Journal of Educational Psychology
2018, Vol. 110, No. 4, 518–543

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000225

Reading Demands in Secondary School: Does the Linguistic Complexity of Textbooks Increase With Grade Level and the Academic Orientation of the School Track?

Karin Berendes
University of Tübingen

Sowmya Vajjala
Iowa State University

Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein
University of Tübingen

An adequate level of linguistic complexity in learning materials is believed to be of crucial importance for learning. The implication for school textbooks is that reading complexity should differ systematically between grade levels and between higher and lower tracks in line with what can be called the *systematic complexification assumption*. However, research has yet to test this hypothesis with a real-world sample of textbooks. In the present study, we used automatic measures from computational linguistic research to analyze 2,928 texts from geography textbooks from four publishers in Germany in terms of their reading demands. We measured a wide range of lexical, syntactic, morphological, and cohesion-related features and developed text classification models for predicting the grade level (Grades 5 to 10) and school track (academic vs. vocational) of the texts using these features. We also tested ten linguistic features that are considered to be particularly important for a reader's understanding. The results provided only partial support for systematic complexification. The text classification models showed accuracy rates that were clearly above chance but with considerable room for improvement. Furthermore, there were significant differences across grade levels and school tracks for some of the ten linguistic features. Finally, there were marked differences among publishers. The discussion outlines key components for a systematic research program on the causes and consequences of the lack of systematic complexification in reading materials.

Educational Impact and Implications Statement
In our study, we examined whether German textbooks used in secondary school (Grades 5 to 10, vocational and academic tracks) are constructed in a systematic way with respect to their text complexity. Moreover, we looked at differences between publishers. Our results provided only partial support for a systematic increase in text complexity with regard to grade levels and school tracks. Furthermore, there were marked differences among publishers. Thus, it would be worthwhile for the publishers and authors of school textbooks to more carefully consider the readability characteristics of the learning materials they provide.

Keywords: reading demands, secondary school, textbooks, linguistic complexity, academic language

Teaching materials have a substantial effect on learning outcomes (e.g., Nicol & Crespo, 2006; Pyburn & Pazicni, 2014). Even in times of growing digitalization, textbooks still comprise teachers' primary type of teaching material (Ebner & Schön, 2012). The medium of learning is language, and learning and language are closely interlinked (Halliday, 1993). "Building knowledge by reading, building knowledge of reading, and engaging in reading are always co-occurring events" (Alexander, 2012, p. 262). Because it is not possible to separate school subjects from the language they are presented in, the readability of school texts is essential not only for language lessons but for all other specialized classes as well (e.g., geography). Moreover, for some time now, there has been international broad agreement that reading should be promoted in all subjects (e.g., The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany; KMK, 2012).

If texts are constructed according to the zone of proximal development proposed by Vygotsky (1978), gains in learning will

This article was published Online First November 9, 2017.

Karin Berendes, University of Tübingen; Sowmya Vajjala, Iowa State University; Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein, University of Tübingen.

The project was supported by the LEAD Graduate School & Research Network (GSC1028), which is funded by the Excellence Initiative of the German federal and state governments.

Correspondence concerning this article should be addressed to Karin Berendes, University of Tübingen, Hector Research Institute of Education Sciences and Psychology, Europastraße 6, 72072 Tübingen. E-mail: karin.berendes@uni-tuebingen.de

be more pronounced. If reading demands are too high or too low, students' ability to concentrate on the comprehension of sentence and text content will be negatively affected (Scheerer-Neumann, 1997). Moreover, a reader may become frustrated, bored, or confused when the complexity of a text is not aligned with his or her zone of proximal development. As a result, readers might tune out, and their minds might wander (Feng, D'Mello, & Graesser, 2013). Thus, an adequate level of linguistic complexity is of crucial importance for learning. Allington, McCuiston, and Billen (2015) explained that:

> evidence accumulated suggests that texts that can be read with 95% or greater accuracy are directly, and in some studies causally, related to improved reading achievement. Texts that are read with either significantly lower or higher levels of accuracy fail to produce positive effects as large as the "just right" texts. (p. 499)

Seals (2013) used a control group, pretest-posttest design to evaluate the effectiveness of leveled book programs on reading fluency and reading comprehension and found that "leveled books are effective in increasing student oral reading fluency and comprehension level" (Seals, 2013, p. 3).

The linguistic complexity of textbooks should be expected to vary as a function of readers' reading competence, a pattern that we call the *systematic complexification assumption*, yielding systematic differences across grade levels and school tracks. There is quite a lot of research on educational texts, indicating that the written contents of textbooks are often not adapted to the school grade in which they are used or to students' abilities (e.g., Robison, Roden, & Szabo, 2015). However, existing studies have relied on just a few texts or did not systematically study the characteristics of the written contents of textbooks across several school grades and tracks. Expanding on previous approaches and using a computerized linguistic approach[1] and a large sample of 2,928 texts, we systematically assessed whether the complexity level of textbooks systematically increases with grade level and the academic orientation of the school track.

Complexity Levels in Reading Materials

Text complexity is the "inherent difficulty of reading and comprehending a text combined with consideration of reader and task variables" (NGACBP & CCSSO, 2010, Appendix A, Glossary of Key Terms, p. 43). Absolute text complexity concerns the language system and the linguistic domains (phonology, lexicon, morphosyntax). It can also be called "grammatical complexity" or "linguistic complexity" (we use the term *linguistic complexity*). Relative text complexity takes into account the difficulty of mental processes and the particular language user and depends on a user's language experience (Miestamo, 2008).

Text complexity was underrepresented in research before 2010 (Hiebert & Pearson, 2014), but since the release of the Common Core State Standards for English Language Arts, text complexity has been an important focus of research (Valencia, Wixson, & Pearson, 2014). Moreover, the Common Core Standards Initiative resulted in a push in research into computational extraction and analysis of linguistic features of text complexity. Several systems that can analyze the complexity of English texts were created, for example, SourceRater (Educational Testing Service; Sheehan, Kostin, Futagi, & Flor, 2010), Pearson Reading Maturity Metric (Landauer, Kireyev, & Panaccione, 2011), and CohMetrix (Graesser, McNamara, & Kulikowich, 2011). Nelson, Perfetti, Liben, and Liben (2011) compared the performance of such systems in a collection of different text sets that included test passages used in standard tests. They concluded that the best performing systems considered a broader range of linguistic features that were strongly correlated with the grade levels that Common Core State Standards exemplar texts were designed for. Collins-Thompson (2014) provided a detailed survey of the features used in the development of text complexity systems, mostly for English language texts.

In order to discuss different levels of complexity in reading materials, the processes involved in reading and text comprehension should be considered first. Reading, the process of obtaining meaning from print, is a complex cognitive process (e.g., McNamara & Magliano, 2009). It involves the coordination of lower order processes (decoding, word recognition) and higher order cognitive processes (thinking, analyzing, reasoning, reflecting, connecting; Pressley, 1998).

At the word level, reading requires the decoding of visual input and the use of different strategies that lead to word identification: (a) sequential decoding (letter-sound correspondences); (b) use of spelling patterns or analogy; (c) use of morphemic elements; and (d) automatic recognition (sight word recognition; see, e.g., Chard, Pikulski, & Templeton, 2000; Westwood, 2001). At the sentence and text level, syntactic context cues (e.g., grammatical role of a word in sentence, cohesive devices between words, phrases, and sentences) and semantic context cues (e.g., comparison clues, contrast clues) are also used for comprehension. Advanced readers use these different cues simultaneously and interactively in order to comprehend what they have read. What makes reading difficult is also determined by working memory capacity (the longer the units that have to be processed, the harder it is to process them) and previous knowledge. The coherence and structure of the text and the number of ideas expressed in it affect the reading process as well (Kintsch, 1974).

Returning to text complexity, factors that influence text complexity are mostly classified into three dimensions: (a) quantitative measures (e.g., word and sentence length); (b) qualitative measures (language features, structure, purpose and meaning, knowledge demands, and the layout of a text; e.g., Klare, 1963); and (c) the matching of the reader to the text and task (e.g., NGACBP & CCSSO [National Governors Association Center for Best Practices & Council of Chief State School Officers], 2010).

First, quantitative surface measures such as word frequency and sentence length are typically implemented in readability formulas (e.g., The Flesch-Kincaid Grade Level Readability Formula, The Gunning's Fog Index, SourceRater, Pearson Reading Maturity Metric, CohMetrix). Readability formulas provide a numerical score that ranks reading materials according to their difficulty. These formulas use the length of words as a proxy for semantic complexity, and sentence length is used as a proxy for syntactic complexity. The implication of using these features is that the shorter the words and the shorter the sentences, the easier the text.

[1] The broad range of German complexity measures employed in this study will be made readily accessible through a web application using the recently open-sourced Common Text Analysis Platform (Chen & Meurers, 2016).

These assumptions have been criticized because some research has shown that using simpler, shorter words does not automatically result in better text comprehension (Anderson & Davison, 1988; Urqhart, 1985), and it can be argued that shorter sentences are not necessarily easier to understand than longer ones (see Perera, 1980; but see also Rezaee & Norouzi, 2011). However, overall, these measures are good proxies for the complexity of a text (e.g., Nickel, 2011), and in our study, we examined sentence and word length.

Second, in research using qualitative measures, there is often a focus on one qualitative measure (e.g., on layout measures). In our study, we focused on linguistic features because, from our perspective, it is the most important group of qualitative measures (although the other ones are important as well). The term complexity then refers to text characteristics that are related to the different language subsystems of phonology, morphology, syntax, and semantics (Fenk-Oczlon & Fenk, 2008). Because we examined texts for advanced readers, where phonology is less important, we focused on morphological, syntactical, and semantic features.

Third, the matching of the reader to the text and task involves a consideration of the readers' cognitive capabilities, reading skills, motivation, engagement with the task and text, prior knowledge, and experience and how these qualities are related to the contents, themes, and complexity of the associated tasks. If the matching of the reader to the text fails, this will have negative consequences for the whole reading process: Working memory is overloaded, the capacity to construct a coherent mental representation of the text is not available, meaningful connections between text elements and relevant prior knowledge cannot be constructed, and as a consequence, the reader is not able to comprehend the text (Kendeou, van den Broek, Helder, & Karlsson, 2014). Reading motivation and reading engagement are likely to decrease as well (Guthrie et al., 2007). In the long run, it can be expected that reading frequency (time on task), and therefore reading experience and reading growth, will be much lower compared with the scenario in which texts that reflect a reader's optimal level of challenge are provided.

To meet the optimal level of challenge for a reader, books should be neither too easy nor too hard (Pearson, 2013). Thus, a match between the reading material and a certain readership has to take place. This equating of the reading material and a particular reader is a very complex task (Rog & Burton, 2001). For instance, this is difficult because the development of students' ability to read complex texts might not be linear (NGACBP & CCSSO, 2010, Appendix A). Moreover, in addition to other factors, the calibration of linguistic difficulty requires a comprehensive knowledge base about how reading skills develop over time and about the appropriateness of different levels of text complexity during the different phases of that development (Williamson, Fitzgerald, & Stenner, 2013). The challenge of using an adequate complexity level of reading material and evaluating the fit between texts and readers exists across the globe. In our study, we tested whether the texts were in accordance with the systematic complexification assumption. This assumption states that—as a prerequisite for a good match between text complexity and students' reading competence—reading complexity should systematically differ between grade levels and between higher and lower tracks. To the best of our knowledge, there is no empirical study that has examined the systematic complexification assumption in a broad sample of written material from German textbooks that are actually used in school.

Complexity Level in Textbooks

Most of the written contents of textbooks are written in academic language. Academic language, the so-called language of schooling (Schleppegrell, 2004/2010), is designed to be precise and concise, to refer to complex processes, and to express complicated ideas. For this reason, academic language uses complex grammatical constructions and sophisticated words that can disrupt reading comprehension and consequently block learning (Snow, 2010).

Of course, some features of language complexity differ across different languages, and there are differences in the extent to which certain languages are similar to each other. However, the overall differences tend to be rather small (Fromkin, Rodman, & Hyams, 2011), and the general question of how systematically the complexification assumption is implemented is of interest in every school system. Whereas in most countries, the complexity level of the written contents of textbooks is generated and assessed in a rather unsystematic way and is based on implicit knowledge, there are a few countries that have begun to adopt a more systematic approach. Most notably, the US had become a pioneer in the systematic complexification of the written materials presented in textbooks by implementing the Common Core State Standards (CCSS; NGACBP & CCSSO, 2010). The CCSS call for a staircase of increasing text complexity in what students read. They are based on quantitative as well as qualitative indicators of text complexity, but the tools used to categorize the texts "should be considered only provisional" and should be replaced with more precise, more accurate, and easier-to-use tools (NGACBP & CCSSO, 2010, p. 5). Unfortunately, in addition to some criticism concerning the theoretical and methodological bases of these standards (see Gamson, Lu, & Eckert, 2013; Pearson, 2013; Williamson et al., 2013), there is still a gap with regard to a systematic evaluation of the complexity of typical learning materials used in schools in both the US and other countries.

Whereas there is a large amount of research on the development of text complexity prediction methods, there is not much work on the application of these methods to textbook materials. Typical text complexity analyses are performed on texts that are read by students in a given grade or at a certain age and that are not necessarily (or specifically) textbooks (Graesser et al., 2014). Recently, some researchers have conducted longitudinal analyses of text complexity in textbooks used in the US in terms of lexical diversity and difficulty and have applied quantitative measures (e.g., word length and sentence length; Gamson et al., 2013; Lu, Gamson, & Eckert, 2014; Stevens et al., 2015). These studies have focused on a limited set of features and grades (third and sixth grades). Their historical analyses of change in text complexity and lexical difficulty in reading textbooks from 1905 to 2004 (Gamson et al., 2013; Lu et al., 2014) and text difficulty from 1910 to 2000 (Stevens et al., 2015) indicated that text complexity has increased steadily over the past 70 years (Gamson et al., 2013, p. 388). Moreover, the results showed an increase in lexical diversity and text difficulty from the 1970s to the 2000s (Gamson et al., 2013, p. 111; Stevens et al., 2015, p. 611). In our research, we focused

on different but related questions: We focused on a different language, worked with a broader range of linguistic features covering other aspects of language beyond words, and analyzed differences between different grade levels, types of schools, and publishers.

The Present Study

In Germany, the quality of textbooks (including adequate difficulty levels) is scrutinized by state officials before the books are allowed to be sold to schools and students, but the assessment is primarily based on the implicit knowledge of these officials rather than explicit standards for text complexity. Therefore, in the present study, we tested the systematic complexification assumption for textbooks used in German schools. Using a unique data set, we tested whether the textbooks were constructed in such a way that the language demands of the texts were in line with the systematic complexification assumption across three potential sources of systematic complexification (i.e., grade level, school track, and publisher).

