German Grammar Profile for Learners: Pedagogical Feature Definition and Automated Extraction

Denise Löfflad[1,3], Benedikt Beuttler[1,2,3], Detmar Meurers[1,3],
[1]Leibniz-Institut für Wissensmedien Tübingen, Germany,
[2]University of Education Ludwigsburg, Germany
[3]LEAD Graduate School and Research Network
d.loefflad@iwm-tuebingen.de
d.meurers@iwm-tuebingen.de
benedikt.beuttler@ph-ludwigsburg.de

Abstract

We introduce the German Grammar Profile (GGP), a system of 150 criterial grammatical features designed to support pedagogically meaningful analysis of reading material for learners of German as a second language. Drawing on the functional grammar framework of the Österreichisches Sprachdiplom (ÖSD), the GGP defines grammatical constructions across CEFR levels A1 to B2.

We also present the corresponding extraction system *Pedagogically Oriented Linguistic Feature Extraction* (PALME, *Pädagogisch Ausgerichtete Linguistische Merkmalsextraktion*), which automatically identifies a growing subset of these features in authentic texts. PALME combines standard Natural Language Processing (NLP) tools with efficient finite state rule based processing to annotate the criterial features and is integrated into a web platform we developed, where users can upload or generate texts and receive visual feedback on the detected criterial grammatical features. Evaluation of a growing subset of currently 33 features shows high precision and recall of the approach.

Ongoing work focuses on extending the coverage to the full criterial feature set. By automatically identifying the features for authentic learner data, our approach will make it possible to empirically validate the expert-defined features of the ÖSD *Profile Deutsch* as the central reference spelling out the CEFR for German. In future work, we plan to extend our tool development by integrating the approach with large language models for level-aware text simplification.

1 Introduction

For real-life foreign language teaching and learning – whether in language classrooms or language testing – specific lexical and grammatical material is relevant to support the acquisition of communicative functions and linguistic structures. The functionally driven Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) is widely recognized in Europe as the standard for informing language pedagogy and applied linguistics (Wisniewski et al., 2013; Yancey et al., 2023; Ribeiro-Flucht et al., 2024). One of its strengths is its language-agnostic design: instead of relying on language-specific grammar or syntax, the CEFR uses abstract "can-do" descriptors to define abilities across six levels (A1–C2), which enables its use across all European languages. However, the lack of detailed linguistic specifications also poses challenges for practical implementation, as the CEFR alone cannot fully support proficiency assessment, readability analysis, or automated classification. Due to the lack of language-specific descriptors, teachers and testing institutions need to determine appropriate linguistic markers and align them with grammar and vocabulary at each level.
The quality of available language input plays a pivotal role in Second Language Acquisition (SLA), with reading being a crucial source of this input. Effective reading materials are those that challenge learners just beyond their current proficiency level, a concept often referred to as the Zone of Proximal Development (Vygotsky, 2012) or i+1 input (Krashen, 1985). To enhance motivation, materials should also be engaging, personally relevant, and aligned with learners' current interests. Finding such texts poses immense challenges to teachers, especially when teaching heterogeneous groups, as it is time consuming to find and label potentially appropriate texts.

To address this, Automatic Readability Assessment (ARA) aims to predict the readability of a

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 212–223, 2025

text for a pre-defined target group and can support teachers in finding appropriate texts faster. Different approaches to ARA have been implemented in the past, including approaches based on surface-based readability formulas (Collins-Thompson, 2014; Kincaid et al., 1975; Björnsson, 1983), using neural networks (Weiss et al., 2021), or using individual linguistic measures such as word frequencies (Chen and Meurers, 2016). However, there is only little data available in languages other than English, and only few studies have performed ARA for German L2 data (for exceptions, see Weiss et al., 2021; Weiss and Meurers, 2021, 2022). Given the nature of the available data, these readability classifications do not correspond to the commonly used CEFR levels, making them unsuitable for learning environments where teachers rely on the CEFR scale for guidance.

This issue is further intensified by traditional machine learning approaches, which often rely on linguistic complexity measures derived from SLA research. While these measures provide valuable insights into the linguistic domains of the data, they are not always intuitive for teachers or learners. For instance, a measure might calculate the standard deviation of SUBTLEX-DE Logarithmic Word Frequency for lexical word types (Brysbaert et al., 2011) - a highly interesting linguistic indicator, but one that lacks accessibility. This lack of clarity can make it challenging for teachers to understand how an ARA algorithm categorizes texts into specific levels. Moreover, abstract or technical linguistic measures provide limited practical support for teachers in adapting or handling texts effectively for language teaching purposes.

