Criterial Features in German: Towards Interpretable NLP in Readability Assessment

Denise Löfflad[1], Sofia Kathmann[1], Heiko Holz[2], Detmar Meurers[1,3]
[1]Leibniz-Institut für Wissensmedien Tübingen, Germany
[2]University of Education Ludwigsburg, Germany [3]University of Tübingen, Germany,
d.loefflad@iwm-tuebingen.de, s.kathmann@iwm-tuebingen.de
heiko.holz@ph-ludwigsburg.de, d.meurers@iwm-tuebingen.de

Abstract

This paper presents an empirical evaluation of the German Grammar Profile (GGP), a CEFR-aligned resource of criterial features, and its corresponding extraction system PALME.

We design a systematic test suite in which each feature extractor is evaluated on controlled positive and negative examples. The results show that PALME achieves high precision and recall across all CEFR levels, with over 90% of features achieving scores above 0.8. Qualitative analysis shows that lower performance primarily results from morphological ambiguity in noun and adjective case marking.

To evaluate the usefulness of the criterial features of the GGP for CEFR-aligned readability assessment, we assess their predictive power using Explainable Boosting Machines on graded readers. The model achieves strong performance (precision: 0.75, recall: 0.73). Our qualitative analysis shows that features related to specific verb constructions follow patterns consistent with developmental stages predicted by Processability Theory. These findings underline the value and relevance of criterial features for modeling language development in readability assessment.

1 Introduction

The quality of available language input plays an important role in second language acquisition (SLA), with reading serving as a crucial source of input. Automatic Readability Assessment (ARA) aims to predict the readability of a text and can support teachers in selecting appropriate texts more efficiently. Different approaches to ARA have been implemented in the past, including approaches based on surface-based readability formulas (Collins-Thompson, 2014), or using machine learning methods based on aggregated language characteristics (Weiss and Meurers, 2021; Vázquez-Ingelmo et al., 2026). These approaches typically assign a level to an entire text, allowing for a holistic view of readability. However, while such methods often achieve strong predictive performance, they offer limited insight into the specific language constructions in a text that contribute to its difficulty. This limitation is particularly relevant in language education, where on the one hand input needs to be comprehensible to be effective, but it should also systematically offer developmentally proximal language structures, as pointed out by SLA approaches since the i+1 Input Hypothesis of Krashen (1985) and Pienemann's Teachability Hypothesis (Pienemann, 1989, 2015).

Criterial features (CFs; Hawkins and Buttery, 2010; Hawkins and Filipović, 2012) have been proposed as more linguistically and pedagogically grounded measures for proficiency and ARA. They play a central role in operationalizing the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2020). As CEFR descriptors define purely communicative competences in the form of language-agnostic Can-Do statements, they deliberately refrain from specifying concrete linguistic realizations. Reference level descriptors aim to capture links between linguistic properties of texts or learner production and CEFR levels, and CFs are linguistic properties that are characteristic of specific stages of proficiency. CFs can thus be understood as an attempt to characterize the communicative purposes of CEFR descriptors linguistically, making them measurable and analyzable using computational linguistic methods.

For English, the English Grammar Profile (EGP, O'Keeffe and Mark, 2017) offers a resource of CFs that link linguistic aspects to CEFR levels. The EGP has been used as starting point to build computer-based systems to automatically extract those features (Sagirov and Chen, 2025). For German, comparable resources such as the German Grammar Profile (GGP) and the corresponding extraction system PALME exist, but they have not yet

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 63–77

been systematically evaluated or empirically validated (Löfflad et al., 2025). However, the lack of systematic evaluation leaves open questions about both the reliability of feature extraction and the validity of the underlying features. This raises two key questions: How can such systems be evaluated, and how can the underlying features be validated?

In this paper, we address these gaps by proposing (i) systematic methods for evaluating and validating CF-based system from both technological and pedagogical perspectives, (ii) a dedicated resource for evaluating German CF extraction systems, including potential applications for Large Language Models (LLMs), and (iii) a contribution to the theoretical understanding of CFs through a comparison with developmental trajectories (Pienemann, 1998). More broadly, we aim to provide a framework that can be transferred to other languages, supporting the extension of CEFR- and CF-related resources.

We investigate the computational operationalizability and empirical validity of CFs as indicators of CEFR readability in German graded readers and ask two research questions:

RQ1 How can criterial features reliably be operationalized and automatically extracted?

RQ2 To what extend do automatically detected criterial features predict CEFR levels and link to developmental trajectories?

To address the first question, we construct a systematic test suite of example sentences to evaluate the automatic extraction system PALME. For the second question, we use PALME to extract all 153 CFs in the GGP from a graded reader corpus and train an Explainable Boosting Machine (EBM) to predict readability levels. The EBM further enables us to analyze developmental trajectories and examine how CFs evolve across readability levels.

2 Background

2.1 Criterial Features and Developmental Trajectories

Criterial features reflect 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) and are typically aligned with the CEFR. Recent research has increasingly turned to criterial features as a concrete, developmentally informed and transparent possibility to measure linguistic phenomena (Salamoura and Saville, 2010; Hawkins and Filipović, 2012; Gaillat et al., 2022). The goal of criterial feature research is to identify the linguistic phenomenons that formalize underspecified descriptors as part of a more systematic resource.

Developmental trajectories (also stages or themed sequences) assume that certain grammatical aspects are acquired in predictable sequences. One influential theoretical framework, the Processability Theory (PT), proposes that the acquisition of grammatical structures follows a hierarchical order (Pienemann, 1998). Although criterial feature research aims to capture developmental trajectories, the relationship between trajectories and CEFR descriptors is still under-researched. Previous studies have analyzed the acquisition trajectories of individual grammatical constructs in detail (e.g., Wisniewski, 2020). Importantly, developmental trajectories are primarily based on learner language production. Nevertheless, they should also be examined in reading input such as graded reading materials, which are designed to slightly exceed the learner's current competence (i+1, Krashen, 1985). Since graded readers are pedagogically designed to align with learner proficiency levels, the distribution of criterial features in such texts should, in principle, reflect the developmental stages associated with those levels.