We tested each hypothesis twice, once with a text classification approach, which is a method that is frequently used in computer-based linguistic research, and once with a regression analysis, which is often used in psychological research. For the classification models, we used a wide range of linguistic features simultaneously, whereas for the regression models, we focused on 10 linguistic features individually.

First, we examined whether the linguistic complexity of the texts increased from Grades 5/6 to Grades 7/8 to Grades 9/10. According to the systematic complexification assumption, text complexity should increase with students' age/competence levels. As students progress through school, they have to deal with increasingly complex learning contents, and such input cannot—or can only to a limited extent—be conveyed without complex linguistic structures. Therefore, students need to be introduced to and familiarized with academic language. If the demands are not aligned with the students' abilities and do not increase systematically, it is inevitable that students will become overstrained at some point. Moreover, if the reading demands remain about the same across secondary school, students will not be well-prepared for their later careers.

Second, we tested whether more advanced students were given more complex texts. In Germany, students are placed in—typically—three different tracks after Grade 4 on the basis of their achievement levels. We were able to compare textbooks that are used in the academic track with those from the vocational track. If the systematic complexification assumption held in our sample, the textbooks in the academic track would generally be more difficult to read than the textbooks in the vocational track.

Third, we assessed whether the linguistic complexity of the texts differed between publishers. Generally, textbooks are "cleared" for certain grade levels and tracks and are expected to be tailored to this specific student population (and not a subpopulation thereof). Thus, according to the complexification assumption, variability in the difficulty level across publishers should be small compared with variability in the difficulty level across grade levels and tracks.

Method

Texts

We compiled a collection of 35 geography textbooks that were officially approved in Baden-Württemberg, one of the largest states in Germany. These textbooks cover Grades 5 to 10 and were selected from the academic and vocational tracks on the basis of the textbook regulations in Germany (LS, 2013a, 2013b). The books were published by four different publishers. Thus, this corpus enabled us to study the effects of different factors (e.g., grade level, school track, publisher) on measures of text complexity together as well as separately.

The textbooks were scanned and digitized with Nuance OmniPage Ultimate Optical Character Recognition software (http://www.nuance.de/for-business/by-product/omnipage/ultimate/index.htm). This was followed by a manual inspection phase to ensure that there were no spelling errors due to scanning. To ensure that only relevant information was kept, each reading unit file was cleaned and manually coded with labels. All reading units that were lower than the sentence structure (i.e., no punctuation marks) were left out. Every chapter and its sections were labeled separately. Given that our interest was in the linguistic features of the information presented in the main body of text, we excluded other material (instructions, summaries, interviews, exercises, primary sources, definitions, picture captions, and miscellaneous material such as the table of contents and publisher information). As Gamson, Lu, and Eckert (2013) did, we will refer to each individual reading unit as a text.

Because some of the textbooks were intended to be used for two grades, we grouped the textbooks into three categories, each comprised of two consecutive grades—Grades 5/6, 7/8, and 9/10. Altogether, we considered 2,928 texts in our analysis. Appendix A shows the sample sizes for the subsamples separately for each grade level, school track, and publisher.

Assessment of the Linguistic Features of the Texts

We calculated 165 features that encoded lexical, syntactic, and morphological characteristics of language and discourse cohesion. Moreover, the features covered *surface measures* such as average sentence length and average number of syllables per word, both of which have been used in research on text complexity for several decades now.

The *lexical features* were comprised of several measures of lexical diversity (e.g., type-token ratio), variation (e.g., verb variation), and lexical density from the literature on English corpora, reimplemented in German. We also included word-usage-frequency-based features obtained from dlexDB (Heister et al., 2011) and semantic-relatedness features from GermaNet (Hamp & Feldweg, 1997), which rely on German-specific resources.

The *syntactic features* were comprised of measures that were based on both the phrase structure and dependency representations of sentences. Whereas most of them encode the occurrences and lengths of specific constructions (e.g., noun phrases, dependent clauses, etc.), others encode the dependencies between words in the sentence (e.g., average number of dependents per verb).

Morphological features encode the verbal and nominal inflection (e.g., passive participle, genitive nouns, etc.) and the usage of

various suffixes and compound nouns in German. These features were shown to be very useful for distinguishing between texts intended for young versus adult German readers (Hancke, Vajjala, & Meurers, 2012) but have not been explored in research on textbook complexity before.

Whereas all the features mentioned so far refer to individual words or sentences, *cohesion features* model the relations between sentences. We implemented 27 features for encoding word overlap between sentences, the usage of various kinds of pronouns, the usage of connector words in the texts, and the transformation of entities between sentences (e.g., the subject of one sentence becoming the object of the next sentence).

All the features were extracted after we preprocessed the texts by applying state-of-the-art natural language processing software—OpenNLP (https://opennlp.apache.org) for sentence segmentation, Stanford parser (Rafferty & Manning, 2008) for phrase-structure tree extraction, MATE parser (Bohnet & Kuhn, 2012) for morphological tagging and dependency parsing, and JWordSplitter for compound word splitting (https://github.com/danielnaber/jwordsplitter).

Our analyses consisted of two major steps. The entire feature set, comprised of 165 features, was used to train the classification models by applying supervised machine learning methods (see below). In addition, we computed an in-depth set of multilevel regression models for a number of features that have received a great deal of attention in the (theoretical) literature. For these analyses, we chose two features each on the surface, syntactic, lexical, morphological, and coherence levels, yielding a total of 10 features. We picked these 10 features (a detailed description follows) on the basis of theoretical considerations because "there is still no consensus on which features are actually the best predictors of readability" (De Clercq & Hoste, 2016, p. 458).

Our rationale for choosing these 10 features was the following: First, we picked the most common ones from readability/complexity research, namely, sentence and word length. These two features have been used in traditional readability formulas for several decades now (see Benjamin, 2012) and are good indicators of syntactic and lexical complexity. Second, we picked the most important ones for the language register under study, namely, academic language. All 10 features are expected to have a significant impact on the comprehension of texts written for educational/academic contexts. Third, we picked features that we would expect to differ between grade levels and school tracks for the texts under study. For instance, we did not pick passive voice as a feature because we would expect it to play only a minor role in geography texts. Sentences such as "The melting point is 217°C," "The volcano is erupting," or "The river flows into . . ."—just to name a few—are not predestined for passive voice. This is probably different for history texts (e.g., "Rome was not built in a day," "The fortress was conquered," or "Wilhelm was crowned emperor"). Fourth, we picked features for which a more frequent appearance increases the complexity of a text, or rather, an increase in difficulty can be expected. For example, we picked only certain connectors because some connectors (e.g., "and") would not be expected to increase in difficulty. Fifth, we picked features that have been shown to place special reading demands on students. The pronoun, for instance, is a feature that poses a well-known hurdle for readers (e.g., see Fang, 2016, p. 202f.).

Thus, the 10 features we chose to use in the current study each met one or more of the five criteria described above. However, because the criteria leave room for interpretation, it could be argued that another research group may have rated other features as more important. Therefore, we also analyzed the other 155 features, and these results can be found in Appendix B. Moreover, for each of the 10 features, Table 1 shows its *information gain ranking list* number. The information gain (IG) of a feature refers to the extent to which the feature could be used to split the given data set into the different categories (grades, schools, or publishers). IG ranking is commonly used in classification models to identify the best features from a larger group of features. Hence, a list of features ordered by their IG will essentially provide a list of features ordered by their importance for the classification task. The IG of a feature is calculated by estimating the difference in the entropy of the data set with and without the feature (Frank et al., 2005). Although the value by itself is not useful, it is useful for comparing one feature with another and for ranking features by their importance for the given classification task. The numbers in Table 1 are based on the ordering of the 165 features according to their impact on the classification of the texts. As examples, we present the results for two classifications models, one that classifies by grade level and one by school track.[2] As can be seen from the numbers, eight of the 10 features that we selected on the basis of theoretical considerations appeared in the top 20 at least once. On average, the best ranking emerged for the surface and syntactic levels and the worst ranking for the lexical and cohesion levels. This is mostly in accordance with findings for texts from other studies (e.g., Plakans & Bilki, 2016, for beginning, intermediate, and advanced reading textbooks for English as a second language).[3]

Surface/classical features that are used in readability formulas.

Average sentence length (Feature 1). This feature is measured by the average number of words per sentence. Sentence length is a good proxy for syntactic complexity and is the most general complexity measure (Norris & Ortega, 2009; Vyatkina, 2012). Based on its high validity and reliability, sentence length pertains to the most meaningful features with regard to the readability of a text, regardless of the language under study (Nickel, 2011). Moreover, an increase in academic language structures goes hand in hand with an increase in sentence length (Heppt, Dragon, Berendes, Stanat, & Weinert, 2012). In general, longer sentences are harder to understand than shorter ones (Bamberger & Vanecek, 1984). This is due to the fact that sentences that are longer overall create a higher load on working memory, and a larger number of different pieces of information and concepts must be integrated.

Average word length (Feature 2). This is a measure of the average number of syllables in the words in a text. "[A]t least in languages with clear syllabic boundaries, syllables are functional

[2] The IG ranking results for the other 155 features are presented in the Appendix B.

[3] Alternatively, we could have chosen the features on the basis of the IG ranking list. However, if we had picked our features according to the IG ranking, we would not have had a balance between the different linguistic levels because publishers can be expected to be more aware of the surface level than of the morphological and cohesion levels, for instance.

Table 1
Information Gain Rank of the 10 Features Selected for In-Depth Analyses for Classification by Grade Level and School Track

Variables	Information gain rank for classification by	
	Grade level	School track
1. Average sentence length (in words)	9	1
2. Average word length (in syllables)	2	29
3. Average length of longest dependency	10	13
4. Average number of complex nominals per clause	16	24
5. Root type-token ratio	75	6
6. Modifier variation	37	91
7. Ratio of derived nouns to all nouns	3	60
8. Ratio of genitive nouns to all nouns	11	59
9. Adversative and concessive connectors	45	69
10. Third-person personal pronouns	41	14

sublexical units during reading" (Barber, Vergara, & Carreiras, 2004, p. 545), and word length is one of the most commonly used measures of lexical complexity in traditional readability research. It is expected "that word length has a direct effect on the ease with which a text can be read: The longer a word is, the more difficult it is to comprehend" (Lenzner, 2014, p. 681). More syllables require the processing of more input and—overall—the longer a word, the longer the eye-fixation duration (Kliegl, Grabner, Rolfs, & Engbert, 2004).

Features on the syntactic level.

Average length of longest dependency (Feature 3). This feature refers to the distance between a word and its dependent in a sentence. A displaced dependent poses a challenge to the sentence processor because the first element of the dependency must be held in working memory until the related element can be linked to it. The feature reflects the central idea of Gibson's dependency locality theory (DLT) that "the cost of integrating two elements (such as a head and a dependent [. . .]) depends on the distance between the two" (Gibson, 2000, pp. 95–96). Thus, it can be assumed that longer dependencies pose greater processing demands than shorter ones (Temperley, 2007).

Average number of complex nominals per clause (Feature 4). Complex nominals are defined as comprised of one of the following three conditions (Cooper, 1976): (a) nouns with an adjective, possessive, prepositional phrase, relative clause, participle, or appositive; (b) nominal clauses; or (c) gerunds and infinitives in the subject position. The number occurrences of these three conditions were calculated by counting the number of occurrences of respective patterns in the syntactic parse tree. A clause is defined as a syntactic structure consisting of a subject and a finite verb. This feature is important to consider when studying the reading demands of textbooks because complex noun phrases use various demanding syntactic possibilities and therefore pose a considerable challenge to less experienced readers (Schmidt, 1993). Moreover, complex nominal groups "enable information to be presented in one clause that might otherwise take several clauses to express" (Fang, Schleppegrell, & Cox, 2006, p. 260) and therefore are a key contributor to lexical density. This entails greater processing demands for working memory, and thus a sentence or text is more difficult to process. This can result in comprehension limitations.

Features on the lexical level.

Root type-token ratio (Feature 5). The type-token ratio measures how many different words are used in a text and is a good proxy for its lexical diversity. Thus, next to lexical density, lexical sophistication, and number of errors, it is a good measure of lexical richness (Read, 2000). The calculation of this feature is based on the ratio of the number of unique words (types) in a text to all words (tokens). However, this measure is known to be sensitive to the length of the text, and several alternatives have been proposed to consider this limitation. The root type-token ratio (RTTR; Guiraud, 1960) is one such alternative measure, which is defined as the ratio of the number of types to the square root of the number of tokens. A higher type-token ratio makes a text more demanding because the vocabulary that must be known is richer.

Modifier variation (Feature 6). Modifier variation refers to the ratio of the total number of unique adjectives and adverbs in a text to the total number of lexical words. Adjectives and adverbs are typical modifiers. Metaphorically speaking, they are embellishing ornaments that contribute to the linguistic elaboration of nominal and verbal structures, as is characteristic of academic language. They are not necessary and are not as predictable as other constituents of a sentence. To build a complete sentence, a lexical verb is needed along with one or more constituents that satisfy the requirements of that particular verb. Besides these obligatory constituents (arguments), a sentence often contains optional elements (modifiers). From the psycholinguistic literature on ambiguity resolution, it is well known that the human sentence parser finds it easier to process arguments than modifiers (e.g., Clifton, Speer, & Abney, 1991). We expected this to hold for nonambiguous contexts as well.

Features on the morphological level.

Ratio of derived nouns to all nouns (Feature 7). This is the ratio of the number of nouns with derivational suffixes to all nouns in a text. We focused on the derivational process of nominalization because it belongs to the distinctive characteristics of academic language (Hinkel, 2004). Following morphemic rather than whole-word or full-listing theories of lexical representation (Marslen-Wilson, Tyler, Waksler, & Older, 1994), a derived noun is more complex than a simple noun because the parsing of derived polymorphemic words necessitates decomposition, which results in additional processing costs (Solomyak & Marantz, 2010). Therefore, a high ratio of derived nouns to all nouns should increase reading demands. Moreover, a "nominalization allows an extended explanation to be condensed into a complex noun phrase" (Schleppegrell, 2001, p. 443). Therefore, students have to process more ideas per clause when reading texts with nominalizations, and students who are unfamiliar with this linguistic structure may have trouble constructing the underlying meaning (Fang et al., 2006).

