MWE for Essay Scoring English as a Foreign Language

Rodrigo Wilkens*, Daiane Seibert†, Xiaoou Wang*, Thomas François*
*Cental, IL&C, UCLouvain, †KU Leuven,
rodrigo.wilkens@uclouvain.be, daiane.seibert@student.kuleuven.be,
{xiaoou.wang, thomas.francois}@uclouvain.be

Abstract

Mastering a foreign language like English can bring better opportunities. In this context, although multiword expressions (MWE) are associated with proficiency, they are usually neglected in the works of automatic scoring language learners. Therefore, we study MWE-based features (i.e., occurrence and concreteness) in this work, aiming at assessing their relevance for automated essay scoring. To achieve this goal, we also compare MWE features with other classic features, such as length-based, graded resource, orthographic neighbors, part-of-speech, morphology, dependency relations, verb tense, language development, and coherence. Although the results indicate that classic features are more significant than MWE for automatic scoring, we observed encouraging results when looking at the MWE concreteness through the levels.

Keywords: multiword expression (MWE), MWE feature analysis, MWE concreteness, automatic essay scoring

1. Introduction

Mastering a foreign language has become increasingly important in everyday life. English proficiency, for example, is correlated to higher salaries (e.g., Boyd and Cao (2009; Pendakur and Pendakur (2007; Adamchik et al. (2019)). The increase of foreign language learners also implies an increasing number of participants in the proficiency tests, such as TOEFL and IELTS, which may impact the test cost (e.g. including the need for training new evaluators). Automated scoring makes assessing language proficiency more viable for large-scale tests, which may be mandatory if one wants to study abroad (Weigle, 2013). In addition, the feedback provided by automated scoring based on linguistic features can also provide valuable insights to facilitate language learning (Srichanyachon, 2012).

For English, various tools have been used to support the development of research on foreign language writing development. Some examples are Coh-Metrix (Graesser et al., 2004), L2 Syntactic Complexity Analyzer (Lu, 2010), CTAP (Chen and Meurers, 2016) and TAASSC (Kyle, 2016). Although these tools provide a myriad of functional language descriptors, they are hardly extensible. Also, they are usually based on token units or n-grams as words to build features. However, multiwords expressions raise numerous challenges in natural language processing, descriptive linguistics and foreign language acquisition due to their formulaic structure (Wray, 1999; Wray, 2002), unit at some level of description (Calzolari et al., 2002), and interpretation crossing word boundaries (Sag et al., 2002). MWEs include several subcategories, such as verb-noun combinations (e.g. *rock the boat* and *see stars*), verb-particle constructions (e.g. *take off* and *clear up*), lexical bundles (e.g. *I don't know whether*) and compound nouns (e.g. *cheese knife* and *rocket science*). Targeting English as a foreign language, MWE's importance is undeniable when considering its ubiquity in the discourse produced by native speakers. Moreover, a learner may be considered handicapped in a language without knowledge about MWE (Muraki et al., 2022). Glucksberg (1989) estimated that English native speakers produce about four multiwords per minute and Jackendoff (1997) identified that they likely have the same order of magnitude as a single word in the mental lexicon of native speakers.

Given the prevalence of MWEs in native speakers' speech, we investigate their impact on learners' proficiency prediction. We compare MWE metrics with classic linguistic ones commonly used to identify learner proficiency to achieve this goal. In particular, we focus on MWEs and their concreteness (i.e., degree of concreteness/abstraction of an MWE). The main contributions of this paper are the following: (1) profile of MWE concreteness usage across the different levels of the Common European Framework of Reference for Languages (CEFR); (2) analysis of the capacity of MWE scores to individually identify the level; and (3) comparison of these scores with classic scores used to predict learners' level.

This work is organized as follows: first, we shortly review the literature concerning the essay scoring focusing on English and linguistic descriptors in Section 2. In Section 3, we present the linguistic descriptors and corpus used in this work. Next, in Section 4, we evaluate the impact of MWE descriptors on the prediction of learners' proficiency. Finally, we conclude by discussing the results in Section 5.

2. Related Work

Approaches for automatic prediction of language proficiency are mostly based on machine learning. These can be broadly divided into deep learning-based and feature-based, the latter being more interpretable. We thus focus on feature-based approaches for facilitating the comparison with the MWE descriptors.