In recent years, the use of criterial features, i.e. the identification of distinguishing properties at the level of grammar and the lexicon based on CEFR's functional descriptors, has become more popular (Gaillat et al., 2022; Hawkins and Buttery, 2010). The descriptors used for the CEFR levels are usually formulated as "Can-Do statements" (e.g. "The learner can understand the key concepts of a text.") which are not directly translatable to measurable features.

Despite the growing interest in criterial features, an important gap remains in the current research landscape. For English, the best known resource in this domain is the English Grammar Profile (EGP) (O'Keeffe and Mark, 2022), however, no comparable set of criterial features currently exists for German. This makes it difficult for teachers and assessment tools to link grammatical phenomena to CEFR levels in a systematic and data-driven way. Addressing this gap is essential for developing ARA tools that are both theoretically grounded and practically useful for German L2 education.

We present PALME, a web application that automatically detects German criterial features and allows teachers and students to quickly see potentially difficult grammatical constructs in a text. We furthermore present a German Grammar Profile (GGP) including 150 criterial features based on the CEFR descriptors and the *funktionale Grammatik* (*functional grammar* framework) as defined by the *Österreichisches Sprachdiplom Deutsch* (ÖSD).

2 Related Work

2.1 Automatic Readability Assessment

Text readability refers to degree of ease with which a text can be understood. The most direct, and most used, approach for ARA is to assign a text with a single readability score representing the overall level of comprehensibility. Such scores are so-called readability formulas, and typically rely on easily computable surface-level textual features, such as sentence length (measured in number of words) or word length (measured in number of syllables or number of characters) (DuBay, William H., 2004). Readability formulas such as the Flesch-Kincaid formula (Kincaid et al., 1975) or the LIX index (Björnsson, 1983) have been widely and successfully applied across various domains (Vajjala, 2022) including the assessment of legal texts (Han et al., 2024), medical communication (Kiwanuka et al., 2017; Paul et al., 2021), or to control automatically generated or simplified texts (Marchisio et al., 2019; Picton et al., 2025; Srinivasan et al., 2024).

Although reasonably accurate for many tasks, readability formulas rely on surface level measures. As a result, they fail to capture underlying linguistic phenomenon and provide only limited interpretability. Research has thus increasingly shifted towards feature-based machine learning approaches that extract (more or less extensive) sets of linguistic complexity features (Feng et al., 2009; Deutsch et al., 2020; Weiss, 2024). These approaches enable a more nuanced modeling of text difficulty by incorporating syntactic, lexical, semantic, morphological, and discourse-level

information. Certain machine learning approaches such as Random Forests, Support Vector Machines or LSTMs allow for more interpretability by providing insights into the importance of individual features contributing to readability predictions.

This is especially useful for educational research domains (Collins-Thompson, 2014), where research often aim to understand proficiency development or control for text readability for a certain target group. ARA plays a crucial role in SLA, as it supports educators in aligning instructional and reading materials with learners' proficiency levels. By helping to match text complexity to learner ability, ARA enables more targeted and effective language instruction. A comprehensive literature review by Weiss (2024) revealed, however, that the majority of existing ARA research has focused on L1 language acquisition and primarily on English texts.

Nevertheless, there has been work on ARA in SLA for different languages (Imperial, 2021; Ribeiro et al., 2024; Uçar et al., 2024) and target groups (Feng et al., 2009; Abedi et al., 2012). For example, Weiss et al. (2021) established a multi-level readability algorithm for German L2 that classifies texts for beginner, intermediate an advances learners for German based on linguistic complexity. Moreover, as highlighted by Weiss (2024), most existing ARA research has focused on modeling the readability of entire texts. More recent work had explored sentence-level readability prediction for a more granular perspective (Štajner et al., 2017; Weiss and Meurers, 2022).

Yet, readability is not a property of a text alone. It results from the interaction of text, reading goals and reader characteristics such as prior knowledge, interests and also task context (Collins-Thompson, 2014; Vajjala, 2022; Valencia et al., 2014).

As it is not possible to account for all possible reader characteristics, we propose a shift in focus: rather than classifying texts or sentences by overall difficulty, we introduce a tool that highlights grammatical constructs along with their corresponding CEFR levels. By refraining from assigning readability scores and instead visually identifying potentially challenging linguistic constructs, we aim to support teachers to quickly evaluate whether a text is suitable for their specific teaching goals. This also allows teachers to decide whether targeted adaptations of text segments are needed.

