Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1983-4675
ISSN on-line: 1983-4683
Doi: 10.4025/actascilangcult.v39i4.31811

Identification of the most common phraseological units in the English language in academic texts: contributions coming from corpora

Eduardo Batista da Silva[1*], Adriane Orenha-Ottaiano[2] and Maurizio Babini[2]

*[1]Universidade Estadual de Goiás, Rua Quatorze, 327, 75650-000, Morrinhos, Goiás, Brazil. [2]Universidade Estadual Paulista, São José do Rio Preto, São Paulo, Brazil. *Author for correspondence. E-mail: eduardo.silva@ueg.br*

ABSTRACT. Academic-scientific phraseological units in the English language play a key role in the communication of/to experts, once they reproduce frequent and expected expressions in varied disciplines. This paper aims at identifying and analyzing the 100 non-specialized academic-scientific phraseological units (constituted of 4 words) in the English language, present in eight major fields of knowledge. The theoretical background referred to Phraseology and Corpus Linguistics. Regarding methodology, an academic corpus was compiled with more than 120 million words. The software WordSmith Tools was used for the linguistic-textual process. Through the Juilland dispersion coefficient and use coefficient, the most frequent phraseological units were identified in the academic texts. The list was eventually validated by the Wilcoxon rank sum test ($\alpha = 0.05$), indicating that the phraseological units identified show a higher use in the academic communication when compared to the use in the general language. The most relevant units are 'the case of', 'as a result of' and 'at the end of'. The list with the most functional phraseological units in the English language might provide a valuable pedagogical linguistic reference for the study of the academic genre.

Keywords: specialized communication, frequent expressions, linguistic-textual processing, statistical analysis.

Identificação das unidades fraseológicas em língua inglesa mais comuns em textos acadêmicos: contribuições advindas de *corpora*

RESUMO. As unidades fraseológicas acadêmico-científicas em língua inglesa desempenham um papel importante na comunicação de/para especialistas, uma vez que reproduzem expressões frequentes e esperadas em variadas disciplinas. Este trabalho tem como objetivo identificar e analisar as 100 unidades fraseológicas (compostas de 4 palavras) acadêmico-científicas não especializadas em língua inglesa, presentes em oito grandes áreas do conhecimento. A fundamentação teórica recorreu à Fraseologia e à Linguística de *Corpus*. Com relação à metodologia, foi constituído um *corpus* acadêmico com mais de 120 milhões de palavras. O *software* WordSmith Tools foi utilizado para o processamento linguístico-textual. Por meio do coeficiente de dispersão de Juilland e do coeficiente de uso, foram identificadas as unidades fraseológicas mais recorrentes dos textos acadêmicos. A lista foi posteriormente validada pelo teste de soma de postos de Wilcoxon ($\alpha = 0,05$), indicando que as unidades fraseológicas identificadas apresentam um uso superior na comunicação acadêmica quando comparado ao uso da língua geral. As unidades mais relevantes são 'the case of', 'as a result of' e 'at the end of'. A lista com as unidades fraseológicas mais funcionais em língua inglesa pode fornecer uma referência linguística pedagógica valiosa para o estudo do gênero acadêmico.

Palavras-chave: comunicação especializada, expressões frequentes, processamento linguístico-textual, análise estatística.

Introduction

Given the relevance of the theme and the insertion of Brazilian researches in international scenario in the English language, noted by the increase of the number of versions or texts written by Brazilians, it is necessary to study and provide supporting material to researchers, teachers, students, and translators. When studying academic texts, Babini and Silva (2012) show that Brazilian researchers (and professionals who translated texts into English) produce texts in the English language that, from a lexical perspective, are characterized by the overuse and/or by the lack of certain expected linguistic items in scientific articles. As a consequence, scientific articles written by Brazilians tend to feature some inadequacies, and may sound odd or different to the academic community's eyes,

more specifically to the community that uses the written English language as means of communication.

