# Exploring Verb-Noun collocations in learner English 

Jen-Yu Li, Cyriel Mallart, Thomas Gaillat, Elisabeth Richard

## To cite this version:

Jen-Yu Li, Cyriel Mallart, Thomas Gaillat, Elisabeth Richard. Exploring Verb-Noun collocations in learner English. Deep learning for language assessment (DLLA) closing event, Nov 2023, Paris, France. hal-04321727

HAL Id: hal-04321727
https://hal.science/hal-04321727
Submitted on 4 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. publics ou privés.


## Motivation

## Collocations:

- Subset of phrasemes

Component of lexical competence[3]

- Limited research with respect to proficiency of L2 learners[1].


## Research Questions:

RQ1: What is the distribution of Verb-Noun (VN) pairs in English essays with respect to L1, CEFR levels, and topics?
RQ2: Is the Point-wise Mutual Information (PMI) score of VN pairs an indicator of texts' CEFR levels?

## Data collection

EF-Cambridge Open Language Database (EFCamDat) [6]
L1 French and Mandarin
Common European Framework of Reference for Languages (CEFR) A1 to C1


Figure 2. Pair per text by L1-CEFR levels

## Processing pipeline



RQ1: Variation by topics
There are 24 topics per level, for a total of 120 topics.


Figure 3. Distribution of texts by topic
Top3 topics: Giving instructions to play a game [B1], Writing a campaign speech [C1], Writing about what you do[A2]


Figure 4. Distributions for VN pairs per text by topic
Peaks: Complaining about chores [A2], Giving instructions to a house-sitter [Mn-A2], Studying online [Fr-B1], Giving advice about budgeting [C1]
Correlation coefficient between the L1 groups:

- For texts: $r=0.91$
- For VN pairs/text: $r=0.97$

VN pairs/text:

- increases with CEFR levels
- varies largely by topic
- [5] proposed a control for potential effects of topic/prompt

RQ2: PMI score distribution


Figure 5. Box-plot of PMI scores across CEFR levels
Median reaches the lowest value at the B1 level while the lowest interquartile range (IQR) also seen at the B1 level.


Figere 7 Distirn Figure 1 Distribution of PMIS
CEFR levels in French group
Median and IQR in French group are lower than that in Mandarin group.


Figure 8. Distribution of PMIs
CEFR levels in Mandarin group

More heterogeneous distributions are presented at lower levels in both L1 groups; may be related to the observed phenomenon of overuse and underuse collocations in [2].

Next steps: native speakers; statistical tests; negative PMI score exploration [4]; log-likelihood ratio

| References |  |  |
| :---: | :---: | :---: |
| [1] Yu-Hua Chen and Paul Baker. Investigating criterial discourse features across second language learner essays, CEFR b1 b2 in rated Applied Linguistics, 37:849-880, 2016. | comparing, and interpreting the evidence |  |
|  | Longuage Learning, 67:155-179, 2017. |  |
|  | [4] Magal Paquot, Dana Gablasov Vaclav Brezinaza and Hubert $N$ Na | Sert Le Brunn and Magai Pad |
|  | Phraseological complextr in EFL | Second Language Acausistion, |
| [2] Philip Durrant and Norbert Schmitt. To what extent do native and non-native writers make use of collocations? <br> IRAL - International Review of Applied Linguistics in Language Teaching, 47(2), 2009. | learners' spoken production across proficiency levels. | ambridge Applied Linguistics, pas 22-147. Cambridge University |
|  | In Agniesza Leenko-Syymaíska a |  |
|  | ata Gotz, editors, Complexity, | enning and modifiring the |
|  |  | Les |
|  |  |  |
| (33] Dana Gablasova, Vaclay Brezina, and |  |  |
| Collocations in corpus-based language | [5] Magali Paquo | International Journal of Learner C |

