A Phraseology Approach in Developmental Education Placement

Preprint · September 2022

DOI: 10.26615/978-954-452-080-9_011

2 authors:

Miguel Da Corte
Universidade do Algarve
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Jorge Baptista
Universidade do Algarve
186 PUBLICATIONS **656** CITATIONS
SEE PROFILE

A Phraseology Approach in Developmental Education Placement*

Miguel Da Corte[1][0000-0001-8782-8377] and Jorge Baptista[1,2][0000-0003-4603-4364]

[1] University of Algarve, Faculty of Human and Social Sciences
mlloveradacorte@gmail.com
[2] INESC-ID Lisboa, Human Language Technology Lab
jbaptis@ualg.pt

Abstract. This study focuses on an automatic classification task aiming at placing community college students into the appropriate level (Level 1 and 2) of Developmental Education (DevEd) courses, according to their English L1 proficiency. DevEd courses are designed to remediate and support students' communication skills in reading and writing before they can fully participate in college-level or college-bearing courses. This paper uses machine-learning methods to investigate the impact of considering multiword expressions (MWE) as entire tokens on the automatic classification task. Since many MWE are often non-compositional in meaning and constitute a large percentage of the textual units of many texts, they are likely to have a relevant role in the data representation of texts and, hence, improve subsequent classification task. Information is scarce regarding the tokenization of MWE and how this affects automatic placement. To this end, a random, balanced corpus of 186 sample texts (93 from each level) was used. Experiments compared the performance of a set of classifiers on the plain text corpus and on a version of the same corpus annotated for MWE. Results showed that using MWE as lexical features improved the classification accuracy by 8.1% above the baseline.

Keywords: Developmental education · machine-learning classification · multiword expressions · text mining.

1 Introduction and Objectives

The study of MWE continues to provide an opportunity for the enhancement of linguistic analysis. Several authors consider the lexical variety and repetition of MWE to be key in the process of language acquisition and mastery of language proficiency and fluency [9, 12, 15]. Understanding the lexical and syntactical patterns of community college students paves the way to understanding

* Research for this paper has been supported by University of Algarve, Language Sciences doctoral program, and by national funds through Fundação para a Ciência e a Tecnologia (Proj.Ref. UIDB/50021/2021).

issues related to language that impede effective communication both verbally and in writing. For community college students, the target population of this study, the use of these "prepackaged," "bundled" expressions facilitates communication and provides a sense of fluency, even if their writing skills require improvement to enter higher education. This is one of the main purposes of Development Education (DevEd) [4, 5, 16]. On the other hand, placement in DevEd courses often resorts to machine-learning based systems, such as accuplacer[3], compass[4], act[5], which use textual linguistic features to assist the placement strategies followed by higher education institutions. The Coh-Metrix [10] [6] has been used to provide descriptive textual features that help capture Text Complexity and Readability, e.g., word count, paragraph length, word length; Text Easability; Referential Cohesion; Lexical Diversity; Connectives, Syntactic Complexity; Syntactic Pattern Density; Word Information; and Readability metrics (Flesch-Kincaid). These categories can be used as items for a linguistic and discourse representation to enrich corpora with linguistic data as evidenced in a previous study [1]. However, there is scarce information on how these systems deal with MWE (if at all), or at least what types of MWE are considered. The prevalent use of MWE in students' writing samples from entrance exams motivates the study of the impact of phraseology on machine-learning classification tasks pertaining to DevEd course placement, thus, succinctly defining the focus of this paper.

2 Related Work

Vocabulary in discourse and the connection of words within the larger discourse has been examined by [14], who particularly focused on how meaning is carried through *formulaic sequences*. Particular emphasis is placed on the formation of multiword units as a whole and common categories of these units such as compound words, phrasal verbs, fixed phrases, and lexical phrases are exemplified. The author asserts [14, p.101] that "language production stems from the precept that native speakers tend to use language that is formulaic in nature."