Ratio of genitive nouns to all nouns (Feature 8). This is the ratio of the number of nouns with genitive case markers to all nouns in a text. We selected the genitive for different reasons: Compared with the other three cases, the genitive is less frequently used. Moreover, in colloquial German, some functions of the genitive have been taken over by the dative. The genitive therefore falls within the domain of written academic language and is

perceived as an indicator of high education. Not surprisingly, the genitive is acquired relatively late, with the exception of case-marked proper names (Kemp & Bredel, 2008). In addition, the genitive is subject to ongoing processes of language change such as the substitution of the long affix by the short one (des Fluges—des Flugs) and an overuse of the genitive in written texts after causal prepositions, which can be interpreted as an attempt to counteract the expansion of the dative in colloquial German (Szczepaniak, 2014). Against the background of the abovementioned considerations, genitive constructions can be expected to cause some difficulties in reading.

Features on the cohesion level.

Connectors (Feature 9). Connectors are one of the central characteristics of academic language (Dragon, Berendes, Weinert, Heppt, & Stanat, 2015). Overall, in secondary school, it has been argued that having more connectors makes a text easier to comprehend (Breindl & Waßner, 2006). However, "the potential benefits from connectives in text are not the same for all readers and are dependent on knowledge" (Cain & Nash, 2011, p. 439). Good, experienced readers are better able to use connectors to construct the meaning of a text (e.g., Cain & Nash, 2011), whereas poor readers benefit least from them (e.g., Becker & Musan, 2014). Besides, there is evidence that younger children tend to ignore connectors (Dragon et al., 2015). Moreover, some connectors signal demanding semantic relations, so the more they occur, the more complex the text. We focused on these kinds of connectors. We counted adversative and concessive connectors as listed by the Dudenredaktion (2009) and calculated their average number per sentence. Regardless of the language under study, these two groups are the most complex connectors and the last ones to be acquired (see *cumulative conceptual complexity*, Evers-Vermeul & Sanders, 2009).

Pronouns (Feature 10). This feature was measured by the average number of third-person pronouns per sentence except for the neuter form "es" (it), which we left out because of its various nonreferential functions in the grammatical system. Third-person personal pronouns are fundamental for references and belong to the group of referential expressions that create cohesive links within a text. The interpretation of pronominal references is a complex process that demands the integration and evaluation of various (often conflicting) types of information, and texts "are more difficult to comprehend when there is a higher density of pronouns, all else being equal" (Graesser, McNamara, Louwerse, & Cai, 2004, p. 197). Whereas adults use the whole range of grammatical and discourse-related features to resolve the reference between the pronoun and the potential antecedent, children rely first and foremost on deterministic cues such as gender (e.g., Arnold, Brown-Schmidt, & Trueswell, 2007) and pass through development stages in which more cues are gradually considered (e.g., position and grammatical role of the antecedent), requiring more cognitive effort (Klages & Gerwien, 2015). There are a few aspects that make the German pronoun system extremely complex, and consequently, reference tracking can become quite demanding. First, the assignment of a noun to one of the three grammatical gender classes (feminine, masculine, neuter) is to a large extent semantically opaque (cf. die Lösung—sie (feminine) [the solution—it], der Beweis—er (masculine) [the evidence—it], das Ergebnis—es (neuter) [the result—it]). In addition, the German pronoun system is characterized by a comparably large number of pronoun types with different referential capacities and partly overlapping functions (e.g., Bittner & Kühnast, 2012; Bryant & Noschka, 2015).

Statistical Analyses

Our research questions were all related to the systematic complexification assumption. Thus, the question was whether the texts would differ in linguistic complexity by the criteria grade level, school track, or publisher. To test the systematic complexification assumption, we ran two different sets of analyses. First, we ran classification models that involved the whole set of 165 linguistic features. This is the most typical approach used in text classification research. Second, we created multilevel regression models with the 10 selected features, as this is one of the standard approaches used in educational psychology.

Text classification models. The basic idea of text classification is to "classify" a given text (i.e., to assign the text to a predefined group or category). This is done by developing mathematical models to classify texts on the basis of automatically extracted features (in our study, a total of 165 features) from a large collection of text documents. There are two stages in this process. First, in the "learning (or training) phase," relevant features are extracted from texts and are fed into a classification algorithm. The algorithm then "learns" which features (or a combination of features) are characteristic of the texts from a specific category. This results in the creation of a classification model. In the second phase, called the "classification phase," the classification model created during the learning phase is used to assign new texts to the categories. Typically, the evaluation of the performance of a classification model is done by analyzing the percentage of "correctly" classified texts in a set of test documents for which the actual category is known. This is known as the classification accuracy of the model. The higher the accuracy, the better the model is at distinguishing between the different categories. This test set is not used during the training phase, and its purpose is only to evaluate the classification model to test its prediction accuracy for new texts that are not part of the training.

To train the text classification models, we used a popular text classification algorithm called sequential minimal optimization (SMO; Platt, 1998). We used the implementation of this algorithm in the WEKA (Waikato Environment for Knowledge Analysis) machine learning toolkit (Witten & Frank, 2009). Model performance was evaluated in a 10-fold cross-validation setup. In 10-fold cross-validation, the data are divided into 10 similarly sized partitions, and in each fold of the analysis, one partition serves as a test set, whereas the other nine partitions are used to train the model together. This model is then used to classify the data in the test set. The whole procedure is then repeated 10 times so that all data are classified independently of the training sets. The average accuracy of these 10 folds was used to judge the quality of the classification model. For the classification model, we used a subset of the corpus so that all the prediction categories consisted of an equal number of texts in any predictive model. That is, for comparisons between grade levels, all the grade levels were represented equally (same N) in the analysis. The same was true for the comparisons performed between tracks. This was done to eliminate any bias toward the majority class in the classification model. For the publisher-based classifications, the number of texts was

chosen on the basis of the publisher that had the smallest number of texts per category so that better results for one publisher could not be interpreted as being due to the presence of more training examples. The selection of balanced training data was performed with the SpreadSubSample method in WEKA.

To the best of our knowledge, no study has previously compared texts from Grades 5 to 10 or from different school tracks.[4] Therefore, we did not have a cut-off value or any good comparative values that we could use to judge whether any particular accuracy rate was reasonable or not. However, what we could expect in any case were accuracy rates that were statistically significantly higher than chance. Moreover, even without knowing a determined cut-off, it was interesting to see the differences between the classification rates.

Multilevel regression models for the 10 selected features. In addition to the classification modeling strategy—where the linguistic features were taken as a whole to classify the different texts in our sample regarding their respective grade level, school track, or publisher—we analyzed feature-specific differences that were based on the targets of the books. We used two different analytic approaches focusing on differences between grade and track levels for each of the 10 selected features (as well as publisher effects and interaction effects). First, with regard to grade-level- or school-track-specific differences for each of the selected features, multilevel regression models were applied with book as the cluster variable (i.e., the clustering of texts within books was modeled as a random effect) and the specific linguistic marker as the outcome predicted by a single dummy-coded variable. In these analyses, the selected books in this study were treated as a random sample of "typical" books from different publishers, for different tracks, and for different grades. For school track, a single variable (0 = vocational track, 1 = academic track) was used. Concerning the grade-level-specific comparisons, three different dummy-coded variables were used (0 = Grades 5/6, 1 = Grades 7/8; 0 = Grades 7/8, 1 = Grades 9/10; 0 = Grades 5/6, 1 = Grades 9/10), each referring to subsets of the data (e.g., for the first comparison of Grades 5/6 vs. 7/8, all texts from books for Grades 9/10 were excluded). These models were estimated in SAS (SAS Institute, Inc., 2013) with the mixed procedure and robust maximum likelihood estimation (which is available as the EMPIRICAL option of the MIXED procedure) for adjusted standard errors of fixed effect parameters based on the "sandwich estimator" (Huber, 1967).

Second, to address the research question regarding differences between publishers—"controlling" for grade level and school track—semipartial η^2 coefficients for unbalanced designs based on general linear models with Type III sums of squares (Maxwell & Delaney, 2004) were estimated (SAS GLM procedure). Multilevel models were not feasible here because each book would be uniquely identified by a specific dummy or a combination of dummy variables. Thus, for each of the 10 linguistic features, a full model with all three factors (grade level, school track, and publisher) and all interactions (three two-way interactions and one three-way interaction) was specified. In these models, each book ("cluster" in the multilevel models) was represented by a fixed effect of a dummy variable on the basis of a factor or an interaction between factors. Therefore, the estimates and the statistical inferences refer to the specific books in our sample (in contrast to the above-described multilevel models, where books are treated as a random sample of "typical" books, e.g., for academic school tracks). As the total number of texts in each book was part of our sample, the statistical inference can be interpreted as a potential generalization to prior or future editions of these books. In order to provide robust statistical inferences, we used the SAS Glimmix procedure (https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#glimmix_toc.htm) with robust maximum likelihood estimation (EMPIRICAL option) to estimate the *p*–values (from an identical model). However, the Glimmix procedure does not supply semipartial η^2 coefficients (which, therefore, were estimated with the GLM procedure).

Results

Text Classification

We first present text classification models that used the whole set of 165 features. In accordance with our research questions, we start with the grade-level-based classifications, continue with the school-track-based classifications, and end with the publisher-based classifications. For the grade-level-based and school-track-based classification models, a higher classification accuracy rate would be in accordance with the systematic complexification assumption, whereas high accuracy rates for the publisher-based classifications would be at odds with this assumption.

Classifications by grade level. Our first research question was whether the linguistic complexity of the texts increased from Grades 5/6 to Grades 7/8 to Grades 9/10. If the complexity increased, then it should be possible to use classification models to distinguish between the texts used in Grades 5/6, Grades 7/8, and Grades 9/10. To test whether this was really the case, we chose a random sample of 873 texts from each of the three grade categories (5/6, 7/8, 9/10). Using this sample of texts, we trained a classification model using all the 165 features and the SMO classification algorithm. Given that there was an even distribution of all three grade groups in the training data, the random baseline for the classification accuracy was 33%, and the model achieved a classification accuracy of 53.7%. Thus, it offered an improvement of 20.7% over the random baseline ($p < .001$). However, the results imply that most of the texts would be misclassified when taking the baseline into account.

To determine whether the classification accuracy between the grade levels was different for the two school tracks, we split the corpus into two parts (i.e., academic and vocational tracks) and chose an equal number of texts from each grade level and from each school track. This resulted in a sample of 1,236 texts each for the academic track and the vocational track (412 texts per grade category). Then, we trained two grade-level-based classification models with these two training samples. Whereas the classification of texts from the academic track resulted in an average accuracy of 55.7% (baseline 33%), the classification of texts from the vocational track resulted in an average accuracy of 53.4% (baseline 33%). These accuracies were slightly better

[4] A comparison of German texts targeting children versus adults resulted in classification accuracy rates of 90% (Hancke, Vajjala, & Meurers, 2012). However, texts from different grade levels and school tracks are not comparable to texts written for children and adults.

than the model in which the two school tracks were combined, but this did not translate into a substantial increase in real-world terms. The differences in the classification accuracy rates between the three models (the two school tracks together, academic track only, vocational track only) was not statistically significant. To gather further details, we looked at the grade levels separately (see Table 2).

For the three grade levels, classification was most accurate for distinguishing between Grades 5/6 and 9/10 for both the academic-track texts (76.7%, the baseline was 50% because two groups with an even number of texts were compared) and the vocational-track texts (74.2%, baseline: 50%). This led us to conclude that with the feature set, we were able to identify a pattern such that Grades 7/8 were located between Grades 5/6 and 9/10. The accuracy rates that were greater than chance spoke in favor of the systematic complexification assumption. However, there was still room for further improvements concerning the distinction between the three grade categories, irrespective of school track.

Classifications by school track. To address the second question about the complexity of texts from different school tracks, we considered an equal number of texts per school track, resulting in a data set consisting of 1,461 texts per school track (a total of 2,922 texts). The classification accuracy for the academic track versus the vocational track did not differ much at any grade level. It was 76.8% for Grades 5/6, 78% for Grades 7/8, and 77.9% for Grades 9/10. These findings were between 26.8% and 28.0% higher than the random baseline (which was 50% here because there were only two school tracks). Given the fact that we were looking at texts for the same grade levels, the improvement seemed rather good and was in line with the systematic complexification assumption. However, as for the classifications by grade level, we should note that the accuracy rates could be higher.

Classifications by publisher. We next examined the degree of variation between grade levels and school tracks across the various publishers. Because the data were unevenly distributed between publishers for individual grade levels and school tracks, it was not possible to develop predictive models for the grade-level-based classifications for each school track or for the school-track-based classifications for each grade level. This would result in too little data for some publishers (fewer than 50–100 texts per category), which would make it difficult for the predictive models to "learn" anything. Hence, we built two models per publisher—one to perform the grade-level-based classifications (considering both school tracks) and one for the school-track-based classifications (considering all grades). We excluded Publisher D from these models because it contained texts for only one school track. Table 3 (see first line per publisher) shows the classification results for three publishers for the grade-level-based classifications, and Table 4 (see first line per publisher) shows them for the school-track-based classifications. To avoid differences in classification accuracies due to unequal sample sizes (with higher accuracies expected for larger sample sizes), we chose the numbers of texts on the basis of the publisher with the smallest number of texts per category. All the grade-level-based models were built on 167 (randomly selected) texts per category, and the school-track-based models were built on 256 (randomly selected) texts per category.

In the grade-level-based classification (see Table 3), the differences in accuracies between publishers was not statistically significant (55.5%, 52.3%, and 56.7%). However, the results from Table 4 showed clear differences between publishers for the school-track-based classification (66.9%, 77.9% and 77.7%). The

Table 2
Classification Accuracy (Percentage of Correctly Classified Texts) for Two-Way Classification (Baseline: 50%) by Grade Level

Grade level	Texts from the academic track	Texts from the vocational track
5/6 vs. 7/8	67.5%	63.6%
7/8 vs. 9/10	70.1%	70.8%
5/6 vs. 9/10	76.7%	74.2%

Note. A baseline of 50% means that one would expect 50% of the texts to be classified correctly by chance. Results were based on a total of 1,236 texts. 10-fold cross-validation (CV) was used, that is, the classifier was always tested on texts not seen during training.

Table 3
Classification Accuracy (Percentage of Correctly Classified Texts) for Three-Way Classification (Baseline: 33%) by Grade Level

Training data	Test set	Grade-level-based classification accuracy
Publisher A	Publisher A (CV)	55.50%
	Publisher B	44.30%
	Publisher C	37.80%
Publisher B	Publisher B (CV)	52.30%
	Publisher A	43.40%
	Publisher C	46.10%
Publisher C	Publisher C (CV)	56.70%
	Publisher A	40.60%
	Publisher B	44.30%

Note. The tool was trained with texts from one publisher (training data) and tested on the texts from the other publishers (test set), except for the 10-fold cross-validation (CV) cases, where cross-validation was performed on the single publisher data. Publisher D was not included in these analyses because it had no texts for the academic track.