The EGP offers a well-established resource of criterial features for English (O'Keeffe and Mark, 2017), expanding on CEFR descriptors by mapping them onto fine-grained grammatical structures. There is work for other languages to develop reference level descriptors (Mohamed, 2023; Cuberos Vicente and De Cock, 2023), however, the EGP is a unique resource of language-specific grammatical features. Nevertheless, a similar yet smaller resource for German was developed by Löfflad et al. (2025) as described in the following.

2.2 The German Grammar Profile

The GGP (Löfflad et al., 2025) is a resource of German criterial features developed on the basis of the English Grammar Profile (EGP, O'Keeffe and Mark, 2017) and the *Profile Deutsch* (Glaboniat, 2010), a functional grammar for German as a second language learners and teachers. Following a theory-driven approach, the GGP defines a set of 153 criterial features mapped to CEFR levels based on the functional grammar provided by Glaboniat (2010), and not derived from corpus data (example features can be seen in Appendix A). Of these 153 features, 47 are assigned to level A1, 49 to

level A2, 35 to level B1, and 22 to level B2. The distribution of features across levels is somewhat imbalanced. This is partly due to fewer new grammatical features being introduced at higher levels and partly because higher level features tend to be usage-based, which is not in the scope of the GGP. Nevertheless, this can impact subsequent data analyses, as A level features might be more prominent.

Moreover, the theory-driven approach used to create the GGP, in contrast to the data-driven approach used in the development of the EGP, may lead to mismatches between the levels assigned in the GGP and the levels observed in graded readers or learner data. Such mismatches are plausible given that previous research on the EGP has already found limited agreement between statistically derived feature mappings and the mappings proposed by the EGP (Verratti-Souto et al., 2025).

2.3 The PALME System

To automatically extract the measures defined in the GGP, Löfflad et al. (2025) developed the PALME system, the *Pedagogically Oriented Linguistic Feature Extraction* (*Pädagogisch Ausgerichtete Linguistische Merkmalsextraktion*). PALME relies on rule-based NLP components. While designing such components is labor-intensive, it offers advantages in comparison to Generative AI in terms of reliability and transparency. Once the system has been systematically tested, its behavior is predictable and performance does not depend on changes in black-box AI models.

The system is related to other approaches supporting automatic grammatical feature extraction and annotation. The POLKE system (*Pedagogically Oriented Language Knowledge Extractor*, Sagirov and Chen, 2025), for example, extracts more than 600 measures from the EGP using a comparable rule-based framework that includes Java and RUTA-based preprocessing. POLKE primarily focuses on the large-scale implementation and evaluation of EGP-based measures. For Portuguese, the SABER system (*Sistema de Análise e Busca de Estruturas Relevantes*, Akef et al., 2025) provides a related approach to grammatical feature extraction.

2.4 Test Suite vs. Corpus-Based Evaluation

Test suite-based evaluations inherently differ from corpus-based evaluations: While corpus-based evaluations assess systems on naturally occurring language data, test suites consist of carefully constructed, focused examples of language use (Lehmann et al., 1996; Prasad and Sarkar, 2000). Both methods come with advantages and disadvantages. Corpus-based evaluations facilitate testing a system's robustness against authentic language and are well-suited for frequent linguistic constructs. However, many linguistic constructs occur only rarely in natural data. Test suites address this limitation by supporting a controlled evaluation that systematically evaluates the coverage of specific constructs (Lehmann et al., 1996).

For a similar extraction system for English, Sagirov and Chen (2025) used a corpus-based evaluation approach. They randomly selected 15 features for each CEFR level and automatically extracted sentences containing the respective features from the Corpus of Contemporary American English (Davies, 2009). In addition, they randomly selected sentences that did not contain the feature as negative examples.

3 Methods

3.1 System Evaluation

To evaluate the system and answer RQ1, we opted for a test suite-based approach. This allowed us to control the evaluation sentences, particularly the negative examples, enabling us to construct challenging cases that stress-test the system. Moreover, our goal was to evaluate every measure extracted by PALME. As mentioned previously, some features occur only rarely in natural language data, which makes them difficult to capture reliably through corpus-based sampling. Although there are, to the best of our knowledge, no widely accepted guidelines on writing test suites, Balkan et al. (1994) recommend to include both positive and negative examples of the target feature. They further suggest covering a range of instantiations and to intentionally add misleading inputs and traps for the system. We thus constructed a systematic test suite consisting of positive and negative examples for each feature in the GGP. For every feature, the test suite contains ten sentences: five positive examples that include the target feature and five negative examples that do not.

To ensure variability and linguistic diversity, the sentences were developed collaboratively by three experts: the first author, the second author as a natural language processing (NLP) expert, and a pedagogical expert in second language teaching. The positive examples were informed by reading and learner corpora, but were adapted or constructed

where necessary to ensure controlled evaluation, for example by avoiding multiple instances of the same feature or by including theoretically motivated constructions. We deliberately included different structural and lexical realizations of each feature to capture its range of possible realizations. For the negative examples, we constructed sentences that were closely related to the target feature, either structurally, linguistically, or pedagogically, but did not actually instantiate it. For example, in the case of adjectival case marking without an article, accusative constructions were contrasted with structurally parallel sentences in other cases. Similarly, for verbal features, negative examples were designed to resemble the relevant constructions while differing in tense, agreement, or syntactic realization. This design ensured that the evaluation tested the system's ability to distinguish subtle contrasts. After constructing the test suite, we evaluated the system's performance on all features. We computed precision and recall for each feature and identified those with scores below a threshold of 0.8. As each feature is represented by five positive and five negative examples, a score of 0.8 or higher corresponds to at least four correctly classified instances within a set of five (i.e., at most one error). These features were subsequently revised and improved, and the sentences targeting the revised features were updated with new examples to prevent overfitting to the original test sentences. Following each revision cycle, the system was re-evaluated on the test suite. Features that continued to fall below the performance threshold were further refined. This iterative process was conducted for two improvement cycles and concluded with a final evaluation. Although additional cycles might have led to further performance gains, we stopped after the third evaluation round to minimize the risk of overfitting the system to the test suite.