This way, our investigation aims to bring awareness concerning academic linguistic production, and highlight that, besides terminological and lexical common units, the academic text is also constituted of phraseological units, which potentially give authenticity to the scientific article in English, fulfilling the target reader's expectations – thus, having common vocabulary, terminological vocabulary and typical structures of academic-scientific communication.

To our knowledge, there are no existing studies in Portuguese, in the Brazilian variant, that have been developed under this methodological-theoretical framework, listing non-specialized phraseological units coming from texts of specialty produced by the academic community, common to all fields of knowledge. This work is characterized by a computational-linguistic approach, accompanied by statistical calculations.

In face of this subject, three research questions were formulated: (1) which are the most frequent 100 expressions from an interdisciplinary standpoint?, (2) Do the phraseological units of the sample occur at the same frequency of the general non-specialized language?, and (3) Is there any difference between the phraseological units use in the English language between natives and non-natives?

Considering what was previously exposed, we aim at identifying and describing 100 academic-scientific phraseological units in the English language, present in texts of eight great areas of knowledge, organized by the Coordination for Higher Education Staff Development (CAPES), namely: Exact and Earth Sciences; Biological Sciences; Engineering; Health Sciences; Agrarian Sciences; Applied Social Sciences; Human Sciences and Linguistics, Literature and Arts.

Regarding the research outlined, we will present the theoretical foundation based mostly on Phraseology and Corpus Linguistics. Soon after, we will describe the methodological procedures linked to the collection and processing of the corpus, as well as the statistical calculations used. In the third section, we will analyze the obtained data and the results that we have found. In the fourth section, we will discuss the final considerations.

Theoretical foundation

The theoretical foundation resorts to Phraseology and Corpus Linguistics.

Ellis (2008) explains that from the 1950s on, structural patterns started to be called 'constructions' or 'phraseologisms'. In comparison to the past century, a considerable amount of studies has been developed in the Phraseology purview, with contributions in a prominent position coming from studies affiliated with cognition, description, acquisition, teaching of native and foreign language, and also with Terminology, as phraseological units also occur in specialized texts.

In order to differ the theoretical line from the object of study, we will adopt the term Phraseology - with an initial capital letter - to refer to the discipline which studies the phraseological units and also the term phraseology - with an initial lower case letter - to allude to the object of study of this discipline, the phraseological units. In this context, the object of study tackled here may be discussed under varied names, depending on the theoretical strands: multi-word expressions, statistical phrases, chains, formulaic sequences, multi-word units, grouping, combinations of recurring words, lexical package, n-grams etc. Orenha-Ottaiano (2009) mentions that there is a broad conceptual and definitional range when it refers to phraseological units, for example: conventional expressions, lexicalized formulaic structures, prefabricated blocks, multi vocabulary lexical units, phrasemes, set phrases etc. Due to length constraints and the scope of our work, we will not discuss the different conceptions, approaches and definitions of the terms presented above.

To conceptualize our understanding regarding Phraseology, we bring along the contribution of Corpas Pastor (1996, p. 20 apud Orenha-Ottaiano, 2009) who defines the phraseological units as units formed by more than two graphic words in its inferior limit - likely to reach, in its superior limit, the compound sentence. Other significant characteristics, according to the author, are identified by their high frequency in use and co-occurrence of their constituent elements; by the sense of fixedness as well as semantics specialization; by their idiomaticity and the degree of manifestation of all these aspects. According to Bevilacqua (2004-2005), it is possible to affirm that phraseology of common language covers the study of several different units: proverbs, sayings, idiomatic expressions, collocations, for example.

Hunston (2010) provides some examples that illustrate phraseology in use: collocations (words that take place together); expressions (proverbs, popular sayings, set phrases); tendencies of use (verbs in passive voice, negative clauses, plural form

etc) and complementation patterns (verbal and nominal regency).

