Automatic Generation of Vocabulary Lists with Multiword Expressions

John S. Y. Lee and **Adilet Uvaliyev**
Department of Linguistics and Translation
City University of Hong Kong
jsylee@cityu.edu.hk, uvaliyevadilet@gmail.com

Abstract

The importance of multiword expressions (MWEs) for language learning is well established. While MWE research has been evaluated on various downstream tasks such as syntactic parsing and machine translation, its applications in computer-assisted language learning has been less explored. This paper investigates the selection of MWEs for graded vocabulary lists. Widely used by language teachers and students, these lists recommend a language acquisition sequence to optimize learning efficiency. We automatically generate these lists using difficulty-graded corpora and MWEs extracted based on semantic compositionality. We evaluate these lists on their ability to facilitate text comprehension for learners. Experimental results show that our proposed method generates higher-quality lists than baselines using collocation measures.

1 Introduction

Effective processing of multiword expressions (MWEs) is critical for many natural language processing (NLP) applications. In addition to intrinsic evaluation on the quality of extracted MWEs, researchers have conducted extrinsic evaluation to measure their impact on syntactic parsing, machine translation and other tasks (Constant et al., 2017). However, MWE extraction methods have not yet been evaluated in generating vocabulary lists, even though the importance of MWEs, which may require idiosyncratic interpretations, is well established for language learning (Bahns and Eldaw, 1993; Paquot and Granger, 2016).

Graded vocabulary lists recommend a language acquisition sequence for language learners and teachers, in order to optimize learning efficiency of the target language. These lists help prioritize words and expressions that are more likely to be encountered by learners, so that they can understand more texts within a shorter period of study. According to Sag et al (2002), the number of MWEs in a speaker's lexicon has been estimated to be of the same order of magnitude as the number of single words (Jackendoff, 1997). It is no surprise, then, that a significant number of MWEs are included in prominent vocabulary lists such as English Vocabulary Profile (EVP)[1] and the Pearson Global Scale of English (GSE).[2]

We investigate the selection of MWEs for graded vocabulary lists, assuming only a graded corpus for n-gram statistics and large general corpora for MWE extraction. To the best of our knowledge, this is the first evaluation on corpus-based methods for generating vocabulary lists with MWEs. The rest of the paper is organized as follows. After reviewing previous research (Section 2), we present our datasets (Section 3) and evaluation metrics (Section 4). We then describe our approach (Section 5) and report experimental results (Section 6).[3]

2 Previous work

The research most closely related with ours is EFLLex, a vocabulary list for learners of English as a foreign language (Durlich and François, 2018). It contains both single words and MWEs, including compounds and phrasal verbs. A rule-based method identifies the MWEs by considering the dependency labels and verb particles in parse trees of sentences in a large collection of English corpora, followed by manual checking. While CEFRLex resources have been found to be effective in predicting the CEFR levels of the EFLLex entries (Graën et al., 2020), MWEs have not been evaluated. Several other popular vocabulary lists, such as the New General Service List[4] and the Oxford lists[5], do not feature MWEs and therefore are not comparable

[1] https://www.englishprofile.org/wordlists/evp
[2] https://www.english.com/gse/teacher-toolkit/user/lo
[3] Data available at https://github.com/Adilet33709
[4] http://www.newgeneralservicelist.org/
[5] https://www.oxfordlearnersdictionaries.com

Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 81–86

List	# single words	# bigram MWEs	# trigram MWEs
EVP	6,749	993	839
GSE	18,391	2,821	1085
EFLLex	10,019	3,745	106

Table 1: Number of single words and MWEs in the graded vocabulary lists in our experiments

with ours.

In addition to EFLLex, we also evaluate a recently proposed MWE extraction method based on unsupervised measurement of semantic compositionality (Pickard, 2020). This method first identifies bigrams and trigrams as MWE candidates using the Poisson collocation measure (Quasthoff and Wolff, 2002). It then ranks these candidates according to the average cosine similarity between the word vector of the MWE candidate and the word vector of each of its constituent words. Experimental results show that the use of `word2vec` embeddings can achieve substantial correlation with human judgment.

