https://doi.org/10.11606/issn.2317-9511.v37i0p64-87

Do-it-Yourself Corpora: what are they and who are they for?

DIY Corpora: o que são e para quem são?

Carolina Tavares de Carvalho[*]
Luana Aparecida Nazzi Laranja[**]
Paula Tavares Pinto[***]

[*] Mestranda do Programa de Estudos Linguísticos do Instituto de Biociências, Letras e Ciências Exatas da Unesp, Campus de São José do Rio Preto - SP. E-mail: carolina.tavares@unesp.br

[**] Mestre do Programa de Estudos Linguísticos do Instituto de Biociências, Letras e Ciências Exatas da Unesp, Campus de São José do Rio Preto - SP. E-mail: luananazzi@gmail.com

[***] Professora Assistente Doutora do Programa de Estudos Linguísticos do Instituto de Biociências, Letras e Ciências Exatas da Unesp, Campus de São José do Rio Preto - SP. E-mail: paula@ibilce.unesp.br

Abstract: The use of corpora in translation, teaching, and academic writing has been increasing significantly over the last years. Do-it-yourself (DIY) corpora are particularly useful to those areas. This paper presents a theoretical review and discusses the use of DIY corpora in translation, in language teaching, and in academic writing. Evidence shows that, with the aid of computational tools, DIY corpora provide translators with better options for translating texts and extracting terminology. At the same time, it can be a resourceful tool for teachers planning their language for general or specific purposes classes, or for researchers who aim to observe the structure and language patterns of their own areas of research. The studies described show that DIY corpora are resourceful tools that can be used in different but correlated contexts.

Keywords: DIY corpora; Corpus linguistics; Language teaching; Translation; Academic writing.

Resumo: O uso de *corpora* na tradução, no ensino de línguas e na escrita acadêmica aumentou significativamente nos últimos anos. O do-it-yourself (DIY) corpus é especialmente útil nessas áreas. Este artigo apresenta uma revisão teórica e discute o uso de DIY *corpora* na tradução, no ensino de línguas e na escrita acadêmica. Evidências mostram que, com o apoio de ferramentas computacionais, o DIY corpus fornece ao tradutor melhores opções de tradução e de levantamento de terminologia, assim como pode ser um recurso valioso para professores ao prepararem aulas para fins gerais e específicos, ou para pesquisadores que queiram observar a estrutura e os padrões de linguagem de suas respectivas áreas. Os resultados das diferentes propostas descritas mostram que os DIY corpora são recursos poderosos utilizados para atingir diferentes metas em contextos distintos, mas correlatos.

Palavras chaves: DIY corpora; Linguística de Corpus; Ensino de línguas; Tradução; Escrita acadêmica.

1. Introduction

The use of computerized corpora, which are collections of texts analysed by lexical tools for research and teaching, is already a well-established area of Applied Linguistics (TRIBBLE; JONES 1990; MAIA 1997; BOWKER; PEARSON 2002; BERBER SARDINHA 2004; REPPEN 2010; FRANKENBERG-GARCIA ET AL. 2011; VIANA; TAGNIN 2011). The interdisciplinary nature of Corpus Linguistics has turned out to be an approach widely used in different areas, such as Lexicography, Terminology, Translation Studies and Teaching (BAKER 1996; OLOHAN; BAKER 2000; TAGNIN; BEVILACQUA 2013; FRANKENBERG-GARCIA ET AL. 2019).

When English teachers use a corpus to develop teaching material they usually rely on previously compiled corpora, such as the Corpus of Contemporary American English (COCA), the Michigan Corpus of Upper-Level Student Papers (MICUSP) or the British National Corpus (BNC). In this paper we argue for the use of do-it-yourself (DIY) corpora, or recyclable corpora, as you wish, which are collections of texts that can be used to answer current questions or test hypotheses and, afterwards, be disposed of or enlarged with new texts.

One of the research papers that discuss the idea of using DIY corpora was presented by Lee and Swales (2006), in which graduate students of different areas used their own “personalized self-compiled corpora” (p. 61) to analyse parts of their academic texts. This study also served as a basis for Charles’ research (2012) in which the scholar discusses a methodology for enabling students and researchers to work on their own tailored material to answer specific research needs. This work will be discussed again in another section of this paper.

This article will discuss how DIY corpora can be used as a powerful tool to help (i) translators during the translation process and to extract terminology; (ii) English teachers when preparing their classes, and (iii) researchers who need to know more about the structure and language patterns when writing their own research papers. Therefore, this paper is divided into three main topics: 1. The use of DIY corpora in Translation; 2. DIY corpora in Language Teaching and 3. DIY corpora for Academic Writing, which will be presented as follows.

2. The use of DIY corpora in Translation

Translation practice has been changing over the years, especially because of the internet´s evolution and, consequently, the tools and information which are available online to facilitate the process of translating from language to another. Among a huge variety of resources such as TM – translation memories - and dictionaries – virtual or print -, the use of corpora has increased considerably and there are good reasons for it. Zanettin (2002) reminds us that a corpus can offer diverse translation strategies when the professional is faced with some problems in his/her work, especially because the corpus approach provides information that a dictionary does not usually contain. As Tagnin, Teixeira and Santos (2009: 3) state, “the majority of them do not provide any information on how the terms are actually used in real texts”. When provided with a specialized corpus the user can improve his/her efficiency in choosing a word or an expression that works better in specific context (GRANGER; TRIBBLE 1998).

According to Kübler and Aston (2010), the translation process involves three phases: documentation, drafting and revision, and corpora can be useful in all of them. Firstly, during the documentation phase, the authors state that most translators may spend a lot of time reading about some specific topics, while “corpora of documents from that domain can provide readily consultable collections of reading materials, and a basis for acquiring conceptual and terminological knowledge” (KÜBLER; ASTON 2010: 501).

Besides being an authentic source, corpora also provide data about the terminology of a certain domain. Therefore, if a translator is searching for a Dentistry term and does not find it in a dictionary, a corpus about this subject can be more helpful than searching on the internet, which may also be useful, but not always reliable.

In the second phase, the professional starts to formulate and evaluate possible translations based on a corpus, selecting terms and words he/she believes to be the most suitable to the text that is being translated. Varantola (2000) also highlights how helpful corpora can be as a resource when the translator is not confident about his/her choices.

In the third and last phase, Kübler and Aston (2010: 502) argue that corpora can reassure the "readability, comprehensibility, coherence grammaticality, terminological consistency etc." of a translated text. This way, when using a corpus or several corpora, the professional who will translate the text will have many tools at his/her disposal to use in order to adapt the text into a more appropriate way, by means of lexicon and language adequacy.

Moreover, another key point in using corpora when translating a text is the relevance of searching for information in authentic sources. In other words, the translator does not need to rely on his/her intuition or depend on consulting a specialist in the area. Tagnin (2002) defends the relevance of using corpora in translation, especially to find the correct term in authentic contexts of use, besides providing the translators with more natural and fluent data.

In this context, corpora

> can provide data which is not pre-digested but comes in the shape of samples of actual texts, allowing translators to acquire and apply skills which are, after all, central to their trade – ones of the text interpretation and evaluation (KÜBLER; ASTON 2010: 503).

It is important to mention that general reference corpora are also very helpful when translating a text, even if they are used in a broad perspective. General corpora were designed "to provide information about the text as a whole, showing how it is generally used in speech and writing of various kinds" (KÜBLER; ASTON 2010: 504). Therefore, when the translator's doubts are about the language as a whole (grammatical aspects, for instance) a general reference corpus may be very useful once it will show authentic patterns, as mentioned before.

