Measuring the Influence of Text Normalization on the Prediction of Text Complexity of Learner Productions

Jeanette Bewersdorff **Josef Ruppenhofer** **Torsten Zesch**
CATALPA – Center of Advanced Technology for Assisted Learning and Predictive Analytics,
FernUniversität in Hagen, Germany
Correspondence: jeanette.bewersdorff@fernuni-hagen.de

Abstract

Text complexity prediction typically relies on features extracted from written texts, but most work assumes error-free input. We investigate how learner language errors affect such features, using the German subset of the MERLIN corpus, which provides both learner productions and manually corrected target hypotheses. We compare features computed on learner texts, on corrected texts, and in mixed "split" settings in an automatic CEFR level prediction task. Overall performance is very similar across all settings and largely driven by simple length features such as sentence count, which are robust to errors. By contrast, OOV- and error-related syntactic features behave quite differently depending on whether they are computed on learner or corrected texts, an analytical split that assigns spelling-related OOV to learner texts and lexical OOV to corrected texts works best in OOV-only experiments. Our results suggest that manually corrected target hypotheses are particularly useful for designing error-sensitive text complexity features. All experiments and code are made publicly available.

1 Introduction

Text complexity refers to the degree to which a written text is challenging to read and understand. Predicting text complexity relies on measuring a set of text characteristics that (usually taken as a whole) are indicative of low or high complexity. A very large number of such text characteristics (often called features) have been proposed (Graesser et al., 2004; Schwarm and Ostendorf, 2005; Vajjala and Meurers, 2012; François and Fairon, 2012) and their interactions been studied (Feng et al., 2010; Vajjala and Meurers, 2014). An understudied area is how language errors made by the learners impact the measurements. A feature that works very well for an edited, error-free reading text might yield unacceptable results when applied to a learner text.

Zipf Value	Frequency Band	Learner Production
0.00	OOV	Teh
0.00	OOV	lectur
6.82	F6	was
0.00	OOV	ful
7.40	F7	of
0.00	OOV	ultraceprn
4.86	F4	claims

Zipf Value	Frequency Band	Target Hypothesis
7.73	F7	The
4.14	F4	lecture
6.82	F6	was
5.54	F5	full
7.40	F7	of
0.00	OOV	ultracrepidarian
4.86	F4	claims

Figure 1: An example of an learner production with corresponding target hypothesis. Zipf value means $\log_{10}$ frequency per billion words.

For example, Figure 1 shows a sentence and the values assigned from a feature measuring corpus frequency and using the log probability to establish *frequency bands*. Depending on the calculated Zipf value, the token is categorized in one of seven frequency bands or in the out of vocabulary category (OOV), meaning that the word was not recognized, either because of spelling mistakes or because of very specific vocabulary that is not covered by the underlying dictionary. If backed by a suitable corpus, we might see results as on the right side where tokens like 'the' are correctly identified as being high frequency, while 'lecture' or 'claims' are marked as less frequent and hence more complicated. The feature works as intended. Now compare that to the left side, where the same feature is applied to error-prone learner text. Rare misspellings are not found in the corpus at all. Moreover, in the learner production, there is no distinction between spelling errors and real OOV tokens like 'ultracrepidarian', which might even be correctly spelled but are not recognized by the feature.

In this paper, we first analyze how vulnerable established features are to language errors as made by language learners. For this purpose, we use the German subset of the MERLIN corpus, which

includes both L2 learner productions and corresponding manual target hypotheses.

Based on this analysis, we then conduct a study to automatically assign CEFR levels to the learner productions. This helps us understand the impact of our analytical results on a practical task.

We make our experiments and code publicly available at https://github.com/JeaBew/text-normalization-complexity.

2 Text Complexity Features

The range of features that have been proposed for use in text complexity measurement is large (Graesser et al., 2004; Schwarm and Ostendorf, 2005; Feng et al., 2010; Vajjala and Meurers, 2012; François and Fairon, 2012; Collins-Thompson, 2014; Chen and Meurers, 2016). However, for our intended goal of investigating the impact of error-prone texts on feature results, we utilize a selection of features from the most important feature groups. We discuss each group in detail in the remainder of this section.

