Quantifying word complexity for *Leichte Sprache*: A computational metric and its psycholinguistic validation

Umesh Patil[1], Jesús Calvillo[1], Sol Lago[2], Anne-Kathrin Schumann[1]

[1]t2k GmbH, Dresden, Germany, [2]Goethe University Frankfurt, Germany

umesh.patil@text2knowledge.de j.calvillo@text2knowledge.de

sollago@em.uni-frankfurt.de ak.schumann@text2knowledge.de

Abstract

Leichte Sprache ("Easy Language" or "Easy German") is a strongly simplified version of German geared toward a target group with limited language proficiency. In Germany, public bodies are required to provide information in *Leichte Sprache*. The initial rules for *Leichte Sprache* were developed instinctively by non-linguists, without grounding in linguistic research or cognitive science, and lacked precise criteria for assessing the complexity of linguistic structures (Bock and Pappert, 2023).[1] Although more recent rulebooks have introduced scientifically grounded guidelines for *Leichte Sprache* (Bredel and Maaß, 2016), there remains a need for a computational metric to evaluate language complexity. In response, this paper proposes a model for determining word complexity by training an XGBoost classifier using word-level linguistic features, corpus-level distributional data, frequency information from an in-house *Leichte Sprache* corpus, and human-annotated complexity ratings. We psycholinguistically validate our model by showing that it captures human word recognition times above and beyond traditional word-level predictors. Moreover, we discuss a number of practical applications of our classifier, such as the evaluation of AI-simplified text and detection of CEFR levels of words. To our knowledge, this is one of the first attempts to systematically quantify word complexity in the context of *Leichte Sprache* and to link it directly to real-time word processing.

1 Introduction

1.1 German *Leichte Sprache*

Text Simplification (TS), Complex Word identification (CWI) and Lexical Complexity Prediction (LCP) are popular NLP tasks that have attracted widespread attention due to increased awareness regarding the importance of making information easily accessible to diverse audiences. In the European Union, this awareness has led to legislation, for instance, in the form of the European Accessibility Act (Eur, 2019) and the German *Behindertengleichstellungsgesetz*:[2] In certain scenarios, texts are required to be translated intralingually from standard German into *Leichte Sprache* (Hansen-Schirra et al., 2020). *Leichte Sprache* is a strongly simplified version of German that uses a reduced inventory of German linguistic forms and structures (Maaß, 2020; Maaß et al., 2021; Bock and Pappert, 2023). For illustration, examples (1) and (2) show two versions of the same text: a standard German input text and its translation in *Leichte Sprache*, created by our (*t2k GmbH's*) simplification model. As can be seen, sentences in *Leichte Sprache* are shorter and avoid abstract nouns (such as "Paradigmenwechsel", meaning 'paradigm change') or complex nominalisations, and also avoid complex syntactic structures. Remaining compound nouns are visually split ("Lebens-Bereich" instead of "Lebensbereich", both meaning 'sphere of life') to further facilitate processing.

1. Der mit der Konvention verbundene **Paradigmenwechsel** weg von Fürsorge und Integration hin zur Inklusion betrifft alle Menschen und nahezu jeden Lebensbereich.

2. Alle Menschen mit und ohne Behinderung sollen sich besser um die Menschen mit Behinderung kümmern. Und das in fast jedem **Lebens-Bereich**.

Given the difficulty of generating *Leichte Sprache* translations, an automated complexity met-

[1]The DIN Institute's DIN SPEC 33429:2025-03 provides an overview and discussion of these rules, see Deutsches Institut für Normung (DIN) (2025).

[2]Literally meaning "law to prevent discrimination against people with disabilities", see https://www.gesetze-im-internet.de/bgg/__11.html

Proceedings of the 1st Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts (AI & EL/PL), pages 94–107, June 23, 2025

ric is a vital requirement not only for model development and tuning, but also for output evaluation. It is also required for data curation and quality-checking AI- or human-generated simplifications. Naturally, this metric needs to be psycholinguistically valid to actually measure processing difficulty in human comprehenders. None of these specific requirements has been covered by existing research in text simplification and complex word identification.

1.2 Word Complexity

Word complexity is considered as the perceived difficulty of a word by language users, and is typically assessed from the perspective of a target group with limited language proficiency, consisting of individuals with cognitive impairments, second language learners or children (North et al., 2023). The text simplification process, geared towards such target groups, requires the identification of complex words that can then be substituted in the final simplified text (North et al., 2024).

1.2.1 Word Complexity in NLP

After determining CWI as a stand-alone task in Shardlow (2013), it has been researched through various shared tasks that focused on classifying words or expressions as complex or non-complex, for example CWI–2016 at SemEval (Paetzold and Specia, 2016), CWI–2018 at BEA (Yimam et al., 2018) and ALexS–2020 at SEPLN (Ortiz-Zambrano and Montejo-Ráez, 2020). Later on, CWI has been extended with the Lexical Complexity Prediction (LCP) task (e.g. LCP–2021 at SemEval (Shardlow et al., 2021)) which denotes the complexity of a word or phrase on a continuous scale rather than assigning a binary "complex" or "non-complex" label. Both CWI and LCP tasks have primarily focused on English (in terms of the dataset size and the frequency of being part of such tasks), but at times also included French, German, and Spanish as parallel tasks. German was part of the CWI-2018 shared task, which involved a binary classification task (predicting whether a target word was complex or simple) and a probabilistic classification task (predicting the probability of a target word being complex). The best performing system for German, which was submitted by Kajiwara and Komachi (2018), used a random forest classifier and regressor, and features such as two types of word frequency estimates, and the length of the word or phrase.

So far, the primary resources for German word complexity analysis–both datasets and models–have predominantly come from the CWI-2018 shared task. As a result, research specifically targeting German remains limited, which emphasizes the importance of our current effort.