When dealing with phraseology in the academic context, Coxhead (2008) stresses the importance of typical words in the academic communication as well as stresses the importance of the groups to which these words belong. The author believes that we are still in a preliminary phase regarding the understanding of formulaic sequences nature in the academic context.

Hereafter, the sequences of words will be referred to as 'phraseological units', in accordance with the concept of Corpas Pastor (1996, p. 20 apud Orenha-Ottaiano, 2009).

Chen and Baker (2010) affirm that 4-word sequences are the most explored in written language studies. We will analyze 4-word sequences in written academic communication that commonly occur in academic scientific texts, but do not have the status of a specialized phraseological unit. For instance, the sequence constituted by 4 words 'scientific name of the animal' is likely to occur in the Biology field, whereas, in the Mathematics field, its occurrence may be rare or even null. Our purpose is to identify the units that may be, at the same time, common in several areas of expertise. Therefore, the phraseological unit focused here is not identified *a priori* as pertaining to a particular area of knowledge.

Taking into account the relevance of academic Phraseology knowledge and mastery, we agree with Bevilacqua (2004-2005, p. 75), when she states that

> [...] knowing phraseological units implies a linguistic competence relating to linguistics resources used in texts of specific learning contexts. Furthermore, it is assumed a degree of the subject or the themes knowledge covered by these texts, because they constitute, with the terms, transmitting units of specialized knowledge.

For that matter, Biber (1999 apud Martinez & Schmitt, 2012) points that about 21% of his academic corpus consisted of phraseological units. This finding leads to believe that the theme in question has an important role in academic texts, making them relevant for reception and production. In the academic field, Paquot (2008) brings to light the English as a foreign language students' ability of producing phraseological units, identifying an excessive use of structures in their native language. The author suggests that more teaching materials should be published so as to incorporate information based on language observation in scientific texts, for example. Biber, Conrad and Reppen (2004) believe that pattern descriptions in different linguistic genres identify characteristics in the genres, through similarities or differences. However, there is a need of a large-sized corpus to detect important events.

Besides the possibility of language study through isolated words, there is another path to be followed along with phraseological units. In relation to the language field, the addition of these items in pedagogical materials can bring benefits to learners, based on empiric data collected in actual usage. Since the use of any idioms by its users provide an infinite generation of data (Biderman, 2001), it can be inferred that such amount of data can be analyzed according to researchers' interests in many areas, including Corpus Linguistics.

Also, supporting Phraseology, Corpus Linguistics is an area that studies the language with the aid of computational resources in order to process great quantities of actual linguistic data (spoken and/or written texts) coming from natural communication, that is to say, the language in use. In other words, this line of research, according to Berber Sardinha's (2004) definition, is an approach that favors the observation of large amount of authentic data collected based on electronic corpora (collection of empiric data, of one or more languages, or varieties of a language, digitally stored), with research purposes, based on computational tools, carefully collected textual linguistic data.

Corpus Linguistics, on the one hand, provides theoretical and methodological subsidies to the description of natural language, for foreign language teaching and Translation studies, and on the other hand, works in a conceptual framework formed by an empiric approach and a vision of the language as a probabilistic system (Berber Sardinha, 2004). By this concept, we can determine the importance of the corpus as a source of information, since it corresponds to the register of natural language used by the language users in real situations. In the academic context, such research resource sets itself as essential, mainly when there exists hundreds of millions of texts or millions of words to analyze.

In a new linguistic construct, not only does it offer a set of computerized techniques to the verification of traditional phenomena connected, for example, to the lexicon, semantics or syntax; on the contrary, the analysis of a corpus can revel facts about a language which may have never been thought. This way, such line of thought proves to have an exploratory orientation –being not only a new methodology of language studying, but also a new way of research.

Sinclair (2004a e b) argues that the observation of real language turns out to be a safe way of describing a language, providing the study of a range of patterns, many of which unexpected. Without empiric data, the task of indicating if certain linguistic phenomenon happens naturally in a context becomes hampered.

