l, concatenated with the variables (text type and learner background), which are encoded into one-hot vectors, is transformed through a feed-forward neural network. The feed-forward neural-network consists of three fully connected layers, whose output dimensions are 512, 256, 1 (German) and 3 (English). The Rectifier Linear Unit (ReLU) was used as activation function. Two dropout layers were added between fully connected layers 1 and 2 and between layers 2 and 3, both with a dropout rate of 0.3. Before the final output, a sigmoid layer was applied. For the mean-based model, we used the same neural network as in the contour-based model, except that the network was trained on vectors of text-average complexity scores. For the purpose of comparison, we also built two baseline models based on the control variables and the prior probability distribution. The first one is a statistics-based baseline model. We trained this model by grouping the instances in the dataset by the control variables and computed the empirical distribution over grades for each group. For prediction, we classified instances of the test set into grades by

[4]The lengths of the feature vector sequences depends on the number of sentences of the texts in our corpus.

$$p(y|x,c) = p(y|c) = \frac{N_{c,y}}{N_c}$$

where y is the class label, i.e. grade, x the features of an instance from the test set, and c is a control variable. $N_{c,y}$ denotes the number of instances in the training set, for a control variable c and class label y, while $N_c = \sum_y N_{c,y}$ is the total number of instances in the training set, which has c as their control variable. The second baseline model is a neural network model that has the same structure of the upper part of the RNN model which is a feedforward neural network. The input of this model is one-hot encoded control variables and the output stay the same as the RNN model.

Since the task is to classify instances of the dataset into a set of ordered categories, i.e. grade $2 < 6 < 9 < 11$ for English and grade $5 < 9$ for German, our task can be treated as an ordinal classification problem. To adapt the neural network classifier to the ordinal classification task, we followed the NNRank approach described in (Cheng et al., 2008), which is a gerneralization of ordinal perceptron learning in neural networks(Crammer and Singer, 2002) and outperforms a neural network classifier on serveral benchmark datasets. Instead of one-hot encoding of class labels and using softmax as the output layer of a neural network, in NNRank, a class label for class k is encoded as $(y_1, y_2, \ldots, y_i, \ldots, y_{C-1})$, in which $y_i = 1$ for $i \leq k$ and $y_i = 0$ otherwise, where C is the number of classes. For the output layer, a sigmoid function was used. For prediction, the output of the neural network $(o_1, y_2, \ldots, o_{C-1})$ is scanned from left to right. It stops after encountering o_i, which is the first element of the output vector that is smaller than a threshold T (e.g. 0.5), or when there is no element left to be scanned. The predicted class of the output vector is the index k of the last element, whose value is greater than or equal to T.

We use ten-fold cross-validation, using a 90%–10% split into training and testing sets. As the loss function for training, binary cross entropy was used:

$$\mathcal{L}(\hat{Y}, c) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}) + (1-y_i)\log(1-\hat{y}))$$

in which $c = (y_1, y_2, \ldots, y_N), N = C - 1$ is the true class label of the current observation encoded in accordance with the NNRank method, where C is the number of classes and $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ is the output vector of the sigmoid layer. For optimization, we used Adamax with a learning rate $\eta = 0.001$ and weight decay$= 1 \times 10^{-6}$. The minibatch size is 32, which was shown as a reasonable value for modern GPUs (Masters and Luschi, 2018). All models were implemented using PyTorch (Pytorch, 2019).

