INNOVATION IN LANGUAGE LEARNING International Conference

Corpus Linguistics and the Identification of Linguistic Patterns and Meanings: Insights from Learners' Practices – An Exploratory Study

Joana Aguiar[1]

[1]Centro de Investigação Transdisciplinar em Educação e Desenvolvimento do Instituto Politécnico de Bragança, Portugal

Abstract

This paper explores the potential applications of corpus linguistics in English Language Teaching (ELT), specifically in resolving collocations and identifying instances of semantic change. The application of corpus linguistics to language learning has been proven to be successful and to have a positive impact on L2 students. Corpus analysis also draws attention to words or phrases that students might not have accessed via intuition or direct grammatical transposition from L1 [1]. Prior research has identified the use of collocations as a strong indicator of L2 proficiency [2]. In order to assess which online resource best suits students' needs when using a specific collocation or structure, while simultaneously helping them write and speak English more naturally, students were asked to resort to online resources to complete five tasks. The five online resources/platforms used were: Cambridge Learner's Dictionary [3], OZDIC [4], SKELL [5], UrbanDictionary [6], and the British National Corpus website [7]. Students also evaluated which platform was more efficient in problem resolution. Results show that students tend to rely more on traditional online dictionaries despite the task. Overall, the activities involving corpus linguistics are evaluated very positively, as students are given the opportunity to explore online resources autonomously and integrate explicit knowledge through hands-on experience rather than memorisation.

Keywords: *English Foreign Language, corpus linguistics, linguistic patterns, frequency, metalinguistic knowledge.*

1. INTRODUCTION

This paper explores the potential of corpus linguistics and corpus-based online tools in English language classes, namely in the resolution of collocations and in the identification of instances of semantic change. The integration of corpus linguistics into the pedagogical practices of English Language Teaching (ELT) may be a powerful tool. In fact, according to [8], "CBLT [corpus-based language teaching] in teaching aspects of English grammar is more result-oriented than the use of traditional teaching methods.". The same study reveals that corpus linguistics activities improve the understanding of grammatical concepts, such as verb tense and aspect, as well as modal verbs. Resolution of subject-verb agreement and parts of speech analysis also yield better results when using corpus linguistics.

Considering that one of the most common strategies L2 learners adopt is transposing structures from L1, in this case, Portuguese, into L2, corpus linguistics can be used to bring students' attention to words or phrases that might not be accessed via intuition or direct grammatical transposition from L1. Although in some cases, this strategy results in well-formed structures in English, in other cases, the direct translation results in ill-structured phrases and sentences. Another strategy learners use is to resort to the first hit or the first dictionary entry when looking for the meaning of a specific word or phrase, which can be misleading when it comes to collocations, idioms, or words whose meaning is subject to variation or is under a language change process.

To test the potentialities of online tools based on corpus linguistics, a pilot study with five tasks was conducted in 10th-grade classes. By resorting to corpus linguistics and online tools fed by real data, students answered the following questions: What patterns are associated with specific lexical or grammatical features? Do these patterns vary in terms of register or sociopragmatic contexts? How can I make my discourse more natural? At the same time, these tasks aim to develop students' metalinguistic awareness.

2. THEORETICAL BACKGROUND

According to [9], "a corpus is a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis". Among other potential applications in language teaching contexts, corpus linguistics can be used to study language variation and change, quantify linguistic phenomena, and identify linguistic patterns. Moreover, in English Language Teaching (ELT) settings, corpus linguistics provides valuable data-driven insights into how English is used in real-world contexts. As corpora provide authentic language use, they can be used in classroom contexts to identify word usage patterns, frequency, and collocation patterns. In this respect, [10] highlights the importance of corpus-based word lists in selecting high-frequency words and common collocations that are most useful for learners.
The study of phraseology, rather than single-unit analysis, is central to corpus linguistics, as [11] refers. Also, [12] emphasises the relevance of phraseology study in second language proficiency. Nonetheless, phraseology is often given a minor role in language classes and supporting teaching materials. According to [13], the fact that sometimes dictionary definitions do not provide information about the unit of meaning may be misleading to foreign speakers of English, who sometimes use some phrases in a way that sounds amusing and unnatural to native speakers. In this respect, corpus linguistics may be helpful. Ultimately, exploring corpora improves learners' proficiency and develops metalinguistic awareness. By aligning classroom activities with authentic language data, corpus linguistics bridges the gap between theoretical knowledge and practical language use, making learning more relevant and effective [14]. Ultimately, by developing their grammatical competence, corpus-based activities also boost their communicative strategy. In this respect, see [15], according to whom "[Grammatical competence] serves as the linchpin of effective communication, a catalyst for academic and professional success, a guardian of linguistic richness, and a facilitator of aesthetic expression.".
Considering that "collocation is the statistical tendency of words to co-occur" [16], this study departed from four collocations and five tasks associated with them. The objective is for students to explore the potential of corpus linguistics analysis in the classroom, become familiar with various online resources and platforms, and evaluate which one best meets their needs in completing the assignment.