Table 4
Classification Accuracy (Percentage of Correctly Classified Texts) for Two-Way Classification (Baseline: 50%) by School Track

Training data	Test set	School-track-based classification accuracy
Publisher A	Publisher A (CV)	66.90%
	Publisher B	62.90%
	Publisher C	68.70%
Publisher B	Publisher B (CV)	77.90%
	Publisher A	58.60%
	Publisher C	70.30%
Publisher C	Publisher C (CV)	77.70%
	Publisher A	59.90%
	Publisher B	73.82%

Note. The tool was trained with texts from one publisher (training data) and tested on the texts from the other publishers (test set) except for the 10-fold cross-validation (CV) cases, where cross-validation was performed on the single publisher data. Publisher D was not included in these analyses because it had no texts for the academic track.

performance difference between Publishers B and C was not statistically significant, but the classification accuracy for Publisher A was statistically lower than it was for the other two publishers.

To investigate the publisher differences more directly, we trained our tool on one publisher and tested it on the other publishers. For example, the texts from Publisher A served as training data, and the texts from Publisher B as well as the texts from Publisher C served as a test set. If the same (implicit) rationale for a systematic complexification of texts regarding the 165 features in this study were to apply to all publishers, it would not matter which texts from a specific publisher were chosen as the training data. In this case, all classification accuracies reported in Table 3 would be identical (the same applies for Table 4). Table 3 (see the second and third lines per publisher) shows the results for the grade-level-based classification and Table 4 (see the second and third lines per publisher) for the school-track-based classification. The results show that the grade-level-based classification accuracy across publishers was lower *between* publishers (accuracy rates between 37.8% and 46.1%) than *within* publishers (accuracy rates between 52.3% and 56.7%; see Table 3). The school-track-based results show that the classifier trained on Publisher A was better at distinguishing between the school tracks for Publisher C's data than for its own data. This means that Publisher A was not very successful at distinguishing between school tracks and that the linguistic complexity differences that Publisher A had in its texts (accuracy: 66.9%) were even better in Publisher C's texts (accuracy: 68.7%; see Table 4). Overall, this set of analyses provided only limited support for systematic complexification in that the publishers showed differences in how text difficulty varied across school tracks and grades.

Differences in 10 Linguistic Features Between Grade Levels, School Tracks, and Publishers

In our next analytical step, in order to better understand the differences between grade levels, school tracks, and publishers, we created multilevel models for the 10 linguistic features that are particularly important. The intercorrelations and descriptive statistics for these features are depicted in Table 5. The low to moderate correlations show that the features are relatively independent from each other. Moreover, the descriptive statistics are visually represented in Figures 1–10.

With regard to the expected increasing complexity of texts from books developed for higher grades, we estimated multilevel models with each of the 10 selected linguistic features as the outcome and a single dummy-coded variable for the respective comparison of grade levels (Table 6, columns 2–7).

A comparison of the texts from the lowest and highest grades in this study (Grades 9/10 vs. 5/6) showed statistically significantly higher text complexity for Grades 9/10 (i.e., positive regression coefficients) for seven out of the 10 features. All statistically significant differences found in adjacent grade groups (7/8 vs. 5/6, 9/10 vs. 7/8) referred to four out of these seven features (also with positive effects for higher grades) with one exception (for the feature modifier variation, statistically significant differences were found only for Grades 7/8 vs. 5/6). Three features (*word length*, *ratio of genitive nouns to all nouns*, and *ratio of derived nouns to all nouns*) showed significant differences for all grade comparisons.

'The school-track-specific comparisons (*academic track* vs. *vocational track*; columns 8–9 in Table 6) showed statistically significant differences for seven features, indicating higher text complexity in books edited for the *academic track* (i.e., positive regression coefficients).[5] Contrary to our expectations, the feature *third-person personal pronouns* showed a higher occurrence in books edited for the vocational track. It should be noted that, on the one hand, for two of the features with statistically significant differences regarding school track, no statistically significant differences emerged in the grade-level comparisons (*root type-token ratio, third-person personal pronouns*). On the other hand, for each of the three features where no statistically significant school-track differences were found, at least one statistically significant effect showed up in the grade-level-specific comparison. However, for all 10 selected features in this study, at least one statistically significant difference between school tracks or grade levels was detected. For two of the 10 features, *word length* and *ratio of genitive nouns to all nouns,* all differences (grade-level and school-track comparisons) were statistically significant.

In order to investigate publisher-specific book characteristics regarding the selected linguistic features while controlling for general grade-level- and school-track-specific effects, ANOVAs for unbalanced (or nonorthogonal) designs with grade level, school track, and publisher as well as all two- and three-way interactions between the factors were estimated. The results in Table 7 present statistically significant coefficients regarding the explained variance in the full model for all of the 10 features ($.034 \leq \eta^2 \leq .194$). Regarding the additional amount of the total variance explained by a single factor or interaction effect compared with a model without the respective factor or interaction (semipartial η^2), the results showed statistically significant estimates—besides the grade-level and school-track factors—for the publisher factor as well as several interactions involving the publisher factor. These effects indicate "idiosyncrasies" of publishers that may be due to general higher or lower values on the respective feature as in the case of the average word length feature (where none of the interaction effects were statistically significant) or more grade-specific differences between publishers (e.g., the feature *third-person personal pronouns* with a statistically significant Grade Level × Publisher effect but no other publisher effects). For the features *average number of complex nominals per clause* and *adversative and concessive connectors*, all of the publisher-related effects were statistically significant. For each feature, at least one statistically significant effect involving publisher was found.

Discussion

In the present study, we explored the complexity of German geography textbooks for secondary education (Grades 5 to 10) in different school tracks (academic track, vocational track). To our knowledge, this study was the first to explore the systematic complexification assumption using a large data set of German secondary school textbooks. We examined three research questions all related to the linguistic complexity of the texts. In the following, we

[5] A series of identical models based on the subset of data from the three publishers that provided books for both tracks revealed a comparable pattern of statistically significant effects with one additional statistically significant effect for modifier variation ($b = .01$, $p = .022$).

Table 5
Intercorrelations (Pearson) as Well as Descriptive Statistics for the 10 Linguistic Features

Variables	1	2	3	4	5	6	7	8	9	10
1. Average sentence length (in words)	1									
2. Average word length (in syllables)	.26**	1								
3. Average length of longest dependency	.81**	.27**	1							
4. Average number of complex nominals per clause	.37**	.38**	.40**	1						
5. Root type-token ratio	.06*	.13**	.08**	.03	1					
6. Modifier variation	.13**	.20**	.11**	.21**	.01	1				
7. Ratio of derived nouns to all nouns	.18**	.43**	.18**	.18**	−.01	.08	1			
8. Ratio of genitive nouns to all nouns	.18**	.23**	.17**	.22**	.02	.09	.24**	1		
9. Adversative and concessive connectors	.23**	.03	.21**	.10	.05	.27**	.09**	.02	1	
10. Third-person personal pronouns	.02	−.14**	.00	−.20**	−.01	−.08**	−.06*	−.11**	.03	1
M	14.35	1.94	8.72	0.64	8.39	0.25	0.14	0.09	0.21	0.13
SD	4.82	0.23	2.96	0.40	2.20	0.07	0.10	0.06	0.24	0.18

Note. Results were based on a total of 2,928 texts.
* $p < .05$. ** $p < .01$.

will briefly summarize the main results according to our three research questions and will then discuss the results as a whole.

Grade-Level-Based Comparison

Our first research question was related to the grade-level-based classification. We asked whether the linguistic complexity of the texts would be found to increase from grade level to grade level as they would be expected to do according to the systematic complexification assumption.

The classification models for 165 features (see Table 2) showed a grade-level-based classification accuracy of around 75% (Grades 5/6 vs. 9/10, baseline 50%); the multilevel regression estimates (Table 6, columns 2–7) showed significant differences between Grades 5/6 and 9/10 for seven of the 10 features; and the results of an ANOVA (Table 7, column 3) showed that all 10 features were statistically significant predictors of the grade-level factor. Overall, these results indicate that a certain grade-level-based complexification has taken place. This is reassuring because the phase from age 10 to age 16 provides large gains in competence in the comprehension of demanding texts and requires texts of increasing complexity. However, the classification results and the results of the ANOVA showed that the complexification we observed was not all that systematic. There was clearly room for further improvement in correct classification rates. Moreover, we will discuss later how it would be misleading to interpret these results in isolation.

School-Track-Based Comparison

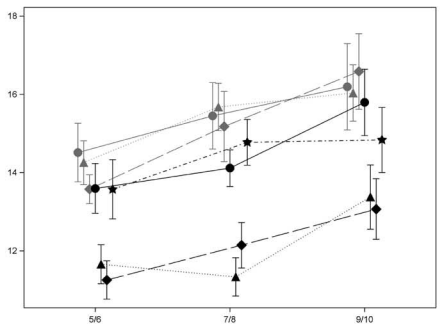
Our second research question addressed differences between school tracks. To be in accordance with the systematic complexification assumption, the complexity of the texts from the academic track would need to be higher in general than the complexity of the texts from the vocational track.

The classification models for 165 features showed a school-track-based classification accuracy of about 77.6% (baseline 50%); the multilevel regression estimates (see Table 6, columns 8–9) showed significant differences for seven of the 10 features; and the results of an ANOVA (Table 7, column 4) showed that all 10 features were statistically significant predictors of the school-track factor. But surprisingly, the occurrence of *third-person pronouns* was higher in the vocational track than in the academic track. Perhaps this can be traced back to more low-frequency synonyms, hypo- and hypernyms,[6] and complex nominal phrases referring to the antecedent instead of third-person pronouns in the texts used in the academic track. Low-frequency synonyms, hypo- and hypernyms, and complex nominal phrases can all be expected to be more difficult than third-person pronouns. Besides, the lower occurrence of third-person pronouns in the academic track might reflect the possibility that good writers are very likely to use fewer pronouns and more nouns to help increase the cohesion of the text when the topic becomes more difficult.

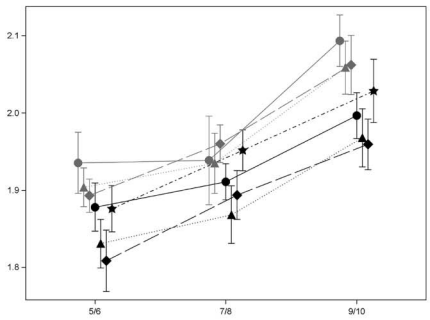
But irrespective of these ideas about why third-person pronouns appeared more in the vocational track, our results indicate that third-person pronouns do not necessarily belong to the important features that should be considered explicitly in geography texts because they appear only rarely. Third-person pronouns occurred only in every 10th sentence (overall mean: 0.13, see Table 5). This can probably be traced back to the fact that "school-based texts do not tend to introduce a referent and then say many things about that referent" (Schleppegrell, 2001, p. 443). Instead, clause by clause, new information is added using the resources of noun phrases (Schleppegrell, 2001). Therefore, third-person pronouns were probably of minor importance for the sample of texts we studied, although the information gain ranking number was not too bad for this feature (41 for the grade-level-based classification and 14 for the school-track-based classification). For other genres or text types, this is probably quite different (e.g., novels or stories are often written in the third person, and the third person is often used in argumentative essays in order to present facts and arguments in an objective tone).

Overall, as already concluded for the grade-level-based comparison, the results of the school-track-based comparison indicate that a certain complexification has taken place. However, again, the classification results and the results of the ANOVA show that there is still room for improvement. Because the competence levels of students in the vocational track were found to differ considerably from the levels of students in the academic track, an important goal should be to

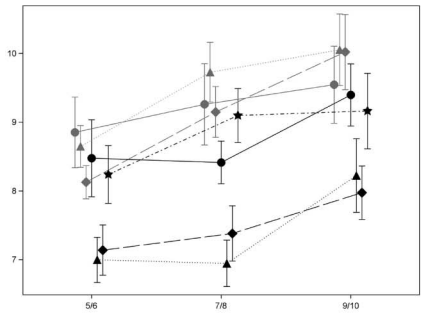
[6] A hyponym is a word or phrase that stands in a "type-of" relationship with its hypernym, which is a superordinate. For example, rain, snow and hail are all hyponyms of precipitation, which is, in turn, a hyponym of weather.



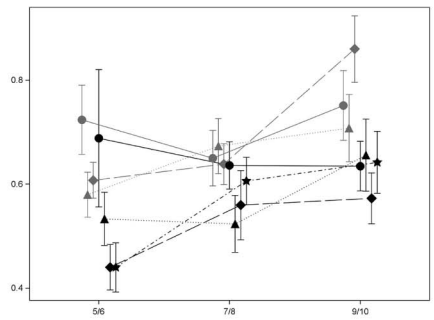
1. Average sentence length in words.



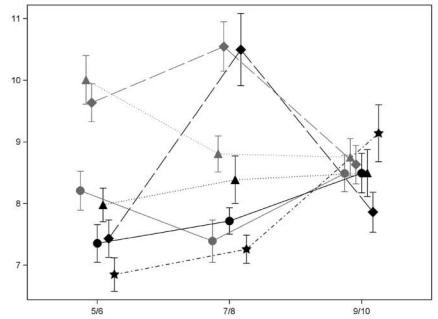
2. Average word length in syllables.



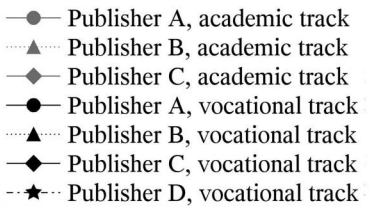
3. Average length of the longest dependency.



4. Average number of complex nominals per clause.



5. Average root type token ratio.



Figures 1–10 Descriptive data for the four different publishers separately for the two school tracks and Grades 5/6, 7/8, and 9/10. Results were based on a total of 2,928 texts. The figures include the 95% confidence intervals around the averages of each feature for all books from the given publisher for the given school track and grade. Whereas the averages can be viewed as population parameters, we included the confidence intervals to indicate the plausible range of values for books produced by the same publisher targeting the same track and grade.

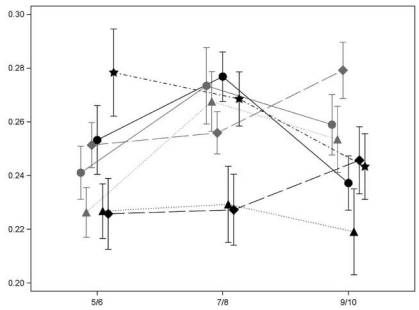
provide additional improvements to students' zone of proximal development.

