Automatic Generation of Vocabulary Lists with Multiword Expressions

John S. Y. Lee, Adilet Uvaliyev
Department of Linguistics and Translation
City University of Hong Kong

香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative For The World

Introduction

- A *vocabulary list* prioritizes learning of words and expressions that are more likely to be encountered in text
- E.g., English Vocabulary Profile (EVP) and the Pearson Global Scale of English (GSE) are widely used by language learners and teachers
- Multiword expressions (MWEs) are important for language learning and are often included in these lists
- We investigate the selection of MWEs for graded vocabulary lists, using semantic compositionality and difficulty-graded corpora
- The proposed method generates lists that facilitate text comprehension more effectively than baselines using collocation measures

Data and Metrics

Graded text corpora
- Training: OneStopEnglish; WeeBit
- Test: Articles from Cambridge English Exams, labeled at CEFR levels A2, B1, B2, C1, C2 (Xia et al., 2016)

Evaluation set-up
- A simulated learner follows the vocabulary list to learn one word per time unit
- The learner "understands" a text if s/he knows at least 90% of the words and MWEs in the text, based on 5,722 MWEs taken from EVP, GSE, and existing MWE datasets
- The learner "graduates" from a CEFR level when s/he can understand 80% of the texts at that level

Evaluation metrics
- **Study time**: Time units needed for the learner to graduate from a CEFR level
- **Text Comprehension**: Average number of texts that can be understood by the learner during the period of simulation

EnglishProfile
The CEFR for English

GSE
Global Scale of English

Image credits: englishprofile.org, pearson.com

Funding: We gratefully acknowledge support of the General Research Fund (11207320), and of the Language Fund from the Standing Committee on Language Education and Research (EDB(LE)/P&R/EL/203/14)

Approach

Rank unigrams and MWEs that appear in training corpora according to their frequency, weighted with a dispersion coefficient (Juilland's D)

Algorithms for identifying MWE candidates
- **Collocation**: Extract top 500K bigrams and trigrams as candidates from English Wikipedia based on Poisson collocation measure (Pickard 2020)
- **Compositionality**: Retrieve top 75% of these 500K candidates with the highest semantic compositionality score (Pickard 2020)

Method	**Study time**					**Text Compre-hension**
	A2	B1	B2	C1	C2	
Collocation	**4,536**	6,007	**10,323**	**25,184**	26,326	87.42
Compositionality	4,984	**5,712**	11,253	25,983	**25,983**	**90.10**
EVP (Human)	2,502	3,610	4,805	/	/	158.95
GSE (Human)	3,728	3,956	6,165	11,157	11,175	135.69

Vocabulary list produced with Collocation method generally yields shorter Study Time at lower levels

Vocabulary list produced with Semantic Compositionality method maximizes Text Comprehension; and minimizes Study Time at highest level