3.2 Feature Validation

After completing the system optimization phase, we validated the extracted CFs on graded reader data to answer RQ2. We used texts provided by *Deutsch Perfekt*[1], a magazine that publishes reading materials for second language learners.

While this dataset does not consist of learner production, graded reading materials provide an approximation of proficiency levels and are therefore a suitable starting point for evaluating the usefulness of criterial features in readability classification. Given that this study represents an initial step toward a broader validation of CFs, we focus here on input texts, which play a central role in language development. In line with pedagogical principles such as the i+1 hypothesis (Krashen, 1985), reading materials should be designed to be just above the learner's current proficiency level, thereby reflecting increasing linguistic complexity across levels.

The corpus is categorized into three proficiency levels: easy (approximately CEFR A2), intermediate (approximately B1), and advanced (approximately B2/C1). In total, the dataset consists of 4,906 texts spanning these proficiency levels. Table 1 shows a descriptive overview of the corpus.

Level	Num. Texts	Num. Tokens
Easy	2,535	447,287
Intermediate	1,463	830,158
Advanced	908	461,446
Total	4,906	1,738,891

Table 1: Overview of the Deutsch Perfekt corpus

The classes are imbalanced, with most tokens in the corpus being part of texts in the intermediate class. We conducted a predictive evaluation by training an Explainable Boosting Machine (EBM) on the extracted CFs to predict CEFR levels using an 80–20 train–test split. We normalized raw feature counts by calculating feature occurrences per 1,000 tokens to control for differences in text length. EBMs are generalized additive models that learn each feature function using gradient boosting and maintaining additive structures, thus enabling high interpretability (Nori et al., 2019; Lou et al., 2013). EBMs train each feature function one at a time, allowing for each feature function to be examined individually. They belong to so-called "glassbox" models providing transparent feature effects, in contrast to post-hoc explanation methods (e.g., LSTMs or SVMs). Using EBMs instead of other machine learning models allowed us not only to assess predictive performance but also to interpret the contribution and shape of individual features in relation to proficiency level.

4 Results

4.1 System Evaluation

We first report system performance aggregated of CFs by CEFR level in order to assess overall robustness across proficiency bands. Table 2 presents

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

precision, recall, and F1 scores for A1–B2. Performance is consistently high across proficiency levels, with precision ranging from 0.956 (A1) to 0.981 (B1). Recall shows a slight decrease with increasing proficiency, ranging from 0.932 at A1 to 0.855 at B2.

Macro-averaged performance across levels yields a precision of 0.968, recall of 0.886, and an F1 score of 0.925. Micro-averaged performance across all 1,530 instances is similarly high. Our results are in line with those reported by Sagirov and Chen (2025) for the POLKE system, indicating state of the art performance.

Level	Precision	Recall	F1	N
A1	0.956	0.932	0.944	470
A2	0.978	0.894	0.934	490
B1	0.981	0.863	0.918	350
B2	0.959	0.855	0.904	220
Macro Avg.	0.968	0.886	0.925	–
Micro Avg.	0.969	0.893	0.929	1530

Table 2: Precision, recall, and F1 scores per CEFR level. Macro averages are computed across levels; micro averages are computed across all instances.

Across the 153 evaluated measures, 138 measures (90.1%) achieved both precision and recall values of at least 0.8. Table 3 lists the 15 measures (9.9%) that fell below this threshold in at least one of the two metrics. As described in Section 3, the threshold of 0.8 corresponds to at most one error within the five positive or five negative test instances per feature. Table 5 in Appendix A shows feature descriptions from the GGP and examples sentences in German and English for all features discussed in this paper.

Overall, the results indicate robust and reliable detection performance for the large majority of measures. The measures with lower scores are largely explainable. Several involve case marking (e.g., adjectival inflection without an article or genitive marking), which depends on accurate morphological analysis. We found that the taggers are currently not able to reliably annotate the accusative case, as the form is often similar to the nominative case, which leads to subsequent errors in the rule-based system.

Feature 32148 (subjunctive II) is primarily form-based and involves subtle semantic distinctions that are difficult to operationalize computationally. The comparatively lower recall for *indefinite pronouns* (3168) is less immediately expected and warrants further inspection. Importantly, precision is consistently high across measures, with the large majority reaching a value of 1.0. This indicates that false positives are rare. For subsequent corpus-based analyses, this is particularly desirable, as it minimizes the risk of overestimating the frequency of linguistic structures. A full table of performance for all features can be seen in Appendix B – each feature is assigned a construct ID that encodes the CEFR level: IDs beginning with 21 correspond to A1 features, 22 to A2 features, 31 to B1 features, and 32 to B2 features.

ID	Level	Feature	Precision	Recall
3132	B1	Superlative of B1 level adverbs	1.00	0.20
2264	A2	Adjectives in the accusative after zero articles	1.00	0.40
2268	A2	A2 level Indefinite Pronouns	1.00	0.40
32148	B2	Subjunctive II	1.00	0.40
2111	A1	Negation article ‘kein” in the accusatives	1.00	0.60
2113	A1	Personal pronouns in the accusative	0.75	0.60
2263	A2	Adjectives in the nominative after zero article	1.00	0.60
2265	A2	Adjectives in the dative after zero article	1.00	0.60
2281	A2	Past Participle	1.00	0.60
3168	B1	B1 level Indefinite pronouns	1.00	0.60
31120	B1	Articles in genitive case	1.00	0.60
3134	B1	Adverbial temporal determinations	1.00	0.60
3256	B2	Genitive after indefinite article	1.00	0.60
3257	B2	Genitive after zero article	1.00	0.60
2282	A2	Perfect Tense	1.00	0.60

Table 3: Precision and recall scores for features below the 0.8 threshold in at least one metric.