Publisher-Based Comparison

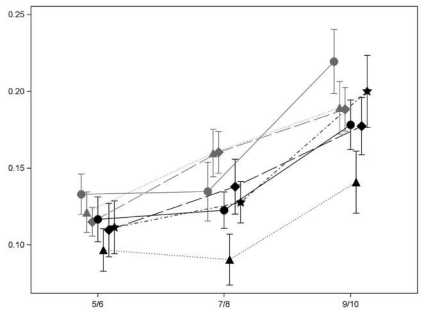
Third, we asked whether the linguistic complexity of the texts differed between publishers. In this case, *no* differences should occur if the texts were perfectly matched to their target readership. The classification results showed that there were meaningful differences between the publishers in the study and that the differences between publishers were higher than the differences within publishers. Upon closer inspection, we identified that one publisher was significantly less successful at varying the texts between school tracks than the other ones. Besides, the ANOVAs showed statistically significant differences between publishers for nine of the 10 features. Overall, these results indicate that the publishers, or at least some of them, were only partially successful in aligning their texts with the intended readership (discussed later).

Interactions Between the Different Factors (Grade Level, School Track, Publisher)

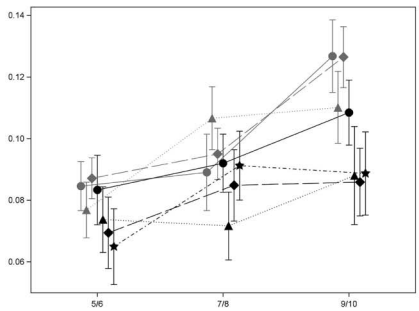
Up to this point, we have considered the results for the different factors separately, but we also examined the extent to which the different factors interacted. In particular, we wanted to determine how the differences we found between grade levels and the dif-



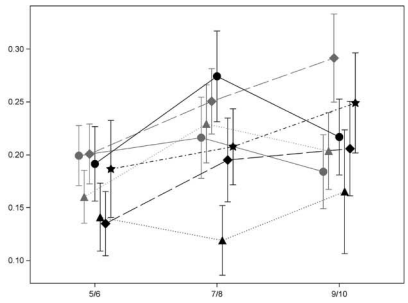
6. Modifier variation.



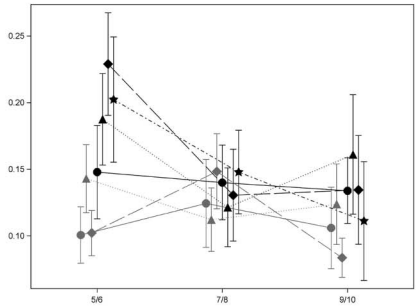
7. Average ratio of derived nouns to all nouns.



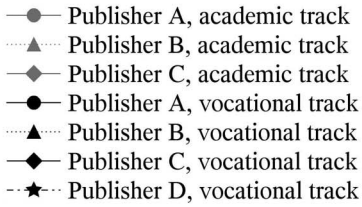
8. Average ratio of genitives to all nouns.



9. Average number of adversative and concessive connectors.



10. Average number of third-person personal pronouns per sentence.



Figures 1–10 (continued)

ferences we found between school tracks interacted with the differences we found between publishers.

Looking at the two-way interaction effects for the 10 features (Table 7, columns 6–8), the results showed four significant interactions for Grade Level × School Track, six significant interactions for Grade Level × Publisher, and eight significant interactions for School Track × Publisher. The three-way interactions (Table 7, column 9) showed six significant interactions for Grade Level × School Track × Publisher. These results indicate that many of the significant results for the isolated factors were moderated by the publisher and thus the author group. This does not speak for a good adjustment to the intended readership, although, overall, the results indicate that a certain complexification from grade level to grade level and between school tracks has taken place. The results indicate that the complexification has been made on the basis of the wisdom of the practice rather than on a thorough and consistent systematic approach. The finding of significant differences between publishers is in line with results from Obermayer (2013), who studied the academic language content in elementary school texts and found large differences between the seven publishers in her study. Such results are not unexpected because it is ultimately individuals (e.g., teachers and academic specialists for the specific subject) who write the textual material presented in school textbooks.

Speaking on behalf of the publishers, it must be noted that appropriately aligning the reading materials with a particular readership is a very complex task (Rog & Burton, 2001), and it requires comprehensive evidence concerning the "theoretical understanding of how reading ability develops over time and the role of text complexity challenge level during different phases of that development" (Williamson et al., 2013, p. 61). For instance, this is particularly difficult because the development of students' ability to read complex texts is not linear (NGACBP & CCSSO, 2010, Appendix A).

Table 6
Grade-Level- and School-Track-Specific Differences for the Selected Linguistic Features as Multilevel Regression Estimates Based on Dummy-Coded Variables for Grade-Level and School-Track Comparisons (With Book as a "Cluster" Variable to Account for the Nested Data Structure)

	Grade level						School track	
	7/8 vs. 5/6		9/10 vs. 7/8		9/10 vs. 5/6		Academic vs. vocational track	
Linguistic feature	*b*	*p*	*b*	*p*	*b*	*p*	*b*	*p*
Average sentence length (in words)	0.93	.053	0.66	.145	1.62	**.004**	1.71	**<.001**
Average word length (in syllables)	0.04	**.017**	0.08	**<.001**	0.13	**<.001**	0.04	**.049**
Average length of longest dependency	0.57	.054	0.42	.135	1.00	**.003**	0.99	**<.001**
Average number of complex nominals per clause	0.03	.206	0.05	.063	0.08	**.035**	0.09	**.003**
Root type-token ratio	−0.11	.418	0.07	.441	−0.04	.465	1.01	**.005**
Modifier variation	0.02	**.006**	−0.01	.095	0.01	.167	0.01	.142
Ratio of derived nouns to all nouns	0.02	**.037**	0.04	**.001**	0.06	**<.001**	0.02	.113
Ratio of genitive nouns to all nouns	0.01	**.002**	0.01	**.041**	0.02	**<.001**	0.01	**.019**
Adversative and concessive connectors	0.04	**.007**	0.00	.490	0.04	**.015**	0.01	.184
Third-person personal pronouns	−0.02	.143	−0.01	.171	−0.02	.087	−0.03	**.004**

Note. The degrees of freedom referring to the cluster level sample size (number of books) were as follows: *df*(7/8 vs. 5/6) = 22, *df*(9/10 vs. 7/8) = 21, *df*(9/10 vs. 5/6) = 21. The *p*-values refer to one-tailed tests. Statistically significant *p*-values are printed in bold. Results were based on a total of 2,928 texts.

Overall, the results provide some very important initial information about the reading demands of secondary textbooks. Moreover, they can be used as a starting point for future analyses that will test the systematic complexification assumption with different sets of data. The combination of educational science and computational linguistics opens up new possibilities, which we have just explored and which can be further elaborated on.

Limitations

The present study offers good insights into the reading demands of secondary school texts. We used advanced modeling techniques to analyze the data and had a large set of texts (a total of 2,928 texts). However, there are several issues that need to be considered when interpreting our results. For instance, our study was restricted to geography texts from one state, and therefore, it is unclear how generalizable the results are. However, text complexity, at least implicitly, can be assumed to be taken into account by publishers with regard to the specific readership on the basis of age and school track. Moreover, to our knowledge, text complexity in school textbooks as measured in our study is usually not examined by applying computational linguistic tools. Therefore, publisher-specific "idiosyncrasies" regarding text complexity features can be assumed to be a general phenomenon. The specific pattern of these idiosyncrasies, however, may be different for school subjects or regions that differ from the ones used in the present study.

Moreover, with our data, it was not possible to differentiate all grades from each other because some publishers in our sample

Table 7
Variance Explained in Linguistic Features (Text Level) by Grade Level (5/6, 7/8, 9/10), School Track (Vocational Track, Academic Track), Publisher (Four Levels), and All Two- and Three-Way Interactions (η^2 and Semipartial η^2 Based on Type III Sums of Squares)

	η^2	Semipartial η^2						
Linguistic feature	Full model	G	S	P	G × S	G × P	S × P	G × S × P
Average sentence length (in words)	**.096**	**.021**	**.044**	**.018**	**.001**	.002	**.010**	.002
Average word length (in syllables)	**.100**	**.058**	**.020**	**.006**	.001	.002	.000	.000
Average length of longest dependency	**.088**	**.020**	**.041**	**.015**	**.002**	.003	**.011**	.002
Average number of complex nominals per clause	**.065**	**.013**	**.013**	**.006**	.001	**.011**	**.004**	**.003**
Root type-token ratio	**.194**	**.005**	**.019**	**.045**	.022	**.063**	**.006**	**.002**
Modifier variation	**.066**	**.006**	**.012**	**.017**	.004	**.013**	**.006**	**.002**
Ratio of derived nouns to all nouns	**.111**	**.071**	**.014**	**.006**	.002	**.005**	**.003**	**.002**
Ratio of genitive nouns to all nouns	**.064**	**.025**	**.011**	**.003**	**.003**	.003	**.003**	**.003**
Adversative and concessive connectors	**.034**	**.006**	**.004**	**.006**	.000	**.004**	**.007**	**.003**
Third-person personal pronouns	**.036**	**.006**	**.008**	.001	**.005**	**.005**	.001	.003

Note. G = Grade Level (5/6, 7/8, 9/10); S = School Track (Vocational Track, Academic Track); P = Publisher (four levels). Bold-faced effects are statistically significant ($p < .05$). The *p*-values were estimated with robust maximum likelihood models (SAS Glimmix procedure with EMPIRICAL option). η^2 and semipartial η^2 (based on Type III sums of squares) were estimated with the SAS GLM procedure. Degrees of Freedom: *df*(G) = 2, *df*(S) = 1, *df*(P) = 3, *df*(G × S) = 2, *df*(G × P) = 6, *df*(S × P) = 2, *df*(G × S × P) = 4. Results were based on a total of 2,928 texts.

published one book for two grades. Therefore, we combined two grades into one variable (Grades 5/6, 7/8, 9/10). With another data set, a more detailed look at different grade levels would probably be possible. Future research could pick books for which this kind of detailed assignment is possible.

We also did not have any reading competence data from students, and therefore, it is just a theoretical assumption that the features we chose actually make the texts more difficult for the intended readership. However, results from studies that performed competence assessments in addition to considering the theoretical assumptions from reading research tend to support such effects.

Furthermore, our analyses focused on linguistic text characteristics, and thus, we did not take into account any characteristics of the readers (e.g., cognitive capabilities, motivation, knowledge, and experience) or tasks (e.g., purpose of reading, intended outcome). Reader and task characteristics could be the focus of other studies, or ideally, all three domains could be brought together in one large research project.

Moreover, in our multilevel analyses, we did not consider the interplay among text characteristics. However, the important text characteristics usually make unique contributions to text complexity. Nevertheless, looking at the interplay would be another step toward expanding the understanding the reading demands in secondary school.

Finally, we considered textbooks that were written at one point in time for one subject in one state in Germany. A recent study explored the differences in textbooks published across several decades in English (Stevens et al., 2015). It would be interesting to pursue this strand of research and collect textbooks across timeframes, subjects, and different parts of Germany. It may give us more insights into the relations between text complexity and grade levels, schools, publishers, and subjects.

Conclusions and Practical Implications

The results of the present study contribute to answering the question of whether language issues and linguistic complexity are taken into account when German textbooks are developed for science subjects. Overall, our results provide only partial support for systematic complexification. They indicate that the geography textbooks we studied were not constructed totally systematically with regard to grade levels and school tracks in terms of a comprehensive set of features of text complexity. It would be worthwhile for publishers and authors of school textbooks to more carefully consider the readability characteristics of the learning materials they provide. To do so, they need a sound understanding of what makes texts more or less complex for students at different age and proficiency levels. At this point, research on reading is needed to provide publishers with sound information supported by strong evidence on the reading competencies and trajectories of different student groups. The information should be so detailed that publishers can decide *when* (i.e., for which age groups or grade levels) to give *what* (i.e., the kind and complexity of linguistic features) to *whom* (i.e., which school track or for good vs. poor readers).

References

Alexander, P. (2012). Reading into the future: Competence for the 21st century. *Educational Psychologist, 47,* 259–280. http://dx.doi.org/10.1080/00461520.2012.722511

Allington, R. L., McCuiston, K., & Billen, M. (2015). What research says about text complexity and learning to read. *The Reading Teacher, 68,* 491–501. http://dx.doi.org/10.1002/trtr.1280

Anderson, R. C., & Davison, A. (1988). Conceptual and empirical bases of readability formulas. In A. Davison & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 23–54). Hillsdale, NJ: Erlbaum. http://dx.doi.org/10.2307/415234

Arnold, J. E., Brown-Schmidt, S., & Trueswell, J. (2007). Children's use of gender and order-of-mention during pronoun comprehension. *Language and Cognitive Processes, 22,* 527–565. http://dx.doi.org/10.1080/016909600845950

Bamberger, R., & Vanecek, E. (1984). *Lesen–Verstehen–Lernen–Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache* [Reading–comprehension–learning–writing. The levels of difficulty of texts in German language]. Wien, Germany: Jugend und Volk. http://dx.doi.org/10.2307/3530491

Barber, H., Vergara, M., & Carreiras, M. (2004). Syllable-frequency effects in visual word recognition: Evidence from ERPs. *Cognitive Neuroscience and Neuropsychology, 15,* 545–548. http://dx.doi.org/10.1097/01.wnr.0000111325.38420.80

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics, 34,* 1–34. http://dx.doi.org/10.1162/coli.2008.34.1.1

Becker, A., & Musan, R. (2014). Leseverstehen von Sachtexten: Wie Schüler Kohärenzrelationen erkennen [Reading comprehension of expository texts: How students recognize coherence relations]. In M. Averintseva-Klisch & C. Peschel (Eds.), *Aspekte der Informationsstruktur für die Schule* (pp. 129–154). Hohengehren, Germany: Schneider.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24,* 63–88. http://dx.doi.org/10.1007/s10648-011-9181-8beschluesse/2012/2012_10_18-Initiative_Sprachfoerderung_Programmskizze.pdf

Bittner, D., & Kühnast, M. (2012). Comprehension of intersentential pronouns in child German and child Bulgarian. *First Language, 32,* 176–204. http://dx.doi.org/10.1086/458890

Bohnet, B., & Kuhn, J. (2012). The best of both worlds: A graph-based completion model for Transition-based parsers. *Proceedings of the 13th Conference of the European chap. of the Association for Computational Linguistics*. Retrieved from http://dl.acm.org/citation.cfm?id=2380828

Breindl, E., & Waßner, U. H. (2006). Syndese vs. Asyndese. Konnektoren und andere Wegweiser für die Interpretation semantischer Relationen in Texten [Syndese vs. Asyndese. Connectors and other sign points for the interpretation of semantic relations in texts]. In H. Blühdorn, E. Breindl & U. H. Waßner (Eds.), *Text–Verstehen. Grammatik und darüber hinaus* [Text–comprehension. Grammar and beyond], (pp. 46–70). Berlin, Germany: Walter de Gruyter. http://dx.doi.org/10.1515/9783110199963

Bryant, D., & Noschka, N. (2015). Personal-und Demonstrativpronomen im Sprachverstehensprozess: Untersuchungen zum Erwerb funktionaler Anapherndistribution bei DaM, DaF und DaZ [Personal and demonstrative pronouns in the process of language comprehension: Investigations on the acquisition of functional anaphor distribution in German as native language, German as foreign language and German as second language]. In H. Klages & G. Pagonis (Eds.), *Linguistisch fundierte Sprachförderung und Sprachdidaktik* (pp. 17–46). Berlin, Germany: Walter de Gruyter. http://dx.doi.org/10.1515/9783110355109.17

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology, 103,* 429–441. http://dx.doi.org/10.1037/a0022824

Chard, D. J., Pikulski, J. J., & Templeton, S. (2000). *From phonemic awareness to fluency: Effective decoding instruction in a research-based reading program*. Boston, MA: Houghton Mifflin Company.