When the general corpus is not so supportive, the DIY or disposable corpus can help as long as it is specialized and compiled for a specific task. As Kübler and Aston point out (2010: 505), "specialized corpora do not grow on trees", so, the researcher/translator needs to build their own corpus and this process is mentioned by the authors:

> [...] creating a corpus involves putting together a collection of relevant documents. Ideally these should already be in form, and this usually means locating and downloading them, converting them to a format which query software can handle, and cleaning them of unnecessary parts such as tables and images, HTML links and formatting instructions (KÜBLER; ASTON 2010: 507-508).

A specialized corpus which is compiled by the translator/researcher/teacher can also be called DIY corpus – do-It-yourself corpus – and it has already been studied and explored by many researchers (VARANTOLA 2000; MAIA 1997, 2002; ZANETTIN 2002).

According to Zanettin (2002: 242), a DIY web corpus has the following specific features:

- it is a collection of Internet documents, or more precisely of web pages in HTML.
- it is created ad hoc as a response to a specific text to be translated.
- it is an open corpus. More material can be added as the need arises.
- it is disposable (VARANTOLA, 2000) or virtual. It is not destined to be part of a more permanent corpus and can be disposed of as soon as the translation is completed. Copyright permissions are not required.
- like "parallel texts" it can be either bilingual comparable or target monolingual.

As we previously stated, the first characteristic Zanettin (2002) cites about a DIY corpus is the use of Internet documents or web pages. However, it is essential to explain that, as the author works with the World Wide Web as a resource to build a DIY corpus in his paper, he describes a DIY as a web corpus. Nevertheless, the translator does not necessarily have to compile texts from the web. Articles from the translators' computer (.doc, .pdf, .txt etc.) can be compiled as a specialized corpus. For instance, the professional who has to translate a text in the area of computer science will choose reliable texts from this domain and compile a corpus.

This view is supported by Kübler (2011) who writes that

> In understanding the source text, not only are language problems at stake, but also cultural ones. In texts that deal with specialized subjects it is necessary to get acquainted with the domain. An expert in the domain can help the translator understand it, but it is not always possible to have an expert at hand, and here corpora can play an important role (KÜBLER 2011: 14).

In this way, this professional will have a large range of texts that, together, may help him/her to find better options for his/her translation, including, especially, the terminology about the subject, as if there was an expert in the specific domain helping him/her to translate the text.

Zanettin (2002) states that it is common for translators who use corpora to use the bottom-up approach when translating a text, which means that they go from words to texts. This approach can be complemented by the top-down approach, that is, from texts to words. In other words, the professional does not have a pre-existing corpus to be examined, and "by compiling their DIY corpus prior, during or even after the translation task [...] students (and translators) can get a first acquaintance with texts and take full advantage of web pages prior to word prompted analysis" (ZANETTIN 2002: 246)

Another key point in translating a text, in the perspective of the professional who makes a living with this practice, is the time spent in the process of translating. According to Varantola (2002: 173), "it has been estimated that translators may spend up to 50% of their total time performing a particular translation task when trying to find relevant lexical information". One question that researchers may have is: "Is compiling a corpus time-consuming?". It will certainly depend on some aspects such as size, specialty and tools used to do so. As far as size is concerned, Koester (2010) holds the view that small corpora:

> allow a much closer link between the corpus and the contexts in which the texts in the corpus were produced. [...] With a small corpus, the corpus compiler is often also the analyst, and therefore usually has a high degree of familiarity with the context." (KOESTER 2010: 66-67)

Thus, if the professional will translate a text about a specific domain, as it has been discussed here, it is not necessary to have a large amount of text. If the translator or researcher selects highly specialized and reliable texts in the area he is working on, small corpora will be enough to generate specific terminology in that area. It always depends on the objective of the translator (BERBER SARDINHA 2004). Furthermore, it is important to say that if the professional always translates texts about that domain, a DIY corpus may be such a relevant tool, since the translator can add texts from the corpus whenever he/she considers necessary, making the corpus gradually larger. Varantola (2002) argues that

> disposable do-it-yourself corpora are highly adaptable because the user decides what is included in them. They are also adaptable in the sense that the user defines what type of information he or she wants to extract from the corpus (VARANTOLA 2002: 185).

Moreover, this process of developing a DIY corpus is becoming much easier because of modern technology (Maia 1997). In 2005, the European project MeLLANGE[1], whose aim was to prepare better translators for a globalized market, carried out a survey which compiled 623 questionnaires answered by UK, French, Italian and German professional translators. The survey showed that 41.9% of the interviewees had never heard of corpora before. Some years later, Kübler (2011: 2) stated that "professional translators are still not very keen on using corpora for translation". It means that with so many options of corpora and even with the option of creating their own corpus, many translators do not know how to use this source of data, its advantages, or even the existence of it. Furthermore, Kübler (2011: 04) argues that "not all training syllabuses include corpus use as a skill to be taught". In this way, the trainees are not aware of the advantages of a corpus, and in some cases, the trainer does not as well. In the same vein, Varantola (2002) notes that:

> translator trainees need to be taught to understand the various uses of corpora, how to compile them and apply them in an intelligent way. Translator training should thus include courses in the compilation and use of corpus information (VARANTOLA 2002: 184).

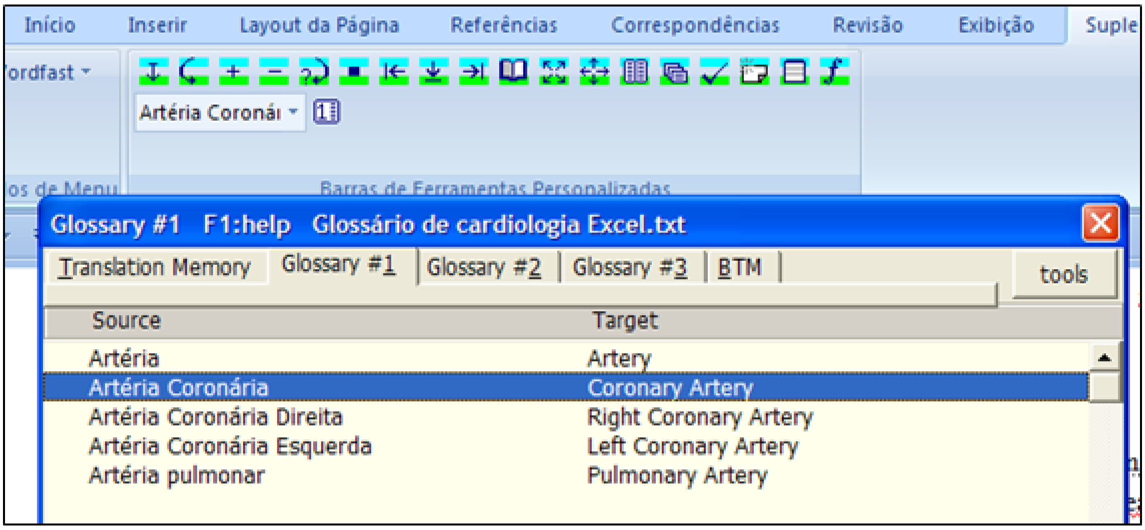
Therefore, it is very important to find ways to introduce the use of corpora to show future professionals how useful this tool can be as a general and specialized reference, and, at the same time, teach them how to create their own corpora. This action will enable them to become more autonomous in their practice by discovering the variety of possibilities for using a corpus when translating a text.