Understanding common categories of MWE requires a deep examination of their properties and the impact of these properties on NLP applications. According to [8], collocation and discontiguity phenomena are heavily exploited features, among the many properties MWE present, that have aided parsing tasks and, thus, enhanced syntactic analysis [3]. Studying the impact of discriminated tokens versus *single syntactic constituents*, as coined by [8], on the structural and functional aspects of developing writers' skills can aid in a more accurate placement of students in DevEd [8].

[3] https://www.accuplacer.org/ (Last access: January 26, 2023; all URL in this paper were check on this date.)
[4] https://www.compassprep.com/practice-tests/
[5] https://www.act.org
[6] http://141.225.61.35/CohMetrix2017/

The incidence of multiword expressions was studied by [11] in 746 argumentative essays with 121,638 tokens from Korean-university students with the goal of exposing reocurrent structural sentence patterns and their frequency. These patterns were compared to those exhibited by American-university students in two large corpora (LOCNESS; MICUSP), where the incidence of phrased-based expressions exhibited by L2 compared to the incidence of noun phrases by L1 emerged as a theme. Additionally, L2 student showed a wider variety of MWE in their writing, suggesting that the use of these lexically-bound expressions facilitates communication among developing writers by adding a level of functionality to the writing process. Even though this study focused on L1 and L2 students, it supports the investigation of "MWE across proficiency levels or novice and professional academic writing [...] in an academic context." [11, p.11]

The ongoing variability of MWE, particularly verbal ones, and the challenges it poses for automated machine learning identification tasks are recognized by [13]. The authors focused on strategies to improve the identification of verbal MWE (VMWE) and used the PARSEME[7] corpora as a starting point. They completed three tasks comprised of a training and development phase, a prediction phase, and an evaluation phase all aided by a simple [candidate VMWE potential] extraction of filtering (*Seen*2020) techniques for precision [13]. Promising results were evidenced in boosting the identification of global MWE by using a combination of morphosyntactic filters. Suggestions for further work emphasize the representation of "VMWE as multisets of (lemma-POS) pairs rather than lemma and POS multisets separately." [13, p.3342]

A framework to further investigate the lexical and syntactical features of MWE is provided by [8, 11, 13, 14], among others. However, it is unclear whether automatic placement classification systems, i.e., accuplacer, compass, and act, take MWE into consideration or at least what types of MWE are considered. To the best of our knowledge, these systems provide a holistic score of writing samples based on the purpose and focus of the essay; organization and structure; development and support of ideas; sentence variety and style; accurate use of Standard Written English; and how one communicates and connects ideas while addressing a topic [2]. However, [16] emphasizes the need to improve college readiness by more accurately assessing students' performance in writing tasks. We aim at detecting MWE and MWE types to aide in the assessment of lexical features towards language proficiency.

3 Methods

To assess the impact of MWE in our classification task, a small corpus of 186 essays was collected from community college students, in the State of Oklahoma, U.S.A., placed in DevEd after completing their writing placement exam during years 2020-2021. The sample texts, consisting of a short multiparagraph essay of 300-600 words, prompted by a short excerpt with guiding questions, were

7 https://typo.uni-konstanz.de/parseme/

retrieved from the accuplacer platform of the higher education institution where this study took place. The corpus was then balanced for placement level (93 essays from each level: Level 1 and Level 2), and for sample size (keeping only up to 100 tokens per text). All the essays addressed the same topics.

Two experiments were devised, where the Classification Accuracy (CA) performance of a set of classifiers of the raw corpus and on a version of the same corpus annotated for all types of MWE was compared. We first devised a classification criteria for identifying MWE. Then, we independently annotated the corpus and compared the annotations in order to validate the selected MWE. The Orange[8] data-mining toolkit [6] was chosen for analysis and modeling since it provides, in a practical way, several NLP tools (mainly for preprocessing and data representation) within the same machine-learning (ML) toolkit and a large set of ML algorithms and data visualization tools.

