Revista Crátilo, 10(1): 87-96, ago. 2017
© Centro Universitário de Patos de Minas
http://cratilo.unipam.edu.br

Corpus Linguistics as a tool for language teachers[1]

A Linguística de Corpus como ferramenta para professores de línguas

Mateus Emerson de Souza Miranda
English undergraduate student at Universidade Federal de Minas Gerais.
E-mail: mateusesm@gmail.com

Abstract: A corpus (plural corpora) refers to a large collection of linguistic data which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language (CRYSTAL, 1991). Currently, corpora are collected for various purposes, including materials for English teaching, such as grammars, books or dictionaries. These materials illustrate how the language is actually used. The use of corpora in the classroom can be an effective tool for language teaching. In this paper, we will present corpora-based activities for vocabulary teaching in the context of English as a Foreign Language (EFL).
Keywords: Corpus Linguistics. Tool. Teaching.

Resumo: Um corpus (plural corpora) refere-se a uma grande coleção de dados linguísticos que pode ser usado como um ponto de partida para a descrição linguística ou como um meio de verificar hipóteses sobre uma língua (CRYSTAL, 1991). Atualmente, corpora são coletados para vários fins, inclusive materiais para o ensino de Inglês, como gramáticas, livros ou dicionários. Esses materiais ilustram como a língua é realmente usada. O uso de corpora em sala de aula pode ser uma ferramenta eficaz para o ensino de línguas. Neste artigo, apresentaremos atividades baseadas em corpora para o ensino de vocabulário no contexto Inglês como Língua Estrangeira (EFL).
Palavras-chave: Linguística de Corpus. Ferramenta. Ensino.

1 Introduction

DATA-DRIVEN LEARNING AND THE USE OF CORPORA IN LANGUAGE TEACHING

Corpus Linguistics (CL) has been used in the language classroom since the 1980s through concordance lines, which are all the instances of a word or *node* in a corpus. A corpus (plural: corpora) refers to "a large collection of linguistic data, either written texts, or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language" (CRYSTAL, 1991, p 86). Therefore, Corpus Linguistics deals with the analysis of language use through corpora. Currently, corpora are collected for various purposes, such as materials for English teaching, like grammar books or dictionaries. These materials illustrate how a language is really used.

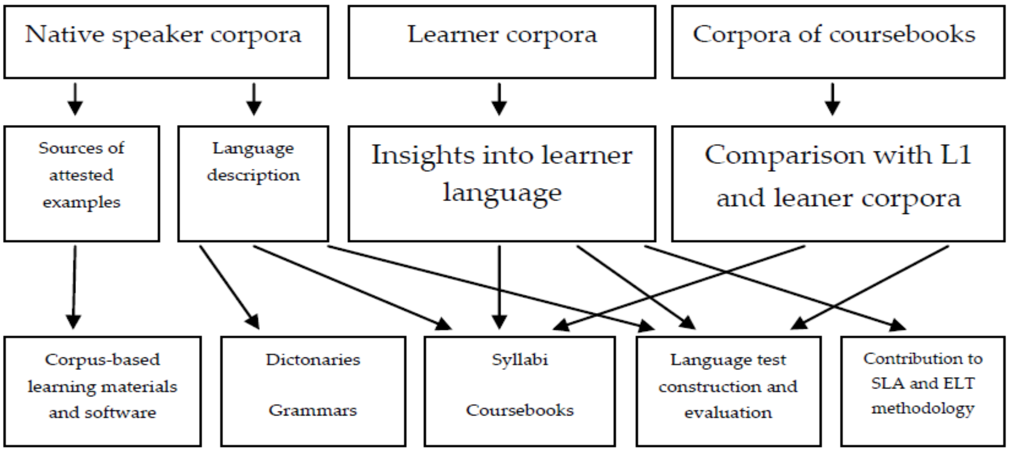
[1] This research was carried out as part of *Seminários de Pesquisa* subject requirements. A preliminary version of this paper was presented at VII Semana da Letras da UFMG.