The features have been drawn from explorations of linguistic patterns in corpora. For example, Lan et al. (2022) showed that there is an association between the use of noun phrases and whether the author is an L1 or L2 user of English. The first language plays a vital role in the developmental trajectories, characterizing behavior, as discussed by Chen et al. (2021), who observed different developmental trajectories in learners whose L1 has clause subordination structures distinct from English. They may overuse or underuse certain grammatical structures depending on their CEFR level (Zilio et al., 2018). Errors, such as punctuation, spelling and verb tense, are significant in predicting specific CEFR levels (Ballier et al., 2019). Jung et al. (2019) demonstrate relevance regarding the conceptual similarity between paragraphs when comparing with the lexical diversity, familiarity and abstractness of the word. Some works also combined properties such as part-of-speech and n-grams (Yannakoudakis et al., 2011), the edit distance between errors and their corresponding target hypothesis (Tono, 2013), and syntactic, lexical, discourse and error features (Vajjala, 2018).Jung et al. (2019) showed that length-based features, specifically the number of words, are stronger predictors than the cohesion and syntactic complexity. However, they also emphasize that text length alone cannot be considered a good predictor of writing quality.

Moreover, despite the variety of language-based features studied, only a few studies have tried to test multi-dimensional models with several features to investigate how they are comparable (e.g. (Tack et al., 2017)). Corpus specificities may also bias studies. In EFCAMDAT (Geertzen et al., 2013), the task (i.e., the prompt[1]) presented in the test might drive the learner to use different skills, as discussed by Alexopoulou et al. (2017) and by Michel et al. (2019), who identified task influence by exploring lexical and syntactic features.

Despite the amount of work on language assessment, there is still a comparability gap in the results. In this sense, Ballier et al. (2020) called for solutions for predicting CEFR levels for written productions using only the French part of the EFCAMDAT. Competitors used a variety of machine learning approaches with different processes including feature engineering, data representation and classification. The winner, Balikas (2018), used Gradient Boosted Trees and compared the use of language models, part-of-speech, bag-of-words (BoW) and Latent Dirichlet Allocation (LDA) as features. Interestingly, their results of both BoW and LDA models were close. Arnold et al. (2018) use a multi-dimensional feature representation of written essays exploring LSTM and dense layers achieving an accuracy of 70%. Using EFCAMDAT texts written by French and Spanish learners, Gaillat et al. (2021) achieved an accuracy of 82% when exploring microsystems, identifying lexical and syntactic features as the more significant.

[1]Prompts are the proposed topics for the writing.

Focusing on MWE, the literature has reported different effects depending on their type. Römer (2019) and Römer and Berger (2019) studied the verb-argument construction (VCP) repertoire of English learners, remarking an increase in vocabulary, productivity and complexity according to learners' level. Du et al. (2022) studied collocation usage by English learners, using a list of 2,501 *make/take+noun (the direct object)*. They observed that proficient learners tend to use collocations containing more semantically complicated and abstract nouns. Garner (2016) examined the use of p-frames[2] by L1 German learners of English as a foreign language, observing that p-frames in texts from higher proficiency learners are more variable, less predictable, and more functionally complex. Arnon and Snider (2010) explored the perceived transparency affected by multiword phrases (MWP; the specific combinations of words that occur together more than would be predicted by chance). For that, they compared *verb+object* phrase[3] knowledge among intermediate and advanced L2 English learners in comparison to monolingual L1 speakers, observing that intermediate learners performed less accurately and advanced learners performed comparably with native English on transparent and semi-transparent items but were less accurate for non-transparent items. Moreover, both intermediate and advanced learners answered non-transparent items less accurately than transparent items. Exploring MWE validity, Dahlmann and Adolphs (2007) studied pauses in various instances of very frequent extracted MWE candidates (i.e. n-grams) from a learner corpus. Arnon and Snider (2010) studied the frequency of four-word phrases using the distributional information, identifying an association between frequency and the identification as a valid MWE. Based on n-grams statistics, Jung et al. (2019) identified a correlation between their frequency and essay score.