3. METHODOLOGY

Collocations are part of the English (second language) syllabus. Instead of providing random examples of collocations and having traditional drills, I challenged 10th-grade students (CEFR B1 level) to explore the potential of online tools (dictionaries and other online resources fed by corpora) to develop their grammatical competence. At the same time, students improve their digital competence. This section describes the online tools consulted and the methodology used to collect data.

3.1 Online tools under analysis

This subsection provides a brief description of the online tools used by students to complete the proposed tasks.
- Cambridge Learner's Dictionary [3] is an online dictionary for intermediate learners of English (CEFR B1-B2 levels). According to the website's description, it contains simple definitions and thousands of carefully chosen example sentences from the Cambridge English Corpus, a database of over two billion words.
- OZDIC [4] is an online English collocation dictionary based on the British National Corpus. It displays all the words frequently used in combination with each headword, including nouns, verbs, adjectives, adverbs, prepositions, and common phrases. It also provides examples and elementary grammatical information. According to the information available on the website, "[it] is designed to help language learners and users write and speak natural-sounding English" [4].
- SKELL [5] stands for Sketch Engine for Language Learning. It is a free, user-friendly tool designed for English students and teachers. By typing a given word or expression, users access selected sentences taken from existing corpora, which allows them to quickly check how a specific word or phrase is used by speakers in various contexts. SKELL retrieves sentences from Wikipedia articles, selected parts of the English Web 2013 corpus and the BNC corpus, and English websites crawled by the WebBootCat tool [17], totalling 57 million sentences. However, the tool does not provide the source from which the example was taken. In addition to the tab that provides corpus examples, users

can access the "Word Sketch" and "Similar Words" tabs. Word Sketch displays the frequency of occurrence according to the syntactic context. The "Similar Words" tab presents the most frequent synonyms in a word cloud format.
- Urban Dictionary [6] is a crowdsourced online dictionary of slang words and phrases. The dictionary entries, definitions, and examples are proposed by users and are subject to approval.
- The British National Corpus (BNC) [7] is a 100-million-word collection featuring samples of written and spoken language from a diverse range of sources from the late 20th century. It is freely accessible online, although registration is requested.

3.2 The informants

This study was conducted in five 10th-grade classes from a secondary school in the Northeast of Portugal. A total of 59 surveys were collected. All students are enrolled in general secondary programmes: 22 students in Sciences and Technologies, 6 in Socio-Economic Sciences, 11 in Social and Human Sciences, and 20 in Visual Arts.

3.3 The tasks

To conduct this pilot study, students were asked to complete five tasks involving collocations (cf. Appendix section) plus an additional tool assessment task.
The collocations were selected from the syllabus and vocabulary content to be covered in the 10th grade. The only exception is Task 1, "come to terms with". This task was based on [13]. The objective is to explore the semantic prosody of this phrase. Although most dictionaries list the object of this phrase as non-human, the examples from the BNC corpus pair the expression with human entities.
In Task 2, students must search for the expression "be fond of" and identify whether the subject is commonly a human entity (somebody) or something. Resorting to the online tools, they should also determine the grammatical category of "fond."
Task 3 involves the expression "be sick" (students should conjugate the verb if needed). The objective is to explore whether the connotations of "sick" are negative or positive. Although the expression *it is sick* or *simply sick* has been used with a positive connotation to refer to *cool things*, this definition is not yet dictionarised. Considering the platforms under assessment, at the time of this article, only UrbanDictionary listed this meaning.
In Task 4, students must search for the phrase "let alone" and identify the most frequent context of use. The BNC website offers information on the type of corpus (written or oral), and Cambridge's Learners Dictionary provides information on the sentence semantic context (*used after a negative statement to emphasize how unlikely a situation is because something much more likely has never happened* [3]).