The term data-driven learning (DDL) was first introduced by Tim Johns in 1990 to describe an inductive approach, where learners are considered detectives by analyzing patterns through computer-generated concordance lines in the target language to form generalizations (JOHNS, 1991). According to Granger and Gilquin (2010), data-driven learning consists of using the tools and techniques of corpus linguistics for pedagogical purposes. By using this approach, teachers can provide learners direct access to a corpus so they can explore concordance lines to explore patterns. According to Johns (1991: 30):

> What distinguishes the DDL approach is the attempt to cut out the middleman as much as possible and give direct access to the data so that the learner can take part in building his or her own profiles of meanings and uses. The assumption that underlines this approach is that effective language learning is itself a form of linguistic research, and that the concordance printout offers a unique resource for the stimulation of inductive learning strategies – in particular, the strategies of perceiving similarities and differences and of hypothesis formation and testing.

CL tools and the empirical studies provide contributions to increasing language access to learners and language teaching (Figure 1).

Figure 1: Interconnecting ways in which corpora are relevant to language teaching


Native speaker corpora
Learner corpora
Corpora of coursebooks
Sources of attested examples
Language description
Insights into learner language
Comparison with L1 and leaner corpora
Corpus-based learning materials and software
Dictonaries
Grammars
Syllabi
Coursebooks
Language test construction and evaluation
Contribution to SLA and ELT methodology

Source: McEnery; Gabrielatos, 2006, p. 51.

Granger and Gilquin (2010) state that, among other advantages, the DDL approach brings authenticity into the language classroom and, including the element of discovery, it makes learning more fun, in addition to motivating the students. Corpus-based activities are thought to increase learner autonomy, "as students are taught how to observe language and make generalizations rather than depending on a teacher" (CONRAD, 2005, p. 402). Therefore, the use of corpora in the language classroom can be effective in exploring language aspects such as grammar (categories such as words, phrases, clauses and sentences), discourse (organization of speech and writing), lexis (describes the typical environment of a word), pronunciation (describes accents and explores intonation), and many others.

There are three realms (BARLOW, 2002) in which Corpus Linguistics can be applied to teaching: syllabus design, materials development, and classroom activities. Firstly, syllabus design organizes the teacher's decisions regarding the focus of a class with respect to an individual student's needs. By conducting an analysis of a corpus which is relevant to the purpose of a particular class, the teacher can determine what language items are linked to the target register. Secondly, materials development relies on a developer's intuitive sense of what students need to learn. With the help of a corpus, a materials developer could create exercises based on real-life examples, which provide students with a more interactive opportunity to discover unique features of language use. Finally, classroom activities consist of hands-on, student-conducted language analyses in which the students use a concordance program and a deliberately chosen corpus to make their own discoveries regarding language use (BARLOW, 2002). The teacher can guide a predetermined language investigation, which will lead to predictable results, or the teacher can have the students perform their own investigations, leading to less predictable findings.

This paper will present a corpus-designed activity teachers can use in the classroom. More specifically, we will present a practical technique using an online corpus to teach phrasal verbs to increase students' communicative competence by manipulating corpus data and interpreting the information it provides.

2 Phrasal verbs as a problem for EFL students

A very common situation in the English as a Foreign Language (EFL) classroom is students lacking vocabulary when trying to communicate with classmates. It is also very common to listen to their complaints regarding a lack of strategies for learning vocabulary. Larsen-Freeman states that communicative methods are concerned with basing course content on activities that are contextualized, by focusing on the discourse level rather than the sentence level, and by providing students with opportunities to develop strategies for interpreting and using the language as it is used by native speakers (LARSEN-FREEMAN, 2000).

McCarthy (1990) states that students can master second language (L2) grammar and phonology, but communication in an L2 will not happen in a meaningful way without the knowledge of vocabulary in order to express such meanings. When it comes to learning vocabulary, phrasal verbs have been an especially difficult issue for non-native speakers of English (SCHMITT and SIYANOVA, 2007). L2 learners usually find phrasal verbs hard to both master and apply in conversations. In addition, many English teachers also have difficulty finding a suitable approach to teaching phrasal verbs. According to Brown (2000, p 377):

> The best internalization of vocabulary comes from encounters (comprehension or production) with words with the context of surrounding discourse. Rather than isolating words and/or focusing on dictionary definitions, attend to vocabulary with a communicative framework in which items appear. Students will associate new words with a meaningful context to which they apply.

