Exploring Verb-Noun collocations in learner English

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

Exploring Verb-Noun collocations in learner English

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

LIDILE, Université Rennes 2

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**

L1	Texts	Percentage
French	32,519	4.5%
Mandarin	129,588	17.92%
Others	561,175	77.59%

Table 1. Distribution of texts per L1

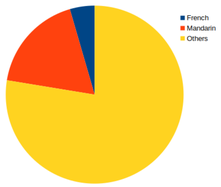


Figure 1. Distribution of texts per L1

L1-CEFR	Texts	VN pairs	Pair/text
French-A1	12,236	17,584	1.44
French-A2	10,008	35,351	3.53
French-B1	6,876	34,043	4.95
French-B2	2,770	17,743	6.41
French-C1	629	5,061	8.05
Mandarin-A1	63,286	104,426	1.65
Mandarin-A2	43,368	178,198	4.11
Mandarin-B1	18,047	94,715	5.25
Mandarin-B2	4,242	30,191	7.12
Mandarin-C1	645	5,762	8.93

Table 2. Distribution of texts per L1-CEFR levels

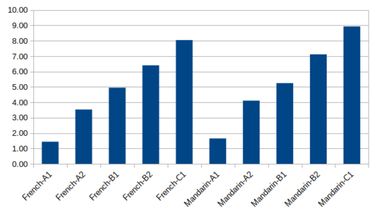
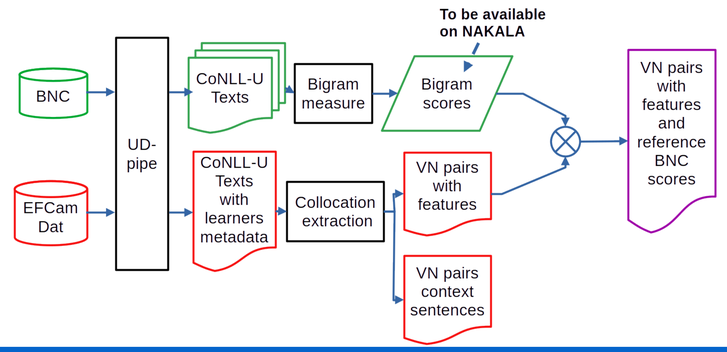


Figure 2. Pair per text by L1-CEFR levels

Processing pipeline



RQ1: Variation by topics

There are 24 topics per level, for a total of 120 topics.

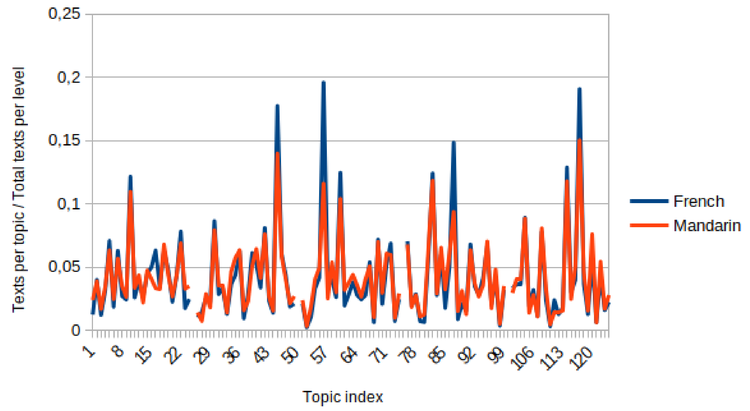


Figure 3. Distribution of texts by topic

Top3 topics: *Giving instructions to play a game [B1], Writing a campaign speech [C1], Writing about what you do[A2]*