By taking into consideration what was stated, the approximation between Phraseology and Corpus Linguistics to study academic texts provides other paths to linguistic exploration, adding empiric approach to the theoretical reflection.

Methodology

Based on its objectives, this is anexploratory and descriptive research. During data analyses, a software of linguistic analysis was used, and, afterwards, the statistical calculations were performed. The software used and the computational and statistical procedures are described in this section.

In terms of software aid, version 5 developed by Scott (2011) of the linguistic-statistical tool WordSmith Tools (WST) was used to process the corpora textual content.

Regarding the methodological procedures, all the selected textual content was in English. All texts were converted to the format 'simple text', with the extension 'txt' to simplify the manipulation of the data by the software. It was impossible to consider the formulas, charts, figures and graphs in this conversion. The subtitles were, however, retrieved.

There was a need to compile a corpus with academic content in English that represented the academic discourse to a large extent. The corpus in question contains only the written form of eight major areas of knowledge. Articles, journals and reference works collected from the internet and available for free were used. Such decision is justified because, in this case, there was no influence of other foreign languages in the texts. The corpus specially constituted for this research was named Academic Corpus of English (ACE).

In order to include all the areas in the corpus and to follow a coherent parameter for the insertion of sub-areas to the larger ones, the Capes table was adopted as a parameter with the division of knowledge areas. The first large area collected was 'Exact and Earth Sciences' (containing 7 sub-areas). The second major area was 'Biological Sciences' (with 14 sub-areas). Followed by the 'Engineering' (including 12 sub-areas). The fourth major area was 'Health Sciences' (with 9 sub-areas). Successively, the 'Agrarian Sciences' (6 sub-areas). In the wide area 'Applied Social Sciences', 12 sub-areas were designated. The corpus 'Human Sciences' has 10 sub-areas. Finally, the corpus 'Linguistics, Literature and Arts', which includes the sub-areas with the same name.

It is important to mention that the unequal dimensions between the areas and sub-areas do not impair our analysis, once the statistical procedures used take into account divergences and 'correct' the distortions.

According to Sinclair (2004a), certain criteria underlying the choice of texts that constitute a corpus should be established. The ACE can be described as follows: text mode (written and originated from the electronic medium); type of text (reference books and scientific articles); knowledge of the text (academic); variety of language (specialty language); region of texts (mainly United States and United Kingdom) and date of texts (years 2000 to 2012).

The specialized corpus constituted for this research shows 122,464,043 tokens, or occurrences. According to the categorization levels of a corpus size, as proposed by Berber Sardinha (2004), our corpus can be considered as a large corpus.

The identification of the most common intra- and interdisciplinary phraseological units of academic communication required a series of computational and statistical procedures.

The first step in organizing the files relied on the creation of a table for each major area. The table is a file generated by WST that records the position of all the words in the study corpus, allowing the query of n-grams or concordance lines. A table was generated for each sub-area. For that, the 'Compute Clusters' feature was used, in which the option of searching for 4-word n-grams with a minimum frequency equal to 5 was chosen (smaller numbers could also be chosen, but would imply a greater number of data without relevance to the analysis). The option to ignore numbers and symbols (both represented by #) was selected, as shown in Figure 1.

Next, the Detailed Consistency List allowed all phraseological units in the corpora to be identified, computed and compared. In other words, this tool catalogues, in descending order, all the expressions and their occurrence in every corpora. A list of detailed consistency was created for each sub-area so as to produce a new list of detailed consistency (with sub-areas) in order to obtain a list of the larger area. Figure 2 presents the final phase of creation of the last lists. All information in this list was exported to an electronic chart in MS Excel 2016 package.