3.1 Corpus Annotation and Classification Guidelines

In order for the Orange data-mining toolkit to tokenize a MWE as a single token, its elements had to be joined in the input text (an underscore was used to this end). This task was done manually. One of the reasons for this has to do with the large number of typos and spelling errors found in the essays, which we could only ignore with a manual inspection of the text. For the manual identification of MWE, the set of linguistic criteria adopted derived from the literature [3]. The MWE categories of [7] and the caveats by [8] were particularly insightful. Based on these criteria, the annotation of MWE in the corpus included multiple categories, based on traditional classification guidelines and lists available in dictionaries, e.g., **compound nouns**: *golden rule*, *refresher course*; **compound adjectives**: (stay) *tall and straight*, (be) *on* [one's] *feet* ; *strong minded* ; **compound conjunctions and prepositions**: *as long as*, *in spite of* ; **compound adverbs**: *by the way*, *back in the old day*; **verbal idioms**: *see the bigger picture*, *knock at your door* ; *weigh the pros and cons*; **phrasal verbs**: *fit in*, *mess up*; **support verb constructions**: *have no clue*, *make mistakes*; among others. In the case of **nouns**, these also included named entities, both person names (*Colin Powell*), locations (*Guatemala City*), and organizations (*United Nations*).

For the annotation proper, the MWE were joined following these guidelines: Sequential strings of words were just joined by an underscore (*phone_call*). Productive inserts, as in phrasal verbs, were marked by a double underscore and the inserted word permuted (*bring you down* → *bring__down you*). Long-distance elements of the same MWE were joined in the first element, and the slot left void marked by a hashtag '#': The more *success* <*sic*> *you are in your job* the more *you will make* → *The_more_the_more success you are in your job # you will make*. Prepositions introducing complements of predicative elements (adjectives, nouns, and verbs, including phrasal verbs) are not taken into account, e.g. *to get_out of something* where *of something* is just a complement of the phrasal verb *get_out* (joined). At the end, 1,260 MWE were annotated.

[8] https://orangedatamining.com/

3.2 Data Processing and Modeling

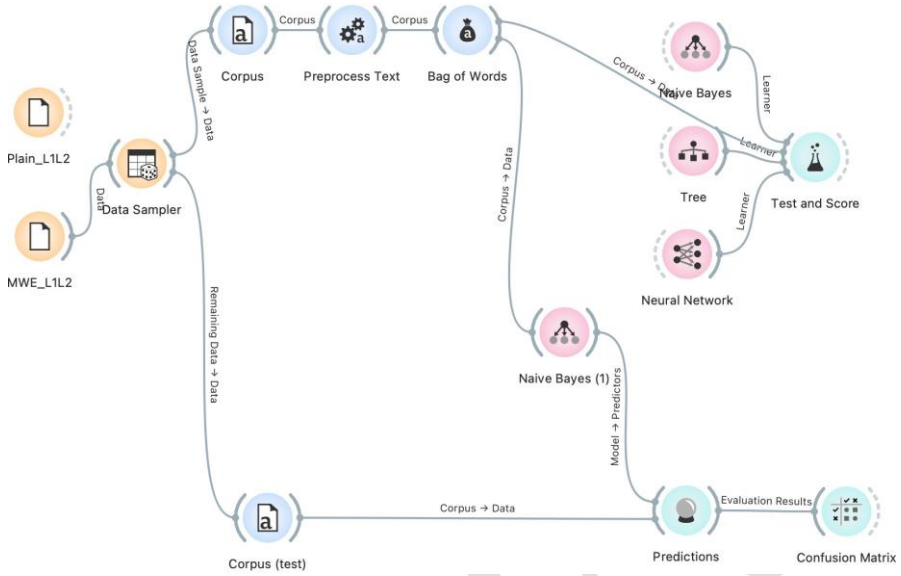

Fig. 1. Orange Text Mining Environment

The workflow adopted for this study is shown in Fig. 1. The datasampler widget was used to partition the data for a 3-fold cross-validation, leaving 2/3 for training and 1/3 for testing purposes. In the Preprocessing stage, basic tokenization options were selected, capturing both words and punctuation as tokens. Experiments were then carried out introducing different part-of-speech (PoS) taggers. The data representation chosen was Bag-of-Words (using the count of term frequency). For the training step, and to assess the models, the Test&Score widget was used. The models used were Naive Bayes, Decision Tree (forward pruning) and Neural Network (multi-layer perceptron with back propagation). The data representation and the learning models were chosen because they had already proven to perform well with this type of data in previous experiments. For the Predictions (testing) step, the best performing model in each configuration setting was assessed.