3 Data

3.1 Graded corpora

Training set OneStopEnglish (Vajjala and Lučić, 2018) consists of 189 aligned texts, each written at three difficulty levels.[6] WeeBit (Vajjala and Meurers, 2012) consists of 3,125 documents from WeeklyReader and BBC-Bitesize, each labeled at one of five age groups, with 625 documents per group.

Test set The Cambridge corpus (Xia et al., 2016) contains articles for various Cambridge English Exams, labeled at five CEFR levels, A2, B1, B2, C1, and C2.

3.2 Human benchmarks

As human benchmarks, we used two large-scale graded vocabulary lists (Table 1):

English Vocabulary Profile (EVP) EVP is an online vocabulary resource with containing words, phrases, phrasal verbs and idioms (Capel, 2015), all labeled according to the Common European Framework of Reference (CEFR, 2001).[7]

[6]https://github.com/nishkalavallabhi/

[7]https://www.englishprofile.org/wordlists/evp

Pearson Global Scale of English (GSE) The GSE Teacher Toolkit is an online database containing English vocabulary items labeled on a proficiency scale from 10 to 90, and also aligned to the CEFR scale based on psychometric research (De Jong et al., 2016).

4 Evaluation methodology

Our evaluation focuses on MWEs up to trigrams only, since longer ones are not available in the dataset from Pickard (2020). Let $S = \{S_1, \ldots, S_k\}$ represent a graded vocabulary list with k grades, where S_i is the set of n-grams ($n \leq 3$) that are recommended for learners at Grade i. All n-grams are in lemma form.

The benchmark vocabulary lists adopt different numbers of grades and lemmas. We transform each list into a single ranked list (Section 4.1) to facilitate a fair evaluation (Section 4.2).

4.1 Transformation to ranked list

To transform a graded vocabulary list into one ranked list, we first rank the n-grams within each set S_i. Let L_i represent the ranked list derived from the set S_i by decreasing order of the n-gram frequency in the test set (Section 3.1). The final list L is then constructed by concatenating L_1, ..., L_n. In other words, *within each grade*, the more frequent n-grams are ranked higher towards the top.

4.2 Evaluation metrics

Suppose user u learns one lemma at a time, following the order prescribed by $L = [w_1, ..., w_l]$. Let u_i represent the user at time unit i, i.e., when s/he has learned all n-grams $w_1, ..., w_i$.

We define a text to be "understood" by user u_i if the percentage of known words exceeds 90%, using the minimum threshold suggested in second language acquisition literature (Laufer, 1989).[8] When a test passage contains a gold MWE (Section 4.3) that has not yet been learned, the MWE is considered unknown even if its constituent words has been learned separately. We evaluate the quality of L in two metrics:

Study Time We define "graduate from grade N" to mean the user understands at least $m\%$ of

[8]The calculation of the percentage of known words in a text excluded tokens tagged as NUM, PROPN, PUNCT, SPACE, SYM, or X by SpaCy (Honnibal and Johnson, 2015); and those consisting of digits and punctuation only. American and British spelling were both accepted.

Gold MWE	# MWEs
In EVP only	1,127
In GSE only	3,386
In both EVP and GSE	697
Added from MWE datasets	626

Table 2: Breakdown of the set of gold MWEs used in our experiments

Training set		Test set	
MWE	freq	MWE	freq
to do so	3,378	go to	150
web browser	1,356	the first	129
to date	1,118	the same	100
go to	1,004	part of	89
the first	891	a lot	87
at the moment	828	come to	71
such as	821	a few	67
the same	767	be so	66
look at	567	out of	61
for example	519	for example	60

Table 3: Top ten most frequent gold MWEs in the training set and test set

the texts included in grades $1, ..., N$ in the test corpus. This metric measures the time required, i.e., the minimum i required for u_i to graduate from level N. We report results for $m = 80$.

Text Comprehension The number of texts that can be understood by u_i, averaged over times $i = 1, ..., j$. We set j to the size of the shortest benchmark vocabulary list, i.e., EVP.