1.2.2 Word Complexity in Psycholinguistics

A major limitation in current CWI and LCP research is the lack of psycholinguistic validation. The primary objective of simplification is to facilitate readability and comprehension for low language proficiency groups (Shardlow, 2014; Al-Thanyyan and Azmi, 2021). Most CWI and LCP models are trained and evaluated with annotations collected from participants who indicated which words or phrases they found difficult for themselves or for a specific low language proficiency group. These annotations are untimed, however, it is key to evaluate complexity models with real-time comprehension data, ideally collected from participants with low language proficiency.[3]

In the domain of real-time word recognition, psycholinguists seek to identify the variables that determine processing effort. A variety of methods can be used to measure word recognition time, with the most common being lexical decision, word naming, and reading eye-tracking (Ferrand et al., 2011; Kliegl et al., 2010; Kuperman et al., 2013; New et al., 2006). Of specific relevance for this study is the lexical decision task, in which participants see strings of letters and press different keys depending on whether they think that a string corresponds to a word in their language or not. Their response times and accuracy are recorded.

The main properties that affect response times in word recognition tasks are word length (New et al., 2006; Barton et al., 2014), word frequency (Brysbaert et al., 2016, 2018; Kuperman and Van Dyke, 2013; Ferrand et al., 2011; Kliegl et al., 2010), and the size of a word's orthographic neighborhood (Mathey, 2001; Yarkoni et al., 2008; Schröter and Schroeder, 2017; Chen and Mirman, 2012). While these factors are often studied individually, a contribution of the CWI word complexity metric proposed here is that it combines them into a single metric to quantitatively describe word complexity.

[3]Although, in some cases, real-time lexical comprehension data are available from psycholinguistic studies with low language proficiency groups (e.g. the lexical decision task in Pappert and Bock, 2020), they are typically smaller in size than the data required for training CWI or LCP models.

Recent psycholinguistic work also shows that properties of speakers (i.e., their language experience) can affect word recognition (Brysbaert et al., 2016; Keuleers et al., 2015; Kuperman and Van Dyke, 2013; Davies et al., 2017). For example, it has been shown that corpus-based (objective) word frequencies are worse at predicting lexical decision times than subjective ratings, especially with less skilled readers (Kuperman and Van Dyke, 2013). Similarly, frequency effects differ between university students with larger vs. smaller vocabularies, as well as between native vs. non-native (second language) speakers, which suggests that differences in language experience affect word recognition (Keuleers et al., 2015; Cop et al., 2015). Because of this, it is important to create complexity measures that are informed by the different types of text that readers might have access to, including simplified texts. To do this, the CWI model reported in this article incorporated word frequency estimates based on an in-house proprietary dataset of *Leichte Sprache*, which may better capture the type of linguistic input available to second language learners, as well as individuals with lower literacy and/or language impairments.

1.2.3 Application: CEFR Level Detection

Psycholinguistic research shows that simplifying text at different levels of proficiency may help comprehension in second language learners (Crossley et al., 2014; Rets and Rogaten, 2021). This idea aligns naturally with the Common European Framework of Reference for Languages (CEFR), a widely accepted standard that categorizes second language proficiency into six levels (A1–C2). These levels help instructors design materials and courses, and institutions/employers to understand candidates' linguistic proficiency.

Intuitively, the CEFR levels lie between CWI and LCP, classifying language proficiency into distinct levels yet maintaining a progressive continuity of complexity (A1 $<$ A2 $<$ B1 $<$ etc.). Regarding *Leichte Sprache*, we can expect that the vocabulary range at the initial levels (A1, A2) complies with its lexical requirements, while the middle levels (B1, B2) would require more careful assessment, and vocabulary from the advanced levels (C1, C2) is likely to be avoided. While complexity can differ between second language learners and native speakers, one may expect a considerable overlap between these two groups (North and Zampieri, 2023). Moreover, *Leichte Sprache* is also intended to help second language learners with limited proficiency (BMAS, 2014).

Despite the wide acceptance of CEFR levels, the classification of a linguistic unit into a level is usually done manually based on somewhat vague guidelines that can lead to inconsistencies. Some efforts have been made to automatically classify text as per CEFR levels in many languages (e.g., François and Fairon, 2012; Santucci et al., 2020; Velleman and van der Geest, 2014; Branco et al., 2014). However, to our knowledge, only a few studies have been carried out for German (Hancke and Meurers, 2013; Vajjala and Rama, 2018), which were mainly targeted towards classifying bigger segments of text (e.g. essays). This shows the need of having a word-level classifier for German CEFR levels.

1.3 Approach and Summary of Contributions

Language complexity can be conceptualized as both a continuum and a multidimensional construct, spanning various levels of linguistic analysis (e.g. pragmatic, syntactic, lexical). Correspondingly, the task of language simplification needs to be approached at different points along this continuum and across different linguistic levels, depending on the needs of the target audience (Maaß, 2020).

The main objective of this work is to develop a word complexity metric tailored to the requirements of the target groups for *Leichte Sprache*—the "Easy Language target groups" as defined in Maaß (2020), which includes individuals with dyslexia, cognitive disability, dementia, prelingual hearing impairment, aphasia, functional illiteracy, and learners of German as a second language. However, the development and evaluation of such a tool is inherently constrained by the availability of relevant resources. In our case, these resources include: (i) the CWI dataset, annotated considering the target group involving children, language learners, and individuals with reading impairments; (ii) the CEFR wordlists, developed primarily for second language learners; and (iii) the DeveL dataset, compiled using data from young and adult speakers.

Although the metric is constructed using data from diverse target groups, we propose that its quantifiable nature helps progress in mapping word complexity along this complexity continuum. Given that different target groups have distinct complexity requirements, this metric holds potential for broader applicability—not only for *Leichte*

Sprache, but also for other simplified language contexts. This perspective aligns with the "chest of drawers" approach proposed by Maaß (2020), which advocates for differentiated simplification strategies tailored to specific audiences.