4.2 Feature Validation

We evaluated the predictive performance of the extracted criterial features for regression as well as classification models, using EBMs for both approaches. Results are reported in Table 4. In addition, we report global and feature-specific results from the EBM. In the regression setting, the model achieved a Root Mean Squared Error (RMSE) of 0.501 and a Mean Absolute Error (MAE) of 0.396. To enable comparison with the classification setup, regression predictions were rounded to the nearest CEFR level. Under this discretized evaluation, the model achieved an accuracy of 0.704, a macro-averaged F1 score of 0.667, a macro-averaged precision of 0.692, and a macro-averaged recall of 0.656. In the classification setting using an EBM, the model achieved an overall accuracy of 0.769. Macro-averaged F1 was 0.740, with a macro-averaged precision of 0.750 and a macro-averaged recall of 0.733. Direct comparison of the regression results and the classification model shows higher performance for the classification approach across all reported metrics. Subsequent analyses are therefore based on the classification model.

Category	Metric	Value
Regression	RMSE	0.501
	MAE	0.392
Classification	Accuracy	0.769
	F1 (macro)	0.740
	Precision (macro)	0.750
	Recall (macro)	0.733
Regression	Accuracy	0.711
	F1 (macro)	0.678
	Precision (macro)	0.724
	Recall (macro)	0.665

Table 4: Model performance metrics for regression, classification EBM, and rounded regression outputs.

Figure 1 shows the confusion matrix for the classification model. Consistent with previous findings (e.g., Lagutina et al., 2023), most errors occur between adjacent levels, especially at the intermediate level. This is especially interesting considering the class imbalance: As most data exists for the intermediate level, it would be expected that the model tends to overestimate this level, which is not the case for this model. Misclassifications spanning two levels are comparatively rare. Interestingly, recall for A1 level is surprisingly high, and the confusion matrix shows that, if misclassified, intermediate texts are more frequently classified as easy texts, meaning that the models has a slight tendency to underestimate CEFR levels.

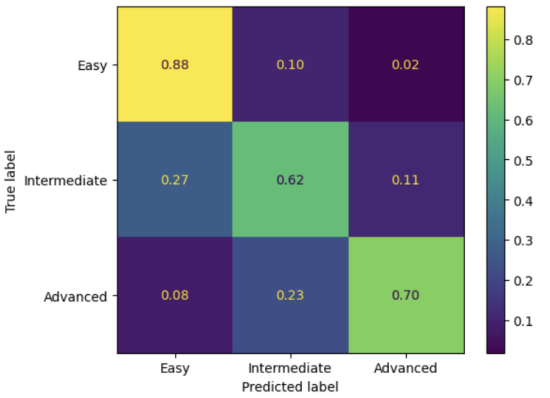

Figure 1: Confusion Matrix showing results in percent for the classification model.

It is important to note that there currently is no work on using German CFs as features for machine learning classifiers, however there is work using linguistic complexity measures. Our results are lower than classification results reported in previous work on linguistic complexity for German on the same corpus (Weiss et al., 2021), but are comparable to studies using criterial features for English classification tasks (Gaillat et al., 2022).

Figure 2 shows the development of the ten most strongly changing B1/B2 features across the three readability levels. Overall, most features show a

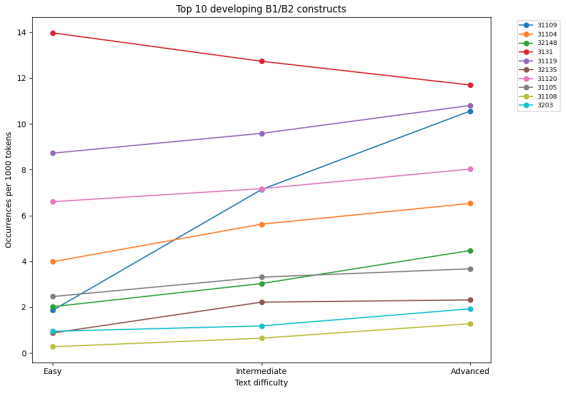

Figure 2: Developmental trajectories of B1 and B2 features with strongest development.

clear increase from easy to advanced texts, indicating that higher-level grammatical constructions tend to occur more frequently in more difficult texts. The overall trend suggests that the frequency of B1/B2 features increases with text difficulty, as

would be expected. The strongest increase is observed for feature 31109 (simple past), which rises substantially from easy to advanced texts. One interesting observation is the development of feature 3131 (comparative of B1-level adverbs), which shows a slight decrease across levels.

As can be seen in Figure 3, global term importances derived from the EBM (mean absolute weighted scores) indicate that features 31109 (simple past) and 2182 (perfect tense of A1 verbs) were the most influential predictors. These were followed by features 2116 (A1 level verbs in present tense), 2118 (*to be* in present tense), 2112 (nominative of personal pronouns), 2105 (nominative of A1 nouns), 3203 (B2 level conjunctive adverbs), 2274 ("that" clauses), 2117 (*to have* in present tense), 2206 (relative pronouns in nominative case), 32135 (*to become* in passive present tense), 32148 (subjunctive II), and 2281 (past participle).

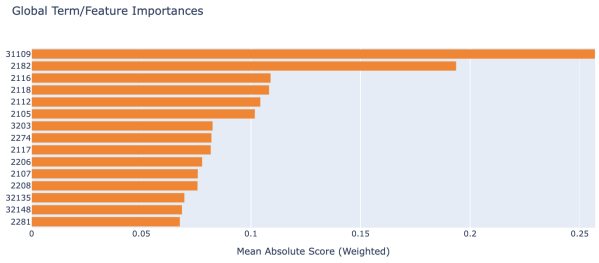

Figure 3: Global term importances for ten most informative features derived from the EBM.