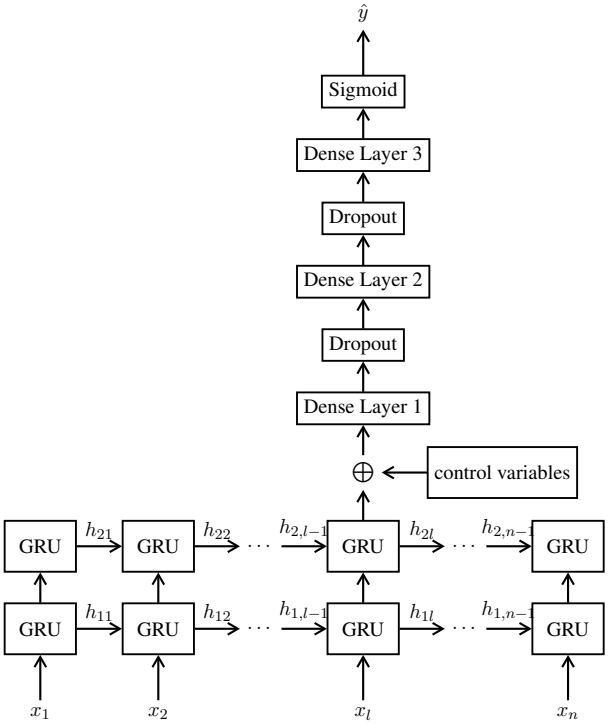

Figure 2: Roll-out of the RNN model based on complexity contours

6.2 Feature Importance

To determine the relative importance of the complexity features, we conducted feature ablation experiments for the contour-based RNN. Classical forward or backward sequential selection algorithms that proceed by sequentially adding or discarding features require a quadratic number of model training and evaluation in order to obtain a feature ranking (Langley, 1994). In the context of neural network model training a quadratic number of models can become prohibitive. To alleviate this problem, we used an adapted version of the iterative sensitivity-based pruning algorithm proposed by Díaz-Villanueva et al. (2010). This algorithm ranks the features based on a 'sensitivity measure' (see, (Moody, 1994; Utans and Moody, 1991)) and removes the least relevant variables one at a time. The classifier is then retrained on the resulting subset and a new ranking is calculated over

the remaining features. This process is repeated until all features are removed (see Algorithm 1). In this fashion, rather than training $\frac{n(n+1)}{2}$ models required for sequential algorithms, the number of models trained is reduced to $\frac{n}{m}$, where m is the number of features that can be removed at each step. We report the results obtained with m = 1, i.e. the results after the removal of a single feature at each step. At step t, neural network models $M_{t,n}, n \in \{1, \ldots, k\}$ are trained on the training sets of a 10-fold cross-validation, where n is the fold ID. The training sets at step t consist of instances with feature set $F_t = \{f_1, f_2, \ldots, f_{D_t}\}$ where $f_1, \ldots f_{D_t}$ are the remaining features at the current step, whose importance rank is to be determined. We define $X_{t,n}$ as the test set of the nth fold with feature set F_t and $X^i_{t,n}$ as the same dataset as $X_{t,n}$ except we set the i^{th} feature f_i of each instance within the dataset to its average. Furthermore, we define $g(X)$ as the classification accuracy of $M_{t,n}$ for a dataset X. The sensitivity of feature f_i on the nth fold at step t is obtained from:

$$S_{i,t,n} = g(X_{t,n}) - g(X^i_{t,n})$$

The final sensitivity for a feature f_i at step t is:

$$S_{i,t} = \frac{1}{k} \sum_{n=1}^{k} S_{i,t,n}$$

The most important feature at step t can be found by:

$$f_{\hat{i}} : \hat{i} = \underset{i: f_i \in F_t}{\arg\max}(S_{i,t})$$

Then we set the rank for feature $f_{\hat{i}}$:

$$Rank_{\hat{i}} = t$$

In the end, feature $f_{\hat{i}}$ is dropped from F_t and the corresponding columns in training and test dataset are also dropped simultaneously:

$$F_{t+1} = F_t - \{f_{\hat{i}}\}$$

This procedure is repeated, until $|F_{t'}| = 1$. To increase the robustness of the feature importance rank order, 10-fold cross-validation was applied.

7 Results

We report the results of classification with 10-fold cross-validation (see Figures 5 and 6 in the Appendix for a visualization of model accuracy for the means-based and contour-based models over 200 epochs across the 10 cross-validation folds). We first present the results of the experiments on the English data, before moving to results for the German data. The performance metrics of the classification models for English (global accuracy, precision, recall and macro F1 scores per grade level) are presented in Table 2. Both the means-based and the contour-based models achieved grade-level-based classification accuracy of $> 75\%$, a substantial improvement over the baseline model (28%, see Table 4 in the Appendix for details). These findings indicate that text complexity increases with children's age/competence level and provide further empirical evidence in support of grade-level-based complexification assumption (see the study by Berendes et al, 2018 described above). The contour-based model outperforms the means-based model in terms of both precision and recall across all classes, resulting in an increase in global classification accuracy of 6%, from 76% (means-based model) to 82% (contour-based model). Precision and recall rates are found to be highest for grade 2, followed by grades 6 and 11, and lowest for grade 11. Inspection of the confusion matrix for the contour-based model (see Table 7 in the Appendix) indicates that misclassified samples are close to the actual class, indicating that the model was sensitive to the grade ordering.[5] These results suggest that the change in complexity was most pronounced in the earlier grades and decreased with increasing grade levels. The results of the feature importance analysis reveal that classification is mainly driven by features related to vocabulary (the feature importance statistics for the top 30 measures can be found in Table 6 in the Appendix). The top 14 of the 57 measures are related to lexical sophistication, word length and the use of register-based n-grams. These findings are consistent with the available body of research suggesting that the development of children's writing during adolescent years is primarily characterized by higher proportions of unusual/advanced words and words of greater surface length (compare *same* vs. *equal* vs. *identical* vs. *tantamount*) (see, Berman 2007) and replicate and extend the findings reported in Durrant and Brenchley (2019) that the shift towards more academic vocabulary can also be observed in the use of multi-word sequences. The information theoretic measures are

[5] We also examined all pairwise classification errors among the four grades (see Table 8).