The contributions of our article are as follows. First, we train a novel word complexity classifier for German and evaluate it in comparison with earlier work reported in Yimam et al. (2018). Second, since we are interested in CWI in the context of *Leichte Sprache*, we extend traditionally used CWI features using information derived from *Leichte Sprache* data to better account for the specific needs of our target group. Third, we demonstrate the psycholinguistic validity of the model. Fourth, by integrating various features into the model, we effectively produce a unified psycholinguistic measure of word complexity. Finally, we show that the model can be extended to detect CEFR levels of words.

2 CWI Model

Quantifying word complexity is not a straightforward task. Lexical complexity is subjective and it also depends on the context. For the task of CWI we make the simplifying assumptions that a word has a fixed complexity level and that it can be classified as either complex or non-complex.

2.1 Dataset

We used the CWI-2018 dataset, which was released for the second CWI shared task organized as part of the BEA 2018 workshop (Yimam et al., 2018). The dataset consists of offline responses where participants rated single- and multi-word expressions (MWE) on complexity. Participants were shown 5–10 sentences and asked to annotate words or phrases that could pose difficulty in understanding them for a given target reader such as children, language learners or people with reading impairments. The entire dataset consists of English, German, Spanish and French, but we used only the German part. The German dataset was annotated by a mixture of native and non-native speakers (n=23 out of which 12 were native speakers). This led to 7,905 words and MWEs (6,151 training, 795 development and 959 testing instances).

2.2 (Re-)Define Complex Word Label

In the CWI-2018 task a word or MWE was considered complex if at least one of the annotators annotated it as complex. We deem this definition overly simplistic because: (i) an instance could get classified as complex simply because one of the annotators by mistake labeled it complex—it has been observed that CWI can have low inter-annotator agreement (Zampieri et al., 2017); (ii) many proper names such as 'Wikipedia', 'UNICEF' and 'Hannover' (a city in Germany) were rated as complex.

We followed the following procedure to define which words are complex and which are not.

(a) *Complexity Threshold*: We combined all occurrences of a word and calculated the complexity proportion of the word as the ratio of the number of times it was rated as complex to the number of times it received a rating. We defined a threshold for complexity proportion to consider the word as complex or not; for this we again made use of the information in our *Leichte Sprache* dataset.[4] All words with a complexity proportion value above or equal to the threshold were labeled as complex, and below as non-complex.

(b) *Annotation correction*: We experimented with the classification process with an earlier version of the CWI classifier that used heuristics and only a subset of the final features. In the output of the heuristics-based classifier we found that some misclassified instances could have been labeled incorrectly in the dataset. We manually corrected those labels for further use of the dataset. In total 927 labels were manually corrected.

(c) *Proper names are non-complex*: Although, in theory, some proper names can be more complex than others because of their familiarity, pronunciation or cross-linguistic complexity —e.g. 'Berlin' vs. 'Thiruvananthapuram', the capital of the Indian state of Kerala—, we limited the scope of the model to classifying word classes that were not

[4]For determining the threshold we used the *Leichte Sprache* training dataset which consists of input text in standard German and output text in *Leichte Sprache*. For the CWI-2018 dataset we created two classes of words: words that occur in the target texts (the negative class) and words that occur only in the input texts (the positive class). Using the entire range of the difficulty proportion values as the threshold and the binary class labels as the ground truth we computed the True Positive Rate (TPR) and False Positive Rate (FPR) for the positive class. This was done by using the *roc_curve()* function from the `scikit-learn` library (Pedregosa et al., 2018). We chose the optimal threshold to be the one that maximized the difference between the TPR and FPR.

proper names, and we assumed that all proper names are non-complex even if some participants rated them as complex.

(d) *A1-level words are non-complex*: We referred to two wordlists for second language learners of German at the CEFR A1-level. The first wordlist is published by the Goethe-Institut, a globally recognized cultural institute of the Federal Republic of Germany that offers German language courses, and administers German language exams. The second wordlist is published by telc GmbH, an organization known for its language proficiency exams.[5] We defined all words from the dataset that occur in the wordlists to be non-complex even if some participants rated them as complex.

(e) *Drop MWEs*: Since our goal was to capture word-level complexity using lexical and sub-lexical features, we dropped all MWEs.

2.3 Feature Selection & Engineering

For each word we used the following features.

(a) *POS*: Part-of-Speech tag returned by the `spaCy` library employing the medium-sized German language model `de_core_news_md` (Honnibal et al., 2020). We used the Universal POS, a tagset consistent across languages.

(b) *freq_word*: Word frequency estimate returned by the `wordfreq` library (Speer, 2022).

(c) *freq_lemma_word*: The lemma frequency of the word. For calculating the lemma frequency, we first calculated the lemma of each word using two libraries, `spaCy` and `Stanza` (Qi et al., 2020). Based on the POS of the word we picked the best lemma from the two lemmas: `spaCy` lemma for nominal and punctuations (NOUN, PRON, PROPN, NUM and PUNCT) and `Stanza` lemma for the rest (in an experiments for testing the lemmatization accuracy of `spaCy` and `Stanza`, we found that this strategy lead to more accurate final lemmas). To calculate the frequency of the best lemma, we first lemmatized all words from the `wordfreq` library and added the word frequency values for the same lemma entry. We considered these cumulative frequency values to be the frequency estimates of the best lemma.

(d) *length*: Word length in terms of the number of characters.

(e) *freq_LS_target*: The frequency of the word in the entire target part of the *Leichte Sprache* training dataset. The rationale behind adding these frequencies was that the more often a word occurs in the target translation for *Leichte Sprache*, the more likely it is to be a non-complex word.

(f) *freq_proportion_LS*: The proportion of source to target text frequency of the word in the *Leichte Sprache* training dataset. The rationale behind adding this proportion was that if a word occurs very frequently in the source text but very rarely in the target text, it is probably because it is complex.