Correlation Analysis For detecting interesting features that work differently on the learner production or the target hypothesis, we conduct a correlation analysis on our features between the learner production and target hypothesis. By that, we determine which features do change significantly on the target hypothesis from the learner production. This allows us to slim the feature set down to a smaller set containing only features relevant for our setting.

In all correlation analyses, we report Spearman's ρ, as CEFR levels are ordered categories (A1 $<$ A2 $<$ B1 $<$ B2 $<$ C1 $<$ C2) and are therefore more appropriately treated as ordinal rather than interval-scaled. Moreover, Spearman's rank correlation does not rely on assumptions of linearity or normality and is invariant under monotonic transformations, making it less sensitive to scale differences between views, while still capturing the strength and direction of the association. A high value indicates both views are similar for the respective feature, a low value indicates a greater difference between the views, making the respective feature interesting for our experiments.

2.1 Length Features

Length features capture basic quantitative properties of a text, such as the number of tokens, sentences, or characters. They provide essential information about the structure and complexity of a text and are commonly used in readability assessment and linguistic analysis. As length features, we calculate the token count, sentence count, average token length and the average token count per sentence.

Length Feature	ρ_S
Token Count	1.0
Sentence Count	.95
∅ Token Length	.95
$^{\text{token}}/_{\text{sentence}}$	.91

Table 1: Length features. For each feature, we report the Spearman rank correlation ρ_S between learner production and target hypothesis. Higher correlations indicate greater cross-view similarity.

As shown in Table 1, the correlation of length feature values between learner production and target hypothesis is generally high. However, features including a sentence component show lower values. Analyzing the results, we find that sentence counts on the learner productions are often incorrectly computed by the preprocessor we use (spaCy[1]), especially for lower CEFR levels, where learners frequently omit punctuation or link sentences with commas instead of full stops. Figure 2 illustrates a learner text without punctuation and its manually punctuated target hypothesis. Consequently, the 14 sentences in the normalized target hypothesis are (incorrectly) analyzed as 9 sentences on the learner production. The preprocessor gets similarly confused by texts written completely in uppercase, also leading to results similar to Figure 2.

The difference between the length features is not high when looking at the correlations, but better results should be obtained on the target hypothesis as we have noticed the preprocessor having problems with texts without proper punctuation. Because of that, we decide to keep the Sentence Count feature despite its high correlation value. As a result of our analysis, we conclude:

Length Features

We decide to keep the **Sentence Count** and calculate it on the **target hypothesis**.

[1] https://spacy.io/

L1. Hallo Jens Herzlich Glückwünsch Sie Sint Fata geworden ich freumich fur dich Vie viel wigt das Bebi und wie heist das Bebi istas aien Junge oder ain Mötchen Istes alles

L2. oke wigetes die Mamma gehtes

L3. ir

L4. gut ist ales oke mit

L5. ir

L6. Bitte wen Sie raus ist ruf mich am Ich wolte vor bei kommen

L7. Viele Gruse

L8. Deiner Eva

(a) Learner production, 8 sentences.

T1. Hallo Jens !

T2. Herzlichen Glückwunsch !

T3. Sie sind Vater geworden .

T4. Ich freue mich für dich .

T5. Wie viel wiegt das Baby und wie heißt das Baby ?

T6. Ist es ein Junge oder ein Mädchen ?

T7. Ist alles ok ?

T8. Wie geht es der Mama ?

T9. Geht es ihr gut ?

T10. Ist alles ok mit ihr ?

T11. Bitte , wenn sie raus ist , ruf mich an .

T12. Ich wollte vorbeikommen .

T13. Viele Grüße

T14. Deine Eva

(b) Target hypothesis, 14 sentences.

Figure 2: Example text from MERLIN corpus. In the learner production only 8 sentences are found by the spaCy preprocessor, while in the target hypothesis there are 14.

2.2 Frequency Features

The intuition behind frequency-based features is that advanced learners are more likely to use less frequently occurring words. There are resources that directly encode this through a mapping from word forms (or lemmas) to CEFR levels. As these resources are often derived through corpus counts, an alternative operationalization is to measure the frequency of a word in a learner corpus, usually in the form of *frequency bands*.