3.4 The task-evaluation survey

After completing the tasks, students answered an online survey via menti.com. The online survey aimed to evaluate the impact of using the online tools involving corpus linguistics and the extent to which they would likely use some of them in the future.
Question 1 collected personal information, such as the class in which they are enrolled. Question 2 asked them to indicate how much they agree or disagree with the following statements (rating from 1- Strongly disagree to 5- Strongly agree): (i) Overall, I enjoyed this task; (ii) I learned a lot about English with this task; (iii) I became familiar with new tools for learning English; (iv) I will use these online tools/resources in the future. Questions 3 and 4 asked them which tool they would use to find the definition of a word and to find the collocation of a word, respectively. Finally, in question 5, students had to indicate their favourite tool.

4. FINDINGS AND DISCUSSION

To complete the task, each student was required to use their laptop, tablet, or cell phone. Although students could exchange ideas with colleagues, the task was performed individually. Some students were initially reluctant and felt it was pointless to go through the tasks and compare the results retrieved on the different websites. As BNC is not as user-friendly as other websites and requires registration, I began exploring BNC and showcasing its potential. The hits were projected, and students were asked to collaboratively answer the proposed tasks. In this first moment, it was noticeable how some students could quickly infer occurrence patterns from the random sentences

extracted from BNC. After answering all questions using BNC, students were given time to explore the other websites. This task was performed individually.

When exploring both the Cambridge Learner's Dictionary and the Urban Dictionary, students tend to rely on the first hit or the first information displayed. The fact that the selected collocations were not the most salient uses of the term under analysis meant that it was crucial to develop the ability to search further. In practical terms, this meant scrolling down or using Ctrl+F in some cases. It was clear that, in some cases, students lacked basic digital competencies, namely in finding and retrieving specific information.

According to the survey results (cf. Figure 1), most students enjoyed this task (37.3% agreed and 25.4% strongly agreed with the statement "Overall, I enjoyed this task"). Only seven students did not enjoy this activity. Furthermore, 84% reported that they learned a lot about English through this task, and 83% became familiar with new tools for learning English. Finally, 69.5% indicated they will make use of these online tools in the future.

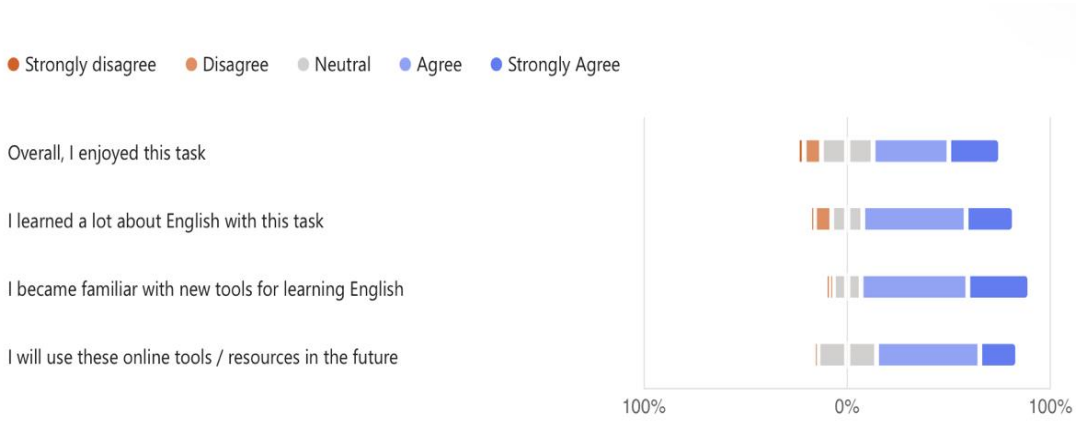

Fig. 1. – Evaluation of the dynamic *Corpus Linguistics and beyond.*

Figure 2 displays the results to the question "Which tool would you use to find the definition of a word?". 51% would use the Cambridge Learner's Dictionary, and 17% indicated SKELL. Although SKELL is mainly used as a corpus engine to retrieve real data, students also explored the tab "Similar Words", which displays similar words in a word cloud format. In order to check possible crosstabs between the answers to this question and the branch of studies, this variable was verified. Most students answering SKELL are enrolled in Arts and Sciences and Technologies. OZDIC, on the other hand, is preferred by students enrolled in Humanities. No statistically significant results were found for the results obtained for other online tools.