2.2 Criterial Features

In response to the challenges associated with interpreting linguistic complexity, particularly its limited pedagogical specificity and lack of transparency in assessment settings, recent research has increasingly turned to criterial features as a more concrete, developmentally informed alternative (Salamoura and Saville, 2010; Hawkins and Filipović, 2012; Gaillat et al., 2022). These features refer to linguistic properties that are reliably acquired at specific stages of second language development and can thus serve as indicators of proficiency (Hawkins and Buttery, 2010). Criterial features are typically aligned with the CEFR. Although the CEFR descriptors provide general guidance for assessment, due to the language-agnostic properties of the CEFR, the descriptors are often underspecified in terms of the actual linguistic features that distinguish levels. Nevertheless, examiners tend to show high agreement in their level judgments (De Clercq et al., 2014). This implies that there must be underlying linguistic regularities that guide their decisions, even if these are not explicitly formulated in the CEFR itself. The goal of criterial feature research is therefore to identify the linguistic cues that examiners implicitly respond to, and to formalize these as part of a more systematic, data-informed approach to L2 proficiency.

Importantly, criterial features are not about linguistic complexity or difficulty per se, but about when certain linguistic forms emerge in the learning process (Bulté et al., 2025). A structure's presence at a particular level does not mean it is inherently more complex or more difficult. It simply means it tends to appear at that point in a typical learner's developmental trajectory. As Bulté et al. (2025) put it, "difficulty can be the cause for a certain developmental order (and this order may be taken as evidence for the construct's difficulty)", but this is not necessarily the case. A structure might not appear earlier because it wasn't needed in relevant genres or communicative situations, or because it wasn't taught or noticed.

The EGP offers a well-established resource of criterial features for English (O'Keeffe and Mark, 2022), expanding on CEFR descriptors by mapping them onto fine-grained grammatical structures. These descriptors are not only measurable and interpretable but also highly suitable for classroom use and automated assessment.

Unfortunately, for L2 German and to the best of our knowledge, no equivalent resource exists to date. While learner corpora offer valuable data, only the MERLIN corpus (Wisniewski et al., 2013) explicitly links learner texts to CEFR levels. A practitioner-oriented tool such as *Profile Deutsch* (Glaboniat, 2010) also attempts to associate grammatical phenomena with CEFR levels, but its categories are based on expert consensus rather than empirical learner language. As Wisniewski (2020) notes, this reflects a major gap in the availability of criterial feature resources for German, especially for educational applications. With the present research, we aim to close this gap by formulating 150 criterial features for German as well as building a web application that can automatically detect those features.

2.3 Related Systems

Several existing systems share similarities with our approach to automatic grammatical feature extraction and annotation. The POLKE system (Pedagogically Oriented extraction of Linguistic Knowledge, Sagirov and Chen (2025)), for example, extracts over 600 measures from the EGP using a comparable framework including Java and RUTA[1] preprocessing. POLKE's primary focus is the large-scale implementation and evaluation of EGP-based measures. For Portuguese, the SABER system offers a related approach to grammatical feature extraction (Sistema de Análise e Busca de Estruturas Relevantes, Akef et al. (2025)).

Highlighting linguistic constructions for learner noticing is a well-established concept. ICALL systems such as VIEW and WERTi provide input enhancement techniques on websites by visually emphasizing selected linguistic features to increase learner awareness (Meurers et al., 2010). While their primary focus is on language acquisition, the underlying idea aligns with our goal of making grammatical constructions salient.

Additionally, the FLAIR system (Form-Focused Linguistically Aware Information Retrieval; Chinkina et al. (2016)) demonstrates the value of integrating grammar knowledge into information retrieval. FLAIR incorporates grammatical patterns derived from an official English L2 curriculum into a content-based search engine, enhancing retrieval by linguistic criteria.

Together, these systems illustrate the relevance and growing interest in grammar-informed linguistic analysis and learner support, contextualizing our work within a broader landscape.

3 The German Grammar Profile

Building on the methodology of the EGP, we manually formulated 150 criterial grammatical features for German, which we refer to as the GGP, examples for the different levels can be seen in Table 1, the original German version can be see in the appendix in Table 2.