2.1 Teaching phrasal verbs with corpus-designed activities

To address the topic of phrasal verbs, we came up with and planned an interactive activity that can be effective when teaching phrasal verbs to intermediate and advanced students. Through this activity with concordance lines, students can interact with authentic native English speaker materials and analyze concordance lines extracted from a native speaker corpus to infer the meanings of the phrasal verbs without using a dictionary, allowing them to increase their vocabulary. More specifically, as an example, we used the phrasal verb 'to come up with' in order to show how the activity we created can be developed.

2.2 The corpus used

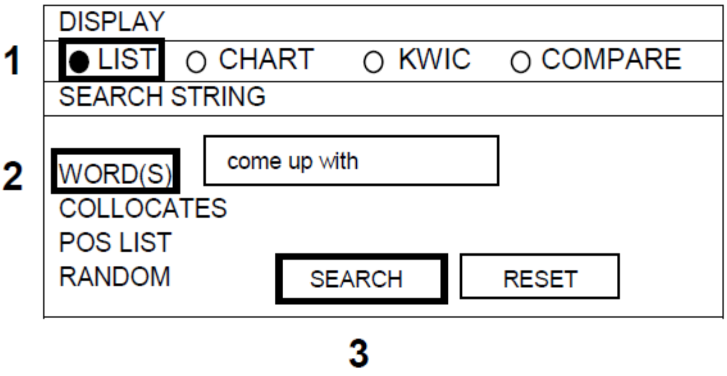
To develop this activity, we used the *British National Corpus (BNC).* This free online corpus was created by Professor Mark Davis at Brigham Young University (BYU), and it contains 100 million words of British English text from a wide range of genres. Ten percent of the corpus is transcribed spoken language. To access the corpus, a school needs access to a computer so students can use the corpus to do the activity in class. Some other free online corpora that teachers can use to design and plan activities like this can be found in Appendix A.

3 Hands-on activity

To begin this corpus-based activity, engage students in a discussion about some common phrasal verbs in English, making sure to evaluate each student's current understanding of them. After a general discussion, divide the students into groups to access the online corpus to analyze the phrasal verbs' concordance lines. Show the students how to access the data by doing the first search with them. Ask students if they know the meaning of the phrasal verb 'to come up with' and elicit their ideas. After the discussion, you will guide them through the following steps:

1. Students will access the BNC, which is available at http://corpus.byu.edu/bnc/.
2. After they access the BNC website, tell students to select LIST (**1**) in DISPLAY on the left of the screen. Then in SEARCH STRING, tell students to write the phrasal verb 'come up with' write in WORD(S) (**2**). Next, they will click on SEARCH (**3**) (Figure 2):

Figure 2: Illustration of how to search for word(s) using the BNC website.



DISPLAY

● LIST ○ CHART ○ KWIC ○ COMPARE

SEARCH STRING

WORD(S) come up with

COLLOCATES

POS LIST

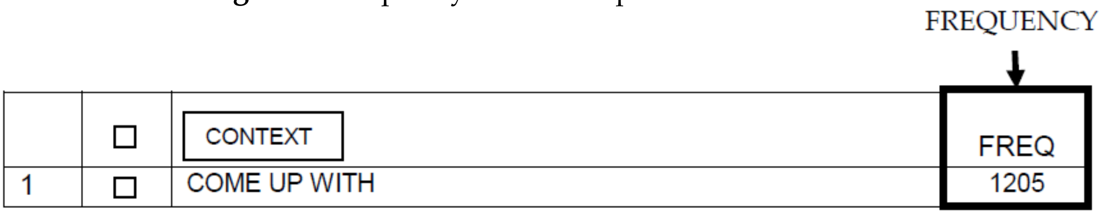
RANDOM

SEARCH

RESET

3. Now, students will be able to see the frequency of "come up with" in the corpus, which appears 1,025 times in the BNC (Figure 3).