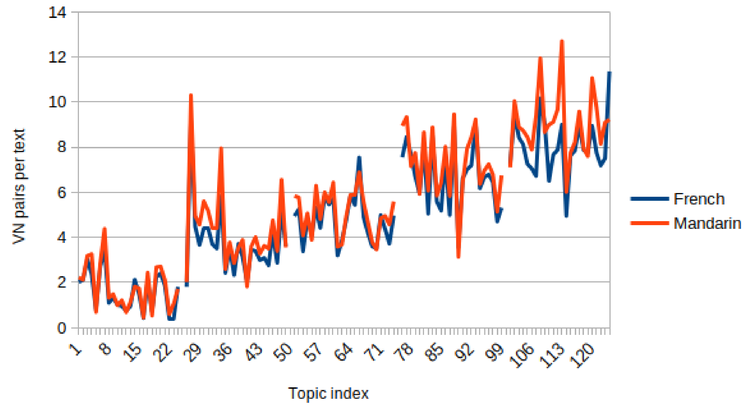


Figure 4. Distributions for VN pairs per text by topic

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

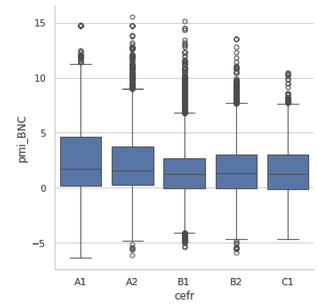


Figure 5. Box-plot of PMI scores across CEFR levels

Median reaches the lowest value at the B1 level while the lowest interquartile range (IQR) also seen at the B1 level.

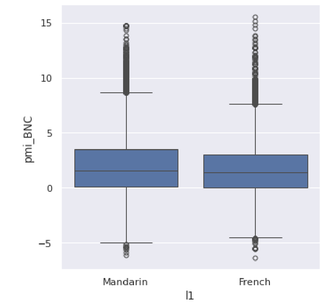


Figure 6. Box-plot of PMI scores across different L1

Median and IQR in French group are lower than that in Mandarin group.

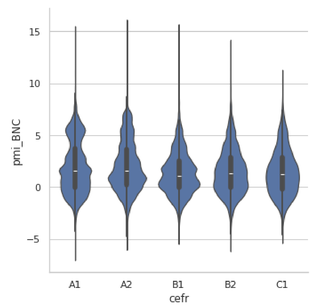


Figure 7. Distribution of PMI scores across CEFR levels in French group

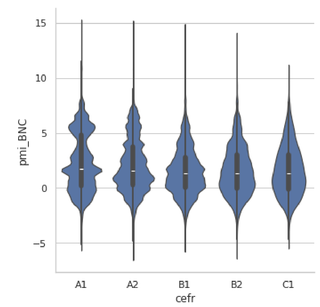


Figure 8. Distribution of PMI scores across CEFR levels in Mandarin group

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 development: Lexical bundles in rated learner essays, CEFR b1, b2 and c1. *Applied Linguistics*, 37:849–880, 2016.

[2] Philip Durrant and Norbert Schmitt. To what extent do native and non-native writers make use of collocations? *IRAL - International Review of Applied Linguistics in Language Teaching*, 47(2), 2009.

[3] Dana Gablasova, Vaclav Brezina, and Tony McEnery. Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, 67:155–179, 2017.

[4] Magali Paquot, Dana Gablasova, Vaclav Brezina, and Hubert Naets. Phraseological complexity in EFL learners' spoken production across proficiency levels. In Agnieszka Leńko-Szymańska and Sandra Götz, editors, *Complexity, Accuracy and Fluency in Learner Corpus Research*, Studies in Corpus Linguistics, pages 115–136. John Benjamins Publishing Company, 2022.

[5] Magali Paquot, Hubert Naets, and Stefan Th. Gries. Using syntactic co-occurrences to trace phraseological complexity development in learner writing: Verb + object structures in LONGDALE. In Bert Le Bruyn and Magali Paquot, editors, *Learner Corpus Research Meets Second Language Acquisition*, Cambridge Applied Linguistics, pages 122–147. Cambridge University Press, 2021.

[6] Itamar Shatz. Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale english learner language database. *International Journal of Learner Corpus Research*, 6(2):220–236, 2020.