Unlike the EGP, which is built in a bottom-up, data-driven way, the GGP follows a top-down approach. Our starting point is the functional grammar framework provided by the ÖSD, which organizes grammatical phenomena according to communicative intentions, relations, and textual elements[2]. This structure identifies which grammatical forms are expected to occur at which CEFR levels, supplemented with descriptive explanations and illustrative examples. Our criterial features are systematically derived from this framework and have been evaluated and ranked by two experts from the ÖSD testing institute for pedagogical relevance and CEFR alignment, currently covering levels A1 to B2. The process was as follows.

To move from descriptive grammar to operationalization, we developed a structured feature schema inspired by the EGP: Each feature is assigned a supercategory, subcategory, guiding keywords (e.g., *Verbs* or *Pronouns* as supercategory, *Inflection* or *Modal verb* as subcategory, *Past Perfect* or *Accusative* as keywords), and example sentences. Every grammatical construct fund in the functional grammar framework resulted in one or several criterial features. Example features can be seen in Table 1. This categorization supports a consistent and interpretable annotation process. The super- and subcategories capture broader grammatical domains and their functional subtypes, while the guiding keywords help to highlight the kinds of linguistic cues that are typically associated with each feature, such as part-of-speech tags, morphological features (e.g. case, tense, mood) and syntactic dependencies. By aligning the conceptual categories with observable patterns, the GGP provides a bridge between pedagogically motivated descriptors and detectable linguistic structures.

[1] https://uima.apache.org/ruta.html

[2] https://www.osd.gr/de/das-osd-prufungssystem/profile-deutsch/ziele-komponenten.html

SuperCat.	SubCat.	Level	Guideword	Example
Main Clause	Interrogative	A1	Yes/No questions	Are you tired?
Negation	Negation	A1	„no“ + nominative	That is not a nice house.
Clause	Subordinate Clause	A2	„so“	I’m saving money so I can buy a car.
Verbs	Inflection	A2	Past Perfect	He had brought cake.
Pronoun	Relative Pronoun	B1	„what“, „where“	This is the city where we met.
Noun	Inflection	B1	Genitive	That’s my father’s car.
Verbs	Conjugation	B2	Conditional 1	If he comes later, we will talk.
Clause	Main Clause	B2	Desiderative	If only I had known!

Table 1: Translated example grammatical features from the GGP across CEFR levels.

Table 3 in the appendix gives a detailed overview of the distribution of features within the GGP by supercategory and subcategory. For instance, the GGP contains 43 verb-related features, including inflectional paradigms and modal verbs, 21 pronoun-related features covering personal, possessive, relative, and demonstrative pronouns, 19 adjective-related features and 12 features related to sentence structures. Other grammatical domains such as negation, connectors, prepositions, articles, nouns, and adjectives are also systematically represented. This structured categorization supports comprehensive coverage of grammatical phenomena across proficiency levels and facilitates targeted automatic detection. The strong focus on verb and tense related features is interesting to note. Previous research on ARA for German as second language using linguistic difficulty measures found that morphological features, which include tense, seem not to be particularly informative for machine learning models (Weiss et al., 2021).

This contrast is particularly interesting, as it illustrates a top-down, tester-centered approach that prioritizes pedagogical relevance over purely data-driven predictive use. Future research will need to explore these discrepancies in more depth, particularly how criterial features relate to linguistic complexity, and where pedagogical expertise and empirical model evidence converge or diverge. Understanding this relationship more precisely is essential for advancing both test design and automatic assessment.

In a later step, we will compare this expert-defined inventory with actual learner data to assess which constructs occur productively in learner language, and which remain rare or are avoided. This comparison will allow us to refine the GGP and better align pedagogical expectations with empirical usage.

4 PALME System Description

We follow the approach implemented by Sagirov and Chen (2025) to process learner texts. As shown in the flowchart (Figure 1), a text input is received via an HTTPS request and first preprocessed using standard Java NLP tools: OpenNLP (v2.2)[3] for tokenization and sentence segmentation, Mate Tools (v2.2)[4] for lemmatization and morphological feature annotation, and Stanford NLP-GPL (v2.2)[5] for dependency parsing and Named Entity Recognition.