Figure 4 shows the learned score contributions of feature 2182 and 31109. For feature 2182 (perfect tense of A1 verbs), the contributions vary across its observed frequency range (0–42). The distribution of this feature shows a concentration of observations in the lower frequency intervals. As the frequency of feature 2182 increases, the model increasingly favors class 1 (easy) while simultaneously decreasing the likelihood of class 3 (advanced), with class 2 (intermediate) exhibiting comparatively small changes. Thus, higher frequencies of feature 2182 are strongly associated with class 1 predictions, whereas low frequencies are more indicative of class 3.

For feature 31109 (simple past), score contributions vary across its observed frequency range (0–155), with most observations concentrated in the lower frequency intervals. In contrast to feature 2182, increasing frequencies of feature 31109 strongly increase the likelihood of class 3 (advanced) while reducing the likelihood of class 1

Figure 4: EBM shape functions and value distributions for the two most informative features. Top: Feature 31109. Bottom: Feature 2182. The upper panels show the learned score contributions across feature values, and the lower panels show the empirical value distributions.

(easy), again with comparatively smaller effects for class 2 (intermediate). High frequencies of feature 31109 therefore act as a strong indicator for class 3. As seen before, of all B1/B2 level features, feature 31109 also showed the strongest increase from easy to advanced texts. The results will be discussed in the following.

5 Discussion

This study addressed two research questions concerning the operationalization and predictive power of CFs for German. Overall, the analyses provide promising results that CFs can be operationalized computationally and that they capture meaningful aspects of language development.

Regarding RQ1, the results indicate that CFs can be reliably operationalized and automatically detected. The results across CEFR levels indicate that the system maintains very high precision across all proficiency levels, meaning that identified structures are highly reliable. The decrease in recall at higher levels likely reflects increasing linguistic complexity, which introduces greater structural variation and thus poses additional challenges. Moreover, 90% of individual features achieve precision and recall above 0.8. This demonstrates that the linguistic constructs defined in the GGP can be translated into computational rules and identified in language data. In addition, the use of a targeted test suite proved to be a useful methodology for

evaluating the performance of the extraction system, as it allowed us to systematically assess the performance of individual features and to improve the system cyclically, incorporating challenging negative examples.

With respect to RQ2, the results show that automatically detected CFs provide meaningful information for predicting CEFR levels. Several CFs were found to be highly informative for the EBM classifier. Notably, the most informative feature was the use of simple past (31109), followed by the perfect tense (2182), which involves the characteristic German split verb construction. Two additional features related to the split verb construction (2281, past participle and 32135, *to become* (*bekommen*) in passive present tense), also appeared among the most informative predictors. These findings suggest that verb-related features play an important role in distinguishing CEFR levels and reflect developmental patterns in second language acquisition.

An interesting observation concerns the distribution of the perfect tense feature of A1 level verbs (2182). While constructions involving the perfect tense and participle formation are theoretically related to the German split verb pattern, the model indicates that feature 2182 is particularly associated with easier texts, corresponding roughly to A2 levels. This appears to diverge from predictions made by PT, where verb-splitting phenomena are expected to emerge only at stage three (out of five stages) of the developmental hierarchy (Pienemann, 1998). One possible explanation is that certain perfect tense forms, particularly with high-frequency verbs such as *to be* (*sein*) and *to go* (*gehen*) (e.g., *I went*, *Ich* ***bin gegangen***), may be introduced early in language instruction as formulaic constructions for communicative purposes. Learners may therefore acquire these forms initially as fixed expressions rather than as fully productive grammatical structures. In this sense, the presence of perfect tense forms in lower-level texts does not necessarily indicate that learners have acquired the underlying verb-splitting mechanism. Instead, the productive use of separable verb constructions may still develop later, in line with the predictions of developmental theories such as PT.

The results also show the importance of case marking (features 2112, 2105 and 2208, respectively, personal pronouns, nouns, and relative pronouns in nominative case), which showed as highly informative features in the classification model. In PT, case marking is treated as a diacritic feature associated with a lexical item, rather than being a central developmental structure. However, previous research has highlighted the importance of morphological development for German L2 learners (Weiss, 2024), and the functional grammar underlying *Profile Deutsch* (Glaboniat, 2010) also assigns a prominent role to case marking. Our results provide further empirical evidence supporting this, suggesting that case marking contributes to distinguishing readability levels in graded readers.

As briefly mentioned in Section 4, classifiers based on linguistic complexity measures achieve higher performance on the same corpus than the approach presented in this paper (Weiss et al., 2021). This difference can be attributed to several factors. First, Weiss et al. (2021) extract a large set of over 300 features capturing multiple linguistic dimensions, including syntactic, lexical, and discourse-level properties. Second, linguistic complexity measures are inherently aggregative, which may be better suited for classification tasks. Finally, differences in model architecture, such as the use of Support Vector Machines (SVMs) compared to EBMs, may also contribute to performance differences. Nevertheless, criterial features offer distinct advantages that make them particularly valuable in educational contexts. In contrast to complexity measures, CFs provide a transparent and linguistically grounded representation of language, as they can be directly linked to CEFR levels and interpreted in terms of developmental trajectories. This allows for a more fine-grained and interpretable analysis of learner-relevant structures, which is not straightforward with aggregative complexity measures.

From a pedagogical perspective, the GGP and the PALME system enable targeted readability assessment by identifying whether a text is suitable for a specific learner group (e.g., B1 learners of German), while also highlighting specific linguistic structures that contribute to its difficulty. This opens up possibilities not only for text selection but also for diagnostic applications, such as identifying particularly challenging constructions within a text. At the same time, the two approaches are not mutually exclusive. Combining criterial features with linguistic complexity measures may lead to more robust and accurate classification systems, while preserving interpretability and pedagogical relevance.