	Means-based		Contour-based	
	M	SD	M	SD
Accuracy	0.76	0.03	**0.82**	0.02
Precision 2	0.87	0.06	**0.90**	0.07
Recall 2	0.85	0.04	**0.90**	0.04
F1 score 2	0.86	0.03	**0.90**	0.03
Precision 6	0.78	0.02	**0.82**	0.04
Recall 6	0.78	0.03	**0.81**	0.04
F1 score 6	0.78	0.01	**0.81**	0.03
Precision 9	0.72	0.04	**0.76**	0.06
Recall 9	0.71	0.05	**0.76**	0.04
F1 score 9	0.71	0.03	**0.76**	0.03
Precision 11	0.71	0.09	**0.82**	0.06
Recall 11	0.75	0.06	**0.82**	0.05
F1 score 11	0.73	0.07	**0.82**	0.04

Table 2: Performance statistics of the means-based (left) and contour-based (right) RNN classifiers aggregated over 10 crossvalidation runs (English data). Baseline classification accuracy was 28%.

situated at ranks 15, 18 and 22. The group of lexical diversity measures (NDW and variants of TTR) is located in the mid-field (ranks 34, 36, 37, 38, 39). Syntactic complexity is found to play only a subsidiary role: with the exception of one measure (VP/T, rank 19) features from this class appeared only after rank 30.

Even with a more restricted features set compared to English, grade-level-based classification accuracy on the German dataset displays considerable - albeit less pronounced - improvement of $\geq$ 19% over the baseline model (51% classification accuracy) (see Table 5 in the appendix). These findings thus provide additional, though somewhat weaker, empirical evidence in support of grade-level-based complexification assumption. As is the case in the English data, the performance of the model based on complexity contours exceeds that of the means-based model on the German data both in terms of precision and recall across the two school grades, leading to an 4% increase in overall classification accuracy from 70% to 74%. Table 3 presents the performance statistics. The feature ablation analysis reveals that the most important features are more evenly distributed across the four groups of CMs (see Table 9 in the Appendix): The top eight features include surface CMs pertaining to the length of production unit (MLWc, MLC, MLS), lexical diversity (NDW, RTTR, CTTR), syntactic complexity (Cl/S), and information density

	Means-based		Contour-based	
	M	SD	M	SD
Accuracy	0.70	0.02	**0.74**	0.02
Precision 5	0.71	0.02	**0.75**	0.02
Recall 5	0.74	0.02	**0.79**	0.02
F1 score 5	0.72	0.02	**0.77**	0.01
Precision 9	0.69	0.03	**0.74**	0.03
Recall 9	0.66	0.03	**0.7**	0.03
F1 score 9	0.67	0.02	**0.72**	0.03

Table 3: Performance statistics of the means-based (left) and contour-based (right) RNN classifiers aggregated over 10 crossvalidation runs (German data). Baseline classification accuracy was 51%.

(KolDef). Within this set of eight CMs, the removal of individual CMs is associated with a relatively minor drop in classification accuracy of less than 1.5%, suggesting that the network is able to compensate for the loss of information from a given feature by relying on the other features. However, when the last feature of the top-8 group is removed, classification accuracy drops by almost 5%, indicating that the remaining features played subsidiary roles in the grade-level classification. These findings nicely complement those reported in the paper by Weiss and Meurers (2019) described above focusing on basic (Hauptschule) and intermediate school tracks (Realschule) by assessing writing skills in the other two tracks of the German educational system: comprehensive school (Gesamtschule) and grammar school (Gymnasium).

8 Conclusion and Outlook

In this paper, we demonstrated how the automatic assessment of text complexity through a sliding window approach in combination with machine learning techniques can provide valuable and unique insights into the development of children's writing as they progress through their school education. Such an approach has the added advantage of capturing the progression of complexity within a text. In classification tasks on two data sets representing children's school writing in L1 English and German, we showed that the inclusion of this sequential information can substantially increase classification performance across grade-levels. We also show that Sensitivity-Based Pruning is a viable complementary approach to other approaches aimed at assessing feature importance to identify

'criterial features' that are characteristic and indicative of language competencies at a given level (Hawkins and Filipović, 2012). More generally, the type of research presented in this paper has the potential to advance our understanding of the development of literacy skills in children during adolescent years, a key stage that is still not well understood. In future work, we intend to extend the approach presented here to larger cross-sectional data sets covering additional school grades in search of valid and reliable benchmarks and norms that can be used to inform school curricula and educational standards.