3. Methodology

Considering the goal of investigating the impact of MWE usage on the prediction of learners' proficiency, we annotated a corpus of essays written by English learners with features describing MWE occurrence and its concreteness. We also annotate the corpus with additional features aiming to assess the importance of MWE features. After we have the annotated corpus, we run the tests described in Section 4.

We used EFCAMDAT (Geertzen et al., 2013), created by the University of Cambridge and Education First (EF) to supply the lack of data for numerous speakers across the proficiency spectrum and the amounts of annotated data. In total, it consists of +1M of essays across the 6 CEFR levels written by learners of

[2]P-frames are a type of semi-fixed word sequence in which fixed words surround an open slot (Stubbs, 2007).

[3]For example, break a bone (Transparent); break the silence (Semi-transparent); break the ice (Non-transparent).

198 nationalities. Levels and nationalities are not balanced (e.g. 40% of all texts are from Brazilians, and 53.04% and 0.16% of the texts are at levels A1 and C2, respectively). Therefore, we selected only the 10 most common nationalities and joined levels C1 and C2 due to their low representation in the corpus. We also truncated the number of essays using the level with the least essays by nationality. Table 3 presents the corpus size employed in this work, identifying the number of essays considered in each level for each nationality.

Nationality	Usage per level	Corpora (%)
Brazil	2469	22.99
Germany	2469	22.99
Italy	1238	11.53
Russia	1195	11.13
France	818	7.62
Mexico	762	7.09
China	555	5.17
Saudi Arabia	468	4.36
Japan	420	3.91
Taiwan	347	3.23

Table 1: Number of used texts for each nationality and its percentage in corpus used in this study.

For studying the impact of MWE on text produced by English learners, we explored 2 features:

1. MWE usage ($\mathbf{MWE_{cnt}}$) a list-based (Muraki et al., 2022) feature that consists of 62 thousand expressions from recommended expressions for learners, stimuli expressions used in language studies, dictionaries and n-grams frequency lists.

2. Concreteness of MWE (Muraki et al., 2022) $\mathbf{MWE_{conc}}$. In other words, how the 62 thousand MWE are perceived as concrete/abstract according to 2,825 participants (all English native speakers). [4] The provided annotation was cleaned by removing participants with less than 33% of the ratings and with low correlation with others. On average, each MWE received 10.4 valid scores (minimum of 10).

Aiming to compare these 2 features with others reported in the literature, we also employed 337 features. As some of them are close in terms of definition and represented phenomenon, we grouped them into 14 families of features.

Length-based features (**LEN**) count the word length (i.e., number of letters in a token and its stem, and the number of syllables) and the number of words per sentence. In total, 4 length-based features.

Graded resource features (**GRD**) contain normalized frequencies of word lemmas divided by level from EFLLex (Dürlich and François, 2018). We use a total of 6 features based on graded resources.

Frequency features (**FRQ**) consider the frequency of words in a reference corpus. In this work, we consider the frequency of all words in a text, only content words (i.e., nouns, proper nouns, verbs, adjectives and adverbs in the text), only functional words, only common nouns, only verbs and only adjective. As the reference corpus, we explored the total normalized frequency (ignoring levels) in EFLLex (Dürlich and François, 2018) and contextual diversity on SUBTLEX (Brysbaert and New, 2009). In sum, 18 frequency-based features.

Features based on orthographic neighbor (**NGH**) measure orthographic or phonetic similarity between words. In this work, we use the mean orthographic and phonologic Levenstein distances (Bartlett et al., 2009) and the absolute and average number of neighbors and their frequency (Brysbaert and New, 2009). Also, the occurrence and cumulative frequency of neighbors with higher frequency than the words in the text are used. In total, 8 features.

Lexical norms (**NRM**) features resort to the MRC database (Coltheart, 1981) to annotate age of acquisition, concreteness, familiarity and imageability of each word. In addition, we also identify the percentage of out-of-vocabulary in each of the four features.

Lexical sophistication (**SOP**) features identify the number of sophisticated tokens and types considering all words, content words, and verbs considering the surface form in Dale and Chall (1948). In sum, 6 features.

Moreover, we use syntactic annotation automatically extracted from the Stanza parser (Qi et al., 2020).[5] Part-of-speech tags (**POS**) are counted using. 17 tags described in the Universal POS tags are considered.