(g) *is_in_LS_source*: A binary value denoting if the word occurs in the source text of the *Leichte Sprache* training dataset.

(h) *is_in_a1_wordlist*: A binary value denoting if the word occurs in A1 wordlists release by the Goethe-Institute and telc GmbH.

2.4 Training & Evaluation

We combined all three splits—training, development and testing—from the CWI-2018 dataset. After applying the data cleaning procedure described above (see 2.2), we were left with 4,892 unique instances (2,316 complex and 2,576 non-complex). We split this dataset into training (80%) and test (20%) sets. Our dataset included a single categorical variable (`POS`) and multiple continuous features. To ensure consistent handling of the categorical feature, we identified all possible `POS` values across the entire dataset and used that set for one-hot encoding in subsequent experiments. Following the results from Hartmann and dos Santos (2018), who found that a feature-engineered XGBoost model outperformed multiple neural network architectures in the CWI domain, we used the `XGBClassifier` in binary classification mode (Chen and Guestrin, 2016). We carried out five-fold cross-validation to discover an optimal set of hyperparameters. The search drew 5,000 random samples from a predefined distributions of these hyperparameters (see Table A1.1 in Appendix B).

[5]The lists are available at https://www.goethe.de/pro/relaunch/prf/de/A1_SD1_Wortliste_02.pdf from the Goethe-Institut and https://www.telc.net/fileadmin/user_upload/Downloads_Verlag/Einfach_gut/Wortschatzlisten/Einfach_gut_A1_Wortschatzliste_alphabetisch.pdf from telc GmbH.

Upon completion of cross-validation, the best hyperparameters were automatically selected according to the highest macro-averaged F1 score. The best model that emerged from the cross-validation process had an F1 score of 0.85 on the held-out test set. For an informal comparison, our classifier performed much better than the best system at CWI-2018 shared task for German, which had an F1 score of 0.75 (Yimam et al., 2018). Since we used a different split of the dataset for testing and adjusted the definition of labels, the performance of our classifier cannot be compared directly with the ones from the CWI-2018 task; nevertheless, it offers an approximate indication of the classifier's performance.

To leverage all available data, we refitted the best model from the cross-validation process on the entire dataset. This model was then used as the final model for further analysis, validation and applications of the classifier.

3 CWI Model: Validation & Applications

3.1 Lexical Complexity Prediction Using The CWI Model

An `XGBClassifier`, after being trained on the dataset, can also generate probability estimates of a data point being of a given class; in our case the probability of a word being "complex" or "non-complex" based on its features. We assume that the predicted probability of a word being "complex" is a proxy of the complexity of the word (0 denotes minimal complexity and 1 corresponds to maximal complexity). We use these word complexity values for further evaluation and applications of the model.

3.2 Psycholinguistic Validation

To provide a psycholinguistic validation of the complexity estimates generated by the CWI model, we re-analyzed a dataset of 1152 German nouns from the Developmental Lexicon Project (DeveL, Schröter and Schroeder, 2017). The DeveL dataset was created by a large scale developmental study conducted with 800 children from school grades 1–6, as well as 43 younger (20–30 years) and 41 older adults (65–75 years). We focused on the adults, because some predictors in our analysis (word frequency and orthographic neighborhood size) were specific to adult populations—a supplementary analysis with the child group can be found in Appendix C. Because all adults were native German speakers with no history of reading or language impairment, they can't be classified as a primary target group for *Leichte Sprache*. However, given the absence of an equivalent dataset with *Leichte Sprache* users, our analysis provides a first step to validating word simplification methods—which should be further validated with psycholinguistic datasets from other populations once they become available.

All groups completed a lexical decision task and a naming task. We analyzed the noun recognition times from the lexical decision task. The DeveL dataset provides the recognition time estimated for each noun in each speaker group. We predicted that nouns with higher CWI complexity should increase processing difficulty and therefore elicit longer recognition times.

As expected, more complex nouns showed longer recognition times (Figure 1). Next, we sought to identify the effect of CWI complexity above and beyond the linguistic variables previously shown to predict recognition times in the DeveL dataset by Schröter and Schroeder (2017). For this purpose, we ran a mixed-effects linear regression model with CWI complexity as a predictor together with the following variables: noun length, trigram frequency, noun type frequency, and orthographic neighborhood size.[6] Note that with the exception of trigram frequency and orthographic neighborhood size, the other variables were used for training the CWI model. Thus, the estimated effect of word complexity in the statistical model incorporating these variables as covariates should reflect the unique contribution of CWI complexity in explaining recognition times, i.e., the contribution of complexity in explaining variance in the data that is not shared with the other variables.

With the exception of CWI complexity, all other variables were taken from the DeveL dataset (Schröter and Schroeder, 2017). Specifically, noun length was operationalized as the number of letters in each noun. Trigram frequency was based on the childLex corpus (version 0.16, December 2015, see Schroeder et al., 2015) and it was the sum of the frequencies of a sequence of three let-

[6]Following Schröter and Schroeder (2017), we initially included two different frequency estimators: noun type (or form) frequency and noun lemma frequency. However, type and lemma frequency were highly correlated (i.e., above 0.93) and caused high collinearity in the statistical model, as evidenced by variance inflation factors above 10 (James et al., 2013). To address this problem, only noun type frequency was kept in the final model—reported in Table 1.