Chen, X., & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. In D. Brunato, F. Dell'Orletta, G. Venturi, T. François, & P. Blache (Eds.), *Proceedings of the workshop on computational linguistics for linguistic complexity* (pp. 113–119). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from http://aclweb.org/anthology/W16-4113.pdf

Clifton, C., Jr., Speer, S., & Abney, S. P. (1991). Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language, 30,* 251–271. http://dx.doi.org/10.1016/0749-596x(91)90006-6

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics, 165,* 97–135. http://dx.doi.org/10.1075/itl.165.2.01col

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research, 69,* 176–183. http://dx.doi.org/10.1080/00220671.1976.10884868

De Clercq, O., & Hoste, V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics, 42,* 457–490. http://dx.doi.org/10.1162/coli_a_00255

Dragon, N., Berendes, K., Weinert, S., Heppt, B., & Stanat, P. (2015). Ignorieren Grundschulkinder Konnektoren? Untersuchung einer bildungssprachlichen Komponente [Do primary school children ignore clause connectors? Investigation of an academic language component]. *Zeitschrift für Erziehungswissenschaft, 18,* 803–825. http://dx.doi.org/10.1007/s11618-015-0640-8

Dudenredaktion. (Ed.), (2009). *Duden–Die Grammatik* [Duden–The grammar] (8. Auflage, Band 4). Mannheim, Germany: Dudenverlag.

Ebner, M., & Schön, S. (2012). Editorial zum Schwerpunktthema Wandel von Lern- und Lehrmaterialien" [Editorial on the key issue "Change of learning and teaching material"]. *Bildungsforschung, 9,* 1–10. Retrieved from http://bildungsforschung.org/index.php/bildungsforschung

Eisenberg, P., Peters, J., Gallmann, P., Fabricius-Hansen, C., Nübling, D., Barz, I., . . . Fiehler, R. (2009). *Duden Bd. 4: Die Grammatik* (8. überarbeitete Auflage). Mannheim, Germany: Bibliographisches Institut AG.

Evers-Vermeul, J., & Sanders, T. (2009). The emergence of Dutch connectives; how cumulative cognitive complexity explains the order of acquisition. *Journal of Child Language, 36,* 829–854. http://dx.doi.org/10.1017/s0305000908009227

Fang, Z. (2016). Text complexity in the U.S. Common Core State Standards: A linguistic critique. *Australian Journal of Language and Literacy, 39,* 195–206. Retrieved from http://www.alea.edu.au/documents/item/1427

Fang, Z., Schleppegrell, M. J., & Cox, B. E. (2006). Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research, 38,* 247–273.

Feng, S., D'Mello, S., & Graesser, A. C. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review, 20,* 586–592. http://dx.doi.org/10.3758/s13423-012-0367-y

Fenk-Oczlon, G., & Fenk, A. (2008). Complexity trade-offs between the subsystems of language. In M. Miestamo, K. Sinnemäki & F. Karlsson (Eds.) *Language complexity: Typology, contact, change* (pp. 43–65). Philadelphia, PA: John Benjamins. http://dx.doi.org/10.1075/slcs.94.05fen

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2005). Weka: A machine learning workbench for data mining. In O. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 1305–1314). New York, NY: Springer. http://dx.doi.org/10.1007/0-387-25465-X_62

Fromkin, V., Rodman, R., & Hyams, N. (2011). *An introduction to language* (9th ed.). Boston, MA: Wadsworth Cengage Learning.

Gamson, D. A., Lu, X., & Eckert, S. A. (2013). Challenging the research base of the Common Core State Standards: A historical reanalysis of text complexity. *Educational Researcher, 42,* 381–391. http://dx.doi.org/10.3102/0013189x13505684

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, MA: MIT Press. Retrieved from http://www2.bcs.rochester.edu/sites/fjaeger/teaching/LabSyntax2006/readings/Gibson_2000.pdf

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115,* 210–229. http://dx.doi.org/10.1086/678293

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223–234. http://dx.doi.org/10.3102/0013189x11413260

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers, 36,* 193–202. http://dx.doi.org/10.3758/bf03195564

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique* [Problems and methods of statistical linguistics]. Dordrecht, the Netherlands: D. Reidel.

Guthrie, J. T., Hoa, A. L. W., Wigfield, A., Tonks, S. M., Humenick, N. M., & Littles, E. (2007). Reading motivation and reading comprehension growth in the later elementary years. *Contemporary Educational Psychology, 32,* 282–313. http://dx.doi.org/10.1016/j.cedpsych.2006.05.004

Halliday, M. A. (1993). Towards a language-based theory of learning. *Linguistics and Education, 5,* 93–116. http://dx.doi.org/10.1016/0898-5898(93)90026-7

Hamp, B., & Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo, & Y. Wilks (Eds.), *Proceedings of the workshop on automatic information extraction and building of lexical semantic resources for NLP applications*. Madrid, Spain: Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W97-0802.pdf

Hancke, J. (2013). *Automatic prediction of CEFR proficiency levels based on linguistic features of learner language* (Master's thesis). University of Tübingen, Tübingen, Germany.

Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for German using lexical, syntactic and morphological features. In M. Kay, & C. Boitet (Eds.), Proceedings of the 24th International Conference on Computational Linguistics (pp. 1063–1080). Mumbai, India: The COLING 2012 Organizing Committee. Retrieved from http://aclweb.org/anthology/C12-1065.pdf

Heister, J., Würzner, K. M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, R., & Kliegl, R. (2011). dlexDB–eine lexikalische Datenbank für die psychologische und linguistische Forschung [dlexDB–a lexical database for psychological and linguistic research]. *Psychologische Rundschau, 62,* 10–20. http://dx.doi.org/10.1026/0033-3042/a000029

Heppt, B., Dragon, N., Berendes, K., Stanat, P., & Weinert, S. (2012). Beherrschung von Bildungssprache bei Kindern im Grundschulalter [Mastery of academic language by children of primary school age]. *Diskurs Kindheits - und Jugendforschung, 3,* 349–356. Retrieved from http://www.ssoar.info/ssoar/bitstream/handle/document/39057/ssoar-disk-2012-3-weinert_et_al-Beherrschung_von_Bildungssprache_bei_Kindern.pdf

Hiebert, E. H., & Pearson, P. D. (2014). Understanding text complexity: Introduction to the special issue. *The Elementary School Journal, 115,* 153–160. http://dx.doi.org/10.1086/678446

Hinkel, E. (2004). *Teaching academic ESL writing: Practical techniques in vocabulary and grammar*. Mahwah, NJ: Erlbaum.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1,* 221–233. Retrieved from http://econ.ucsb.edu/~doug/245a/Papers/Huber%20Behavior%20of%20ML%20Estimates.pdf

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly, 4,* 195–202. http://dx.doi.org/10.2307/3585720

Kemp, R. F., & Bredel, U. (2008). Morphologisch-syntaktische Basisqualifikation [Morphologic-syntactic basic qualification]. In K. Ehlich, U. Bredel, & H. H. Reich (Eds.), *Referenzrahmen zur altersspezifischen Sprachaneignung. Forschungsgrundlagen* [Reference framework for age-specific language acquisition. Research base] (Bildungsforschung Bd. 29/II; pp. 77–102). Berlin, Germany: Bundesministerium für Bildung und Forschung (BMBF).

Kendeou, P., van den Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research & Practice, 29,* 10–16. http://dx.doi.org/10.1111/ldrp.12025

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

Klages, H., & Gerwien, J. (2015). Verstehen anaphorischer Personalpronomina im DaZ- und DaM-Erwerb [Comprehension of anaphoric personal pronouns in German as a second language and in German native language acquisition]. In H. Klages & G. Pagonis (Eds.), *Linguistisch fundierte Sprachförderung und Sprachdidaktik* [Linguistically founded language support and language teaching] (pp. 71–98). Berlin, Germany: Walter de Gruyter. http://dx.doi.org/10.1515/9783110355109.71

Klare, G. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press. http://dx.doi.org/10.1145/344599.344630

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16,* 262–284. http://dx.doi.org/10.1080/09541440340000213

KMK (Kultusministerkonferenz). (2012). *Programmskizze "Bildung durch Sprache und Schrift"* [Programme outlines "Education through spoken and written language"]. Retrieved from http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15,* 92–108. http://dx.doi.org/10.1080/10888438.2011.536130

Lenzner, T. (2014). Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods & Research, 43,* 677–698. http://dx.doi.org/10.1177/0049124113513436

LS–Landesinstitut für Schulentwicklung Stuttgart–Schulbuchzulassung. (2013a). *Zulassungen für allgemeinbildende Gymnasien* [Approval of textbooks for Gymnasium]. Retrieved from http://www.schule-bw.de/service/schulbuchlisten/

LS–Landesinstitut für Schulentwicklung Stuttgart–Schulbuchzulassung. (2013b). *Zulassungen für Haupt- und Werkrealschulen* [Approval of textbooks for Werk- and Werkrealschulen]. Retrieved from http://www.schule-bw.de/service/schulbuchlisten/

Lu, X., Gamson, D. A., & Eckert, S. A. (2014). Lexical difficulty and diversity in American elementary school reading textbooks: Changes over the past century. *International Journal of Corpus Linguistics, 19,* 94–117. http://dx.doi.org/10.1075/ijcl.19.1.04lu

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review, 101,* 3–33. http://dx.doi.org/10.1037/0033-295X.101.1.3

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data. A model comparison perspective. Mahwah, NJ: Erlbaum. Retrieved from http://www.academia.edu/download/30207453/10.1.1.201.8140.pdf

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42,* 381–392. http://dx.doi.org/10.3758/brm.42.2.381

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. H. Ross, B. H. Ross (Eds.), *The psychology of learning and motivation* (pp. 297–384). San Diego, CA: Elsevier Academic Press. http://dx.doi.org/10.1016/s0079-7421(09)51009-2

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In M. Miestamo, K. Sinnemäki & F. Karlsson (Eds.) *Language complexity: Typology, contact, change* (pp. 23–41). Philadelphia, PA: John Benjamins. http://dx.doi.org/10.1075/slcs.94.04mie

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2011). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York, NY: Student Achievement Partners.

NGACBP & CCSSO. (National Governors Association Center for Best Practices & Council of Chief State School Officers). (2010). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects. Appendix A*. Washington, DC: Author. Retrieved from http://www.corestandards.org/the-standards

Nickel, S. (2011). Textschwierigkeit objektivieren: Der Lesbarkeitsindex LIX. Wie schwierig sind Lesetexte in der Alphabetisierung? [Objectifying text difficulty: The readability index LIX. How difficult are reading texts in alphabetization?] *Alfa-Forum, 76,* 30–32.

Nicol, C. C., & Crespo, S. M. (2006). Learning to teach with mathematics textbooks: How preservice teachers interpret and use curriculum materials. *Educational Studies in Mathematics, 62,* 331–355. http://dx.doi.org/10.1007/s10649-006-5423-y

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30,* 555–578. http://dx.doi.org/10.1093/applin/amp044

Obermayer, A. (2013). *Bildungssprache im grafisch designten Schulbuch. Eine Analyse von Schulbüchern des Heimat- und Sachunterrichts* [Academic language in a graphically designed textbook. An analysis of textbooks from social studies instruction]. Bad Heilbrunn, Germany: Julius Klinkhardt.

Pearson, P. D. (2013). Research foundations of the Common Core State Standards in English language arts. In S. Neuman & L. Gambrell (Eds.), *Quality reading instruction in the age of Common Core State Standards* (pp. 237–262). Newark, DE: International Reading Association. http://dx.doi.org/10.1598/0496.17

Perera, K. (1980). The assessment of linguistic difficulty in reading material. *Educational Review, 32,* 151–161. http://dx.doi.org/10.1080/0013191800320204

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In M. Lapata & H. Tou Ng (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 186–195). http://dx.doi.org/10.3115/1613715.1613742

Plakans, L., & Bilki, Z. (2016). Cohesion features in ESL reading: Comparing beginning, intermediate and advanced textbooks. *Reading in a Foreign Language, 28,* 79–100. Retrieved from http://nflrc.hawaii.edu/rfl/April2016/articles/plakans.pdf

Platt, J. (1998). Sequential minimal optimization: A Fast algorithm for training support vector machines. *Tech. Rep. No. MSR-TR-98–14*. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf

Pressley, M. (1998). *Reading instruction that works: The case for balanced teaching*. New York, NY: Guilford Press. http://dx.doi.org/10.5860/choice.36-1108

Pyburn, D., & Pazicni, S. (2014). Applying the multilevel framework of discourse comprehension to evaluate the text characteristics of general

chemistry textbooks. *Journal of Chemical Education, 91,* 778–783. http://dx.doi.org/10.1021/ed500006u

Rafferty, A., & Manning, C. D. (2008). *Parsing three German treebanks: Lexicalized and unlexicalized baselines*. Retrieved from http://dx.doi.org/10.3115/1621401.1621407

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511732942

Rezaee, A. A., & Norouzi, M. H. (2011). Readability formulas and cohesive markers in reading comprehension. *Theory and Practice in Language Studies, 1,* 1005–1010. http://dx.doi.org/10.4304/tpls.1.8.1005-1010

Robison, T., Roden, T., & Szabo, S. (2015). Readability levels show that social studies textbooks are written above grade-level reading. *Journal of Teacher Action Research, 1,* 100–112.