Morphology features (**MOR**) target the morphological components of the words. As they operate in a lower level of the POS, we also use the Stanza parser for annotating the 56 features. Dependency relations (**DEP**) employ the 37 functions proposed by Universal Dependencies[6]. In addition, verb tense (**TNS**) features put together POS and morphology relations to identify the verb tenses as they are commonly taught. We use 19 verb tenses: simple tenses, perfect, continuous, emphatic and conditional tenses, and also the imperative, the tenses. All based on Stanza parser and identified through handcrafted rules. We also explore constituency parser (Kitaev et al., 2019) for extracting phrase (**PRH**) usage, differentiating 25 phrase types. In addition, we also count the number of phrases.

Language development (**DEV**) features include the Yngve index constituency parser (Yngve, 1960), number of words before and after the main verb, and the average phrase and sentence depth in the text. In total, 5 features related to language development.

Lexical diversity features (**DVR**) explore variations of type-token-ratio (TTR) that have been widely used for measuring language proficiency. In this

[4] Unfamiliar MWE were not annotated.

[5] We do not assess parser instabilities stability caused by learner errors, but Berzak et al. (2016) addressed the subject.

[6] https://universaldependencies.org/

work explored the Moving Average TTR (MATTR; (Covington and McFall, 2010)) with a window size of 100 words; Corrected TTR (CTTR; (Carroll, 1964)); Root TTR (RTTR; (Guiraud, 1959)); Bilogarithmic TTR (LogTTR; (Herdan, 1960; Herdan, 1966)); SquaredTTR (Chaudron and Parker, 1990); and UberIndex (Arnaud and Béjoint, 1992). For those, we distinguish between the ratios of lemmas and surface forms as well as all words, content words (i.e. nouns, proper nouns, verbs, adjectives, and adverbs in the text), adjective, adverb, adjective and adverb, nouns and pronouns, and verb. In addition, we specialized the verb features normalizing by the content words and verbs. In sum, we use 112 DVR features.

Coherence features (**COH**) use language models to compare the input text with the language's reference usage. We used ukWaC (Baroni et al., 2009), a 2 billion word corpus that covers a great range of themes, to train our models. Our first model, LSA, has 250 dimensions with stopwords and punctuations being removed and the 100,000 most frequent tokens/lemmas were kept. For the second model, PPMI, the dimension and window size were set to 500 and 2 without removing stopwords (Bullinaria and Levy, 2007). For these models, we calculate the cosine similarity of all pairs of adjacent sentences and the cosine similarity of each sentence with all the other sentences are computed (for the PPMI case, all the word vectors of a sentence are averaged). In total that makes 8 features. We also estimate the probability and perplexity of each sentence by training two 4-gram models on ukWaC (uncased tokens and lemmas) in the third model. This was created using KenLM (Heafield et al., 2013), a language modeling toolkit based on modified Kneser-Ney smoothing (Kneser and Ney, 1995). The n-gram model added 4 features. Finally, the fourth model, 3 features, is a simple n-gram frequency varying n between 2 and 4 on uncased and lemmatized ukWaC using SRILM (Stolcke, 2002), a language modeling toolkit.

4. Results

Following our goal, we analyze the MWE usage on the annotated corpus. We start by describing the MWE usage and concreteness in the corpus. This analysis allowed to draw a general profile of MWE in learners' essays (Section 4.1). Then, we focus on the applicability of MWE features for automatic essay scoring by investigating their correlation with the CEFR level (Section 4.2) and their applicability as features for a machine learning model (Section 4.3). We also compared the proposed features with the classic ones in the last two studies to evaluate their capacity to discriminate the levels.

4.1. Profiling MWE usage

The analysis of MWE usage by learners showed that 5.78% of the essays do not contain MWEs. In A1, A2 and B1 levels, there is an increase in the MWE usage, but they are similarly used at B2 and C.

The use of MWEs along the levels and the 128 prompts were also analysed. Prompts are specific per level, varying between 23 and 31 prompts. Only in the higher levels there are few occurrences of the same prompt shared in different levels (3% of the prompts). The quantity of essays is not the same for each prompt. A normalization considering the average of the prompts that had fewer documents was made to get a reliable result. Considering 2 standard deviations to the prompt to be an outlier, we observe two outlier prompts at A1, none at A2, one at B1 and B2, and three at C. For all levels, it corresponds to less than 10%.