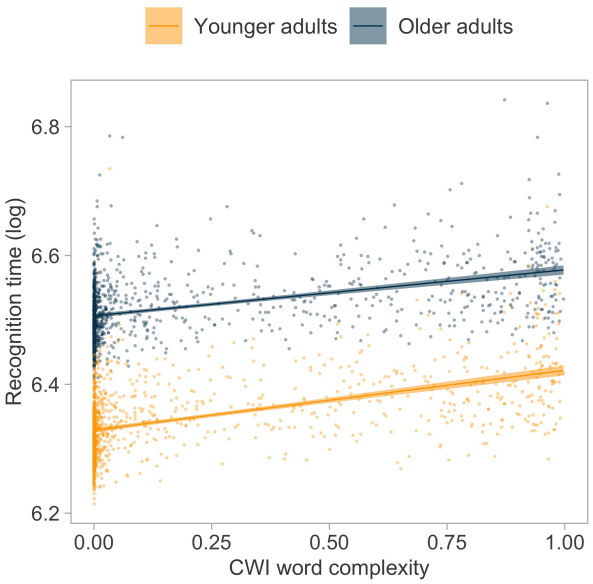

Figure 1: Relationship between CWI complexity and recognition time for the 1152 German nouns in the DeveL dataset (Schröter and Schroeder, 2017). Lines show the effect of word complexity estimated without any covariates in a linear regression model with log-transformed word recognition time as the dependent variable. Ribbons show 95% confidence intervals. Dots correspond to the mean recognition time of each noun in the younger and older adults.

ters within a noun, treating the word beginning and end as separate letters. Noun type (or form) frequency was the number of occurrences of a distinct noun form per million tokens in the DWDS corpus (Digitales Wörterbuch Deutscher Sprache, version 0.4, January 2014; see Geyken, 2007). The orthographic neighborhood size was estimated using the mean Levenshtein Distance from a noun to its 20 closest orthographic neighbors in the DWDS corpus—with this distance being a function of the minimum number of changes, i.e. substitutions, additions and deletions, that are required to turn one word into another (Yarkoni et al., 2008; Schröter and Schroeder, 2017).

All variables mentioned above, together with CWI complexity, were entered in the statistical model as fixed effects nested under the categorical predictor "group" (younger/older adults). This allowed estimating the effect of each variable in the young and old groups separately. Continuous variables were centered. Following Schröter and Schroeder (2017), noun recognition times were log-transformed to account for the right skew of response time distributions. The model included a random intercept by noun, because each noun was seen by both the younger and older group. The data was analyzed using the package lme4 (v.1.1-36; Bates et al., 2015) in R (version 4.5.0, R Core Team, 2025).

The results of the statistical model showed the expected effects of noun length, frequency, and orthographic neighborhood on recognition times (Table 1). Crucially, the effect of CWI complexity was significant after adjusting for these variables: recognition time increased with increasing complexity in both the younger and older adult groups. These results demonstrate that the CWI complexity measure predicted noun recognition difficulties, and that it continued to do so after being adjusted for the effects of frequency, length, and neighborhood size reported in previous research (Schröter and Schroeder, 2017).

3.3 Word Complexity for CEFR Level Detection

In order to address the lack of a German CEFR classifier capable of assigning words to specific levels, we tested the CWI model on this task. The goal was to use the word complexity values to determine the threshold between different CEFR levels. We assume that a word's CEFR level is determined by its complexity value—words from lower CEFR levels should have lower complexity values and complexity values should progressively increase from level A1 (lowest level) to level C2 (highest level). Note that the CEFR framework defines nested levels, meaning that all A1 words are a subset of A2, which in turn is a subset of B1, and so forth. Considering this nested structure, we defined the classification task as follows: for a given word the classifier has to predict the *lowest possible CEFR level* that can be assigned to it. This effectively amounts to first finding out the optimum thresholds for the complexity value that separates the adjacent levels, and then comparing the complexity of a word with the thresholds to determine its level.

To perform this task, we used data from various word lists freely available online that correspond to CEFR levels A1 through C1. Because CEFR levels are nested and also because there are only vague guidelines for defining the levels, these lists initially contained overlapping words. Next, we transformed the lists into mutually exclusive sets by iteratively removing words already assigned to a lower level: first, all words appearing in A1 were removed from the A2 list, then all A2 words were

	Estimate	Std. Error	t-value	p-value
Intercept (younger adults)	6.349	0.001	4723.517	0.000*
Older adults	0.173	0.001	124.187	0.000*
Length: younger adults	0.005	0.001	3.306	0.001*
Length: older adults	0.003	0.001	2.377	0.018*
Trigram frequency: younger adults	0.000	0.000	4.362	0.000*
Trigram frequency: older adults	0.000	0.000	3.945	0.000*
Type frequency: younger adults	–0.013	0.001	–11.768	0.000*
Type frequency: older adults	–0.010	0.001	–9.448	0.000*
Orthographic neighborhood size: younger adults	0.011	0.007	1.656	0.098
Orthographic neighborhood size: older adults	0.023	0.007	3.458	0.001*
CWI complexity: younger adults	0.036	0.005	6.891	0.000*
CWI complexity: older adults	0.020	0.005	3.763	0.000*

Table 1: Output of the statistical model with CWI word complexity as a predictor, together with noun length, trigram frequency, noun type frequency, and orthographic neighborhood size. R model structure: `lmer(log(Noun recognition time) ~ Group / (Length + Trigram frequency + Type frequency + Orthographic neighborhood size + CWI complexity) + (1 | Noun))`. Effects significant at the alpha .05 level are marked with asterisks. Further details of the model: `AIC` = -7781, `BIC` = -7701, `Log Likelihood` = 3905, `Number of observations` = 2304, `Number of groups:Noun` = 1152, `Variance:Noun (Intercept)` = 0.000, `Variance:Residual` = 0.000.

removed from B1, and so on. We did not prepare any list for the C2 level since C2 is essentially the entire lexicon of German; furthermore, we assume that words that are above the B2 level are anyway too difficult for the target group, hence it is sufficient to identify C1–C2 words as being above B2 level. This procedure yielded five distinct lists, each capturing the lowest possible CEFR level for the words in it. From these lists, we extracted a held-out test set of 200 words per level and used the remaining items for training.