As briefly mentioned in the Introduction, we calculate the Zipf value for each token in a text. The Zipf value of a token is the base-10 logarithm of its occurrences per billion tokens. Based on the value, we categorize each token to one of 7 frequency bands, F1 (very rare) to F7 (very common). We also added an eighth category called OOV (out of vocabulary) for tokens not recognized by the algorithm.

We then calculate the distribution of the different frequency bands over the respective text, resulting in our final values used for the classification.

Frequency Feature	ρ_S
$^{F1}/_{token}$	.53
$^{F2}/_{token}$	.55
$^{F3}/_{token}$	.82
$^{F4}/_{token}$	.83
$^{F5}/_{token}$	.86
$^{F6}/_{token}$	.88
$^{F7}/_{token}$	.90
$^{OOV}/_{token}$	.67

Table 2: For each feature, we report the Spearman rank correlation ρ_S between learner production and target hypothesis. Higher correlations indicate greater cross-view similarity.

Table 2 reports cross-view correlations between learner production and target hypothesis for the frequency-based features. The lower bands ($^{F1}/_{token}$, $^{F2}/_{token}$) show only moderate Spearman correlation. In contrast, the higher bands ($^{F3}/_{token}$ to $^{F7}/_{token}$) display strong Spearman correlation, suggesting that their distributions are largely stable across learner production and target hypothesis.

The OOV band shows medium correlation, consistent with the assumption that orthographic errors primarily affect less frequent lexical items and are partially reassigned to regular frequency bands in the corrected view.

As tokens need to be recognized in order to be categorized into a frequency band, we compute all frequency features on the target hypothesis, so that spelling errors in the learner production do not interfere with the frequency categorization. Since the ratios for $^{F3}/_{token}$ to $^{F7}/_{token}$ exhibit high cross-view correlations and little systematic view-specific variation, we exclude these features from the final feature set used in the experiments. We also omit OOV, as we use another OOV feature not bound to the frequency and do not want do have two correlating features.

As a result of our analysis, we conclude:

Frequency Features

We decide to keep **F1/token** and **F2/token** and calculate them on the **target hypothesis**.

2.3 POS Features

Part-of-speech (POS) tags provide a syntactic classification of tokens into grammatical word classes such as nouns, verbs, determiners, and punctuation marks. POS-based features are widely used in automatic readability and proficiency assessment, as they capture aspects of lexical choice and syntactic realization beyond simple length measures. In order to not inflate the feature list and see the influence of POS groups instead of single POS tags, we computed the ratios in a coarse grained manner, meaning we combined all subcategories of a certain POS tag to one superordinate POS tag (e.g. *VAFIN*, *VAIMP* and *VMINF* will all be annotated as *POS_VERB*).

The cross-view correlations of the POS ratios (Table 3) shows moderate to high correlation values for the respective features. For most POS categories, Spearman values are high, indicating that their text rankings are very similar across learner production and target hypothesis and thus largely redundant and view-stable.

Based on the cross-view correlations in Table 3, most POS ratios show very similar behavior between learner productions and target hypotheses and are therefore largely redundant for our purposes.

We retain only four POS ratios in the final feature set: NOUN/token , DET/token , PROPN/token , and PUNCT/token . Other POS ratios, including the catch-all category X/token , are excluded because they are either highly view-stable or too rare and heterogeneous to provide a clear, interpretable signal. We compute all POS ratios on the learner production, not on the corrected target hypothesis. This ensures that POS distributions are based only on words the learner actually writes, and that no additional forms are introduced by correction. Using the target hypothesis could add POS tokens the learner does not yet command and thus distort the POS-based feature profiles.

As a result of our analysis, we conclude:

POS Features

We decide to keep **NOUN/token** , **DET/token** , **PROPN/token** , and **PUNCT/token** and calculate them on the **learner production**.

POS Feature	ρ_S
ADJ/token	.86
ADP/token	.90
ADV/token	.92
CONJ/token	.90
DET/token	.88
NOUN/token	.85
NUM/token	.92
PRON/token	.94
PROPN/token	.89
PUNCT/token	.65
VERB/token	.88
X/token	.78

Table 3: POS features. For each feature, we report the Spearman rank correlation ρ_S between learner production and target hypothesis. Higher correlations indicate greater cross-view similarity.