The MWE' concreteness have a correlation of -0.11 with their usage per level. We observe that beginners are more familiar with more concrete MWEs and get used to more abstracted expressions as they go through the levels (concreteness average scores for A1-C are 3.1603, 3.0151, 2.7119, 2.5263 and 2.6087, respectively). Moreover, C level contains MWE present in the list but without annotated scores. It suggests that these MWEs are truly specific and indicative of a learner's high proficiency.

The skewness and kurtosis of the concreteness were also analysed per level (kurtosis is summarized in Figure 1). The concreteness distribution for A1 is flattened. As the level increases, the distribution approaches a normal distribution. The skewness, on the other hand, has low values for A1 and they increase across the levels, going from 0.0514 (A1) until 0.3966 (C)[7]. This suggests that the data has a positive deviation as the level increase, it means that the weight happens in the direction of the low scores of concreteness.

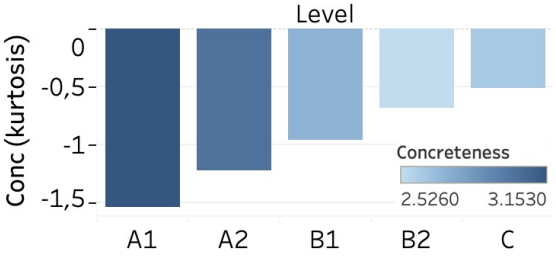

Figure 1: Concreteness kurtosis per level

4.2. Correlation

To study the relationship between the MWE and CEFR levels, we compared the Spearman correlation between MWE features and the level as well as all features described in Section 3. Those are summarized in Table 3 which shows the score most correlated with the level for each family of features presenting their rank and correlation considering the entire corpus and distinguishing by nationality. The table also shows the average rank and correlation of the features by family,

[7] A2 = 0.1572, B1 = 0.2607, B2 = 0.3519

considering the entire corpus, and by nationality; all correlations with p-value$<$0.05.

The top 40 features are predominantly related to lexical diversity. This result goes in the same direction as Jung et al. (2019). We also observed that the top 6 features have different ranks when nationality is considered. However, they are always in the top 6. Moreover, the top 1-3 are based on ratios considering all tokens, while and the top 4-6 are based on ratios of content words only. We also observed a band of features that alternate values between the top 7 and 16. Contrary to the pattern observed in the top 16 features, the features between 17 and 25 have almost constant rank across the nationalities. Below rank 25, we observed a considerable fluctuation in rank. This fluctuation can be seen in the standard deviation of the rank columns in Table 3.

We also analyzed the relation between the feature with the highest correlation and the average correlation for each family. As shown in Table 3, a higher correlated feature does not indicate that most of the features in their family are also highly correlated. For example, the SquaredTTR based on all tokens presented a correlation of 0.81 with the CEFR level, but in average the DVR features presented 0.42 as correlation. This indicates that only a few features are broadly meaningful for level identification. However, it does not mean that the other features may be ignored.

Targeting on MWE, their average concreteness is more correlated with the level than their usage (0.36 v. 0.21). In other words, the use of less concrete MWE is a better indication of a CEFR level than a higher number of MWE, although both features showed weak relationships with the level. Furthermore, we explored 18 statistics descriptors[8] to better describe the MWE usage and concreteness. The correlations between those and the CEFR levels are shown in Table 2 (absolute values lower than 0.26 and those with p-value$>$0.05 are not shown in Table 2). We also highlight that some separation statistical measures, such as minimum (Min) and first quartile (Q1), are better descriptors than the average one for MWEs concreteness. Moreover, we identified that the correlation between the levels and the number of words corrected by the MWE occurrence is 0.82.

MWE	Kurt	Q3	Median	Q1	Min
CONC	0.40	-0.29	-0.35	-0.37	-0.50
CNT	-	-0.02	-	-	-

Table 2: Correlation of MWE features aggregators

[8]Average, sum, minimum, maximum, length and mode as measures of range and tendency. Median, variance, standard deviation, relative standard deviation, dolch, first and third quartile, eighth and ninth percentiles and interquartile range as measures of dispersion and separation. Skewness and kurtosis for description of the curve.