6 Future Work

Several directions for future research emerge from the present study. First, the test suite developed for evaluating the automatic detection of CFs could be used to assess the capabilities of LLMs in identifying and analyzing linguistically motivated grammatical constructs. In particular, LLMs could be explored as a complementary approach in cases where rule-based detection proves difficult or insufficient. Another approach to address limitations of upstream tagging could be to explore the use of constituency parsing. Second, the test suite could be expanded to include authentic learner language data. Incorporating learner corpora would allow us to assess how well the system performs on naturally occurring learner language and to examine how CFs manifest in real learner production. Third, future work could investigate the relationship between automatically detected CFs and CEFR level assignments in more detail. Following approaches such as Verratti-Souto et al. (2025), statistical analyses would contribute to a better understanding of how closely criterial feature mappings correspond to actual proficiency development.

7 Conclusion

In this paper, we presented a systematic, test suite-based method to evaluate rule-based feature extractions systems by analyzing the reliability of the PALME system. We showed that PALME reliably extracts 90% of the extracted features with precision and recall above .8 using positive and negative examples. We moreover used the system to analyze graded reading material, thus contributing to interpretable ARA.

The findings indicate that criterial features capture aspects of developmental trajectories, but they also extend beyond them. PT predicts that certain grammatical phenomena emerge at specific stages of development, such as the split verb pattern associated with separable verb constructions. Our results partially align with these predictions. At the same time, the model also relies on a broader set of linguistic constructs, including morphological features. This suggests that criterial features not only reflect developmental stages but also capture additional dimensions.

8 Limitations

While the present study provides promising results, several limitations point to important directions for future research. First, the evaluation of the PALME system is based on a manually constructed, corpus-informed test suite rather than a fully authentic corpus. Although this allows for controlled testing of individual features, it may not fully capture the variability of real-world language use. In addition, each feature is represented by only a small number of examples, which constrains the range of linguistic realizations covered.

Second, the system depends on upstream preprocessing components such as part-of-speech tagging and parsing. Errors in these components may propagate and affect feature detection, particularly for more complex constructions.

Third, the validation of criterial features is limited to a single corpus of graded readers, raising questions about the generalizability of the findings across different corpora and text types. Future work should therefore investigate cross-corpus performance to assess the robustness of the approach.

Finally, the study does not include learner language data. While graded readers approximate proficiency levels, they do not fully reflect actual learner production. Evaluating the system on learner corpora would therefore provide a more direct validation of criterial features as indicators of language development.

Acknowledgments

We would like to thank the LEAD Graduate School and Research Network for its support. Our research was supported by the WoLKE project (Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg) and the EU Project LATILL (Level-Adequate Texts in Language Learning, Reference number 2021-1-AT01-KA220-SCH-000029604). We also thank Sarina Doster for her contribution to the development of the test suite.

References

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.

Lorna Balkan, Siety Meijer, Doug Arnold, Eva Dauphin, Dominique Estival, Kirsten Falkedal, Sabine Lehmann, Klaus Netter, and Sylvie Regnier-Prost. 1994. Issues in Test Suite Design. *Report to LRE*, 62:089.

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. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing, Strasbourg, France.

Rocio Cuberos Vicente and Barbara De Cock. 2023. Towards a Graded Lexical Inventory in L2 Spanish: Insights from Productive Vocabulary Knowledge. In *Proceedings of the Workshop on Linking Lexicographic and Language Learning Resources (4LR)*.

Mark Davies. 2009. The 385+ Million Word Corpus of Contemporary American English (1990–2008+): Design, Architecture, and Linguistic Insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

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.

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 of *English Profile Studies*. Cambridge University Press.

Stephen D Krashen. 1985. *The input hypothesis: Issues and implications*. Longman, New York.

Nadezhda Stanislavovna Lagutina, Kseniya Vladimirovna Lagutina, Anastasya Mikhailovna Brederman, and Natalia Nikolaevna Kasatkina. 2023. Text Classification by CEFR Levels using Machine Learning Methods and BERT Language Model. *Modeling and Analysis of Information Systems*, 30(3):202–213.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, and Eva Dauphin. 1996. TSNLP - Test Suites for Natural Language Processing. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 711–716.

Denise Löfflad, Benedikt Beuttler, and Detmar Meurers. 2025. German Grammar Profile for Learners: Pedagogical Feature Definition and Automated Extraction. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 212–223.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate Intelligible Models with Pairwise Interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*, pages 623–631.

Salwa Mohamed. 2023. The Development of an Arabic Curriculum Framework based on a Compilation of Salient Features from CEFR Level Descriptors. *The Language Learning Journal*, 51(1):33–47.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223*.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of Learner Competence: Methodology and Key Findings. *International Journal of Corpus Linguistics*, 22(4):457–489. Publisher: John Benjamins Publishing Company Amsterdam/Philadelphia.

Manfred Pienemann. 1989. Is language teachable? psycholinguistic experiments and hypotheses. *Applied Linguistics*, 10(1):52–79.

Manfred Pienemann. 1998. Developmental Dynamics in L1 and L2 Acquisition: Processability Theory and Generative Entrenchment. *Bilingualism: Language and Cognition*, 1(1):1–20.

Manfred Pienemann. 2015. An outline of processability theory and its relationship to other approaches to SLA. *Language Learning*, 65(1):123–151.

Rashmi Prasad and Anoop Sarkar. 2000. Comparing Test-Suite based Evaluation and Corpus-based Evaluation of a wide-coverage Grammar for English. In *Proceedings of the Workshop on Using Evaluation within Human Language Technology Programs: Results and Trends. (LREC)*, pages 7–12, Athens, Greece.

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, Maisa Martin, and Ineke Vedder, editors, *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, volume 1 of *Eurosla Monographs Series*, pages 101–131. European Second Language Association.

Andrea Vázquez-Ingelmo, Denise Löfflad, Alicia García-Holgado, Roberto Therón-Sánchez, and Detmar Meurers. 2026. From CEFR Classification to Generative AI Materials: Designing and Validating the LATILL Platform. *Universal Access in the Information Society*, 25(2):67.