Figure 3: Frequency of "come up with" in the BNC.

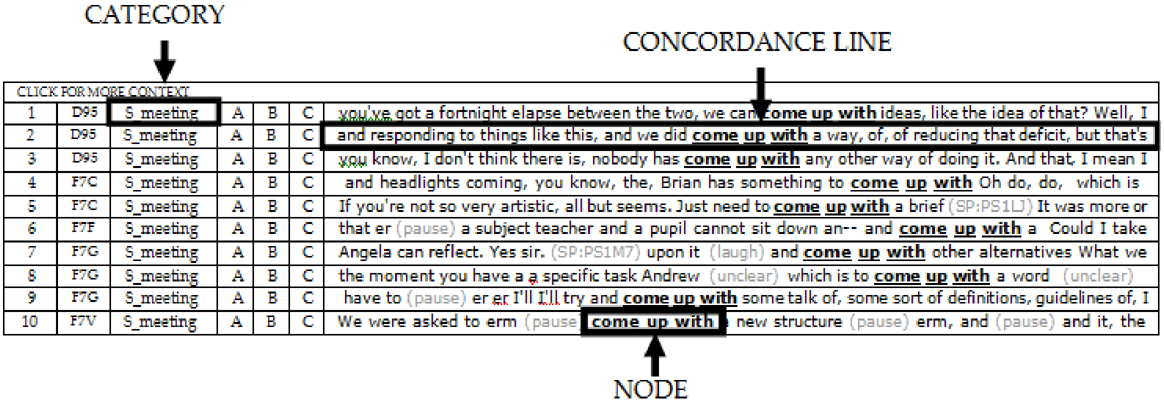


	□	CONTEXT	FREQ
1	□	COME UP WITH	1205

4. Then, students will click on the phrasal verb to access the concordance lines. Now, students have "come up with" concordance lines (Figure 4). The concordance lines and the underlined node word are visible. The node word is the one the students searched for in the BNC. In the first examples below, the examples with "come up with" were taken from spoken interviews which are transcribed in the corpus.

(Note: there is a column with the category from which this sample of language was taken out, e.g. meeting, e-mail, or interview.)

Figure 4: Example of concordance lines generated using "come up with".



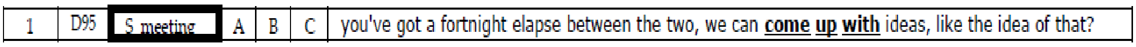
CLICK FOR MORE CONTEXT						
1	D95	S_meeting	A	B	C	you've got a fortnight elapse between the two, we can come up with ideas, like the idea of that? Well, I
2	D95	S_meeting	A	B	C	and responding to things like this, and we did **come up with** a way, of, of reducing that deficit, but that's
3	D95	S_meeting	A	B	C	you know, I don't think there is, nobody has **come up with** any other way of doing it. And that, I mean I
4	F7C	S_meeting	A	B	C	and headlights coming, you know, the, Brian has something to **come up with** Oh do, do, which is
5	F7C	S_meeting	A	B	C	If you're not so very artistic, all but seems. Just need to **come up with** a brief (SP:PS1LJ) It was more or
6	F7F	S_meeting	A	B	C	that er (pause) a subject teacher and a pupil cannot sit down an-- and **come up with** a Could I take
7	F7G	S_meeting	A	B	C	Angela can reflect. Yes sir. (SP:PS1M7) upon it (laugh) and **come up with** other alternatives What we
8	F7G	S_meeting	A	B	C	the moment you have a a specific task Andrew (unclear) which is to **come up with** a word (unclear)
9	F7G	S_meeting	A	B	C	have to (pause) er er I'll I'll try and **come up with** some talk of, some sort of definitions, guidelines of, I
10	F7V	S_meeting	A	B	C	We were asked to erm (pause **come up with** new structure (pause) erm, and (pause) and it, the

5. After having access to the concordance lines, students will have the opportunity to read and analyze them. A useful tip is to guide students by pointing to the examples you want them to read, which are selected prior to the lesson. These are some concordance lines taken from Figure 4:

1. […] we can **come up with ideas**, like the idea of that?
2. […] responding to things like that, and we did **come up with** a way, of, of reducing that deficit
3. we were asked to erm (pause) **come up with** a new structure (pause) erm […]

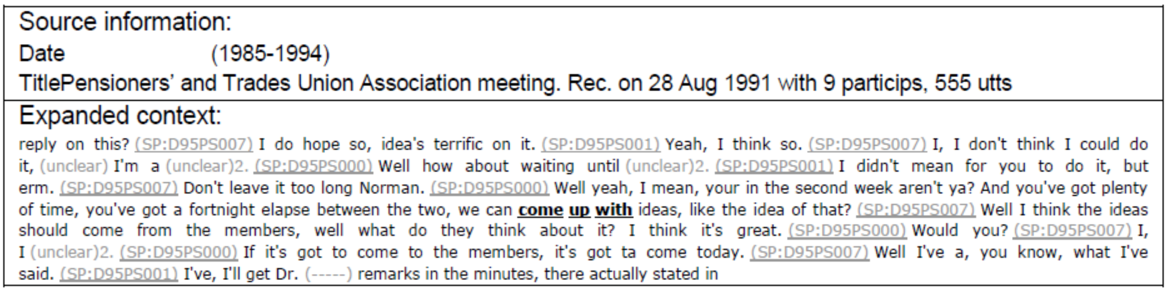
6. In addition to reading the concordance lines, if students need to see the whole context where the word was used in order to interpret the meaning of the phrasal verb, they can click on the context column of the line they are analyzing to check the expanded context of the phrasal verb, as in Figures 5 and 6 below.

Figure 5: Concordance line with context column highlighted



1	D95	S meeting	A	B	C	you've got a fortnight elapse between the two, we can **come up with** ideas, like the idea of that?

Figure 6: Expanded context of "come up with" from a concordance line



Source information:

Date (1985-1994)

TitlePensioners' and Trades Union Association meeting. Rec. on 28 Aug 1991 with 9 particips, 555 utts

Expanded context:

reply on this? (SP:D95PS007) I do hope so, idea's terrific on it. (SP:D95PS001) Yeah, I think so. (SP:D95PS007) I, I don't think I could do it, (unclear) I'm a (unclear)2. (SP:D95PS000) Well how about waiting until (unclear)2. (SP:D95PS001) I didn't mean for you to do it, but erm. (SP:D95PS007) Don't leave it too long Norman. (SP:D95PS000) Well yeah, I mean, your in the second week aren't ya? And you've got plenty of time, you've got a fortnight elapse between the two, we can **come up with** ideas, like the idea of that? (SP:D95PS007) Well I think the ideas should come from the members, well what do they think about it? I think it's great. (SP:D95PS000) Would you? (SP:D95PS007) I, I (unclear)2. (SP:D95PS000) If it's got to come to the members, it's got ta come today. (SP:D95PS007) Well I've a, you know, what I've said. (SP:D95PS001) I've, I'll get Dr. (-----) remarks in the minutes, there actually stated in

7. After students read some lines, they will probably be able to infer the meaning of the phrasal verbs. Since this will be the first example, ask the whole class what they found out during their search. After the discussion, come up with a definition on the board:

Meaning: to develop or discover a solution when a problem is presented.

Other Examples of Phrasal Verb Definitions using Concordance Lines

Here are some more examples of phrasal verbs and their concordance lines, taken from the BNC:

(A) **Bring up**

1. Women are choosing to **bring up** children on their own without close relationships with men. Many

2. What's the problem of returning to work and allowing someone else to **bring up** the child?

3. Our daughter is trying to **bring up** three children on 8,000 a year and has to pay 800 of that in income

Meaning: to educate; raise.

(B) **Put off**

1.He and his wife decided not to **put off** having the second child they had planned, despite financial uncertainty.

2. Bride n' gloom # TV soap star Sue Nicholls has **put off** her wedding to actor Mark Eden.

3. All of us have a tendency to **put off** the difficult tasks or those we dislike. You will need to be honest with yourself

Meaning: to postpone.