4.3. Classification

For exploring the relationship between the scores, we resort to feature-based machine learning. We explored the relation inter-families by combining the different scores that compose each of the 14 families (see Section 3) as features for predicting the CEFR level of an essay. Since some families are strongly related, we also explore the combination of them as features. In other words, we combined *parser* (MOR, POS, DEP, PRH and TNS), and lexical norms-based (NRM and MWE_{conc}) features (NRM_{all}). In addition, for the sake of comparison, we considered the occurrence of MWE and their concreteness as individual features. Finally, we combined all features (*all*) to identify the full prediction capacity of a model trained using all features described in this work. For comparing the impact of the MWE features in this set of all features, we removed the MWE features from the training. Aiming to avoid bias of a specific model, we explored two machine learning models, one based on classification (Random Forest; RF) and the other on regression (Simple Logistic; SL). All these models were trained using stratified cross-validation 10 folds. The average[9] and standard deviation results of these models using the different feature sets are shown in Table 4.

For the SL, the results by feature family indicate that the best results are obtained when using the DVR features, in line with the results of the correlation study (Section 4.2). However, the MOR features seem to be more informative when using the FR. This difference is probably related to the search strategy employed by the RF, which can better divide the search space.

The combination of different families had a remarkable positive effect on the parser-based features (increasing the F1 from 77% to 83% in the RF and the RMSE from 1.065 to 0.857 in the LR). The combination of lexical norms with the MWE concreteness showed a small improvement (p-value$<$0.05). Despite all these improvements by combining new features, the use of only DVR features achieved the best result in the regression. This again points to the need for an intricate search space strategy. Lastly, we did not observe a significant difference between the use of all features and all except the MWE-related features.

5. Conclusions

In this work, we study MWE features to predict essay scores. Concreteness of the MWEs found per level leads us to believe that MWE concreteness has an impact to predict essay scores. However, the correlation and machine learning results do not confirm it. MWE has been studied in other languages, such as French François and Watrin (2011) who observed similar results. In future work, the approach proposed by Wilkens et al. (2022) can be included in the feature's

[9]The standard deviation RMSE is below 0.02 and for the other scores below 0.01.

Family	best score	general				by nationality	
		rank		corr		rank	
		best	family	best	family	best	family
DVR	STTR (all surface tks)	1	138.9 (120.8)	0.81	0.42 (0.25)	2.0 (0.8)	137.9 (117.4)
DEV	depth	25	95.1 (77.5)	0.70	0.48 (0.17)	25.0 (0.0)	97.3 (78.7)
DEP	mark	26	194.5 (126.5)	0.62	0.29 (0.20)	29.7 (4.7)	198.7 (125.1)
POS	punct	35	204.3 (90.0)	0.59	0.27 (0.14)	35.1 (7.0)	214.6 (92.9)
LEN	word per sent.	36	66.8 (28.3)	0.58	0.50 (0.07)	39.5 (12.3)	66.9 (25.6)
NRM	AOA	42	105.6 (66.9)	0.58	0.43 (0.13)	41.2 (5.9)	107.7 (68.1)
FRQ	content words subtlex	44	198.0 (107.8)	0.57	0.28 (0.18)	42.1 (5.1)	199 (107.5)
PRH	SBAR	52	254.1 (103.8)	0.54	0.20 (0.16)	52.5 (6.9)	252.0 (100.8)
TNS	use past	63	266.0 (85.0)	0.51	0.18 (0.12)	64.1 (3.6)	267.7 (84.4)
MOR	finite verb	69	204.4 (94.4)	0.47	0.26 (0.14)	77.5 (14.8)	215.3 (97.1)
NGH	phonologic dist	71	254.3 (118.1)	0.47	0.20 (0.17)	71.5 (8.7)	247.8 (113.2)
SOP	verbs	75	163.8 (88.8)	0.46	0.32 (0.14)	78.7 (12.2)	166.6 (93.5)
MWE	MWE_{conc}	142	-	0.36	-	136.7 (18.9)	-
COH	PPMI (lemma)	183	291.5 (68.3)	0.29	0.14 (0.09)	188.9 (25.4)	288.7 (59.4)
GRD	C1	213	235.6 (28.5)	0.24	0.21 (0.04)	212.2 (14.4)	237.6 (28.8)
MWE	MWE_{cnt}	233	-	0.21	-	239.9 (18.4)	-