This possibility was presented by Paiva (2009). The researcher described a methodology for the compilation of a corpus-based bilingual medical glossary that was inserted in a translation memory tool to be shared with professional translators. The medical-recyclable glossary resulted from a corpus of research papers, with approximately 800,000 words, which was shared and tested by the translators of a professional office, through the translation memory tool Wordfast (CHAMPOLLION 1999). The corpus-based glossary automatically displays the terms

[1] Multilingual eLearning in LANGuage Engineering

inserted in the tool as the translators are working with their texts. We can see how the tool displays the glossaries in the following figure:

Figure 1: Translation Memory Tool Wordfast with a corpus-based medical glossary



Source	Target
Artéria	Artery
Artéria Coronária	Coronary Artery
Artéria Coronária Direita	Right Coronary Artery
Artéria Coronária Esquerda	Left Coronary Artery
Artéria pulmonar	Pulmonary Artery

Source: (Paiva 2009)

This study shows how DIY corpora can be incorporated as a reliable source for organizing shareable glossaries that can constantly be updated and enlarged. Recyclable corpora can be used not only as an aid for translating, but also as teaching tools, as we will discuss in the next section.

3. DIY corpora in Language Teaching

This section will discuss how resourceful a DIY corpus can be for teaching and providing information about learners' production or supplying information about how a specific term is expressed in a specialized field (Millar; Lehtinen 2008). Pinto (2018) highlights the importance of teaching academic English to researchers since it is the language most scientists use nowadays. She highlights corpora can also be a helpful source for preparing teaching material for specific fields.

There are different corpora available online, such as Ensinador (Simões; Santos 2011), COCA -The corpus of contemporary American English - (Davies 2008), CorTrad (Tagnin 2009) and Corpógrafo (Sarmento; Maia; Santos 2004; Maia 2008). Even though some of them do not allow you to create your own corpora, they are valuable tools for teachers. Simões and Santos (2011), for example, propose grammar exercises based on corpora for teaching Portuguese as a foreign language through the tool Ensinador. This tool provides teachers with the creation of pedagogical activities. The activities are also automatically corrected by Ensinador (SIMÕES; SANTOS 2011). The authors argue that since it is an open source code, it can be downloaded and used by any teacher, who can also add her/his own corpora.

Corpógrafo (SARMENTO; MAIA; SANTOS 2004; MAIA 2008) is a free online corpora analysis tool that allows you to create your own corpus. In the platform it is possible to upload different kinds of files, such as PDF; HTML, MS-Word and Ps. This makes the platform more friendly to new users. When creating his/her own DIY corpus using Corpógrafo the teacher will be able to select any of the features to help him/her achieve his/her goals.

Bowker and Pearson (2002) point out that despite not representing language in general, a special purpose corpus can contain a great diversity of terms and expression patterns accepted by researchers from each field. They also argue for the importance of building a specialized corpus with different authors so one can analyze the term distribution inside the corpus and avoid misinterpretation of a term that may have a high absolute frequency, but is used only by a specific author. According to Bowker and Pearson (2002), a specialized corpus used as a resource to teach language patterns should have at least 25,000 words (around 20 different articles from a variety of authors in the chosen field). Another criterion is to avoid extracting parts of an article. When creating a DIY corpus it is interesting to avoid including texts from different subjects. The flexibility of this kind of corpus allows its owner to keep it updated, adding or excluding articles as necessary.

Another important aspect of using DIY corpora as a teaching resource is involving the students in their exploration. It is understood that the learners play an important role in their own learning process. Maia (2002) argues that in order to avoid students' frustration it is important to engage them in the process of creating their own corpora. The author suggests that the main goal should be learning more about the subject being researched. This approach helps learners to understand how relevant the texts are when compiling their own corpora.

Charles (2012) reports that it is relatively easy to adopt a DIY corpus in an English for academic purposes course. The author taught a multi-disciplinary class to advanced-level non-native speakers of English who were enrolled in doctoral and master courses. The students used the tool AntConc (Anthony 2004) to analyze and build their corpus. At the beginning, the professor provided the learners with two corpora that enabled them to learn how to use AntConc. Next, they were taught how to build their own corpora.

According to the author, there were advantages and disadvantages for using DIY corpora. One of the pitfalls appointed by her students was the size and the dirtiness of their corpus. Another challenging aspect was the reliability of the English used in some of the articles, even though they had been chosen from journals of high standards. She argues that some of those journals publish articles by the merit of the research and not necessarily for their language proficiency.

On the other hand, the author also highlights the positive aspects of using DIY corpus, such as the students' sense of accomplishment as they recognized language patterns in their corpora. The learners were able to find specific terms in their field that could have been difficult to find in a general corpus. They reported that it was not difficult to understand the results obtained from the corpus analyses since they were familiar with the context. Based on these data, the author stated that students responded well to the DIY corpus procedures.

Almeida and Prado (2011) developed a syllabus for a course of English targeting the reading of aircraft maintenance manuals. An airline company was having problems with their maintenance employees because the manuals were in English, but the staff did not have the necessary English skills to read them. The authors' challenge was to teach those employees reading abilities in their specific context. To do so they built a DIY corpus based on the manuals available on the internet which are used by the company staff, and based on its content they developed the course syllabus. The authors reported a successful result, since the staff's needs were met.

In addition to teaching language patterns using high quality articles compiled as a DIY corpus, it is possible to identify and analyze students' writing production by building DIY corpora with their texts (Millar; Lehtinen 2008). Maia (2002) states that by exploring texts students are able to not only search for suitable words or expressions, but also learn about the subject.

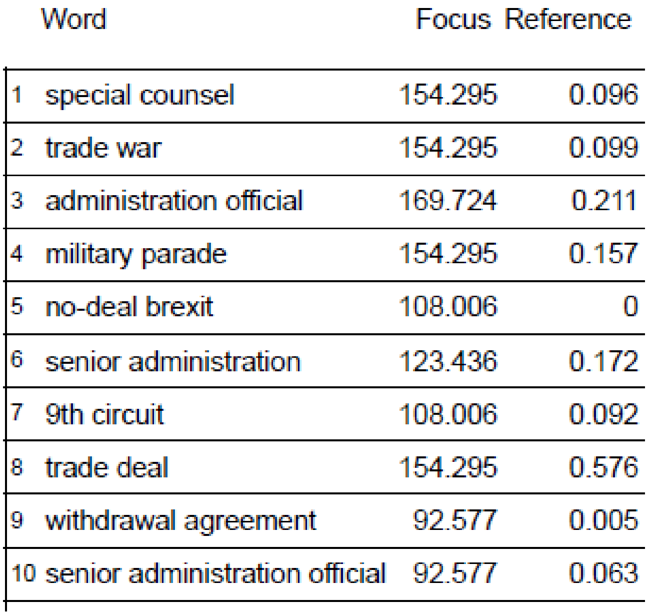
The authors describe some possible analyses of students' texts using two tools available in a corpus analysis software: Wordlist and Concordance. According to the authors, it is possible to identify students' writing patterns, such as how they start and end their sentences, the overuse or underuse of some words and expressions, and their tendency to transfer their mother language structure to the target language, which can cause confusion and an unnatural use of the target language. Therefore, it is possible to redirect the lessons and address those problems with the students by creating tailored classes aiming to fulfill the students' specific needs (MILLAR; LEHTINEN 2008).