(C) **Take up**

1. If you **take up** metaphysics, for example, Ego will delight in trying to convince you that this makes you' better than' other people.

2. she'll be happy to achieve at least one other mission: to encouraged more young people to **take up** science.

3. it was necessary to **take up** mathematics which the headmaster himself taught.

Meaning: to study; start a new activity.

(D) **Cut down**

1. As a safeguard, the group advices pregnant women to eat healthily, **cut down** on drinking and avoid smoking.

2. Yet it's easy to **cut down** on fat without changing your diet completely or giving up all your favourite foods.

3. he Hungarian government is introducing a " green card " system for cars to ensure that they have regular checks designed to **cut down** on emissions of carbon monoxide

Meaning: to reduce.

Alternative Methods in Place of Universal Computer Access

If teachers do not have access to resources like a computer lab with internet access for all students to work together on computers, they can select and print some concordance lines before class begins. After the beginning of the lesson, the teacher can divide the class into groups and give each group a set of lines to analyze. The students can then guess the meanings of the phrasal verbs from the provided lines. If the teachers have only a classroom with a computer and a projector, they can project the screen on the board, search for the phrasal verbs, and show students by working directly with the concordance lines on the projector, using the same methodology.

4 Final remarks

This article presents an idea of an L2 activity using native English speaker corpora, which has high potential for great contributions to teaching and applied linguistics. In some cases, it can be difficult for teachers to teach certain aspects of a language. Therefore, teachers can use corpora as a language teaching technique to create corpus-based activities to incorporate in class, not only making teaching easier but also significantly improving the learning process for the students. By guiding students through the use of concordance lines, they can analyze extracts of real language in different contexts, infer meanings, and share their ideas with fellow classmates.

References

BARLOW, Michael. Corpora, concordancing and language teaching. *Proceedings of the 2002 KAMALL International Conference.* Daejon, Korea, 2002.

BROWN, Douglas. *Teaching by principles:* An interactive approach to language pedagogy (2nd edition). USA: Pearson Education, 2000.

CONRAD, Susan. Corpus linguistics and L2 teaching. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*. Mahwah, NJ: Lawrence Erlbaum, *2005,* p. 393-409.

CRYSTAL, David. *A Dictionary of Linguistics and Phonetics*. 3rd ed. London: Blackwell, 1991.

GILQUIN, Gaetanelle; GRANGER, Sylviane. How can data-driven learning be used in language teaching? In: A. O'Keeffe & M. McCarthy (eds.). *Routledge handbook of corpus linguistics*. London: Routledge, 2010, p. 359-370.

JOHNS, Tim. Should you be persuaded: two examples of data driven learning. *ELR Journal (New Series)*, 1991, p. 1-16.

LARSEN-FREEMAN, Diane. *Techniques and principles in language teaching*: Oxford: Oxford University Press, 2000.

MCCARTHY, Michael. Vocabulary. Oxford: Oxford University Press, 1990.

MCENERY, Tony; GABRIELATOS, Costas. English Corpus Linguistics. In: *Handbook of English Linguistics*, ed. AARTS, Bas and MCMAHON, April. London: Blackwell, 2006, p. 33-71.

SCHMITT, Norbert; SIYANOVA, Anna. Native and nonnative use of multi-word vs. one word verbs. *IRAL*, 45(2), 2007, p. 119-139

Appendix A

Some other free online corpora available for teachers and students

Corpus of Contemporary American English (COCA)

Created by Mark Davis, COCA is the largest English corpus available; it contains more than 450 million words, and it is a reference corpus of English used in the United States. Its distribution takes place in five different categories: spoken, fiction, popular magazines, newspaper articles, and academic journals. COCA is available at http://corpus.byu.edu/coca/.

Michigan Corpus of Academic Spoken English (MICASE)

MICASE, available at http://quod.lib.umich.edu/m/micase/, contains more than 1.9 million words of audio recording and transcripts from university settings.

Time Corpus

Time Corpus is available at http://corpus.byu.edu/time/. Created by Mark Davis at Brigham Young University (BYU), it is a corpus of Time Magazine from 1923 through 2006. It is a useful resource for exploring written academic language.