Table 3: Correlation of different features and families of features considering the entire corpus and the learners' nationalities

Feature set	RandForest		SLogistic	
	ACC	F1	MAE	RMSE
LEN	0.553	0.553	0.897	1.364
FRQ	0.682	0.682	0.739	1.200
GRD	0.490	0.490	1.014	1.487
NGH	0.561	0.560	1.053	1.520
NRM	0.624	0.624	0.744	1.158
SOP	0.498	0.498	0.869	1.294
DVR	0.745	0.745	0.410	0.789
DEP	0.736	0.736	0.630	1.065
PRH	0.645	0.645	0.941	1.406
DEV	0.726	0.726	0.694	1.075
POS	0.745	0.744	0.772	1.235
MOR	0.775	0.775	0.682	1.126
TNS	0.565	0.559	0.731	1.161
COH	0.519	0.519	1.170	1.628
MWE	0.428	0.425	1.455	1.916
MWE_{cnt}	0.454	0.447	1.660	2.121
MWE_{conc}	0.418	0.413	1.499	1.946
Parser	0.835	0.835	0.425	0.857
NRM_{all}	0.640	0.640	0.734	1.153
All	0.843	0.843	0.535	0.697
All-MWE	0.844	0.844	0.534	0.699

Table 4: Results of the machine learning models using different feature sets

creation since we observed different behavior per level that are identified by statistical descriptors other than average. Therefore, it might lead to a better understanding of the learner's usage of MWE and its applicability for essay scoring.

6. Acknowledgements

This research has been partially funded by a research convention with France Éducation International. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

7. Bibliographical References

Adamchik, V. A., Hyclak, T. J., Sedlak, P., and Taylor, L. W. (2019). Wage returns to english proficiency in poland. *Journal of Labor Research*, 40(3):276–295.

Alexopoulou, T., Michel, M., Murakami, A., and Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1):180–208.

Arnaud, P. J. and Béjoint, H. (1992). *Vocabulary and applied linguistics*. Springer.

Arnold, T., Ballier, N., Gaillat, T., and Lissón, P. (2018). Predicting cefrl levels in learner english on the basis of metrics and full texts. *arXiv preprint arXiv:1806.11099*.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, 62(1):67–82.

Balikas, G. (2018). Lexical bias in essay level prediction. *arXiv preprint arXiv:1809.08935*.

Ballier, N., Gaillat, T., Simpkin, A., Stearns, B., Bouyé, M., and Zarrouk, M. (2019). A supervised learning model for the automatic assessment of language levels based on learner errors. In *European Conference*

on Technology Enhanced Learning, pages 308–320. Springer.

Ballier, N., Canu, S., Petitjean, C., Gasso, G., Balhana, C., Alexopoulou, T., and Gaillat, T. (2020). Machine learning for learner english: A plea for creating learner data challenges. *International Journal of Learner Corpus Research*, 6(1):72–103.

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., and Katz, B. (2016). Universal dependencies for learner english. *arXiv preprint arXiv:1605.04278*.

Boyd, M. and Cao, X. (2009). Immigrant language proficiency, earnings, and language policies. *Canadian Studies in Population [ARCHIVES]*, 36(1-2):63–86.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., and Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *LREC*, volume 2, pages 1934–1940.

Carroll, J. B. (1964). Language and thought. *Reading Improvement*, 2(1):80.

Chaudron, C. and Parker, K. (1990). Discourse markedness and structural markedness: The acquisition of English noun phrases. *Studies in second language acquisition*, 12(1):43–64.

Chen, X. and Meurers, D. (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.

Chen, X., Alexopoulou, T., and Tsimpli, I. (2021). Automatic extraction of subordinate clauses and its application in second language acquisition research. *Behavior Research Methods*, 53(2):803–817.