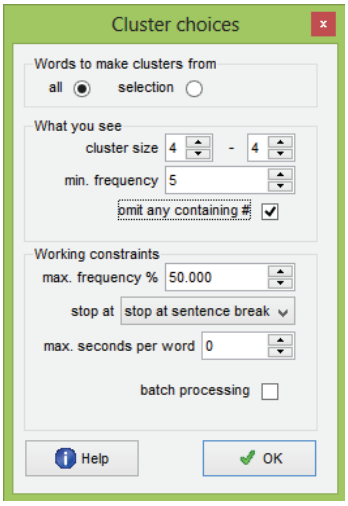

Figure 1. Feature selection of phraseological units.
Source: Scott (2011).

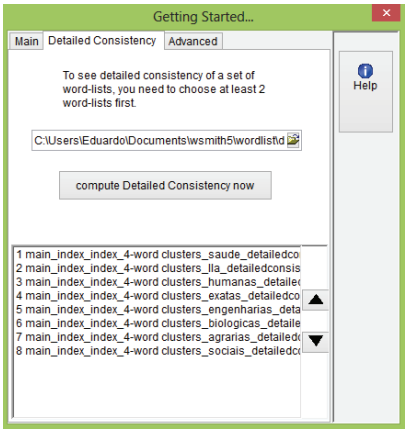

Figure 2. Detailed Consistency List.
Fonte: Scott (2011).

The n-gram frequencies were normalized, once the corpus showed unequal dimensions. The normalization was obtained through this formula: NF = (AF/CS), that is, the normalized frequency (NF) is equal to the absolute frequency (AF) divided by the corpus size (CS). The result is then multiplied by a million. Therefore, all the comparisons are carried out within the same proportions.

After each expression of each research corpus had its frequency normalized, it was necessary to discover which presented the most regular distribution.

The Julliand dispersion coefficient (D) was used, which indicates the distribution of an item. The dispersion coefficient values (D) vary between 1 (very high dispersion) and 0 (no dispersion), according to Oakes (1998, p. 190).

$$D = 1 - \frac{V}{\sqrt{n-1}} \quad (1)$$

Thus, the dispersion coefficient takes into account the variation coefficient (V), which is a measure that provides the data variations regarding the average obtained by comparing the concentration degree around it. The lower the value, the more homogeneous the data. The coefficient of variation (V) corresponds to the results from the division of the standard deviation (*6*) by the arithmetic mean (X), that is $\frac{\sigma}{X}$. The number of the corpus subsections is represented by 'n'.

In order to obtain a more precise measure, which takes into account both the frequency of expression (already normalized) and the dispersion coefficient, it is necessary to use the use coefficient (U), expressed by the following formula:

$$U = F \times D \quad (2)$$

Here, the use coefficient (U) is equal the frequency of expression (F) multiplied by the dispersion coefficient (D). In addition, all the results were placed in an electronic chart from MS Excel 2016 package.

In the next stage, in order to prove if the results were significant, it was necessary to compare the academic corpus data and the general language data. To make it possible, the British National Corpus (BNC) file with phraseological units was used, available for free download at the following URL: <http://www.lexically.net/downloads/BNC_wordlists/downloading%20BNC.htm>. The first 100 expressions of the ACE were searched in the file and the numbers tabulated in a chart for later calculations. Once again, the data were normalized to allow comparison.

To validate the ACE data, we used the Wilcoxon rank sum test (Larson & Farber, 2012) to compare two groups (ACE-BNC) and use frequency of the phraseological units, in search of a significant difference.

Finally, we compared our list of phraseological units with the list made by Chen and Baker (2010).

Results and data analysis

After all procedures described in the methodology, we have the results and data analyses.

It is possible to notice that the normalized frequency indicates very irregular values among the areas. The unit 'at the same time' occurs 26 times per million in the area of Exact Sciences, whereas it occurred 98 times per million in the Social Sciences field – i.e. the frequency is 4 times greater. Such difference imply that certain units are preferred in a given field.

Despite the differences identified, the units show a high dispersion of use, being used in all areas of knowledge analyzed in our sample.