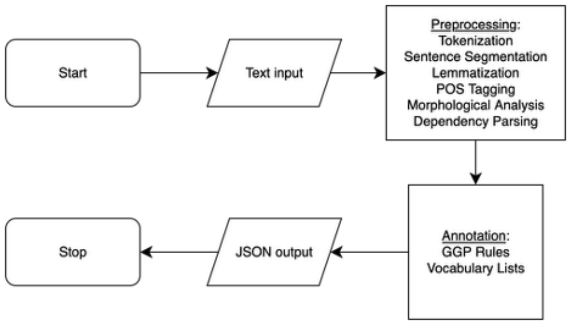

Figure 1: Overview of PALME pipeline and data flow

The GGP is integrated into a prototype web platform that allows users to analyze their own texts or generate new ones via a large language model. We built the web platform with the open-source Python framework Streamlit[7]. All texts are processed through our pipeline and annotated with

[3] https://opennlp.apache.org/news/release-220.html

[4] https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools/

[5] https://universaldependencies.org/treebanks/de_gsd/index.html

[7] https://streamlit.io/

grammatical constructions relevant for learner language.

These annotations provide the linguistic information required for the next step, where the text is annotated with detailed grammatical construction using RUTA, a rule-based language for text annotation. This process also uses specialized word lists based on resources provided by the ÖSD. We focused on the most important 70 features based on our expert rating and implemented those in PALME. The system currently covers 16 features at A1, 24 at A2, 17 at B1, and 13 at B2.

The extraction system is accessible via an API and integrated into a prototype web platform, shown in Figure 2. The platform offers a user-friendly, interactive environment where users can analyze their own texts by typing or pasting them into a text field. In addition to analyzing user-provided texts, the platform allows users to generate new learner-relevant texts on demand. By entering a prompt in a separate text field, users can request text generation via the integrated Gemini API. We decided to use the Gemini API due to its ease of use in this context. The generated texts are automatically analyzed by PALME, providing the same detailed feedback on grammatical features. Once submitted, these texts are automatically processed by PALME, which returns detailed annotations of grammatical constructions corresponding to criterial features in JSON format. Since raw JSON output is not easy to interpret, we developed a clear visualization that highlights annotated constructions in different colors according to their proficiency levels. A legend was added to the visualization, showing the corresponding construct ID. Currently, the visualization references only the construct ID, but we plan to include the supercategory in future versions to help users quickly understand the type of linguistic phenomenon annotated. This visual feedback allows teachers and students to quickly identify parts of a text that may be too difficult or too easy so that texts can be adapted easily.

5 Evaluation

Our extraction system shows promising results. Of the currently 70 implemented features, a subset of 33 randomly selected features was evaluated for precision and recall to establish PALME's algorithmic reliability. Of these 33 evaluated features, four at A1 level, ten at A2, 121 at B1 and seven at B2 level. We found a mean precision of 0.92 and a mean recall of 0.82, showing that the system can reliably extract criterial features.

To evaluate the system, we follow a typical strategy based on standard classification metrics. Precision measures the proportion of correctly identified instances among all extractions made by the tool, while recall assess the proportion of successfully detected instances among all instances to be extracted.

Figure 2: PALME platform interface showing user text input and results from Gemini API. The text on the left is a snippet of a text from the youth magazine *das Biber*[6], both text are about an annual fair about contemporary art in Vienna, the Viennafair.

A practical limitation in the evaluation stems from the lack of sufficiently large, CEFR-aligned corpora of German reading material. While learner corpora such as MERLIN provide CEFR annotations, it consists of learner-produced texts and is thus not ideal for evaluating the system's ability to handle reading input. As a result, we based our evaluation on texts from Spotlight[8], a magazine that publishes graded articles at the levels "Beginner", "Intermediate", and "Advanced", which approximately correspond to A2, B1, and B2/C1 respectively. Although these categories do not perfectly align with CEFR level definitions, they provided a workable approximation in the absence of more suitable corpora. We manually annotated all instances of the 33 evaluated criterial features in this material and matched them to the automatically extracted instances. To ensure sufficient coverage for recall evaluation, we also constructed additional example sentences containing the relevant features.

For precision, the extractor was applied to nine texts from the Spotlight corpus, which contains CEFR-labeled reading materials. These texts were evenly distributed across three proficiency levels: A2, B1, and B2/C1 (three texts per level). Figure 3 presents the number of extracted features per CEFR level, alongside the total word count per group. As expected, the number of higher-level constructs (notably B1 and B2) increased with text level, and no B2 constructs were found in A2-level texts.