An examination of these five wordlists revealed that the A2- and B1-level words share closely related lexical and distributional properties, making it difficult to identify a precise boundary between them. Consequently, we merged A2 and B1 into a single level, thereby reducing the classification task to identifying three thresholds: (1) A1 vs. A2-B1, (2) A2-B1 vs. B2, and (3) B2 vs. C1. We followed the following procedure for determining each threshold: (i) create a balanced set of words from the train split that belonged to all levels, but more for the two adjacent levels on either side of the threshold, (ii) assign them binary class labels based on the side of the threshold they are expected to belong to, (iii) compute the F1 scores of both classes for a range of complexity values as the threshold and the binary class labels as the ground truth, and finally, (iv) select the complexity value that optimizes the performance for the two classes (the point where two F1 scores intersect). We evaluated the performance of these thresholds on the held-out test set.

CEFR levels	F1 score (train)	F1 score (test)
A1	0.78	0.69
A2–B1		0.9
A2–B1	0.68	0.79
B2		0.71
B2	0.56	0.81
C1		0.45

Table 2: Performance of the classification procedure on determining the word complexity thresholds between different CEFR levels. The F1 score (train) is the same for both classes in each group since it is the optimum complexity threshold selected for the two classes.

All F1 scores are listed in Table 2. Based on the F1 scores, the thresholds distinguishing A1 from A2–B1 and A2–B1 from B2 perform well; however, further refinement is needed to improve discrimination between words at the B2 and C1 levels. Overall, these findings indicate that CEFR level classification using word complexity scores effectively identifies words at the A1, A2–B1, and B2 levels, and further show promising potential for distinguishing C1-level words from those at the B2 level.

4 General Discussion

We present a German word complexity classifier and evaluate its performance using existing resources. Given our focus on *Leichte Sprache* ("Easy German"), a strongly simplified version of German for the Easy Language target groups, we complement the standard feature sets for complexity prediction with additional features derived from *Leichte Sprache* datasets. Our results confirm the psycholinguistic validity of the resulting model, and illustrate how the model improves downstream tasks such as text simplification and CEFR-level identification.

Although official guidelines for *Leichte Sprache* do not quantitatively define complexity, making texts accessible critically requires quantitative methods to identify complex words. Our model meets this need by offering a measure of word complexity, validated through word recognition measures in humans, demonstrating its direct impact on readability and comprehensibility. Crucially, once complex words are identified, they can be simplified, which supports both automated text simplification tools and human *Leichte Sprache* translators in tailoring content for less proficient readers. Extending the classifier to map words onto CEFR levels provides additional practical benefits for second language learners of varying proficiency. By aligning text to an appropriate CEFR level, authors and educators can ensure more accessible reading material that is optimally matched to the intended audience.

Limitations

Although the word complexity metric can generate complexity values for all word classes, our psycholinguistic evaluation was restricted to nouns, as the DeveL dataset only contains nouns. It would be informative to extend the evaluation to other word classes, but we are not aware of a dataset with properties comparable to those of DeveL. Furthermore, although our findings suggest that reduced lexical complexity can facilitate reading, this effect is yet to be validated with *Leichte Sprache* users.[7] Again, the absence of suitable datasets currently prevents a direct assessment of whether our results extend to the primary target group of *Leichte Sprache*. Finally, our proposed CEFR classification approach requires additional refinements, particularly for identifying words beyond the B2 level. We see clear potential for improvement, especially by integrating different computational methodologies—such as neural network architectures and word embeddings—and by using larger and/or cleaner datasets.

[7] See Schiffl (2022) who investigated the effects of individual word-level features, such as word length and frequency, comparing a target group of participants with cognitive impairments to a control group. Their study did not find any significant effects for these individual factors.

Data and Code Availability

All non-proprietary data and code used in this paper are publicly available at: https://github.com/text2knowledge/word-complexity-leichtesprache.

Acknowledgments

The work leading to this paper was partially funded by the Federal German Ministry for Labour and Social Affairs through the Civic Innovation Platform[8] and through the MuvAko project,[9] financed by the Sächsische Aufbaubank. We are grateful to our colleague Felix Dittrich for providing technical help and insightful discussion during the development of this work, and also to Johann Seltmann and Tobias Wittig for their contributions to the data collection and curation process. We thank the anonymous reviewers for their insightful comments and constructive feedback, which helped improve the quality of this work.

References

2019. Directive (eu) 2019/882 of the european parliament and of the council on the accessibility requirements for products and services (european accessibility act). https://eur-lex.europa.eu/eli/dir/2019/882/oj. Official Journal of the European Union, L 151, 7.6.2019, pp. 70–115. Accessed: 2025-03-13.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

Jason J. S. Barton, Hashim M. Hanif, Laura Eklinder Björnström, and Charlotte Hills. 2014. The word-length effect in reading: A review. *Cognitive Neuropsychology*, 31(5-6):378–412.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

[8] https://www.knowledgegraph.de/.

[9] https://xr-interaction.com/projects-muvako/.

BMAS. 2014. Leichte Sprache – Ein Ratgeber. https://www.bmas.de/DE/Service/Publikationen/Broschueren/a752-leichte-sprache-ratgeber.html. Accessed: 2025-03-14.

Bettina M. Bock and Sandra Pappert. 2023. *Leichte Sprache, Einfache Sprache, verständliche Sprache*. Narr, Tübingen.

António Branco, Joao Rodrigues, Francisco Costa, Joao Silva, and Rui Vaz. 2014. Rolling out text categorization for language learning assessment supported by language technology. In *Computational Processing of the Portuguese Language: 11th International Conference, PROPOR 2014, São Carlos/SP, Brazil, October 6-8, 2014. Proceedings 11*, pages 256–261. Springer.

U. Bredel and C. Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen ?Orientierung für die Praxis*. Duden - Ratgeber. Duden.

Marc Brysbaert, Paweł Mandera, and Emmanuel Keuleers. 2018. The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, 27(1):45–50. Publisher: SAGE Publications Inc.

Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, 42(3):441–458. Place: US Publisher: American Psychological Association.