Rog, L. J., & Burton, W. (2001). Matching texts and readers: Leveling early reading materials for assessment and instruction. *The Reading Teacher, 55,* 348–356. Retrieved from http://www.jstor.org/stable/20205061

SAS Institute, Inc. (2013). *SAS/STAT® 13.1 user's guide*. Cary, NC: SAS Institute Inc.

Scheerer-Neumann, G. (1997). Lesen und Leseschwierigkeiten [Reading and reading difficulties]. In F. E. Weinert, N. Birbaumer & C. F. Graumann (Eds.), *Psychologie des Unterrichts und der Schule* [Psychology of teaching and school] (pp. 279–325). Göttingen, Germany: Hogrefe.

Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education, 12,* 431–459. http://dx.doi.org/10.1016/s0898-5898(01)00073-0

Schleppegrell, M. J. (2004/2010). *The language of schooling: A functional linguistics perspective*. Mahwah, NJ: Erlbaum. http://dx.doi.org/10.4324/9781410610317

Schmidt, J. E. (1993). *Die deutsche Substantivgruppe und die Attribuierungskomplikation* [The German noun group and difficulties in attribution]. Tübingen, Germany: Niemeyer (Reihe Germanistische Linguistik 138). http://dx.doi.org/10.1515/9783110958515

Seals, M. P. (2013). *Impact of leveled reading books on the fluency and comprehension levels of first grade students* (Doctoral dissertation). Liberty University, Lynchburg, VA. Retrieved from http://digitalcommons.liberty.edu/cgi/viewcontent.cgi?article=1828&context=doctoral

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). Generating automated text complexity classifications that are aligned with targeted text complexity standards. *ETS Research Report Series, 2010*. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02235.x

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science, 328,* 450–452. http://dx.doi.org/10.1126/science.1182597

Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition: A single-trial correlational MEG study. *Journal of Cognitive Neuroscience, 22,* 2042–2057. http://dx.doi.org/10.1162/jocn.2009.21296

Stevens, R. J., Lu, X., Baker, D. P., Ray, M. N., Eckert, S. A., & Gamson, D. A. (2015). Assessing the cognitive demands of elementary school reading curricula: An analysis of reading text and comprehension tasks from 1910 to 2000. *American Educational Research Journal, 52,* 582–617. http://dx.doi.org/10.3102/0002831215573531

Szczepaniak, R. (2014). Sprachwandel und sprachliche Unsicherheit. Der formale und funktionale Wandel des Genitivs seit dem Frühneuhochdeutschen [Language change and language insecurity. The formal and functional change of genitive since the Early New High German]. In A. Plewnia (Ed.), *Sprachverfall? Dynamik – Wandel – Variation* [Language decline? Dynamic – change – variation] (pp. 33–49). Berlin, Germany: De Gruyter. http://dx.doi.org/10.1515/9783110343007.33

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition, 105,* 300–333. http://dx.doi.org/10.1016/j.cognition.2006.09.011

Urqhart, A. H. (1985). The effect of rhetorical ordering on readability. In A. Anderson & A. H. Urqhart (Eds.), *Reading in a foreign language* (pp. 161–180). London, UK: Longman.

Valencia, S. W., Wixson, K. K., & Pearson, P. D. (2014). Putting text complexity in context. *The Elementary School Journal, 115,* 270–289. http://dx.doi.org/10.1086/678296

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal, 96,* 576–598. Retrieved from http://www.jstor.org/stable/23361717

Vygotsky, L. S. (1978). The interaction between learning and development (M. Lopez-Morillas, Trans.). In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: The development of higher psychological processes* (pp. 79–91). Cambridge, MA: Harvard University Press (Original work 1933–1934). http://dx.doi.org/10.1111/j.1540-4781.2012.01401.x

Westwood, P. S. (2001). *Reading and learning difficulties. Approaches to teaching and assessment*. Camberwell, Australia: Acer Press.

Williamson, G. L., Fitzgerald, J., & Stenner, A. J. (2013). The Common Core State Standards' quantitative text complexity trajectory: Figuring out how much complexity is enough. *Educational Researcher, 42,* 59–69. http://dx.doi.org/10.3102/0013189x12466695

Witten, I. H., & Frank, E. (2009). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann Publishers.

(*Appendices follow*)

Appendix A

The Reading Demands Corpus

Grade	Number of texts		
	Academic track	Vocational track	Total[a]
	Publisher A		
5/6	245	156	
7/8	146	223	1,044
9/10	119	155	
	Publisher B		
5/6	116	127	
7/8	147	70	627
9/10	108	59	
	Publisher C		
5/6	202	136	
7/8	150	58	920
9/10	234	140	
	Publisher D		
5/6	0	115	
7/8	0	164	337
9/10	0	58	
Total[b]	1,467	1,461	2,928

[a] Per publisher. [b] Per school track.

(*Appendices continue*)

Appendix B

Description of the 165 Linguistic Features, Information Gain (IG) for the Classification of Texts Regarding Grade Level and School Track, ANOVA Results (η^2, p-value) for each Linguistic Feature (Text Level) With the Factors Grade Level, School Track, Publisher, and All Two- and Three-Way Interactions

No.	Feature name	Feature description	IG		η^2	p
			Grade-level-based class	School-track-based class		
	I. Syntactic features: Features based on parse trees and dependency graphs					
1	SYN_avgLengthOfClause	Average num. of words per clause	.01901	.02578	.07	<.001
2	SYN_avgSentenceLength	Average num. of words per sentence	.02853	.05629	.10	<.001
3	SYN_avgLengthTUnit	Average num. of words per T-unit. A T-unit is "one main clause plus any subordinate clause or non clausal structure that is attached to or embedded in it" (Hunt, 1970).	.02446	.04778	.08	<.001
4	SYN_sentenceComplexityRatio	Average num. of clauses per sentence	.01007	.01239	.04	<.001
5	SYN_TunitComplexityRatio	Average num. of clauses per T-unit	0	.01172	.03	<.001
6	SYN_complexTunitRatio	Ratio of complex T-units (i.e., T-units containing a dependent clause) to all T-units	0	0	.03	<.001
7	SYN_dependentClauseRatio	Ratio of dependent clauses to clauses	0	.00449	.03	<.001
8	SYN_dependentClausesWith ConjToDependentClauses	Ratio of dependent clauses with conjunction to all dependent clauses	0	.0146	.02	<.001
9	SYN_dependentClausesWithOutConjTo-DependentClauses	Ratio of dependent clauses without conjunction to all dependent clauses	0	0	.01	.019
10	SYN_satzwertigeInfinitiveToDependentClauses	Ratio of infinitive clauses to dependent clauses	0	.00877	.02	<.001
11	SYN_interrogativeClausesToDepCWC	Ratio of interrogative clauses to dependent clauses with conjunction	0	0	.01	.001
12	SYN_conjunctionalClausesToDepCWC	Ratio of conjunctional clauses to dependent clauses with conjunction	0	.00981	.02	<.001
13	SYN_relativeClausesToDepCWC	Ratio of relative clauses to dependent clauses with conjunction	0	.00718	.03	<.001
14	SYN_dependentClausesPerTUnit	Ratio of dependent clauses to T-units	0	.00459	.03	<.001
15	SYN_coordinatePhrasesPerClause	Ratio of coordinate phrases to clauses	0	.00604	.02	<.001
16	SYN_coordinatePhrasesPerTUnit	Ratio of coordinate phrases to T-units	0	.00601	.02	<.001
17	SYN_sentenceCoordinationRatio	Ratio of T-units to sentences	0	.00947	.03	<.001
18	SYN_complexNominalsPerClause	Ratio of complex nominals to clauses	.02241	.02634	.07	<.001
19	SYN_complexNominalsPerTUnit	Ratio of complex nominals to T-units	.02472	.03038	.07	<.001
20	SYN_verbPhrasesPerTUnit	Ratio of verb phrases (VPs) to T-units	.00723	0	.03	<.001
21	SYN_avgParseTreeHeight	Average parse tree height per sentence. The height of a parse tree is the length of the longest path from the root to a leaf.	.02971	.03922	.06	<.001

(*Appendices continue*)

Appendix B (*continued*)

No.	Feature name	Feature description	IG: Grade-level-based class	IG: School-track-based class	η^2	p
22	SYN_averageNPFrequency	Average num. of noun phrases (NPs) per sentence	.01684	.02588	.05	<.001
23	SYN_averageVPFrequency	Average num. of verb phrases (VPs) per sentence	.00706	.00513	.03	<.001
24	SYN_averageVZFrequency	Average num. of to- infinitive phrases (zu-infinitive) per sentence	0	.00758	.01	.001
25	SYN_averagePPFrequency	Average num. of prepositional phrases (PPs) per sentence	.0194	.03404	.07	<.001
26	SYN_averageNPFrequencyPerClause	Average num. of noun phrases (NPs) per clause	.0103	.01269	.03	<.001
27	SYN_averageVPFrequencyPerClause	Average num. of verb phrases (VPs) per clause	0	0	.03	<.001
28	SYN_averageVZFrequencyPerClause	Average num. of to- infinitive phrases (zu-infinitive) per clause	0	.01358	.01	.011
29	SYN_averagePPFrequencyPerClause	Average num. of prepositional phrases (PPs) per clause	.01554	.0154	.05	<.001
30	SYN_averageNPLengthInWords	Average num. of words per noun phrase (NP)	.02028	.03856	.07	<.001
31	SYN_averageVPLengthInWords	Average num. of words per verb phrase (VP)	.007	.01029	.03	<.001
32	SYN_averagePPLengthInWords	Average num. of words per prepositional phrase (PP)	.01572	.01332	.03	<.001
33	SYN_depClausesPerSentence	Average num. of dependent clauses per sentence	0	.00478	.03	<.001
34	SYN_complexTUnitsPerSentence	Average num. of complex T-units (i.e., T-units containing a dependent clause) per sentence	0	0	.03	<.001
35	SYN_coordinatedPhrasesPerSentence	Average num. of coordinated phrases per sentence	0	.0063	.02	<.001
36	SYN_passiveVoiceToSentenceRatio	Ratio of passive voice constructions to sentences	0	0	.02	<.001
37	SYN_passiveVoiceToClauseRatio	Ratio of passive voice constructions to clauses	0	0	.02	<.001
The following three features are based on the idea of *nonterminals in a parse tree*. Nonterminals refer to any part of a parse tree except the leaves (words) that indicate different aspects of the syntactic structure of a sentence (phrase boundaries, part-of-speech tag of a word, etc.).						
38	SYN_avgNumNonTerminalsPerSentence	Average num. of nonterminal nodes per sentence	.03411	.048	.09	<.001
39	SYN_avgNumNonterminalsPerClause	Average num. of nonterminal nodes per clause	.01693	.02064	.06	<.001
40	SYN_avgNumNonterminalsPerWords	Ratio of nonterminal nodes to words	.01641	.0162	.04	<.001
41	SYN_avgNumModifiersPerNP	Average num. of noun phrase (NP) modifiers per NP	.01565	.02708	.09	<.001
42	SYN_avgNumModifiersPerVP	Average num. of verb phrase (VP) modifiers per VP	0	.01195	.03	<.001
43	SYN_longestDependency	Longest dependency out of all dependencies in the document	.01529	.04347	.10	<.001
44	SYN_avgLongestDependencyPerSentence	Average longest dependency per sentence (i.e., on average, for each sentence and for all dependencies in that sentence, how many words does the longest of the dependencies contain)	.02726	.03668	.09	<.001
45	SYN_avgNumDependentsPerVerb	Average num. of dependents per verb (for those verbs that have dependents)	0	0	.02	<.001

(*Appendices continue*)

Appendix B (*continued*)

No.	Feature name	Feature description	IG: Grade-level-based class	IG: School-track-based class	η^2	*p*
46	SYN_avgNumDependentsPerVerbExclMods	Average num. of dependents excluding modifiers per verb (for those verbs that have dependents)	0	0	.02	<.001
47	SYN_avgNumDependentsPerNoun	Average num. of dependents per noun (for those nouns that have dependents)	.0115	.01262	.04	<.001
		II. Lexical Features: Features based on word level information				
48	LEX_textLengthBaseline	Num. of words in a text	0	.03733	.21	<.001
Variations of a ratio of all unique words (types) and all words in a text (tokens) indicating the *lexical density and diversity* of text.						
49	LEX_typeTokenRatio	Num. of types/num. of tokens	0	.01344	.09	<.001
50	LEX_rootTypeTokenRatio	Num. of types/square root of num. of tokens	0	.04521	.19	<.001
51	LEX_correctedTypeTokenRatio	Num. of types/square root of (num. of tokens * 2)	0	.04521	.19	<.001
52	LEX_bilogarithmicTypeTokenRatio	Log of num. of types/log of num. of tokens	0	0	.08	<.001
53	LEX_uberIndex	(Log of num. of tokens) ^ 2/log of (num. of tokens/num. of types)	0	0	.02	<.001
54	LEX_YulesK	A measure of vocabulary richness of a text. If N is the number of tokens, V (N) is the number of types, V (m, N) is the number of words that appeared m times in the text, and m_max is the largest frequency of a word, then, YulesK is defined as: $K = C * (-1/N + \sum_{m}^{m\_max}(V(m, N) * ((m/N) ^ 2)))$. C is a constant set to 10 ^ 4 by Yule.	0	0	.01	<.001
55	LEX_HDD	Hypergeometric Distribution of Diversity described in McCarthy and Jarvis (2010). It represents the probability of choosing *N* number of tokens of a particular type from a sample of a particular size without replacement. For each lexical type, HDD is the probability of encountering any of its tokens in a random sample of 42 words from the text. The HDD for the full text is the sum of such probabilities for all lexical types in the text. 42 is the number chosen in the original McCarthy and Jarvis (2010) paper.	0	.01159	.05	<.001
56	LEX_mtld	Measure of Textual Lexical Diversity. It is calculated as the mean length of word sequences in the text that maintain a Type-Token ratio of .72 (McCarthy and Jarvis, 2010).	0	.03391	.07	<.001
For the definitions below, nouns, adjectives, nonmodal and nonauxiliary verbs, and adverbs are considered lexical items. Various *part-of-speech tag distribution ratios*:						
57	LEX_lexicalWordVariation	Num. of lexical types/num. of lexical tokens	0	0	.04	<.001
58	LEX_lexicalDensity	Num. of lexical tokens/num. of tokens	0	.00576	.01	.006

(*Appendices continue*)

Appendix B (*continued*)