3.3 Experiments

Table 1 shows the experimental settings and the breakdown of the results. In these experiments, we compared the non-annotated corpus (PlainText) with the corpus where MWE had been joined (Text-MWE). Since the corpus is small, the training partition corresponds to 2/3 of the corpus and the remaining is

Table 1. Experimental settings and results.

Experiment	Description	NB	NN	TR	best	model
Baseline	PlainText;Tok:W&P. noPoS.BoW	**0.685**	0.629	0.669	NB	0.742
Experiment 01	PlainText; idem+PoS:TB-ME	0.669	**0.718**	0.669	NN	0.677
Experiment 02	PlainText; idem+PoS:AP	0.613	0.661	**0.685**	TR	0.726
Experiment 03	Text-MWE;Tok:W&P. noPoS.BoW	0.707	**0.740**	0.553	NN	**0.823**
Experiment 04	idem+PoS:TB-ME	0.573	0.661	**0.667**	TR	0.742
Experiment 05	idem+PoS:AP	0.540	0.637	**0.653**	TR	0.758

left for testing. The data representation mode was Bag-of-Words with a simple frequency count. Results are provided using the Classification Accuracy metrics (CA). The names of learning models selected for the study are shortened to NB (Naive Bayes), Neural Networks (NN) and Tree (TR). The values reported for each individual model concern the Test&Scoring (training) step. In the rightmost columns, the CA values in the testing step are reported to the best performing model in the previous training step. In the description of the experiments, the two PoS-taggers have been shortened: Treebank - Maximum Entropy (TB-ME) and Average Perceptron (AP). Otherwise, 'noPoS' indicates no PoS-tagging was applied.

We first defined a baseline, using a basic configuration for the preprocessing step. The Word&Punctuation tokenization option was chosen, because it keeps punctuation as is, which might be a factor of placement classification. This is due to the fact that the use of punctuation is one of the writing skills not yet entirely mastered by the students and which is a topic that is addressed by the DevEd courses. No difference was found when the models were trained with different combinations of other tokenization configurations, namely just using white-space or just the words. No PoS-tagger was used in this baseline. For Experiments 1-2, the same baseline configuration was used but now adding each PoS-tagger in turn. For Experiment 3, the initial baseline configuration was enhanced by using the same corpus but with its MWE joined as single tokens. No PoS-tagger was used here. In Experiments 4-5, the same enhanced configuration was used, with joined MWE, but now adding each PoS-tagger.

4 Results and Discussion

Overall, the results from the experiments showed that adding MWE information to single-word tokenized texts improves the classification up to a clear 8.1% above the baseline (Experiment 03). The preprocessing tokenization options available with Orange did not seem to affect results too much. On the other hand, PoS-tagging underperforms the baseline, especially the AP (Experiment 02). With the MWE annotated text, however, AP performs better (Experiment 05) than TB-ME (Experiment 04). When one compares the results from the best performing model in the training step with its results in the testing step, the testing step results usually outperform the training, except in Experiment 01, where testing

results were 4.1% less than training. Otherwise, differences vary from 4.1% (Experiment 02) to 10.5% (Experiment 05). While using the plain text corpus in the training step, and though results are not very different from the baseline, no model proved to outperform the other two, as each in turn occurs as the best model. Using the text with MWE, however, showed that Neural Networks was the best model (Experiment 03), but that Tree outperformed the other two models when PoS-tagging was added to the configuration.

Although the corpus is small, the size of MWE found in it corresponds to 6.75% of the words in the corpus, a non-neglegible percentage. Still, results seem to signal the importance of using information on MWE in this task. This is in line with the literature findings on similar scenarios [7]. A proper PoS-tagging of MWE, combined with an appropriate classification of MWE types [9] may contribute to improving results and providing more robust insights on DevEd students' writing patterns.