Daniela Verratti-Souto, Nelly Sagirov, and Xiaobin Chen. 2025. NLP-Powered Quantitative Verification of the English Grammar Profile's Structure-Level Assignment. *Annual Review of Applied Linguistics*, pages 1–22.

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 Readability Assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54. LiU Electronic Press.

Zarah Weiss and Detmar Meurers. 2021. Analyzing the Linguistic Complexity of German Learner Language in a Reading Comprehension Task: Using Proficiency Classification to Investigate Short Answer Data, Cross-Data Generalizability, and the Impact of Linguistic Analysis Quality. *International Journal of Learner Corpus Research*, 7(1):84–131.

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.

A Appendix

ID	Level	Feature	Example	Translation
2105	A1	Nominative of A1 Nouns	Die **Frau** arbeitet.	The **woman** is working.
2111	A1	Negation article ‘kein” in the accusatives	Er sieht **keinen** Hut.	He has **no** hat.
2112	A1	Nominative of Personal Pronouns	**Ich** wohne in Tübingen.	**I** live in Tübingen.
2113	A1	Accusative of Personal Pronouns	Sie besucht **ihn**.	She visits **him**.
2116	A1	Present Tense of A1 Verbs	Wir **fahren** nach Hause.	We **drive** back home.
2117	A1	*to have* in Present Tense	Wir **haben** das nicht.	We do not **have** that.
2118	A1	*to be* in Present Tense	Es **ist** Donnerstag.	It **is** thursday.
2182	A1	Perfect Tense of A1 Verbs	Ich **habe** das Buch **gelesen**.	I **have read** the book.
2206	A2	*Nominative of Relative Pronouns*	Die Frau, **die** ich sehe.	The woman **that** I see.
2263	A2	Nominative of Adjectives after Zero Article	**Kleine** Kinder spielen dort.	**Small** children play there.
2264	A2	Accusative of Adjectives after Zero Article	Sie sollten **frisches** Obst essen.	They should eat **fresh** fruit.
2265	A2	Dative of Adjectives after Zero Article	**Guten** Freunden gibt man das.	You give that to **good** friends.
2268	A2	A2 Level Indefinite Pronouns	Das kann nicht **jeder**.	Not **everyone** can do that.
2274	A2	"that" Clauses	Ich hoffe, **dass** sie kommt.	I hope **that** she comes.
2281	A2	Past Participle	Wir haben dich **besucht**.	We **visited** you.
3131	B1	Comparative of B1 Level Adverbs	Er geht **öfter** als ich.	he goes there **more often** than me.
3132	B1	Superlative of B1 Level Adverbs	Er geht **am öftestens** dorthin.	He goes there **more often**.
3168	B1	B1 Level Indefinite Pronouns	Es gibt nur noch **wenige**.	There is only **few** left.
31109	B1	Simple Past	Sie **wandte** sich ab.	He **turned** away.
31120	B1	Genitive of Articles	Das Nest **der** Wespen.	The nest of **the** wasps. - The wasps’ nest.
3134	B1	Adverbial Temporal Determinations	Er hat **mit** 21 geheiratet.	He got married **at** 21.
3203	B2	B2 Level Conjunctive Adverbs	**Anschließend** gingen wir weg.	We left **afterwards**.
3256	B2	Genitive after Indefinite Article	Das Nest eines **jungen** Spechts.	The nest of a **young** woodpecker.
3257	B2	Genitive after Zero Article	Der Geruch **frischen** Kaffees.	The smell of **fresh** coffee.
32135	B2	*to become* in Passive Present Tense	Der Brief **wird geschrieben**.	The letter **is being written**.
32148	B2	Subjunctive II for unrealistic Wishes	Wäre ich ein Vogel, **flöge** ich zu dir.	If I were a bird, I **would fly** to you.

Table 5: Feature descritptions from the GGP including an example sentence in German and translated to English.

B Appendix

ID	Pos1	Pos2	Pos3	Pos4	Pos5	Neg1	Neg2	Neg3	Neg4	Neg5	Precision	Recall
2168	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2165	1	1	0	1	1	0	0	0	0	0	1.0	0.8
2166	1	1	1	1	0	0	1	0	0	0	0.8	0.8
2103	1	1	1	1	0	0	0	0	0	0	1.0	0.8
2108	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2111	1	0	1	0	1	0	0	0	0	0	1.0	0.6
2110	1	1	1	0	1	0	0	0	0	0	1.0	0.8
2129	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2130	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2117	1	1	1	1	1	0	0	1	0	0	0.83	1.0
2118	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2119	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2120	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2121	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2122	1	1	1	1	0	0	0	0	0	0	1.0	0.8
2123	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2124	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2125	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2126	1	1	1	0	1	0	0	0	0	0	1.0	0.8
2127	1	1	1	1	1	1	0	0	0	0	0.83	1.0
2128	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2101	1	1	1	1	1	1	1	0	0	0	0.71	1.0
2102	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2131	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2132	1	1	1	1	1	1	0	0	0	0	0.83	1.0
2161	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2162	1	1	1	1	1	0	0	1	0	0	0.83	1.0
2163	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2164	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2129	1	1	0	1	1	0	0	0	0	0	1.0	0.8
2130	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2116	1	1	1	0	1	0	0	0	0	0	1.0	0.8
2146	1	0	1	1	1	0	0	0	0	0	1.0	0.8
2104	1	1	1	1	0	0	0	0	0	0	1.0	0.8
2182	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2107	1	1	1	1	0	0	0	0	0	0	1.0	0.8
2105	1	1	1	1	1	0	1	0	0	0	0.83	1.0
2106	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2133	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2112	1	1	1	1	1	0	0	0	0	1	0.83	1.0
2113	1	1	0	1	0	0	0	0	0	1	0.75	0.6
2114	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2115	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2109	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2169	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2134	1	1	1	0	1	0	0	0	0	0	1.0	0.8
2135	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2200	1	1	0	1	1	0	0	0	0	0	1.0	0.8
2201	1	1	0	1	1	0	0	0	0	0	1.0	0.8
2231	1	1	1	1	0	0	0	0	0	0	1.0	0.8
2254	1	1	0	1	1	0	0	0	0	0	1.0	0.8
2255	1	1	0	1	1	0	0	0	0	0	1.0	0.8
2256	1	1	0	1	1	0	0	0	0	0	1.0	0.8
2257	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2258	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2259	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2260	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2263	1	1	0	0	1	0	0	0	0	0	1.0	0.6
2264	1	0	1	0	0	0	0	0	0	0	1.0	0.4
2265	1	1	1	0	0	0	0	0	0	0	1.0	0.6
2261	1	1	0	1	1	0	1	0	0	0	0.8	0.8
2262	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2267	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2268	0	1	1	0	0	0	0	0	0	0	1.0	0.4
2206	1	1	1	1	1	0	0	0	0	0	1.0	1.0