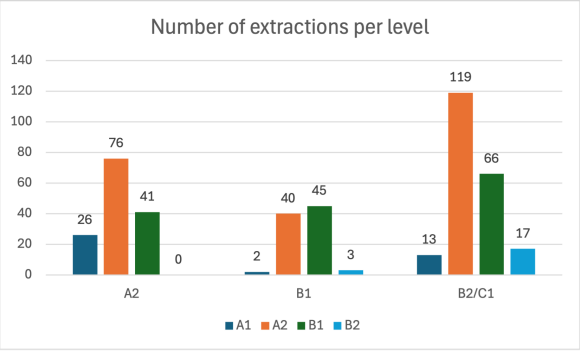

Figure 3: Distribution of occurrences of the 33 implemented criterial feature across CEFR levels

In total, 448 annotations were extracted, of which 416 were correct, resulting in an overall precision of 0.92. Among the 32 incorrect annotations, 28 were attributable to errors in external taggers, while four stemmed from limitations of our RUTA rules. These tagging issues have since been addressed in a revised preprocessing pipeline. Recall was evaluated separately using a controlled set of ten example sentences per feature, in addition to the examples provided in the Grammar Profile. The extractor was applied to this dataset, and recall per feature was calculated based on the proportion of successful detections. All tested features were identified at least four times, indicating robust detection capabilities under ideal conditions. 22 constructs achieved perfect precision. Conversely, a small number of constructs showed lower precision, often due to isolated tagging errors.

Together, these results support the reliability of our extraction method and provide an initial empirical validation of the developmental relevance of our feature set.

Moving forward, we plan to expand evaluation efforts using the MERLIN corpus, which consists of CEFR-annotated learner texts. This will not only verify the technical accuracy of our system but also serve a theoretical purpose: by analyzing the occurrence patterns of criterial features across proficiency levels in authentic learner writing, we aim to empirically assess the alignment of our expert-defined GGP with actual learner language development.

6 Outlook

Future work will focus on expanding and refining the system along several dimensions. We plan to implement the full set of 150 criterial features and extend the evaluation to include precision, recall and learner-level agreement for all features. In addition, we aim to conduct a more detailed analysis of when and how specific grammatical constructions occur in learner texts, based on corpus evidence.

We also aim to explore how traditional NLP components can be integrated into workflows based on large language models (LLMs) to support more informed and effective simplification. A central question will be whether LLMs can produce better, more pedagogically appropriate simplifications when guided by explicit linguistic information. This hybrid approach offers a promising path toward combining the strengths of rule-based analysis and generative flexibility in learner-oriented text generation.

Another important avenue for future research is to investigate the applicability of the system to

[8] https://www.spotlight-verlag.de

learner language, which often includes ill-formed structures. As the current design is based on well-formed input, evaluating its robustness on learner-produced texts will be crucial.

7 Conclusion

We presented the GGP, a system of 150 criterial grammatical features designed to support pedagogical applications in language learning. The GGP is grounded in the functional grammar framework of the ÖSD and systematically adapted to allow for automatic detection in learner-relevant texts. A subset of 70 features, rated as most relevant by experts, has been implemented in our extraction system PALME, which is accessible via an API and integrated into a web-based prototype. The platform allows users to analyze their own texts or generate new ones with a language model via text input fields, providing detailed visual feedback on grammatical constructions across CEFR levels A1 to B2.

The automatic extraction pipeline uses established Java NLP tools and uses a combination of morphological, syntactic, and lexical information to detect relevant criterial features from the GGP. Evaluation of a randomly selected subset of 33 implemented constructs showed high precision and promising recall, supporting the technical validity of the approach.

While the system remains under development, it already provides a practical tool for grammar-aware text analysis. Future work will expand the feature inventory and evaluate feature distribution in larger learner corpora. We also plan to explore whether integration with LLMs can support more targeted text simplification, using the GGP as a guiding framework. In this way, the system aims to contribute to both research and pedagogical practice in the field of learner-oriented grammar profiling.

Limitations

The PALME website currently is a prototype with several limitations. While we have defined 150 criterial features, currently 70 are fully implemented in the extraction system, of which 33 have undergone precision and recall evaluation. The GGP currently also primarily reflects expert judgment and curricular expectations rather than observed learner production. A systematic learner corpus analysis is envisaged to validate and refine the feature set in future work. Moreover, the rule-based approach is ideal for standard language, but its applicability to learner language may be limited. Future research should explore alternative approaches for such contexts.