2.4 Syntax Features

The syntactic complexity features we use quantify, directly or indirectly, the length of production units and the amount of embedding and coordination. The dependency length features target the demands of cognitive processing (Gibson, 1998). The syntactic features are extracted from parses of the target hypotheses using the udapi api (Popel et al., 2017) for dependency trees. Our full syntax feature list with descriptions can be found in Appendix A.

Table 4 reports Spearman correlations between learner production and target hypothesis for the syntactic features. Most general syntactic complexity measures (e.g. dependency-length metrics) show very high cross-view correlations, indicating that they are largely view-stable.

By contrast, several error-related measures (*Proportion of brackets with empty midfields*, *Proportion of canonical brackets*, *Proportion of subjectless finite verbs*) exhibit lower Spearman correlations.

Our results show that error-related syntactic features change strongly between learner production and target hypothesis, while general syntactic mea-

Syntax Features	ρ_S
∅ Dependency Length Left	.89
∅ Dependency Length Right	.83
∅ Dependency Length All	.90
∅ Size Of Lexical NP	.68
Proportion S without finite verb	.83
Proportion S without verb	.84
Proportion of canonical brackets	.67
Proportion of brackets with empty midfields	.66
Proportion of Subj Vfin Inversions	.77
Proportion of subjectless finite verbs	.47
Proportion of coordination between verbs	.80

Table 4: Syntax features. For each feature, we report the Spearman rank correlation ρ_S between learner production and target hypothesis. Higher correlations indicate greater cross-view similarity.

sures are almost identical. We therefore keep only these view-sensitive measures (∅ Size of lexical NP, Proportion of canonical brackets, Proportion of brackets with empty midfields, Proportion of subjectless finite verbs) and drop the view-stable ones. Since syntactic annotation is more reliable on well-formed, correctly spelled sentences, we compute all syntactic features on the target hypothesis.

As a result of our analysis, we conclude:

Syntax Features

We decide to keep ∅ **Size of lexical NP, Proportion of canonical brackets, Proportion of brackets with empty midfields** and **Proportion of subjectless finite verbs** and calculate them on the **target hypothesis**.

2.5 Error Features

We operationalize language errors as words not appearing in a dictionary, i.e. words that are out-of-vocabulary (OOV). However, as no dictionary can ever be fully comprehensive, OOV is a mix of spelling errors (typically more OOV/spelling errors meaning a lower language competency) and rare words (typically attributed to higher language competency). Thus, the OOV feature can work on either the learner language or the target hypothesis.

We calculate the feature as the ratio of OOV tokens to all tokens and use it for the classification. Moreover, we utilize the difference between the views to derieve a sub-OOV feature representing the OOV tokens that are spelling mistakes. As this sub-feature is created using both views, we exclude

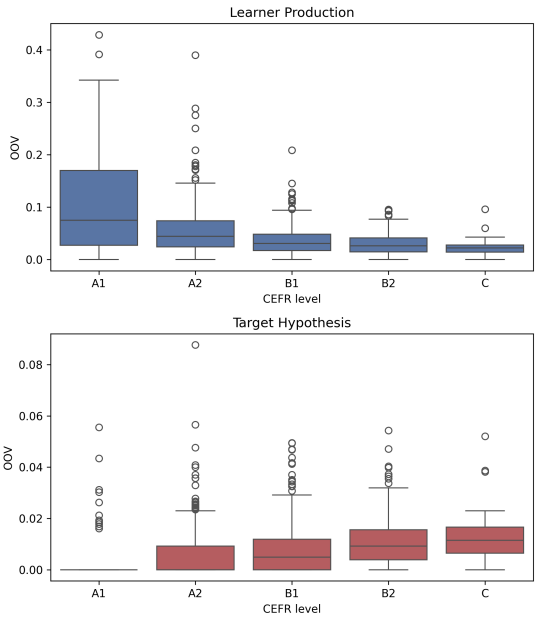

Figure 3: Box plot of spelling mistake OOV tokens on the learner production and real OOV tokens on the target hypothesis for the different CEFR levels.

it from the correlation analysis.