ID	Pos1	Pos2	Pos3	Pos4	Pos5	Neg1	Neg2	Neg3	Neg4	Neg5	Precision	Recall
2207	1	1	1	1	0	0	0	0	0	0	1.0	0.8
2274	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2275	1	1	1	1	1	0	0	0	1	0	0.83	1.0
2276	1	1	0	1	1	0	0	0	0	0	1.0	0.8
2277	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2279	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2280	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2281	0	1	0	1	1	0	0	0	0	0	1.0	0.6
2282	1	1	1	1	0	0	0	0	0	0	1.0	0.8
2283	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2284	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2285	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2205	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2202	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2203	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2294	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2234	1	0	1	1	1	0	1	0	0	0	0.8	0.8
2235	1	1	1	1	1	0	0	0	0	1	0.83	1.0
2278	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2269	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2271	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2270	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2266	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2210	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2208	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2209	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2288	1	1	1	1	1	0	0	0	0	0	1.0	1.0
2286	1	1	1	1	0	0	0	0	0	0	1.0	0.8
2236	1	1	0	1	1	0	0	0	0	1	0.8	0.8
2287	1	1	1	0	1	0	0	0	0	0	1.0	0.8
2295	1	1	1	1	1	0	0	0	0	0	1.0	1.0
31103	1	0	1	1	1	0	0	0	0	0	1.0	0.8
31104	1	1	1	1	1	0	0	0	0	0	1.0	1.0
31105	1	1	0	1	1	0	0	0	0	0	1.0	0.8
3178	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3179	1	1	1	1	1	0	0	0	0	0	1.0	1.0
31108	1	1	1	1	1	0	0	0	0	0	1.0	1.0
31109	1	1	1	1	1	0	0	0	0	0	1.0	1.0
31110	0	1	1	1	1	0	0	0	0	0	1.0	0.8
31111	1	1	1	1	1	0	0	0	0	0	1.0	1.0
31112	1	1	0	1	1	0	0	0	0	0	1.0	0.8
31113	1	1	0	1	1	0	0	0	0	0	1.0	0.8
31115	1	1	0	1	1	0	0	0	0	0	1.0	0.8
31116	1	1	1	1	0	0	0	0	0	0	1.0	0.8
31114	1	1	1	1	1	0	1	0	0	0	0.83	1.0
3167	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3168	1	1	0	0	1	0	0	0	0	0	1.0	0.6
3104	1	1	1	0	1	0	0	0	0	0	1.0	0.8
31119	1	1	1	1	0	0	0	0	0	0	1.0	0.8
31120	0	1	1	1	0	0	0	0	0	0	1.0	0.6
3101	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3102	1	1	0	1	1	0	0	0	0	0	1.0	0.8
3103	1	0	1	1	1	0	0	0	0	0	1.0	0.8
3194	1	1	1	1	1	1	0	0	0	0	0.83	1.0
3134	1	0	1	1	0	0	0	0	0	0	1.0	0.6
3135	1	0	1	1	1	0	0	0	0	0	1.0	0.8
31130	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3105	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3106	1	1	1	1	1	0	1	0	0	0	0.83	1.0
3199	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3198	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3131	1	1	0	1	1	0	0	0	0	0	1.0	0.8
3132	0	0	0	1	0	0	0	0	0	0	1.0	0.2
3133	1	1	0	1	1	0	0	0	0	0	1.0	0.8
3166	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3136	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3201	1	1	1	1	1	0	0	0	0	0	1.0	1.0
32135	1	1	1	1	0	0	0	0	0	0	1.0	0.8

ID	Pos1	Pos2	Pos3	Pos4	Pos5	Neg1	Neg2	Neg3	Neg4	Neg5	Precision	Recall
32132	1	1	1	1	1	0	0	0	0	0	1.0	1.0
32133	1	1	1	1	1	0	0	0	0	0	1.0	1.0
32136	1	0	1	1	1	0	0	0	0	0	1.0	0.8
32137	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3255	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3256	0	1	1	0	1	0	0	0	0	0	1.0	0.6
3257	0	1	0	1	1	0	0	0	0	0	1.0	0.6
3267	0	1	1	1	1	0	0	0	0	0	1.0	0.8
3202	1	1	1	1	0	0	0	0	0	0	1.0	0.8
3203	1	1	1	1	1	0	0	0	0	0	1.0	1.0
32148	1	0	0	0	1	0	1	0	1	0	0.5	0.4
32149	1	1	1	1	0	0	0	0	0	0	1.0	0.8
32150	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3208	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3234	1	0	1	1	1	0	0	0	0	0	1.0	0.8
3235	0	1	1	1	1	1	0	0	0	0	0.8	0.8
3207	1	1	1	1	1	0	0	0	0	0	1.0	1.0
3231	1	1	0	1	1	0	0	0	0	0	1.0	0.8
3232	1	1	1	1	1	0	1	0	0	0	0.83	1.0
3236	0	1	1	1	1	0	0	0	0	0	1.0	0.8