Qi Chen and Daniel Mirman. 2012. Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, 119(2):417–430. Place: US Publisher: American Psychological Association.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Uschi Cop, Emmanuel Keuleers, Denis Drieghe, and Wouter Duyck. 2015. Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review*, 22(5):1216–1234.

Scott A. Crossley, H.S. Yang, and Danielle McNamara. 2014. What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26:92–113.

Rob A. I. Davies, Ruth Arnell, Julia M. H. Birchenough, Debbie Grimmond, and Sam Houlson. 2017. Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 43(8):1298–1338.

Deutsches Institut für Normung (DIN). 2025. DIN SPEC 33429:2025-03 – Empfehlungen für Deutsche Leichte Sprache. Technische Regel [NEU], PAS-Verfahren. 60 pages. Original language: German. Accessible PDF available. English title: *Guidance for German Easy Language*.

Ludovic Ferrand, Marc Brysbaert, Emmanuel Keuleers, Boris New, Patrick Bonin, Alain Méot, Maria Augustinova, and Christophe Pallier. 2011. Comparing Word Processing Times in Naming, Lexical Decision, and Progressive Demasking: Evidence from Chronolex. *Frontiers in Psychology*, 2. Publisher: Frontiers.

Thomas François and Cédrick Fairon. 2012. An "AI readability" formula for french as a foreign language. In *Proceedings of the 2012 joint conference on empirical methods in Natural Language Processing and computational natural language learning*, pages 466–477.

Alexander Geyken. 2007. The DWDS Corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum, editor, *Idioms and collocations: Corpus-based linguistics and lexicographic studies*, pages 23–41. Continuum, New York, NY.

Julia Hancke and Detmar Meurers. 2013. Exploring cefr classification for german based on rich linguistic modeling. *Learner Corpus Research*, pages 54–56.

S. Hansen-Schirra, W. Bisang, A. Nagels, S. Gutermuth, J. Fuchs, L. Borghardt, S. Deilen, A.-K. Gros, L. Schiffl, and J. Sommer. 2020. Intralingual translation into easy language – or how to reduce cognitive processing costs. In S. Hansen-Schirra and C. Maaß, editors, *Easy Language Research: Text and User Perspectives*, pages 197–225. Frank & Timme, Berlin.

Nathan Hartmann and Leandro Borges dos Santos. 2018. NILC at CWI 2018: Exploring feature engineering and feature learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 335–340, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: with Applications in R*, second edition. Springer Texts in Statistics. Springer.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199, New Orleans, Louisiana. Association for Computational Linguistics.

Emmanuel Keuleers, Michaël Stevens, Paweł Mandera, and Marc Brysbaert. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, 68(8):1665–1692.

Reinhold Kliegl, Michael E. J. Masson, and Eike M. Richter. 2010. A linear mixed model analysis of masked repetition priming. *Visual Cognition*, 18(5):655–681.

Victor Kuperman, Denis Drieghe, Emmanuel Keuleers, and Marc Brysbaert. 2013. How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies. *Quarterly Journal of Experimental Psychology*, 66(3):563–580. Publisher: SAGE Publications.

Victor Kuperman and Julie A. Van Dyke. 2013. Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, 39(3):802–823. Place: US Publisher: American Psychological Association.

Christiane Maaß. 2020. *Easy Language – Plain Language – Easy Language Plus*. Frank & Timme, Berlin.

Christiane Maaß, Isabel Rink, and Silvia Hansen-Schirra. 2021. Easy language in germany. In Ulla Vanhatalo Camilla Lindholm, editor, *Handbook of Easy Languages in Europe*, pages 191–218. Frank & Timme, Berlin.

S. Mathey. 2001. The influence of visualization of orthography on the recognition of written words. *Canadian Journal of Experimental Psychology = Revue Canadienne De Psychologie Experimentale*, 55(1):1–23.

Boris New, Ludovic ferrand, Christophe pallier, and Marc brysbaert. 2006. Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, 13(1):45–52.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. MultiLS: An end-to-end lexical simplification framework. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 1–11, Miami, Florida, USA. Association for Computational Linguistics.

Kai North and Marcos Zampieri. 2023. Features of lexical complexity: insights from l1 and l2 speakers. *Frontiers in Artificial Intelligence*, 6.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

Jenny A. Ortiz-Zambrano and Arturo Montejo-Ráez. 2020. Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN. In *Proceedings of ALexS 2020: First Workshop on Lexical Analysis at SEPLN*, volume 2664 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Sandra Pappert and Bettina M. Bock. 2020. Easy-to-read german put to the test: Do adults with intellectual disability or functional illiteracy benefit from compound segmentation? *Reading and Writing*, 33(5):1105–1131.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine learning in Python. *Preprint*, arXiv:1201.0490.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *Preprint*, arXiv:2003.07082.

Irina Rets and Jekaterina Rogaten. 2021. To simplify or not? facilitating english l2 users' comprehension and processing of open educational resources in english using text simplification. *Journal of Computer Assisted Learning*, 37(3):705–717.

Valentino Santucci, Filippo Santarelli, Luciana Forti, and Stefania Spina. 2020. Automatic classification of text complexity. *Applied Sciences*, 10(20):7285.

Laura Schiffl. 2022. *Lexikalische Komplexität in der Leichten Sprache: Effekte von Länge, Frequenz und Wiederholung auf die visuelle Wortverarbeitung einer heterogenen Zielgruppe*. PhD dissertation, Johannes Gutenberg-Universität Mainz, Mainz.

Sascha Schroeder, Kay-Michael Würzner, Julian Heister, Alexander Geyken, and Reinhold Kliegl. 2015. childLex—Eine lexikalische Datenbank zur Schriftsprache für Kinder im Deutschen. [childLex—A Lexical Database for Print Language for Children in German.]. *Psychologische Rundschau*, 66(3):155–165. Place: Germany Publisher: Hogrefe Verlag GmbH & Co. KG.