No.	Feature name	Feature description	IG: Grade-level-based class	IG: School-track-based class	η^2	p
59	LEX_verbVariation2	Num. of lexical verb types/num. of lexical tokens	.01285	.01629	.06	<.001
60	LEX_verbVariation1	Num. of lexical verb types/num. of lexical verb tokens	0	0	.02	<.001
61	LEX_squaredVerbVariation	(Num. of lexical verb types) ^ 2/ num. of lexical verbs	.00713	.03043	.17	<.001
62	LEX_correctedVerbVariation	Num. of lexical verb types/((square root of num. of lexical verbs) * 2)	.00713	.03043	.16	<.001
63	LEX_verbTokenRatio	Num. of lexical verb tokens/num. of tokens	.01342	.01513	.06	<.001
64	LEX_seinToVerbRatio	Num. of "sein" verbs/num. of verbs	0	0	.03	<.001
65	LEX_habenToVerbRatio	Num. of "haben" verbs/num. of verbs	0	0	.02	<.001
66	LEX_nounVariation	Num. of lexical noun types/num. of lexical tokens	0	0	.03	<.001
67	LEX_nounTokenRatio	Num. of noun tokens/num. of tokens	0	.00817	.03	<.001
68	LEX_verbNounRatio	Num. of verb tokens/num. of noun tokens	.01002	.02002	.06	<.001
69	LEX_adjectiveVariation	Num. of adjective types/num. of lexical tokens	.0289	.02316	.08	<.001
70	LEX_adverbVariation	Num. of adverb types/num. of lexical tokens	.01224	.0131	.06	<.001
71	LEX_modifierVariation	(Num. of adjective types + num. of adverb types)/num. of lexical tokens	.01136	.00534	.07	<.001
Word length features:						
72	LEX_numSyllablesPerWord	Average num. of syllables per word	.07413	.02476	.10	<.001
73	LEX_avgWordLength	Average num. of characters per word	.07963	.02735	.09	<.001
Features based on Dlex lexical database, related to *word frequencies*:						
74	LEX_dlex_AnnotatedTypeScore	Average dlex annotated score for words in the text that are in dlex database	0	0	.02	<.001
75	LEX_dlex_TypeScore	Average dlex type score for words in the text that are in dlex database	0	0	.02	<.001
76	LEX_dlex_LemmaScore	Average dlex lemma score for words in the text that are in dlex database	0	0	.02	<.001
77	LEX_dlex_LogAnnotatedTypeScore	Average of the log of annotated score from dlex	0	0	.02	<.001
78	LEX_dlex_LogTypeScore	Average of the log of type score from dlex	0	0	.02	<.001
79	LEX_dlex_LogLemmaScore	Average of the log of lemma score from dlex	0	0	.02	<.001
Dlex lexical database was divided into six frequency bands based on log annotated type frequencies. For more information, see Hancke (2013).						
80	LEX_dlex_ratioOfWordsInDlexAnnotated-TypeLogFrequencyBandOne	Ratio of words in the first log frequency band	0	0	.01	.329
81	LEX_dlex_ratioOfWordsInDlexAnnotated-TypeLogFrequencyBandTwo	Ratio of words in the second log frequency band	0	0	.02	<.001
82	LEX_dlex_ratioOfWordsInDlexAnnotated-TypeLogFrequencyBandThree	Ratio of words in the third log frequency band	0	0	.01	.006
83	LEX_dlex_ratioOfWordsInDlexAnnotated-TypeLogFrequencyBandFour	Ratio of words in the fourth log frequency band	0	.00532	.02	.002
84	LEX_dlex_ratioOfWordsInDlexAnnotated-TypeLogFrequencyBandFive	Ratio of words in the fifth log frequency band	0	.00569	.02	<.001
85	LEX_dlex_ratioOfWordsInDlexAnnotated-TypeLogFrequencyBandSix	Ratio of words in the sixth log frequency band	0	0	.01	.238

(*Appendices continue*)

Appendix B (*continued*)

No.	Feature name	Feature description	IG: Grade-level-based class	IG: School-track-based class	η^2	p
86	LEX_dlex_lexTypesNotInDlexRatioSrict	Num. of words not found in dlex/num. of lexical types	.01393	.03725	.08	<.001
87	LEX_dlex_lexTypesIndelexRatio	Num. of words found in dlex/num. of lexical types	.01409	.05531	.10	<.001
Features based on GermaNet, to assess the *semantic properties of words*:						
88	LEX_gnet_avgNumHypernymsPerWord	Num. of hypernyms per word/num. of words from the text that are found in GermaNet	0	.01929	.06	<.001
89	LEX_gnet_avgNumHyponymsPerWord	Num. of hyponyms per word/num. of words from the text that are found in GermaNet	0	.00849	.03	<.001
90	LEX_gnet_avgNumSynsetsPerWord	Num. of synsets per word/num. of words from the text that are found in GermaNet.	0	.02087	.06	<.001
91	LEX_gnet_avgNumLexUnitsPerSynset	Num. of lexical units per synset/num. of synsets	0	0	.01	<.001
92	LEX_gnet_avgNumRelationsPerSynset	Num. of relations/num. of synsets	0	0	.02	<.001
93	LEX_gnet_avgNumFramesPerVerb	Num. of verb frames/num. of verbs from text that are found in GermaNet.	0	.00577	.02	<.001
III. Morphological features: Features primarily based on word suffixes, usage of compounding, case-marking, etc.						
Various *word suffixes*: for all suffixes listed below, the ratio is calculated between the number of occurrences of that suffix in the text and the total number of tokens in the text.						
94	MORPH_istT	Num. of suffix "ist"/num. of tokens	0	0	.02	<.001
95	MORPH_eiT	Num. of suffix "ei"/num. of tokens	0	0	.01	<.001
96	MORPH_lingT	Num. of suffix "ling"/num. of tokens	0	0	.01	.031
97	MORPH_keitT	Num. of suffix "keit"/num of tokens	.01147	.00651	.02	<.001
98	MORPH_atT	Num. of suffix "at"/num. of tokens	0	0	.01	<.001
99	MORPH_werkT	Num. of suffix "werk"/num. of tokens	0	0	.02	<.001
100	MORPH_schaftT	Num. of suffix "schaft"/num. of tokens	0	.00518	.02	<.001
101	MORPH_tumT	Num. of suffix "tum"/num. of tokens	.00649	0	.02	<.001
102	MORPH_enzT	Num. of suffix "enz"/num. of tokens	0	0	.02	<.001
103	MORPH_astT	Num. of suffix "ast"/num. of tokens	0	0	.01	.011
104	MORPH_eurT	Num. of suffix "eur"/num. of tokens	0	0	.01	.002
105	MORPH_itätT	Num. of suffix "ität"/num. of tokens	.00886	.01032	.02	<.001
106	MORPH_urT	Num. of suffix "ur"/num. of tokens	0	.00786	.01	<.001
107	MORPH_heitT	Num. of suffix "heit"/num. of tokens	0	.00474	.01	.022
108	MORPH_nisT	Num. of suffix "nis"/num. of tokens	0	0	.01	.111
109	MORPH_wesenT	Num. of suffix "wesen"/num. of tokens	0	0	.01	<.001
110	MORPH_atorT	Num. of suffix "ator"/num. of tokens	0	0	.01	<.001
111	MORPH_ismusT	Num. of suffix "ismus"/num. of tokens	0	0	.01	.004
112	MORPH_aturT	Num. of suffix "atur"/num. of tokens	0	.01164	.01	<.001
113	MORPH_entT	Num. of suffix "ent"/num. of tokens	.01964	0	.04	<.001
114	MORPH_antT	Num. of suffix "ant"/num. of tokens	0	0	.01	.289
115	MORPH_ariumT	Num. of suffix "arium"/num. of tokens	.00418	0	.01	.510
116	MORPH_ungT	Num. of suffix "ung"/num. of tokens	.05978	.02322	.11	<.001
117	MORPH_ionT	Num. of suffix "ion"/num. of tokens	.02919	.00987	.04	<.001

(*Appendices continue*)

Appendix B (*continued*)

No.	Feature name	Feature description	IG: Grade-level-based class	IG: School-track-based class	η^2	p
Compounding, derived nouns:						
118	MORPH_derivedNounsToNounsRatio	Num. of derived nouns/num. of nouns	.07248	.01035	.11	<.001
119	MORPH_compoundNounsToNounsRatio	Num. of compound nouns/num. of nouns	0	.01395	.03	<.001
120	MORPH_averageCompoundDepth	Sum of compound depths for all compounds/num. of compounds	0	.02741	.02	<.001
Case markers—nominative, accusative, genitive, dative cases. All the ratios below are calculated with the numerator as the number of nouns with the given case, and the denominator as the total number of nouns.						
121	MORPH_NomRatio	Num. of nominative nouns/num. of nouns	0	.01295	.03	<.001
122	MORPH_AccRatio	Num. of accusative nouns/num. of nouns	0	0	.02	<.001
123	MORPH_GenRatio	Num. of genitive nouns/num. of nouns	.02671	.01154	.06	<.001
124	MORPH_DatRatio	Num. of dative nouns/num. of nouns	0	0	.02	<.001
Verb morphology features:						
125	MORPH_avgNumVerbsPerSentence	Num. of verbs/num. of sentences	.00827	.01019	.04	<.001
126	MORPH_finiteVerbRatio	Num. of finite verbs/num. of verbs	0	.00934	.03	<.001
127	MORPH_infinitiveRatio	Num. of infinite verbs/num. of verbs	0	.01002	.02	<.001
128	MORPH_participleVerbRatio	Num. of participle verbs/num. of verbs	0	0	.02	<.001
129	MORPH_imperativeVerbRatio	Num. of imperative verbs/num. of verbs	0	0	.01	.051
130	MORPH_subjunctiveRatio	Num. of subjunctive verbs/num. of verbs	0	.00474	.01	.005
131	MORPH_fstPersonRatio	Num. of first-person verbs/num. of verbs	.01228	0	.04	<.001
132	MORPH_sndPersonRatio	Num. of second-person verbs/num. of verbs	.01399	0	.02	<.001
133	MORPH_thirdPersonRatio	Num. of third-person verbs/num. of verbs	.00784	.00624	.03	<.001
134	MORPH_allModalRatio	Num. of modal verbs/num. of verbs	0	0	.02	<.001
135	MORPH_allAuxRatio	Num. of auxiliary verbs/num. of verbs	0	0	.02	<.001
IV. Discourse features: Features that model the coherence and cohesion in the text by means of relatively shallow linguistics						
Referential features: features based on word overlaps between sentences in a text. Local overlap refers to the overlap between adjacent sentences, and global overlap refers to the overlap between any two sentences in the text.						
136	Ref_localNounOverlap	Average num. of sentences that have an exact noun overlap with the previous sentence (local noun overlap)	0	0	.01	.043
137	Ref_globalNounOverlap	Average num. of all possible sentence pairs in a text that have an overlapping noun (global noun overlap)	0	0	.02	<.001
138	Ref_localArgOverlap	Local argument overlap	0	0	.02	<.001
139	Ref_globalArgOverlap	Global argument overlap	0	.00539	.03	<.001
140	Ref_localStemOverlap	Local stem overlap	0	.00966	.02	<.001
141	Ref_globalStemOverlap	Global stem overlap	0	.01855	.03	<.001
142	Ref_localContentOverlap	Local content word overlap	0	0	.02	<.001
143	Ref_globalContentOverlap	Global content word overlap	0	.0059	.05	<.001
Descriptive features:						
144	Descr_numSentences	Num. of sentences in the document	.00624	.02387	.20	<.001
145	Descr_numParagraphs	Num. of paragraphs in the document	.00805	.00571	.14	<.001
146	Descr_numWords	Num. of words in the document	0	.03733	.21	<.001

(*Appendices continue*)

Appendix B (*continued*)

No.	Feature name	Feature description	IG		η^2	p
			Grade-level-based class	School-track-based class		
147	Eisenberg_AdversativeConcessiveConnectives	Ratio of num. of adversative and concessive connectives in the text to num. of sentences in the text. The list of adversative and concessive connectives was obtained from Eisenberg et al. (2009).	0	.01032	.03	<.001
Referential cohesion: Features related to types of referring expressions, specifically, third person personal pronouns.						
148	PrepDet_avgProportion3rdPPersonalPronounsPerSentence	Average num. of third-person personal pronouns per sentence	0	0	.04	<.001
149	PrepDet_ratio3rdPPersonalPronounsNoun	Ratio of third-person personal pronouns to nouns	0	0	.06	<.001
Features based on how words transition from one syntactic role to another across sentences in the text (Barzilay & Lapata, 2008; Pitler & Nenkova, 2008). Four roles are identified: subject, object, other, nothing. Thus, all possible combinations of *transition pairs* will result in a total of 16 features.						
150	Tran_probSubSub	Probability that the subject of one sentence will be the subject of the next sentence (transition of subject to subject)	0	.02485	.01	.006
151	Tran_probSubObj	Probability of the transition of subject to object	0	0	.01	<.001
152	Tran_probSubOth	Probability of the transition of subject to *other* entity	0	0	.00	.103
153	Tran_probSubNot	Probability of the subject of one sentence not being present in any role in the next sentence	0	0	.07	<.001
154	Tran_probObjSub	Probability of the transition of object to subject	0	.0224	.01	.045
155	Tran_probObjObj	Probability of the transition of object to object	0	0	.00	<.001
156	Tran_probObjOth	Probability of the transition of object to *other* entity	0	0	.01	.005
157	Tran_probObjNot	Probability of object of one sentence not being present in any role in the next sentence	0	0	.03	<.001
158	Tran_probOthSub	Probability of the transition of *other* entity to subject	.02394	.00786	.01	<.001
159	Tran_probOthObj	Probability of the transition of *other* entity to entity	0	.02572	.01	.015
160	Tran_probOthOth	Probability of the transition of *other* entity to *other* entity	0	.02218	.00	.008
161	Tran_probOthNot	Probability of *other* entity in one sentence not being present in any role in the next sentence	.02254	.00931	.03	<.001
162	Tran_probNotSub	Probability of an entity not existing in a sentence becoming the subject in the next sentence	0	.02033	.07	<.001
163	Tran_probNotObj	Probability of an entity not existing in a sentence becoming an object in the next sentence	.00908	.03438	.04	<.001
164	Tran_probNotOth	Probability of an entity not existing in a sentence becoming *other* entity in the next sentence	.00866	.03226	.02	<.001
165	Tran_probNotNot	Probability of an entity not existing in a sentence not being present in the next sentence	.00811	.0095	.11	<.001

Note. The p-values were estimated with robust maximum likelihood models (SAS Glimmix procedure with EMPIRICAL option). η^2 was estimated with the SAS GLM procedure. Results were based on a total of 2,928 texts. Grades: 5/6, 7/8, 9/10; school tracks: vocational track, academic track; publisher: four levels.

Received November 3, 2016
Revision received July 11, 2017
Accepted July 14, 2017 ■