A recent study has been carried out by Nazzi-Laranja and Pinto (2019) in which DIY corpora are used to help teachers select updated vocabulary from the news to develop reading comprehension exercises for students who are being prepared for college entrance examinations in Brazil (the "Vestibular"). A recyclable corpus makes it possible for teachers to find out the most recent word combinations to produce teaching material. Teaching those groups of words to students can help them get a higher score in the English test for the university admission. Teachers can easily incorporate new articles and exclude old ones from their corpus. This way they will make sure they are teaching the most relevant vocabulary in a chosen context.

The researchers compiled a 55.861-word corpus of World News on "Politics" using texts from the CNN and BBC websites. Three texts were compiled weekly during a five-month period (August to December) of 2018. In total, 66 texts about political news were compiled, and the average length of texts was 500 words. Thus, the researchers worked with texts of the journalistic genre, within a specific theme, in order to select clusters and multi-word keywords using the Sketch Engine tool (KILGARRIFF 2014).

As a result, we observed that one of the most frequent clusters was "military parade", as shown in Figure 2, below. Based on the statistical relevance of the combination it was possible to elaborate some activities using the two words that are important in the political context presented in the DIY corpus. It is key to draw attention to the context and historic moment in which this corpus was created, which is a representation of that period of time.

Figure 2: Corpus Politics

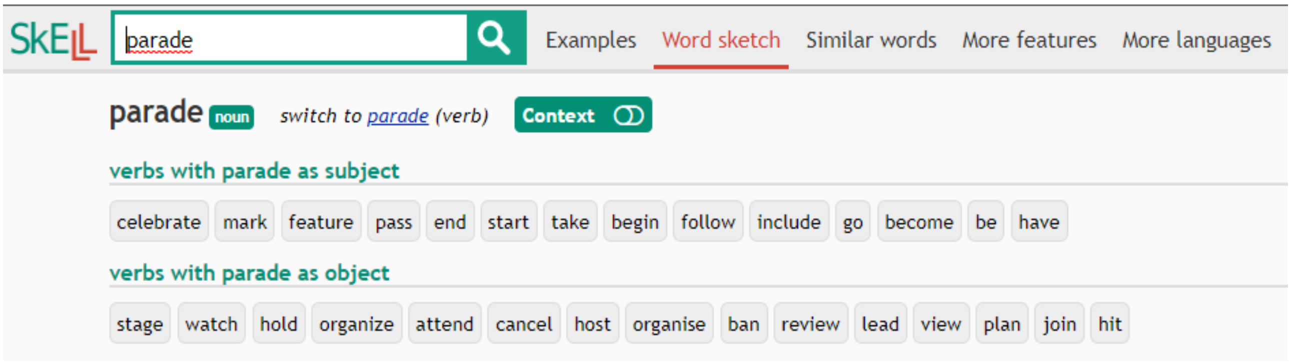


	Word	Focus	Reference
1	special counsel	154.295	0.096
2	trade war	154.295	0.099
3	administration official	169.724	0.211
4	military parade	154.295	0.157
5	no-deal brexit	108.006	0
6	senior administration	123.436	0.172
7	9th circuit	108.006	0.092
8	trade deal	154.295	0.576
9	withdrawal agreement	92.577	0.005
10	senior administration official	92.577	0.063

Source: (NAZZI-LARANJA; PINTO 2019)

After choosing that word combination, the researchers looked for combinations with the word "parade" using the tool SkELL, which stands for Sketch Engine for Language Learning. It showed that the five most frequent verbs used with parade were "stage", "watch", "held", "organize" and "attend", as we see in Figure 3:

Figure 3: Verbs with parade



parade noun switch to parade (verb)

verbs with parade as subject

celebrate, mark, feature, pass, end, start, take, begin, follow, include, go, become, be, have

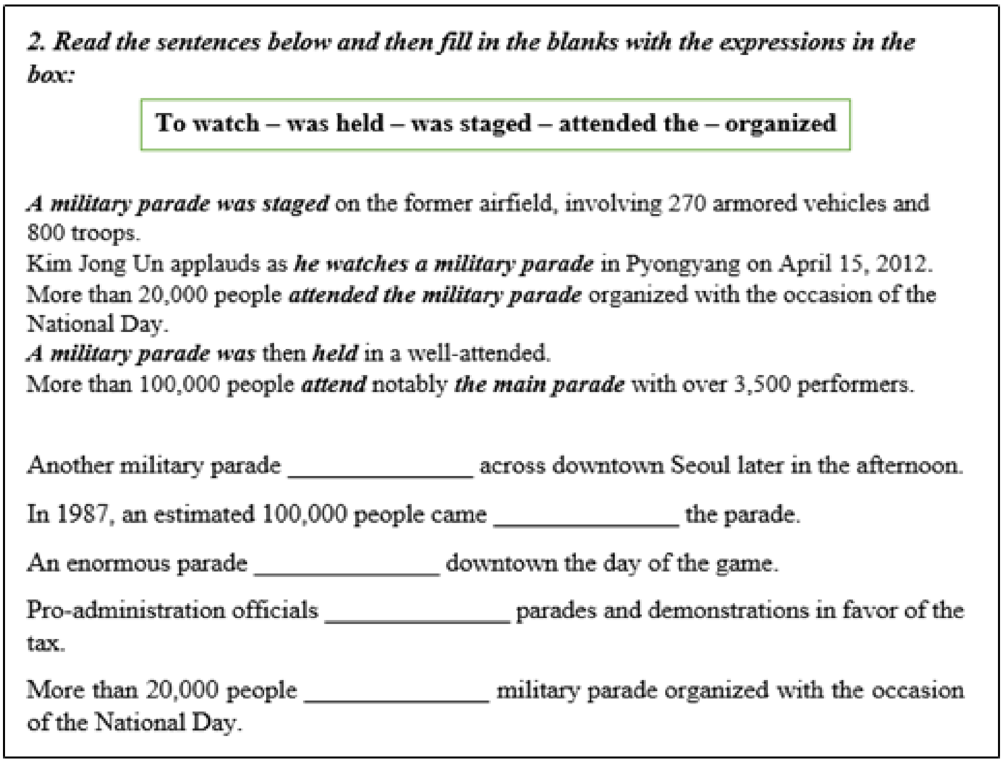
verbs with parade as object

stage, watch, hold, organize, attend, cancel, host, organise, ban, review, lead, view, plan, join, hit

Source: (KILGARRIFF 2014)

Based on this information we developed the exercise below. It is important to mention that activities like the one presented in Figure 4 are important in the context of entrance exams, since students need to identify which vocabulary will complete certain sentences, and even recognize synonyms. Filling the gaps activities are very common in Brazilian entrance exams, and the sentences will usually show themes that are being discussed on the news, as in the exercise shown in figure 4:

Figure 4: Exercise elaborated with the cluster *military parade*



2. Read the sentences below and then fill in the blanks with the expressions in the box:

To watch – was held – was staged – attended the – organized

A military parade was staged on the former airfield, involving 270 armored vehicles and 800 troops.
Kim Jong Un applauds as ***he watches a military parade*** in Pyongyang on April 15, 2012.
More than 20,000 people ***attended the military parade*** organized with the occasion of the National Day.
A military parade was then ***held*** in a well-attended.
More than 100,000 people ***attend*** notably ***the main parade*** with over 3,500 performers.

Another military parade _____________ across downtown Seoul later in the afternoon.

In 1987, an estimated 100,000 people came _____________ the parade.

An enormous parade _____________ downtown the day of the game.

Pro-administration officials _____________ parades and demonstrations in favor of the tax.

More than 20,000 people _____________ military parade organized with the occasion of the National Day.