OOV Feature	ρ_S
OOV ratio	.25

Table 5: OOV feature. For each feature, we report the Spearman rank correlation ρ_S between learner production and target hypothesis. Higher correlations indicate greater cross-view similarity.

Table 5 presents the cross-view correlation between learner production and target hypothesis for the $^{\text{OOV\_token}}/_{\text{token}}$ feature. Spearman correlation is low, indicating a weak association between the OOV ratios in the two views and substantial differences between learner texts and their corrections.

Figure 3 illustrates this pattern by comparing $^{\text{spelling errors}}/_{\text{token}}$ on the learner production and $^{\text{OOV\_token}}/_{\text{token}}$ on the target hypothesis across CEFR levels.

Overall, OOV values are consistently higher in learner productions than in target hypotheses across all CEFR levels. The difference decreases with increasing proficiency, indicating that advanced learners' vocabulary use becomes closer to the target hypotheses.

Within the learner data, the spelling-mistake ratio shows a clear downward trend as proficiency increases. By contrast, OOV rates in the target hypotheses remain comparatively low but rise

Category	Feature	Split
Length	Sentence Count	T
POS	DET/token	L
	NOUN/token	L
	PROPN/token	L
	PUNCT/token	L
Frequency	F1	T
	F2	T
	OOV	T
Syntax	∅ Size Of Lexical NP	T
	Proportion of canonical brackets	T
	Proportion of brackets with empty midfields	T
	Proportion of subjectless finite verbs	T
OOV	spelling_oov_token/token	L
	real_oov_token/token	T

Table 6: Implemented features and their mapping to either learner text or target hypothesis.

gradually from level to level.

The OOV token ratio as per the correlation analysis should be calculated on the learner production, as there is a significant difference visible between the two views. But as we are able to distinguish between spelling mistakes and real OOV tokens, we split the feature for the experiments. The spelling error tokens are calculated on the learner production, and the real OOV tokens are calculated on the target hypothesis.

As a result of our analysis, we conclude:

Error Features

We decide to split the OOV feature into spelling related OOV tokens and real OOV tokens and calculate **spelling error tokens** on the **learner production**, and **real OOV tokens** on the **target hypothesis**.

2.6 Final Feature Set

Table 6 shows our condensed feature set and for each feature whether we think it should be better measured on the target hypothesis or the learner production.

3 CEFR Experiments

So far, we have found that language errors made by the learners impact how accurately single text complexity features capture learner proficiency. We now turn to CEFR classification, as an application combining multiple text complexity measures to extrinsically test how much this influences CEFR classifiation.

	A1	A2	B1	B2	C1	C2	All
original	57	306	331	293	42	4	1033
balanced	46	46	46	46	46	-	230

Table 7: Distribution of data by CEFR level. In the balanced set, we combined C1 and C2 into one level, and downsampled al other classes to 46 in order to match the merged C level.

3.1 Dataset

For our experiments, we used the CEFR annotations from the MERLIN dataset (Boyd et al., 2014) (German subset), also used in the feature analysis. Table 7 shows the distribution of texts over the different CEFR levels. The dataset is clearly imbalanced: most texts are rated A2, B1, or B2, while there are relatively few A1 and C1 texts and only four C2 texts. This imbalance poses challenges for training, as several classes are strongly underrepresented. To mitigate this, we merged the few C2 texts into the C1 class, called C from now on. In addition, we created a balanced subset by downsampling each class to 46 texts, so that all classes contribute equally (230 texts in total).

Balancing the dataset has the advantage that the classifier is not dominated by the majority classes and that performance differences between classes and view conditions are easier to compare. We accept the disadvantage of using fewer training instances in order to focus on the effect of the feature-to-view assignments themselves, rather than on optimising absolute classification performance in a particular data setting.

3.2 Experimental Setup

For our experiments, we use the following four feature configurations.