References

Karin Berendes, Sowmya Vajjala, Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein. 2018. Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, 110(4):518.

Ruth A Berman. 2007. Developing linguistic knowledge and language use across adolescence.

Xiaobin Chen and Detmar Meurers. 2018. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. 2008. A neural network approach to ordinal regression. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1279–1284. IEEE.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

European Commission. 2019. Proposal for a council recommendation on a comprehensive approach to the teaching and learning of languages.

Koby Crammer and Yoram Singer. 2002. Pranking with ranking. In *Advances in neural information processing systems*, pages 641–647.

Scott A Crossley, Rod Roscoe, and Danielle S McNamara. 2014. What is successful writing? an investigation into the multiple ways writers can write successful essays. *Written Communication*, 31(2):184–214.

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.

Peter Deutsch and Jean-Loup Gailly. 1996. Zlib compressed data format specification version 3.3. Technical report, RFC 1950, May.

Wladimiro Díaz-Villanueva, Francesc J Ferri, and Vicente Cerverón. 2010. Learning improved feature rankings through decremental input pruning for support vector based drug activity prediction. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 653–661. Springer.

Zeynep Duran-Karaoz and Parvaneh Tavakoli. 2020. Predicting L2 fluency from L1 fluency behavior: The case of L1 Turkish and L2 English speakers. *Studies in Second Language Acquisition*, pages 1–25.

Philip Durrant and Mark Brenchley. 2019. Development of vocabulary sophistication across genres in english children's writing. *Reading and Writing*, 32(8):1927–1953.

Katharina Ehret and Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in sla production data. *Second Language Research*, 35(1):23–45.

Nick C Ellis. 2019. Essentials of a theory of language cognition. *The Modern Language Journal*, 103:39–60.

Thomas François and Cédrick Fairon. 2012. An ai readability formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477. Association for Computational Linguistics.

John A Hawkins and Luna Filipović. 2012. *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*, volume 1. Cambridge University Press.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Pat Langley. 1994. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, pages 1–5.

Francisco Lorenzo and Oliver Meyer. 2017. Special issue: Languages of schooling: explorations into disciplinary literacies.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *Tesol Quarterly*, 45(1):36–62.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Xiaofei Lu. 2017. Automated measurement of syntactic complexity in corpus-based l2 writing research and implications for writing assessment. *Language Testing*, 34(4):493–511.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Dominic Masters and Carlo Luschi. 2018. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*.

Detmar Meurers. 2020. Natural language processing and language learning. In Carol A. Chapelle, editor, *The Concise Encyclopedia of Applied Linguistics*. Wiley.

John Moody. 1994. Prediction risk and architecture selection for neural networks. In *From statistics to neural networks*, pages 147–165. Springer.

Ildikó Pilán, David Alfter, and Elena Volodina. 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 120–126.

Pytorch. 2019. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. https://github.com/pytorch/pytorch.

Anna N Rafferty and Christopher D Manning. 2008. Parsing three german treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, pages 40–46.

Patrick Emanuel Rebuschat, Meurers Detmar, and Tony McEnery. 2017. Language learning research at the intersection of experimental, computational and corpus-based approaches. *Language Learning*, 67(S1):6–13.

Catherine E Snow. 2010. Academic language and the challenge of reading for learning about science. *science*, 328(5977):450–452.

Catherine E Snow and Paola Uccelli. 2009. The challenge of academic language. *The Cambridge handbook of literacy*, pages 112–133.

Marcus Ströbel. 2014. *Tracking complexity of l2 academic texts: A sliding-window approach*. Master thesis. RWTH Aachen University.

Marcus Ströbel, Elma Kerz, Daniel Wiechmann, and Yu Qiao. 2018. Text genre classification based on linguistic complexity contours using a recurrent neural network. In *MRC@ IJCAI*, pages 56–63.

Marcus Ströbel, Elma Kerz, and Daniel Wiechmann. 2020. The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning. *Language Learning*, n/a(n/a).

Joachim Utans and John Moody. 1991. Selecting neural network architectures via the prediction risk: Application to corporate bond rating prediction. In *Proceedings First International Conference on Artificial Intelligence Applications on Wall Street*, pages 35–41. IEEE.

Zarah Weiss and Detmar Meurers. 2019. Analyzing linguistic complexity and accuracy in academic language development of german across elementary and secondary school. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 380–393.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2019. Text readability assessment for second language learners. *arXiv preprint arXiv:1906.07580*.