Source: (NAZZI-LARANJA; PINTO 2019)

The previous exercise was created for reading comprehension purposes. If an English teacher is teaching graduate students to apply for a language exam in the area of Pharmacy, a corpus about politics will not be useful. In this case, the teacher should compile a specific corpus with texts about Pharmacy to find language features of that specific subject.

The relevance of DIY corpora can be expanded as the needs present themselves. While trying to identify some drawbacks, the researchers previously cited have concluded that the positive aspects of using DIY corpora are more relevant than the pitfalls. DIY corpora, therefore, can be a resourceful tool to help teachers in preparing teaching material and students to achieve their goals and specific needs.

4. DIY corpora for Academic Writing

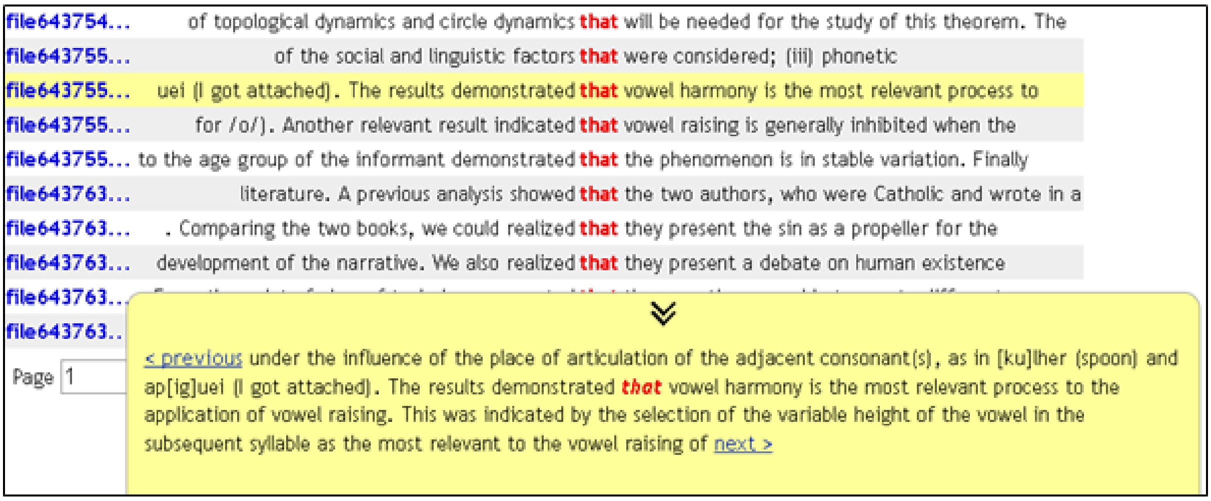
According to Elsevier's World of Research (2015), although Brazil already has a sizable research output, there is still low performance along measures of research impact (citations, highly cited articles). After the support for internationalization of Brazilian universities by national programs, such as Science Without Borders (Aveiro 2014) and Languages Without Borders (Abreu-E-Lima et al. 2016), we saw a growing number of studies concerning university students' writing (Dutra et al. 2014; Dutra et al. 2015; Silva et al. 2018). Although there are some studies that report the use of academic vocabulary in the writing of Brazilian experienced authors, (Aluísio 1995; Dayrell 2007; Paiva 2009; Silva et al. 2017) others are needed.

In order to illustrate recent work on how recyclable corpora can be used in research, we present two studies whose aims were to analyze graduate's academic writing and help scholars to improve their own research papers. The first case was reported by Pinto (2018) in which a class of English for Academic Purposes was being taught to graduate students of Humanities, Biology and Math. There were 23 (twenty-three) students who received instruction on how to compile small corpora of abstracts in their own areas so that, in class, they would point out the similarities and differences of parts of the abstracts, or their moves (cf. Swales; Feak 2009).

Students examined how the area of Literature, for example, did not stress their methods and materials in comparison to abstracts of Biology. Many times, they emphasize the description of the author being analyzed and the theoretical background. At the same time, students could see the similarities between texts within their own corpora, such as specific terminology or recurrent clusters used in their field that could be used as examples when writing their own abstracts.

Following the observation of students' DIY corpora, the teacher asked the students to write their own abstracts and send them to her, so she could compile a small learner corpus of 14,122 tokens to analyze the marked that-clause (Alves; Tavares Pinto 2018) in the students' writing. According to Swales and Feak (2009), "that-clause" is frequently used to point out the new discoveries and results of research. Below we can see the concordance lines with "that" as search word in the research of Alves (2018) with the same corpus:

Figure 5: Concordance lines with "that" as search work (ALVES 2018)



```
file643754...        of topological dynamics and circle dynamics that will be needed for the study of this theorem. The
file643755...                   of the social and linguistic factors that were considered; (iii) phonetic
file643755...     uei (I got attached). The results demonstrated that vowel harmony is the most relevant process to
file643755...       for /o/). Another relevant result indicated that vowel raising is generally inhibited when the
file643755... to the age group of the informant demonstrated that the phenomenon is in stable variation. Finally
file643763...               literature. A previous analysis showed that the two authors, who were Catholic and wrote in a
file643763...     . Comparing the two books, we could realized that they present the sin as a propeller for the
file643763...    development of the narrative. We also realized that they present a debate on human existence
file643763...
file643763..
Page 1
```

< previous under the influence of the place of articulation of the adjacent consonant(s), as in [ku]lher (spoon) and ap[ig]uei (I got attached). The results demonstrated *that* vowel harmony is the most relevant process to the application of vowel raising. This was indicated by the selection of the variable height of the vowel in the subsequent syllable as the most relevant to the vowel raising of next >

Source: (ALVES 2018)

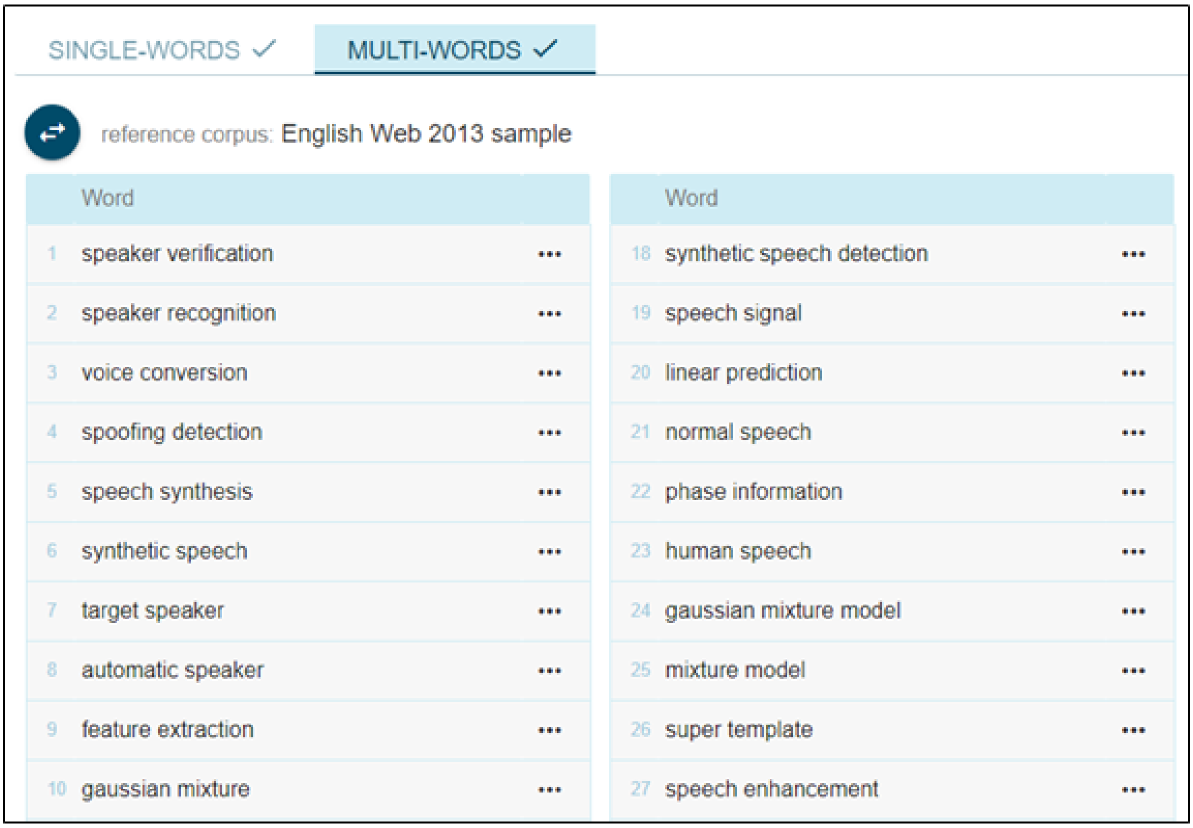
The study showed that EAP students have the tendency to use subject+verb in the simple present+that in their conclusion, as we see in the third line of figure 5.