Fig. 2. – Answers to the question, "*Which tool would you use to find the definition of a word?*"

Figure 3 displays the results to the question, "Which tool would you use to find the collocation of a word?". Results are more dispersed: 34% of the students would use Cambridge Learner's Dictionary, 25% OZDIC, and 25% SKELL. No correlation between the option selected and the area of study was found.

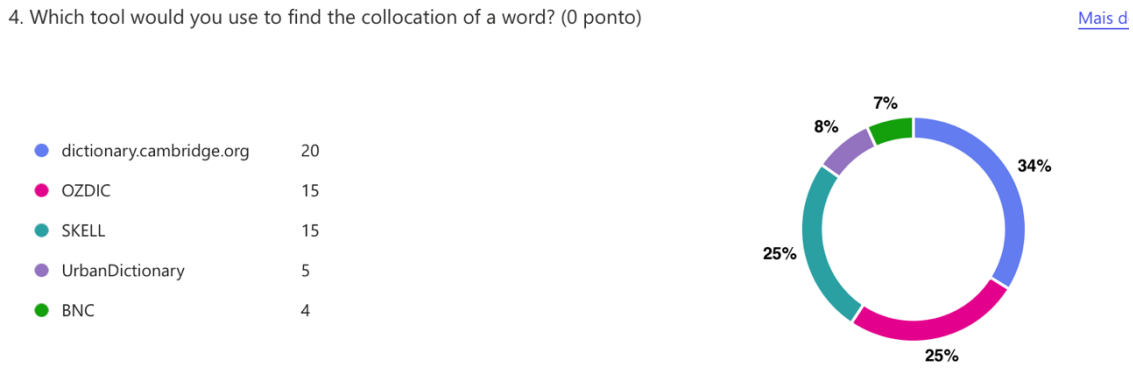

Fig. 3. – Answers to the question, "*Which tool would you use to find the collocation of a word?*"

Finally, question 5 asked students which their favourite tool was. Results (cf. Figure 4) show that although most students (39%) preferred Cambridge Learner's Dictionary, 25% selected Urban Dictionary, and 22% selected Skell. While no relevant differences were found according to the study branch, Urban Dictionary was mostly selected by students enrolled in the Arts and Humanities branches. Only 4 students preferred OZDIC or BNC. Students who selected BNC are enrolled in Arts (2) and Sciences and Technology (2). Students who prefer OZDIC are enrolled in the Humanities (3) and Economics (1) branches.

Fig. 4. – Answers to the statement, "My favourite tool was…"

5. Conclusion

This pilot study is an exploratory work on how corpus linguistics may be used in EFL settings. The task aimed to raise students' awareness that meaning does not exist except in context [18], and that some verbs select specific prepositions or other grammatical categories to form specific meanings. Furthermore, students had the opportunity to explore the meaning and uses of fixed expressions. Semantic change was also tackled with the collocation "to be sick". Overall, applying corpus linguistics to language exploration proved to be a highly productive task. Fundamentally, it was more than a task about a grammar topic. The task introduced some online tools based on corpus linguistics and facilitated the exploration of data-driven materials. It was an incentive for students to develop strategies to "learn how to learn" [19] and to diversify the didactic approaches to grammar teaching. As a teacher, I had a minor role in students' fulfilment of the task. My role was to facilitate their interaction with the online tools and clarify any doubts they may have had. By integrating activities involving corpus linguistics, students become familiar with useful resources in English language learning that may be useful in the future. Furthermore, these activities are student-centred and result-oriented, thus having a positive impact on language learning.

Acknowledgements

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/05777/2023

REFERENCES

[1] Hunston S., "Corpora in Applied Linguistics", Cambridge, Cambridge University Press, 2002.