Acknowledgments

We are grateful to the ÖSD, especially Karen Schramm and Anne Raveling, for the collaboration during the construction of the GGP. We thank our student assisstant Sofia Kathmann for the great work regarding the implementation of the criterial features. We also thank Elina Schnaper for her work during her bachelor thesis which contributed to the present paper.

References

Jamal Abedi, Robert Bayley, Nancy Ewers, Kimberly Mundhenk, Seth Leon, Jenny Kao, and Joan Herman. 2012. Accessible Reading Assessments for Students with Disabilities. *International Journal of Disability, Development and Education*, 59(1):81–95.

Soroosh Akef, Detmar Meurers, Amália Mendes, and Patrick Rebuschat. 2025. Interpretable Machine Learning for Societal Language Identification: Modeling English and German Influences on Portuguese Heritage Language. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 50–62.

C. H. Björnsson. 1983. Readability of Newspapers in 11 Languages. *Reading Research Quarterly*, 18(4):480.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur Jacobs, Jens Bölte, and Andrea Böhl. 2011. The Word Frequency Effect: A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental psychology*, 58:412–24.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2025. Complexity and Difficulty in Second Language Acquisition: A Theoretical and Methodological Overview. *Language Learning*, 75(2):533–574.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing Text Difficulty with Word Frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–94, San Diego, CA. Association for Computational Linguistics.

Maria Chinkina, Madeeswaran Kannan, and Detmar Meurers. 2016. Online information retrieval for language learning. In *Proceedings of ACL-2016 System Demonstrations*, pages 7–12.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325.

Tovly Deutsch, Masoud Jasbi, Masoud Jasbi, and Stuart M. Shieber. 2020. Linguistic Features for Readability Assessment. *Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17.

DuBay, William H. 2004. The principles of readability.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 229–237, Athens, Greece.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2022. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2):130–146.

Manuela Glaboniat, editor. 2010. *Profile deutsch: gemeinsamer europäischer Referenzrahmen; Lernzielbestimmungen; Kannbeschreibungen; kommunikative Mittel; Niveau A1-A2, B1-B2; C1-C2; [CD-ROM Version 2.0 mit Begleitbuch]*, nachdr. edition. Langenscheidt, Berlin München Wien Zürich.

Yu Han, Aaron Ceross, and Jeroen H. M. Bergmann. 2024. The Use of Readability Metrics in Legal Text: A Systematic Literature Review. *arXiv preprint*. Version Number: 1.

John A. Hawkins and Paula Buttery. 2010. Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal*, 1:e5.

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

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

J. P. Kincaid, Jr Fishburne, R. P., R. L. Rogers, and B. S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.

Elizabeth Kiwanuka, Raman Mehrzad, Adnan Prsic, and Daniel Kwan. 2017. Online Patient Resources for Gender Affirmation Surgery: An Analysis of Readability. *Annals of Plastic Surgery*, 79(4):329–333.

Stephen D. Krashen. 1985. *Second language acquisition and second language learning*, reprinted edition. Language teaching methodology series. Pergamon Pr, Oxford.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the Reading Level of Machine Translation Output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18.

Anne O'Keeffe and Geraldine Mark. 2022. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, pages 457–489.

Simon Paul, Marc-Daniel Ahrend, Jan-Christoffer Lüers, Kersten Sven Roth, Peter P. Grimmiger, Florian Bopp, and Babak Janghorban Esfahani. 2021. Systematic Analysis of Readability of Patient Information on Internet Pages from Departments for Trauma Surgery of German University Hospitals. *Zeitschrift für Orthopädie und Unfallchirurgie*, 159(02):187–192.

Bryce Picton, Saman Andalib, Aidin Spina, Brandon Camp, Sean S. Solomon, Jason Liang, Patrick M. Chen, Jefferson W. Chen, Frank P. Hsu, and Michael Y. Oh. 2025. Assessing AI Simplification of Medical Texts: Readability and Content Fidelity. *International Journal of Medical Informatics*, 195:105743.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024. Automatic text readability assessment in European Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, Santiago de Compostela, Galicia, Spain.

Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2024. Explainable ai in language learning: Linking empirical evidence and theoretical concepts in proficiency and readability modeling of portuguese. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 199–209.

Nelly Sagirov and Xiaobin Chen. 2025. POLKE: A system for comprehensively annotating pedagogically-oriented grammatical structure use in language production.

Angeliki Salamoura and Nick Saville. 2010. Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. In Inge Bartning, M. Martin, and Ineke Vedder, editors, *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, volume 1 of *Second Language Acquisition and Testing in Europe Monograph Series*, pages 101–131. European Second Language Association.