A more recent experience of using DIY corpora for writing research papers was carried out during the Academic Masterclasses Supporting the Internationalization of Brazilian Research (FRANKENBERG-GARCIA ET AL. 2019) taught at the Federal University of Rio Grande do Sul (UFRGS) and at the São Paulo State University (UNESP). During these classes, junior and senior researchers were introduced to techniques of Corpus Linguistics and were able to compile their own corpora with research papers from high impact journals from their areas. In this course, they learned how to use the Sketch Engine (KILGARRIFF 2014) to analyze vast amounts of lexicon and see how words were used in specific contexts. In total, there were 53 English teachers who worked together with 72 researchers, so that both professionals could help each other in observing how language and terminology had been used by the authors of the papers in their DIY corpora. There were specialists of Engineering, Agricultural Sciences, Humanities, Social Sciences and Health among others.

Results showed that although scholars were familiar with the terminology of their own areas, the tool pointed out other possibilities of word combinations they had difficulty with, such as verbal collocations and the most common patterns of academic English if compared to Portuguese. At the same time, the English teachers who were participating in the workshops were inspired by the terminology and language to develop teaching activities for their own EAP students.

In figure 6 we see the terminology from a DIY corpus of Speech Recognition which was compiled by a researcher who participated in the Academic workshops. The English teacher who was working in partnership with the researcher was able to prepare an activity of EAP specifically for students of Phonology with terms such as “speaker verification”, “speaker recognition”, “synthetic speech detection”, “speech enhancement”, among others, which we can see in the next figure:

Figure 6: terminology from a DIY corpus of Speech Recognition



reference corpus: English Web 2013 sample

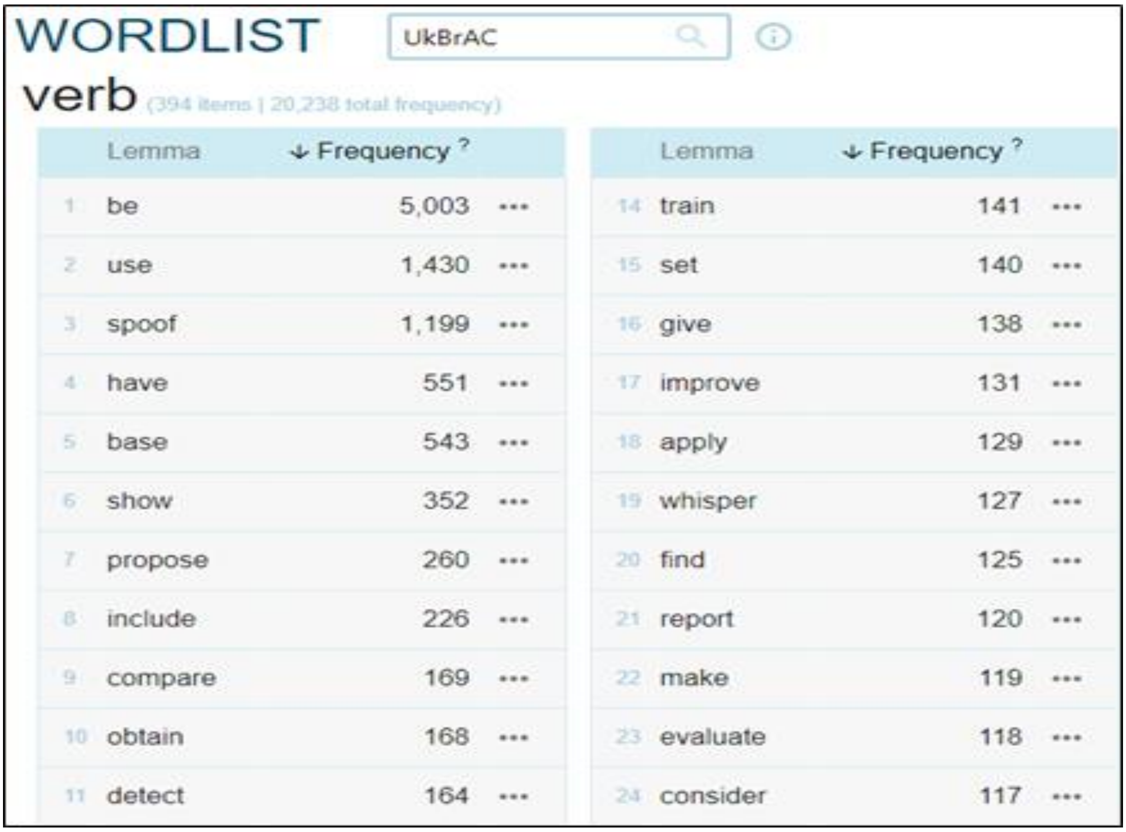
	Word
1	speaker verification
2	speaker recognition
3	voice conversion
4	spoofing detection
5	speech synthesis
6	synthetic speech
7	target speaker
8	automatic speaker
9	feature extraction
10	gaussian mixture

	Word
18	synthetic speech detection
19	speech signal
20	linear prediction
21	normal speech
22	phase information
23	human speech
24	gaussian mixture model
25	mixture model
26	super template
27	speech enhancement

Source: (FRANKENBERG-GARCIA *ET AL*. 2019)

At the same time, the researcher from the area of Speech Recognition was able to observe a broad list of verbs in her DIY corpus. This result was relevant since she became aware of verbs commonly used by that research community. Some of these verbs are shown in the Wordlist in figure 7:

Figure 7: verbs commonly used by international researchers in the area of Speech Recognition