- **All-learner** All features are computed on the learner production.
- **All-target** All features are computed on the target hypothesis.
- **Split-analytical** Each feature is either computed on the learner production or the target hypothesis, depending on what we analytically consider to be more sensible. See Table 6 for an overview.
- **Split-counterfactual** Each feature is computed on the 'wrong' layer, inverting the ana-

	A1	A2	B1	B2	C1	avg
Learner	.64	.52	.27	.48	.55	.49
Target	.64	.36	.33	.49	.48	.46
Analytical	.65	.35	.33	.49	.48	.46
Counterfactual	.65	.52	.29	.49	.55	.50

Table 8: CEFR classification results of logistic regression in term of F_1 for our four feature-to-layer assignment conditions. We show macro averaged results and a result breakdown per CEFR level.

lytical decisions made for the split-analytical configuration.

For the classification, we use logistic regression. It is a simple, well-established linear model that provides stable results on relatively small datasets and allows us to interpret the influence of individual features. Most importantly, its limited complexity helps to isolate effects of the feature-to-view assignments themselves, without introducing additional variability from more powerful but harder-to-interpret models.

We train multinomial logistic regression models using Scikit-learn. We used the default `lbfgs` solver with L_2 regularization (`penalty="l2"`, $C = 1.0$). To ensure convergence, we set the maximum number of iterations to `max_iter=5000`. For reproducibility, we fixed the random seed (`random_state=42`). Evaluation was performed with 5-fold stratified cross-validation (`StratifiedKFold`, `n_splits=5`, `shuffle=True`, `random_state=42`), and overall performance metrics (e.g., accuracy and F_1) were computed from cross-validated predictions obtained via `cross_val_predict`.

To assess the impact of the choice of view on CEFR classification, we trained four models with identical hyperparameters on the described feature splits. In all conditions we used the same set of texts, balanced across CEFR levels, 46 texts per level, 230 texts in total.

3.3 Results

Table 8 shows the results for the classification. With the full reduced feature set, all four view configurations yield very similar CEFR classification performance. The learner-based and counter-split settings perform only slightly better than the target-based and analytical splits, and the qualitative pattern across levels remains stable (A1 and C are easiest to distinguish, B1 is hardest).

	removed feature group:				
	Length	POS	Syntax	Freq.	OOV
Learner	.40	.50	.42	.49	.50
Target	.35	.45	.41	.46	.50
Analytical	.39	.45	.42	.46	.46
Counterfactual	.38	.50	.40	.50	.50

Table 9: Ablation results removing feature groups. Average CEFR classification results of logistic regression in terms of F_1 for our four feature-to-layer assignment conditions.

	A1	A2	B1	B2	C1	avg
Learner	.58	.18	.04	.17	.40	.27
Target	.23	.28	.00	.16	.38	.21
Analytical	.57	.14	.12	.19	.40	.29
Counterfactual	.57	.15	.04	.17	.39	.26

Table 10: CEFR classification results of logistic regression in terms of F_1 when using only OOV-based features under different view configurations.

In order to explain these results, we conducted an ablation study in which we removed one feature category at a time to determine which categories have the greatest impact on the classification.

The results in Table 9 show that removing the length features leads to the strongest drop in performance across all view configurations, whereas removing POS, frequency or OOV features has only a small effect. Syntactic features have a moderate impact, but not as high as length. This indicates that our CEFR classifier is driven mainly by simple length information, while the other feature groups contribute comparatively little. Since these length features are also the ones with the highest cross-view correlation between learner production and target hypothesis, it is expected that the overall classification results change only minimally when switching between view assignments in the full-feature setting.

OOV Experiments Since our feature analysis revealed the largest cross-view differences for OOV-based measures, we examine these features in isolation to assess whether their view differences translate into differences in CEFR classification. Table 10 shows the results. Here, a clear view effect can be seen.

OOV measures computed on the learner view primarily encode orthographic difficulty and thus sharply separate very low and very high proficiency, but provide little resolution for intermediate levels. OOV measures computed only on the corrected target hypotheses shift the signal towards lexical sophistication, making A1 less clearly separable and leaving B1 largely unprofiled. The analytical split, which assigns spelling-related OOV to the learner view and genuine lexical OOV to the target view, combines both aspects and performs best among the OOV-only configurations, particularly for the intermediate levels; the counter split does not achieve the same improvement. Overall, this shows that while view choice has only a moderate effect in the full-feature setting due to dominant, view-stable features, it is crucial for OOV-based measures, and the analytical split is the empirically most effective configuration.