Covington, M. A. and McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Dahlmann, I. and Adolphs, S. (2007). Pauses as an indicator of psycholinguistically valid multi-word expressions (mwes)? In *Proceedings of the workshop on a broader perspective on multiword expressions*, pages 49–56.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Du, X., Afzaal, M., and Al Fadda, H. (2022). Collocation use in efl learners' writing across multiple language proficiencies: A corpus-driven study. *Frontiers in Psychology*, 13:752134–752134.

François, T. and Watrin, P. (2011). On the contribution of mwe-based features to a readability formula for french as a foreign language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 441–447.

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., and Zarrouk, M. (2021). Predicting cefr levels in learners of english: the use of microsystem criterial features in a machine learning approach. *ReCALL*, pages 1–17.

Garner, J. R. (2016). A phrase-frame approach to investigating phraseology in learner writing across proficiency levels. *International Journal of Learner Corpus Research*, 2(1):31–67.

Glucksberg, S. (1989). Metaphors in conversation: How are they understood? why are they used? *Metaphor and Symbol*, 4(3):125–143.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique*, volume 2. D. Reidel.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696.

Herdan, G. (1960). *Type-token mathematics*, volume 4. Mouton.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. Springer Berlin.

Jackendoff, R. (1997). Twistin'the night away. *Language*, pages 534–559.

Jung, Y., Crossley, S., and McNamara, D. (2019). Predicting second language writing proficiency in learner texts using computational tools. *The Journal of Asia TEFL*, 16(1):37–52.

Kitaev, N., Cao, S., and Klein, D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, July. Association for Computational Linguistics.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Ph.D. thesis, Georgia State University, Atlanta, Georgia.

Lan, G., Zhang, Q., Lucas, K., Sun, Y., and Gao, J. (2022). A corpus-based investigation on noun phrase complexity in l1 and l2 english writing. *English for Specific Purposes*, 67:4–17.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Michel, M., Murakami, A., Alexopoulou, T., and

Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: Evidence from a large learner corpus of a1 to c2 writings. *Instructed Second Language Acquisition*, 3(2):124–152.

Pendakur, K. and Pendakur, R. (2007). Colour my world: Have earnings gaps for canadianborn ethnic minorities changed over time?

Römer, U. and Berger, C. M. (2019). Observing the emergence of constructional knowledge: Verb patterns in german and spanish learners of english at different proficiency levels. *Studies in Second Language Acquisition*, 41(5):1089–1110.

Römer, U. (2019). A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics*, 24(3):268–290.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.

Srichanyachon, N. (2012). Teacher written feedback for l2 learners' writing development. *Humanities, Arts and Social Sciences Studies (Former Name Silpakorn University Journal of Social Sciences, Humanities, and Arts)*, pages 7–17.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Stubbs, M. (2007). An example of frequent english phraseology: distributions, structures and functions. In *Corpus linguistics 25 years on*, pages 87–105. Brill.

Tack, A., François, T., Roekhaut, S., and Fairon, C. (2017). Human and automated cefr-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179.

Tono, Y. (2013). Automatic extraction of l2 criterial lexico-grammatical features across pseudo-longitudinal learner corpora: using edit distance and variability-based neighbour clustering. *L2 vocabulary acquisition, knowledge and use*, pages 149–176.

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1):79–105.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, 18(1):85–99.

Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K., and François, T. (2022). Fabra: French aggregator-based readability assessment toolkit. In *Language Resources and Evaluation Conference (LREC)*.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language teaching*, 32(4):213–231.

Wray, A. (2002). *Formulaic language and the lexicon.* ERIC.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

Zilio, L., Wilkens, R., and Fairon, C. (2018). Investigating productive and receptive knowledge: A profile for second language learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3467–3478.

8. Language Resource References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Bartlett, S., Kondrak, G., and Cherry, C. (2009). On the syllabification of phonemes. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 308–316.

Brysbaert, M. and New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Dürlich, L. and François, T. (2018). Efllex: A graded lexical resource for learners of english as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Geertzen, J., Alexopoulou, T., and Korhonen, A. (2013). Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). *Selected Proceedings of the 31st Second Language Research Forum (SLRF), Cascadilla Press, MA*.

Muraki, E., Abdalla, S., Brysbaert, M., and Pexman, P. (2022). Concreteness ratings for 62 thousand english multiword expressions. 03.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.