Pauline Schröter and Sascha Schroeder. 2017. The Developmental Lexicon Project: A behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods*, 49(6):2183–2203.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, 4(1).

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

R Development Core Team. 2025. R: A Language and Environment for Statistical Computing.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

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

Tal Yarkoni, David Balota, and Melvin Yap. 2008. Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex word identification: Challenges in data annotation and system performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 59–63, Taipei, Taiwan. Asian Federation of Natural Language Processing.

A Sustainability Statement

All model development, training, and evaluation were conducted on an `Apple M2` laptop (8 cores), yielding minimal carbon impact beyond ordinary laptop use. Each training run, including the hyperparameter optimization, completed in under 30 minutes.

B `XGBClassifier`: Hyperparameter Space

Hyperparameter	Distribution
`classifier__n_estimators`	$\mathcal{U}\{100, 500\}$
`classifier__max_depth`	$\mathcal{U}\{5, 12\}$
`classifier__learning_rate`	$\mathcal{U}[0.2, 0.5]$
`classifier__subsample`	$\mathcal{U}[0.75, 1]$
`classifier__colsample_bytree`	$\mathcal{U}[0.6, 1]$

Table A1.1: The hyperparameter space used for drawing 5,000 random samples during the five-fold cross-validation of `XGBClassifier`.

C DeveL: Supplementary Analysis

This appendix reports the supplementary analysis of the lexical decision child dataset in DeveL, which includes recognition times from 1152 German nouns collected from 800 children from school grades 1–6 (Schröter and Schroeder, 2017). As shown in Figure A2.1, the noun recognition times from children also showed a positive relationship with the complexity measure generated by the CWI model: more complex nouns elicited longer recognition times.

The statistical analysis of the child data was performed separately from the adults, in order to use co-predictors for the CWI complexity measure that were appropriate for children. As with the adult analysis, we sought to identify the effect of the complexity measure above and beyond the linguistic variables previously shown to predict recognition times in the DeveL dataset by Schröter and Schroeder (2017). For this purpose, we ran a linear regression model with CWI complexity as a predictor together with the following variables: noun length, trigram frequency, noun type frequency, noun lemma frequency, and orthographic neighborhood size.

The predictors noun length and trigram frequency were identical to those used in the analysis of the adult groups. Noun length was operationalized as the number of letters in each noun and trigram frequency was the sum of the frequencies of a sequence of three letters within a noun, treating the word beginning and end as separate letters. But in contrast with the adult groups, the type frequency and lemma frequency predictors, as well as the orthographic neighborhood size predictor, were based on the childLex corpus, which is derived from a set of ten million tokens drawn from 500 popular German children's books (version 0.16, December 2015, see Schroeder et al., 2015). This allowed using frequency estimates that are more reflective of the lexicon of children at earlier stages of reading development.

The dependent measure in the model was the recognition time estimated for each noun in the child group. We predicted that nouns with higher CWI complexity should increase processing difficulty and therefore elicit longer recognition times.

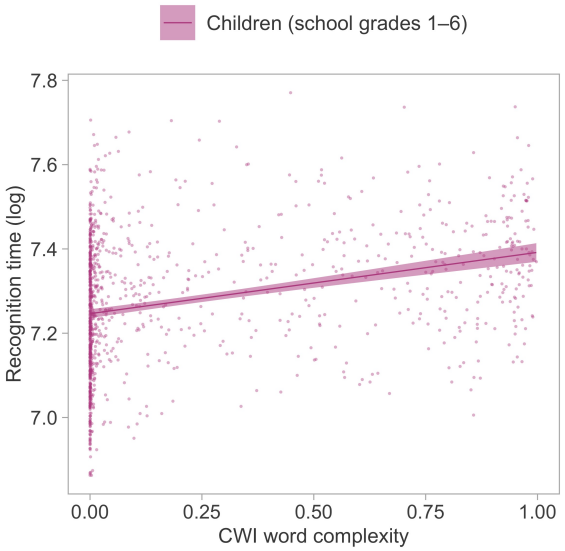

Figure A2.1: Relationship between CWI complexity and recognition time for the 1152 German nouns in the child DeveL dataset (Schröter and Schroeder, 2017). Lines show the effect of word complexity estimated without any covariates in a linear regression model with log-transformed word recognition time as the dependent variable. Ribbons show 95% confidence intervals. Dots correspond to the mean recognition time of each noun in the child group.

The results of the statistical model showed the expected effects of noun length, frequency, and orthographic neighborhood on recognition times (Table A2.1). Crucially, the effect of CWI complexity was significant after adjusting for these variables: recognition time increased with increasing complexity. These results demonstrate that the CWI

	Estimate	Std. Error	t-value	p-value
Intercept (child group)	7.279	0.004	1794.832	0.000*
Length: child	0.045	0.005	8.840	0.000*
Trigram frequency: child	–0.000	0.000	–8.713	0.000*
Type frequency: child	–0.023	0.003	–7.129	0.000*
Orthographic neighborhood size: child	–0.035	0.013	–2.640	0.008 *
CWI complexity: child	0.041	0.015	2.790	0.005*

Table A2.1: Output of the statistical model in the child data. The model used CWI word complexity as a predictor, together with noun length, trigram frequency, noun type frequency, and orthographic neighborhood size. R model structure: `lm(log(Noun recognition time) ~ Length + Trigram frequency + Type frequency + Orthographic neighborhood size + CWI complexity)`. Effects significant at the alpha .05 level are marked with asterisks. Further details of the model: `R`$^2 = 0.26$, `Adjusted R`$^2 = 0.25$, `Number of observations` $= 1152$.

complexity measure predicted noun recognition difficulties in children from different stages of reading development, and that it continued to do so after being adjusted for the effects of frequency, length, and neighborhood size reported in previous research (Schröter and Schroeder, 2017).