4 Conclusion

We have investigated how the choice of annotation view in a multi-view learner corpus, original learner production vs. manually corrected target hypothesis, affects feature-based modelling of text complexity. Using German L2 essays from MERLIN, we represented each text as a multi-view CAS and compared four feature-to-view assignments in an CEFR classification task. Overall, the global prediction performance of a multinomial logistic regression remains remarkably stable across all configurations: whether all features are computed on the learner view, on the target hypothesis, or split between them according to analytical or counterfactual criteria, aggregate scores change only marginally. An ablation study reveals that this stability is largely driven by a small set of very robust, view-invariant length features (in particular sentence count), which already capture most of the signal exploited by the classifier.

At the same time, our focused OOV experiments show that the view choice has a substantial impact on what more fine-grained complexity indicators actually measure, and on which contrasts they emphasise. Learner-based OOV features primarily encode orthographic well-formedness plus rarity and thus act as a coarse indicator of very low vs. very high proficiency, with limited resolution in the middle of the scale. In contrast, OOV features computed on the target hypothesis mainly reflect lexical sophistication, as spelling errors have been removed, and consequently shift the relative separability of individual levels. When we explicitly disentangle spelling-OOV (computed on the learner view) from lexical-OOV (computed on the target view), the analytical split that aligns spelling-related OOV with the learner production and genuine lexical OOV with the corrected text yields the most informative and complementary representation and performs best among the OOV-only configurations.

These findings suggest that manually corrected target hypotheses can be particularly valuable for automated text complexity rating when designing error-related features: they allow us to decouple orthographic accuracy from lexical sophistication and to assign each aspect to the view where it is most meaningfully captured. While the overall prediction performance may appear largely view-robust for a broad feature set, the interpretation and usefulness of individual features depend crucially on how learner and target views are exploited. We therefore argue that future text complexity models built on multi-view learner corpora should make their view choices explicit and theoretically motivated, especially for error-sensitive feature families.

Limitations

Due to the language-specific nature of many features used in text complexity classification, we only report results for the German language. While it is highly likely that similar results hold also for other languages, we do not provide empirical evidence here, but invite future research to investigate this issue.

We rely on manually created target hypotheses to observe the raw effect without compounding errors. For a real-life application, one would have to apply automatic target hypotheses.

Another reason why text complexity measures might be inaccurate are textual borrowings (Shi, 2004; Chan et al., 2015). If learners copy material from the task description or from a reading text, the complexity of the reading text might be correctly measured, but it would be wrong to attribute the corresponding complexity level to the learner.

Acknowledgments

This work was conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany.

References

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sathena Chan, Chihiro Inoue, and Lynda Taylor. 2015. Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, 26:20–37.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 113–119, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284.

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

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 523–530.

Ling Shi. 2004. Textual borrowing in second language writing. *Written Communication*, 21(2):171–200.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2):194–222.

A Syntax Features Description

Short description of the syntax features we examined.

∅ **Dependency Length All/Left/Right** Average absolute difference in token index between a dependent and its head (All); restricted to dependencies where the head precedes the dependent (Left) or follows it (Right).

Lexical NP size Average length (in tokens) of lexically headed noun phrases.

Proportion S without finite verb Proportion of sentences that do not contain a finite (conjugated) verb.

Proportion S without verb Proportion of sentences that do not contain any verb (finite or non-finite).

Proportion of canonical brackets Proportion of sentences in which the verbal bracket appears in canonical (standard) order.

Proportion of brackets with empty midfields Proportion of sentences in which no constituents occur in the middle field between the parts of the verbal bracket.

Proportion of Subj Vfin Inversions Proportion of sentences in which the subject appears after the finite verb (subject–verb inversion).

Proportion of subjectless finite verbs Proportion of finite verbs that occur in sentences without an explicit subject.

Proportion of coordination between verbs Proportion of coordinations (e.g. with *and*, *or*) that link two or more verbs.