WORDLIST UkBrAC

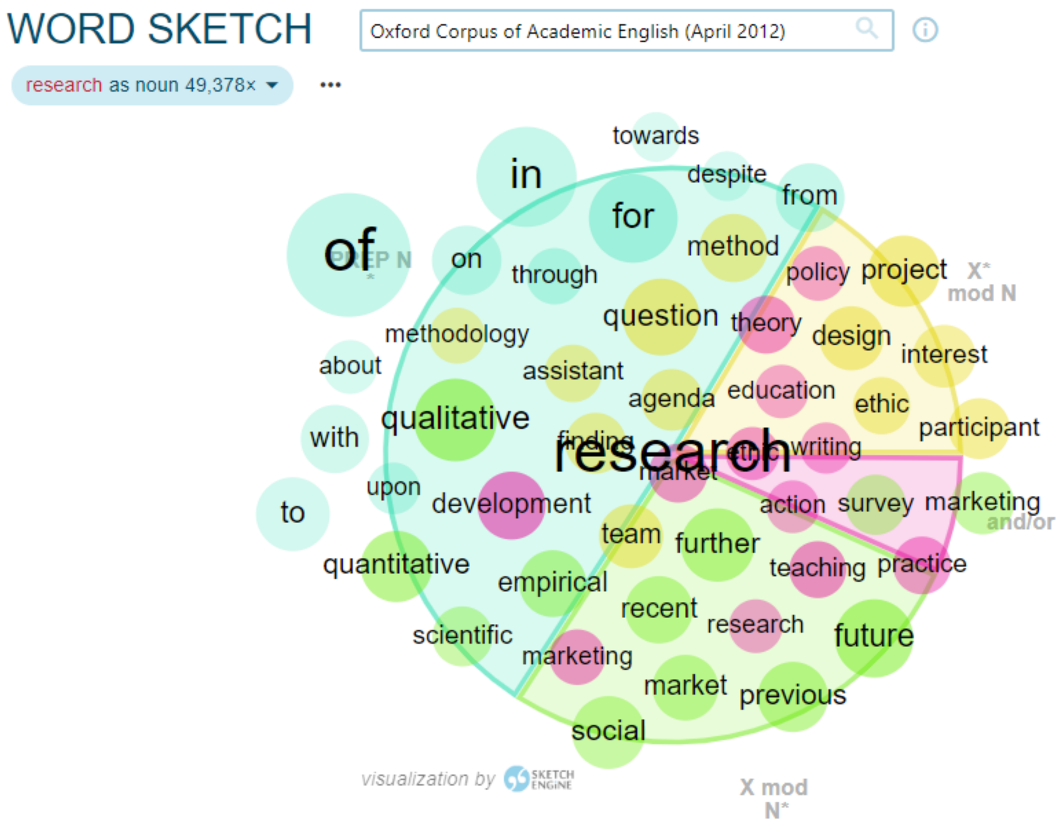
verb (394 items | 20,238 total frequency)

	Lemma	↓ Frequency ?		Lemma	↓ Frequency ?
1	be	5,003	14	train	141
2	use	1,430	15	set	140
3	spoof	1,199	16	give	138
4	have	551	17	improve	131
5	base	543	18	apply	129
6	show	352	19	whisper	127
7	propose	260	20	find	125
8	include	226	21	report	120
9	compare	169	22	make	119
10	obtain	168	23	evaluate	118
11	detect	164	24	consider	117

Source: (FRANKENBERG-GARCIA ET AL. 2019)

Other features pointed out by the tool to the researchers were the lists of adjectives international authors constantly use, such as “state-of-the-art”, “genuine”, “spectral”, which they could use in their own writing to have a more diversified range of vocabulary. At the same time, with access to a corpus of academic English, researchers and EAP teachers who learn more easily with visual aids could see different word combinations using the tool word sketch, as shown in figure 8, where “research” is the search word:

Figure 8: Word sketch of “research” from the Oxford Corpus of Academic English



Source: (FRANKENBERG-GARCIA *ET AL*. 2019)

This tool shows the student different possibilities of combinations with a search word. In this case, the writer can see the frequent prepositions used with the search word, such as “through research”, “with research”; the search word modifying other nouns, such as “research design”, “research education”, “research ethic”, and other words modifying the search word as in “further research”, “marketing research”, “team research”, etc. We believe there is a greater possibility of visual learners to improve their vocabulary repertoire using this tool.

After six months since the first Academic masterclasses, both researchers and English teachers have stated that they are still using their own DIY corpora and Sketch Engine for writing their research papers and preparing their classes.

5. Final Considerations

The theoretical review and discussions presented in this paper suggest that the use of DIY corpora is not only useful, but also a more empirical and reliable practice for translating, teaching and researching. In the area of language teaching, there is an infinite range of possibilities for using a corpus to analyse, describe, and prepare teaching material for English as a foreign language classes through a corpus. We add the fact that the use of authentic language offers real context of communication.

Translators will find specific terminology in a DIY corpus if compiled with specialized texts from the area they are translating to. Although corpus linguistics may not be known by some professional translators, translation training courses have gradually been showing students how to use computational tools to quickly compile specialized corpora in different areas.

At the same time, researchers can use DIY corpora to support their own research writing in a way that it puts aside any speculative characteristics since the information found in the corpus will be more accurate and specifically related to the theme being researched.

Finally, we can state that DIY corpus is an ally to different professionals who become more autonomous when compiling corpora that will meet their specific needs and goals, however, further research focusing on the advantages and disadvantages of the use of this kind of corpus is still necessary to be carried out.

References

ABREU-E-LIMA, D. M.; MORAES-FILHO, W. B.; SARMENTO, S. O programa Idiomas sem Fronteiras. *Do Inglês sem Fronteiras ao Idiomas sem Fronteiras*: a construção de uma política linguística para a internacionalização, 2016: 293-308.

ALMEIDA, D.; Prado, M. Desenvolvendo o conteúdo programático de um curso de inglês para mecânicos de aeronaves com base em um corpus DIY: um estudo de caso. *Aviation in Focus-Journal of Aeronautical Sciences*, v. 2, n. 2, p. 6-20, 2011.

ALUÍSIO, S. M. *Ferramentas de Auxílio à Escrita de Artigos Científicos em Inglês como Língua Estrangeira*, Tese de doutoramento, Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 1995.

ALVES, A L.; TAVARES PINTO, P. A utilização de that-clauses em abstracts escritos por alunos-pesquisadores brasileiros. *Revista Entrepalavras*, v. 8, p. 288-303, 2018.

ALVES, A. L. L. *As That-clauses em abstracts escritos por alunos brasileiros de Universidades Públicas*: uma análise baseada em corpus. 2018. Dissertação (Mestrado em Mestrado em Estudos Linguísticos) - Universidade Estadual Paulista "Júlio de Mesquita Filho" - IBILCE, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, 2018.

ANTHONY, L. *AntConc*: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL*, p. 7-13, 2004.

AVEIRO, T. M. M. O programa Ciência sem Fronteiras como ferramenta de acesso à mobilidade internacional. *Revista de Educação Ciência e Tecnologia*. Rio Grande do Sul: Instituto Federal, v. 3, n. 2, p. 1-21, 2014.

BAKER, M. *Corpus-based translation studies*: The challenges that lie ahead. Benjamins Translation Library, v. 18, p. 175-186, 1996.

BERBER SARDINHA, T. *Linguística de Corpus*. Barueri, SP: Editora Manole, 2004.

BOWKER, L.; PEARSON, J. *Working with specialized language*: a practical guide to using corpora. Routledge: New York, 2002.

CHAMPOLLION, Y. *Wordfast*. 1999. Disponível em: <http://www.wordfast.net/index.php> Acesso em 23 de janeiro de 2020.

CHARLES, M. 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, v. 31, n. 2, p. 93-102, 2012.

DAVIES, M. *The corpus of contemporary American English* (COCA): 520 million words, 1990-present. 2008.

DAYRELL, C. A quantitative approach to compare collocational patterns in translated and non-translated texts. *International Journal of Corpus Linguistics,* v. 12, n. (3), 375-414, 2007.