5 Conclusion and Next Steps

This study focused on an automatic classification task aimed at placing community college students into the appropriate level (Level 1 and 2) of Developmental Education (DevEd) courses, according to their English L1 proficiency. We investigated and presented experimental results on the impact that multiword expressions (MWE), considered as entire tokens, have on an automatic classification task using machine-learning methods. The classification was based on linguistic features derived from small corpus of texts written by native English speakers at the onset of their higher education journey. Issues of orthography, punctuation, lexical usage such as slang, grammatical issues such as agreement, semantic issues, and discursive aspects of the text have been previously explored. Now, the inclusion of MWE provided promising results as they constitute a large percentage of the textual units of many texts and are likely to have a relevant role in the representation of the information in texts. An extended corpus will be required to confirm the hypothesis presented in this preliminary study.

A quantitative evaluation to verify the usefulness of including MWE information was presented and discussed, including the benefits of using different PoS-taggers and the degree to which these tools affect the classification accuracy. Although, at this point in the research study, we are not concerned with the automatic identification of MWE in DevEd students' writing, we are committed to developing, first, a more precise MWE scheme for annotating writing samples to further gain insights into linguistic issues that prevent effective communication for this student population.

[9] The list of MWE is available at: https://www.researchgate.net/project/Linguistic-Aspects-of-Developmental-Education

References

1. Abba, K.A.: Community college students' writing: Lexical, syntactic, and cohesion differences in L1, L2, and Generation 1.5 students and examining knowledge of the writing process. Ph.D. thesis, Texas A&M University, Graduate and Professional Studies (2015)
2. Board, C.: ACCUPLACER program manual. The College Board New York (2018)
3. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Multiword Expression Processing: A Survey. Computational Linguistics **43**(4), 837–892 (12 2017). https://doi.org/10.1162/COLI_a_00302, https://doi.org/10.1162/COLI_a_00302
4. Cormier, M., Bickerstaff, S.: Research on developmental education instruction for adult literacy learners. The Wiley Handbook of Adult Literacy pp. 541–561 (2019)
5. Darkenwald-DeCola, J.A.: 'In College, I'm the One People Go To': Lessons from Successful Developmental Literacy Students About the Transition to College-Level Courses Across Disciplines. Ph.D. thesis, Rutgers The State University of New Jersey, School of Graduate Studies (2021)
6. Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research **14**, 2349–2353 (2013), http://jmlr.org/papers/v14/demsar13a.html
7. Kochmar, E., Gooding, S., Shardlow, M.: Detecting multiword expression type helps lexical complexity assessment. arXiv preprint arXiv:2005.05692 (2020)
8. Laporte, E.: Choosing features for classifying multiword expressions. In: Sailer, M., Markantonatou, S. (eds.) Multiword expressions: In-sights from a multi-lingual perspective, pp. 143–186. Language Science Press, Berlin (2018). https://doi.org/10.5281/zenodo.1182597
9. Martinez, R.: A framework for the inclusion of multi-word expressions in elt. ELT journal **67**(2), 184–198 (2013)
10. McNamara, D.S., Ozuru, Y., Graesser, A.C., Louwerse, M.: Validating CoH-Metrix. In: Proceedings of the 28th annual Conference of the Cognitive Science Society. pp. 573–578 (2006)
11. Nam, D., Park, K.: *I will write about* : Investigating multiword expressions in prospective students' argumentative writing. Plos one **15**(12), e0242843 (2020)
12. Omidian, T., Shahriari, H., Ghonsooly, B.: Evaluating the Pedagogic Value of Multi-word Expressions based on EFL Teachers' and Advanced Learners' Value Judgments. TESOL Journal **8**(2), 489–511 (2017)
13. Pasquer, C., Savary, A., Ramisch, C., Antoine, J.Y.: Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 3333–3345 (2020)
14. Schmitt, N., Schmitt, D.: Vocabulary in language teaching. Cambridge University Press (2020)
15. Thomson, H.: Building speaking fluency with multiword expressions. TESL Canada Journal **34**(3), 26–53 (2017)
16. Zachry Rutschow, E., Edgecombe, N., Bickerstaff, S.: A brief history of developmental education reform (Oct 2021), https://postsecondaryreadiness.org/research/history-developmental-education-reform/