4.3 Gold MWEs

A set of ground-truth MWEs is necessary to apply the automatic metrics defined above. We compiled our gold MWE set from both language learning experts and past MWE research:

- The 5,096 MWEs found in the EVP and/or GSE lists (Section 3.2);
- MWEs that have been assigned an above-average score in the following benchmark MWE datasets: noun compounds (Reddy et al., 2011; Farahmand et al., 2015), adjective-noun compounds (Biemann and Giesbrecht, 2011), verb-particle pairs (McCarthy et al., 2003) and verb-object pairs (McCarthy et al., 2007). These yield an additional 626 MWEs to the gold set.

Method	Text Comp.
Frequency	57.52
Collocation	83.89
Collocation+Disp	87.42
EFLlex+Disp	59.01
Compositionality(50%)+Disp	76.96
Compositionality(75%)+Disp	**90.10**
Compositionality(Gold)+Disp	188.99
EVP	158.95
GSE	135.69
Ceiling	236.28

Table 4: Performance based on the "Text Comprehension" metric: average number of texts understood over the study period

Table 2 shows a breakdown of the 5,722 MWEs in the final set. Table 3 shows the most frequent MWEs in our datasets.

5 Approach

MWEs may include fixed and semi-fixed expressions, syntactically-flexible expressions and institutionalized phrases (Sag et al., 2002). As shown in Table 3, not all entries in vocabulary lists may conform to the standard MWE definition. Nonetheless, their inclusion in these lists by experts suggest that it is useful to treat them as a unit for the purpose of language learning.

Frequency All n-grams ($n \leq 3$) in the training corpora (Section 3.1) are considered as single-word and MWE candidates for the vocabulary list. They are lemmatized and ranked them according to frequency in the training corpora.

Collocation Same as the above, except that the MWE candidates are the top 500,000 n-grams in English Wikipedia based on the Poisson collocation measure (Quasthoff and Wolff, 2002).[9]

Compositionality(N%) Among the 500,000 MWE candidates above, this method retains as candidates only the top $N\%$ according to the semantic compositionality measure (Section 2).[10]

[9] https://github.com/Oddtwang/MWEs
[10] https://github.com/Oddtwang/MWEs

Method	A2	B1	B2	C1	C2
Frequency	7,164	9,054	16,712	58,139	58,139
Collocation	4,980	6,600	10,784	**24,610**	27,045
Collocation+Disp	**4,536**	6,007	**10,323**	25,184	26,326
EFLLex+Disp	/	/	/	/	/
Compositionality(50%)+Disp	8,679	8,679	17,508	/	/
Compositionality(75%)+Disp	4,984	**5,712**	11,253	25,983	**25,983**
Compositionality(Gold)+Disp	2,152	2,853	3,871	7,198	7,198
EVP	2,502	3,610	4,805	/	/
GSE	3,728	3,956	6,165	11,157	11,175
Ceiling	1,685	2,134	2,772	3,915	4,300

Table 5: Performance based on the "Study Time" metric: the number of time units needed for graduation from each level (Shorter time is better; "/" means the learner cannot graduate, as defined in Section 4.2)

EFLLex The MWE candidates are those found in EFLLex (Durlich and François, 2018).

+Disp The raw frequencies are weighted with Juilland's D (Gries, 2020), a dispersion coefficient that measures the degree to which occurrences of the n-gram are distributed evenly in the training set.

In addition, we implemented the following method to gauge the upper limit in performance:

Compositionality(Gold) The MWE candidates are the gold MWEs.

EVP / GSE The expert-crafted lists, transformed into a ranked list using the procedure in Section 4.1.

Ceiling The MWE candidates are the gold MWEs and all n-grams are ranked by frequency in the test set (Section 3.1).

6 Results

Text Comprehension. As shown in Table 4, the Collocation method (83.89) outperformed both the Frequency baseline (57.52) and EFLLex (59.01). The MWE candidates in the Collocation method covered 38% of the gold MWEs; retaining only the best-scoring three-quarters of the MWEs decreased the coverage to 32%, but was compensated by the higher quality among the selected MWEs. This can be seen in the performance of Compositionality(75%)+Disp, which was the best (90.10) of the automatic methods according to the Text Comprehension metric. This result suggests that the semantic compositionality measure was able to reduce the number of superfluous MWEs, and open up the learner's priority for other n-grams that appeared more often in the test set.