[2] Eguchi M., Kyle K., “L2 collocation profiles and their relationship with vocabulary proficiency: a learner corpus approach”, Journal of Second Language Writing, 60, 2023,100975.
[3] Cambridge University Press. (n.d.). Cambridge Learner's Dictionary. https://dictionary.cambridge.org/dictionary/learner-english/
[4] OZDIC. (n.d.). Ozdic English Collocation Dictionary. https://ozdic.com
[5] Sketch Engine. (n.d.). Skell: Sketch Engine for Language Learning. https://skell.sketchengine.eu/#home?lang=en
[6] The Urban Dictionary (n.d.). Urban Dictionary. https://www.urbandictionary.com
[7] The British National Corpus. (n.d.). The British National Corpus (BNC). https://www.english-corpora.org/bnc/
[8] Jacobs S. & Isaac N., “The use and evaluation of corpus-based English language teaching”, International Journal of Literature, Language and Linguistics, 7(1), 2024, 80–104. https://doi.org/10.52589/ijlll-p4aeacwh
[9] Francis W. N., “Problems of assembling and computerizing large corpora”, in S. Johansson (ed.) “Computer corpora in English language research”. Bergen, Norwegian computing centre for the humanities, 1982, 7–24,
[10] Alzeer S., Thompson P., “Toward a tool for evaluating corpus-based word lists for use in English language teaching contexts”, Applied Corpus Linguistics, Article 100098, Advance Online Publication, 2024. Https://doi.org/10.1016/j.acorp.2024.100098
[11] Sinclair J. (Ed.), “Corpus, concordance, collocation”, Oxford, Oxford University Press, 1991.
[12] Howarth P., “Phraseology and second language proficiency”, Applied Linguistics, 19(1), 1998, 24–44. https://doi.org/10.1093/applin/19.1.24
[13] Hunston S., “Semantic prosody revisited”, International Journal of Corpus Linguistics, 12(2), 2007, 249-268.
[14] Ghadessy M., Henry A., Roseberry R. L., “Small Corpus Studies and ELT: Theory and Practice”, John Benjamins Publishing Company, 2001.
[15] Jacobs S., Isaac N., “The use and evaluation of corpus-based English language teaching, International Journal of Literature, Language and Linguistics, 7(1), 80–104, 2024. https://doi.org/10.52589/ijlll-p4aeacwh
[16] Bennett, G. R., “Using corpora in the language learning classroom: corpus linguistics for teachers”, University of Michigan Press, 2010.
[17] Baroni M., Kilgarriff A., Pomikálek J., Rychlý P., “Webbootcat: a web tool for instant corpora”, Proceedings of EAMT: 11th Annual Conference of the European Association for Machine Translation, Oslo, Norway, 2006, 247–252.
[18] Teubert W., “Writing, hermeneutics, and corpus linguistics”, Logos and Language 2, 2003, 1–17.
[19] Johns T., “Should you be persuaded - two examples of data-driven learning materials”, English Language Research Journal, 4, 1991, 1–16.

APPENDIX

1- ASSIGNMENT - EXPLORING CORPUS LINGUISTICS AND BEYOND

Assignment - Exploring corpus linguistics and beyond

NAME:_____**CLASS:**_____**Nº:**_____

1- Please consider the tasks below. For each task, check the information on the following online resources/websites:
SKELL - https://skell.sketchengine.eu/#home?lang=en
OZDIC - https://ozdic.com
Cambridge Learner's Dictionary - https://dictionary.cambridge.org/dictionary/learner-english/
Urban Dictionary - https://www.urbandictionary.com
British National Corpora - https://www.english-corpora.org/bnc/

2- Use the table below to register your answers:

Collocation	Task	*Cambridge Learner's Dictionary*	OZDIC	SKELL	UrbanDictionary	*The British National Corpus*	Are there any differences in the information retrieved from the different sources?
come to terms with	Is the expression followed by a human entity or something?						
be fond of	What is the grammatical category of fond?						
	Is the subject a human entity or something?						
to be sick	Are the connotations of "sick" negative or positive?						
let alone	What is the most frequent context of use?						

Access menti.com (code 4836 6468) to evaluate this assignment.