DUTRA, D. P.; GOMIDE, A. R. Compilation of a university learner corpus. *BELT-Brazilian English Language Teaching Journal*, v. 6, n. p.1, p. 21-33, 2015.

DUTRA, D. P.; ORFANO, B.; BERBER SARDINHA, T. Stance bundles in Learner Corpora. In: ALUISIO, S. M.; TAGNIN, S. E. O. (eds.) *New Language Technologies and Linguistic Research*: A Two-Way Road. Newcastle upon Tyne: Cambridge Scholars Publishing, 2014, p. 2-17.

ELSEVIER WORLD OF RESEARCH AT <https://www.elsevier.com/research-intelligence/campaigns/world-of-research-2015> access in June, 2020.

FRANKENBERG-GARCIA, A.; BOCORNY, A.E.P.; TAVARES-PINTO, P.; SARMENTO, S. (2019). *Supporting the Internationalization of Brazilian Research*. Workshops delivered at the Federal University of Rio Grande do Sul and at São Paulo State University, Porto Alegre and São José do Rio Preto, April-June, 2019.

FRANKENBERG-GARCIA, A.; FLOWERDEW, L.; ASTON, G. (edsED.). *New trends in corpora and language learning*. A&C Black, 2011.

GRANGER, S.; TRIBBLE, C. "Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning". In: Sylviane Granger, S. (ed.), *Learner English on computer*. New York: Longman, 1998: 199-209.

KILGARRIFF, A. ET AL. The Sketch Engine: ten years on. *Lexicography*, v. 1, n. 1, p. 7-36, 2014.

KOESTER, A. Building small specialised corpora. In: O'Keeffe, A.; MccCarthy, M. *The Routledge Handbook of Corpus Linguistics*. Routledge: New York, 2010.

KÜBLER, N. Working with different corpora in translation teaching. In: Ana Frankenberg-Garcia, A; Lynne Flowerdew, L; and Guy AAston, G. *New Trends in Corpora and Language Learning, Continuum*, pp.62-80, 2011: 62-80,, ffhttp://www.bloomsbury.com/uk/ff. ffhal-01134954f

KÜBLER, N.; ASTON, G. Using corpora in translation. In: O'Keeffe, A; McCarthy, M. *The Routledge handbook of corpus linguistics*, Routledge: New York, 2010.

LEE, D.; SWALES, J. A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. In: *English for specific purposes*, v. 25, n. 1, p. 56-75, 2006.

MAIA, B. CORPÓGRAFO V: Tools for educating translators. In: 4. Rodrigo, E.Y. *Topics in Language Resources for Translation and Localisation*, v. 79, p. 57, 2008.

MAIA, B. Do-it-yourself corpora... with a little bit of help from your friends! *PALC'97: Practical Applications in Language Corpora*, 1997, p. 403-410. Available at: https://hdl.handle.net/10216/23471

MAIA, B. Do-it-yourself, disposable, specialised mini corpora-where next? Reflections on teaching translation and terminology through corpora. *Cadernos de tradução,* v. 1, n. 9, 2002, p. 221-235, 2002.

MILLAR, N.; LEHTINEN, B. DIY local learner corpora: Bridging gaps between theory and practice. *JALT CALL Journal*, v. 4, n. 2, p. 61-72, 2008.

NAZZI-LARANJA, L.A.; PINTO, P.T. O desenvolvimento de atividades de compreensão escrita: um corpus de temática política. *LinguaTech*, v.4, n. 2, p. 29-51, nov. 2019.

OLOHAN, M.; BAKER, M. Reporting that in translated English. Evidence for subconscious processes of explicitation?. *Across languages and cultures*, v. 1, n. 2, p. 141-158, 2000.

PAIVA, P. T. P. *Uma investigação de traduções de textos da área médica sob a luz dos estudos da tradução baseados em corpus*. Ph.D thesis. São Paulo State University, Brazil, 2009.

PINTO, P.T. Um curso de Inglês com Fins Acadêmicos baseado em Corpus para alunos universitários das áreas de Humanas, Exatas e Biológicas. In: FERREIRA, M. M.; STELLA, V.C.R. (Org.). *Redação Acadêmica: múltiplos olhares para a produção textual e o seu ensino*. 01 ed. São Paulo: Humanitas, 2018, v. 01, p. 122-136.

REPPEN, R. *Using Corpora in the Classroom*. Cambridge: Cambridge University Press, 2010.

SARMENTO, L.; MAIA, B.; SANTOS, D. The Corpógrafo-a Web-based environment for corpora research. In: quot; In Maria Teresa Lino; Maria Francisca Xavier; Fátima Ferreira; Rute Costa; Raquel Silva (ed) *Proceedings of the 4 th International Conference on Language Resources and Evaluation* (LREC'2004) (Lisboa, Portugal 26-28 May 2004). 2004.

SILVA, E. B.; BABINI, M.; OTTAIANO, A. O. Identification of the most common phraseological units in the English language in academic texts: contributions coming from corpora. *Acta Scientiarum (UEM)*, v. 39, p. 345-353, 2017.

SILVA, L. G.; MATTE, M. L.; SARMENTO, S. Brazilian students’´s use of English academic vocabulary: an exploratory study. In: Finatto, M. J.; Bocorny, Rebechi, R.

R.,; Sarmento, S.,; Bocorny, A. E. P. *Linguística de corpus*: perspectivas. Org:— Porto Alegre: Instituto de Letras - UFRGS, 2018.

SIMÕES, A.; SANTOS, D. Ensinador: corpus-based Portuguese grammar exercises. *Procesamiento del Lenguaje Natural 47*, septiembre de 2011, pp. 301-309.

SWALES, J. M.; FEAK, C. B. *Abstracts and the writing of abstracts*. Michigan: University of Michigan Press, 2009.

TAGNIN, S. E. O.; BEVILACQUA, C. (ED.). *Corpora na terminologia*. HUB Editorial, 2013.

TAGNIN, S.E.O. Os Corpora: instrumentos de auto-ajuda para o tradutor. *Cadernos de Tradução*. n.9, v.1, Florianópolis, 2002, p.191-219.

TAGNIN, S.E.O. TEIXEIRA, E.D.; SANTOS, D. CorTrad: a multiversion translation corpus for the Portuguese-English pair. In: quot; Arena Romanistica 4; 4 (2009; 2009) [The 28 th International Conference on lexis and grammar. 2009.

TRIBBLE, C.; JONES, G. *Concordances in the classroom*: A resource book for teachers. Longman/: London, 1990.

VARANTOLA, K. Disposable corpora as intelligent tools in translation. *Cadernos de Tradução*, Florianópolis, v. 1, n. 9, 171-189, 2002/1.

VARANTOLA, K. Translators, Dictionaries and Text Corpora. In: S. Bernardini, S.; and F. Zanettin, F. (eds) I *Corpora nella Didattica della Traduzione*. Bologna: CLUEB, 2000: 117–133.

VIANA, V.; TAGNIN, S. E. O. *Corpora no ensino de línguas estrangeiras*. Hub Editorial, 2011.

ZANETTIN, F. DIY Corpora: The WWW and the Translator. In: Maia, Belinda / Haller, Jonathan / Urlrych, Margherita (eds.) *Training the Language Services Provider for the New Millennium*, Porto: Facultade de Letras, Universidade do Porto, 2002: 239-248.

Recebido em: 30/04/2020

Aceito em: 25/09/2020

Publicado em janeiro de 2021