Study Time. As shown in Table 5, the learner graduated from the C2 level most quickly with the list generated from the top 75% of the MWEs, a result that is consistent with the Text Comprehension metric. At all lower levels except B1, however, the Compositionality(75%)+Disp method was outperformed by the Collocation method. The collocation statistics appeared to correlate better with the basic gold MWEs (e.g., "a few", "at least", "go out"), but less so with more advanced MWEs, likely because of the more divergent content. At all levels, the best automatic methods still lagged behind the expert-crafted lists, EVP and GSE, by large margins.

7 Conclusion

This paper has presented the first corpus-based evaluation of automatically generated vocabulary lists that incorporate MWEs. Using MWEs extracted by semantic compositionality (Pickard, 2020), we constructed a vocabulary list by ranking both single-word and MWE candidates by frequency and dispersion. Experimental results show that this method outperforms baselines using collocation measures, both in facilitating text comprehension and in shortening the study period. These algorithms can potentially enhance existing human-crafted lists, and compile new ones in resource-poor languages for which no vocabulary list is available.

Limitations

The experiments in this study were limited to MWEs up to three words long, given the dataset provided by Pickard (2020). Future work should

explore the effects of longer MWEs on the results. The evaluation can also be made more accurate by considering part-of-speech information. Finally, the gold MWE set could be expanded by harvesting more human-annotated MWEs.

Acknowledgements

This work was partly supported by the Language Fund from the Standing Committee on Language Education and Research (project EDB(LE)/P&R/EL/203/14) and by the General Research Fund (project 11207320).

References

Jens Bahns and Moira Eldaw. 1993. Should we teach efl students collocations? *System*, 21(1):101–114.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional Semantics and Compositionality 2011: Shared Task Description and Results. In *Proc. Workshop on Distributional Semantics and Compositionality (DiSCo)*.

Anette Capel. 2015. The English Vocabulary Profile. In *English profile in practice*, page 9–27, Cambridge, UK. Cambridge University Press.

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

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892.

John H. A. L. De Jong, Mike Mayer, and Catherine Hayes. 2016. Developing Global Scale of English Learning Objectives aligned to the Common European Framework. In *Global Scale of English Research Series*. Pearson.

Luise Durlich and Thomas François. 2018. EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *Proc. 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A Multiword Expression Data Set: Annotating Non-Compositionality and Conventionalization for English Noun Compounds. In *Proc. 11th Workshop on Multiword Expressions*, page 29–33.

Johannes Graën, David Alfter, and Gerold Schneider. 2020. Using Multilingual Resources to Evaluate CEFRLex for Learner Applications. In *Proc. 12th Conference on Language Resources and Evaluation (LREC)*, pages 346–355, Marseille, France.

Stefan Th. Gries. 2020. Analyzing Dispersion. In *A Practical Handbook of Corpus Linguistics*, pages 99–118.

Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Batia Laufer. 1989. What percentage of text is essential for comprehension? In *Special Language; from Humans Thinking to Thinking Machines*, pages 316–323, Clevedon. Multilingual Matters Ltd.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proc. Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, page 73–80.

Diana McCarthy, Sriram Venkatapathy, and Aravind Joshi. 2007. Detecting Compositionality of Verb-Object Combinations using Selectional Preferences. In *Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, page 369–379.

Magali Paquot and Sylviane Granger. 2016. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32:130–149.

Thomas Pickard. 2020. Comparing word2vec and GloVe for Automatic Measurement of MWE Compositionality. In *Proc. Joint Workshop on Multiword Expressions and Electronic Lexicons*, page 95–100.

Uwe Quasthoff and Christian Wolff. 2002. The Poisson Collocation Measure and its Applications. In *Proc. 2nd International Workshop on Computational Approaches to Collocations*, Wien. IEEE.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proc. CICLing*, page 1–15.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proc. 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proc. 7th Workshop on Innovative Use of NLP for Building Educational Applications*.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications*, page 12–22.