Nitin Srinivasan, Jamil S. Samaan, Nithya D. Rajeev, Mmerobasi U. Kanu, Yee Hui Yeo, and Kamran Samakar. 2024. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3.5, GPT-4, Bard, and online institutional resources. *Surgical Endoscopy*, 38(5):2522–2532.

Suna-Şeyma Uçar, Itziar Aldabe, Nora Aranberri, and Ana Arruarte. 2024. Exploring Automatic Readability Assessment for Science Documents within a Multilingual Educational Context. *International Journal of Artificial Intelligence in Education*, 34(4):1417–1459.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377.

Sheila W. Valencia, Karen K. Wixson, and P. David Pearson. 2014. Putting Text Complexity in Context. *The Elementary School Journal*. Publisher: University of Chicago PressChicago, IL.

Lev S Vygotsky. 2012. *Thought and language*, volume 29. MIT press.

Zarah Weiss. 2024. *An integrative approach to linguistic complexity analysis for German*. PhD Thesis, Universität Tübingen.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for cross-lingual readabili ty assessment. In *Proceedings of the Joint 10th Workshop on NLP for Comput er Assisted Language Learning*, pages 38–54.

Zarah Weiss and Detmar Meurers. 2021. Analyzing the linguistic complexity of German learner langua ge in a reading comprehension task: Using proficiency classification to investi gate short answer data, cross-data generalizability, and the impact of linguist ic analysis quality. *International Journal of Learner Corpus Research*, 7(1):84–131.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153, Seattle, Washington. Association for Computational Linguistics.

Katrin Wisniewski. 2020. SLA developmental stages in the CEFR-related learner corpus MERLIN: Inversion and verb-end structures in German A2 and B1 learner texts. *International Journal of Learner Corpus Research*, 6(1):1–37.

Katrin Wisniewski, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel, and Jirka Hana. 2013. MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data. In *ICT for Language Learning 2013, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni*.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 576–584.

Sanja Štajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. *International Joint Conference on Artificial Intelligence*, pages 4096–4102.

A Appendix

SupKat.	**SubKat.**	**Level**	**Leitword**	**Beispiel**
Hauptsatz	Fragesatz	A1	Ja/Nein-Frage	Sind Sie müde?
Negation	Negation	A1	„kein“ + Nominativ	Das ist kein schönes Haus.
Satz	Nebensatz	A2	damit	Ich spare Geld, damit ich ein Auto kaufen kann.
Verben	Konjugation	A2	Perfekt	Er hat mir den Kuchen gebracht.
Pronomen	Relativpron.	B1	„was“, „wo“	Das ist die Stadt, wo wir uns getroffen haben.
Substantiv	Deklination	B1	Genitiv	Das ist der Wagen meines Vaters.
Verben	Konjugation	B2	Konjunktiv 1	Er sagte, er komme später.
Satz	Hauptsatz	B2	Desiderativsatz	Wenn ich das gewusst hätte!

Table 2: Original example grammatical features from the GGP across CEFR levels.

SuperCategory	No. Constructs	SubCategory	No. Constructs
Verbs	43	Inflection	23
		Separating verbs	10
		"to have"/ "to be"	4
		Modal verbs	2
		Conjugation	2
		Imperative	1
		Valence	1
Pronouns	21	Relative pronouns	6
		Inflection	3
		Demonstrative pronouns	3
		Indefinite pronouns	3
		Interrogative pronouns	2
		Personal pronouns	2
		Impersonal pronoun ("es")	1
		Possessive pronouns	1
Adjectives	19	Inflection	12
		Comparison	4
		Adverbial	2
		Predicative	1
Clauses	18	Subordinate clauses	9
		Main clauses	3
		Comparative clauses	2
		Interrogative clauses	2
		Others	2
Articles	13	Demonstrative articles	3
		Other indefinite articles (e.g. "manche")	3
		Definite articles	2
		Indefinite articles (e.g. "ein")	2
		Inflection	2
		Interrogative articles	1
Connectors	12	Connecting adverbs	5
		Conjunctions	4
		Subjunctions	3
Prepositions	9	Other	4
		Temporal	3
		Local	1
		Changing	1
Nouns	7	Inflection	7
Main clauses	4	Interrogative clauses	2
		Imperative clauses	1
		Negated clauses	1
Negation	2		
Possessive article	2		

Table 3: Overview GGP features by categories.