It was not observed a distribution, which could be considered quantitatively perfect. The unit 'in the case of' occurs very frequently in the corpus of Engineering. The unit 'in the absence of' was not identified in the corpus of Linguistics, Literature and Arts. That does not mean that it is not used. Indeed, in the analyzed corpus its occurrence is not meaningful. The odds of finding it in other knowledge areas are considerably higher.

The 100 most relevant phraseology units:

IN THE CASE OF; AS A RESULT OF; AT THE END OF; THE END OF THE; AT THE SAME TIME; AS WELL AS THE; IN THE FORM OF; ON THE BASIS OF; CAN BE USED TO; ONE OF THE MOST; IT IS IMPORTANT TO; IS ONE OF THE; IN TERMS OF THE; A WIDE RANGE OF; IN THE PRESENCE OF; THE NATURE OF THE; IT IS POSSIBLE TO; IN ADDITION TO THE; IN THE ABSENCE OF; THE SIZE OF THE; IN THE CONTEXT OF; A LARGE NUMBER OF; IS BASED ON THE; WITH RESPECT TO THE; AS A FUNCTION OF; THE FACT THAT THE; AT THE TIME OF; THE REST OF THE; THE USE OF THE; THE BEGINNING OF THE; A WIDE VARIETY OF; AT THE BEGINNING OF; IN THE DEVELOPMENT OF; THE DEVELOPMENT OF THE; CAN BE FOUND IN; TO BE ABLE TO; THE BASIS OF THE; THE TOP OF THE; IN THIS CASE THE; THE USE OF A; AS PART OF THE; THE VALUE OF THE; AN IMPORTANT ROLE IN; IN A VARIETY OF; THE CENTER OF THE; THAT THERE IS A; A RESULT OF THE; IN RELATION TO THE; IN A NUMBER OF; THE STRUCTURE OF THE; IN SUCH A WAY; THE CASE OF THE; TO THE DEVELOPMENT OF; THE EXTENT TO WHICH; AS WELL AS IN; THE TOTAL NUMBER OF; A GREAT DEAL OF; THE PRESENCE OF A; IT IS CLEAR THAT; AN EXAMPLE OF A; REFERRED TO AS THE; OF THE MOST IMPORTANT; TO THE FACT THAT; AN INCREASE IN THE; IN THE SAME WAY; IT IS NECESSARY TO; IN THE COURSE OF; THE ROLE OF THE; IT IS DIFFICULT TO; THE FORM OF A; AS WELL AS A; THE QUALITY OF THE; AND THE USE OF; THE SUM OF THE; IS REFERRED TO AS; OF A NUMBER OF; THE DIFFERENCE BETWEEN THE; THAT THERE IS NO; THE RELATIONSHIP BETWEEN THE; IS AN EXAMPLE OF; IS DETERMINED BY THE; ON THE OTHER HAND; IN THE PROCESS OF; BE FOUND IN THE; SUCH A WAY THAT; FOR THE PURPOSE OF; AS A RESULT THE; WITH THE EXCEPTION OF; THE BOTTOM OF THE; IN CONTRAST TO THE; ON THE ONE HAND; A MEMBER OF THE; IT SHOULD BE NOTED; CAN BE USED FOR; IS LIKELY TO BE; AT THE LEVEL OF; THE RESULTS OF THE; THE LENGTH OF THE; THE WAY IN WHICH; IS RELATED TO THE

The data reflect the usage of the expressions that stand out in the academic scientific communication, that is, the most important expressions in the written academic register, whether for reading or written purposes. Nevertheless, since the academic communication is a subgroup of general language, it is valid to question the significance of our data. In other words, do the most functional expressions identified in the ACE show, in fact, a more expressive use in the specialized communication when compared to the general language? Is there a possibility that these expressions would also play a main role in other genres? If that is the case, we cannot consider them functional in the academic communication *per se*.

By using WST, we processed and identified the 100 phraseological units (with four elements) in the BNC and compared the normalized frequency with those of the ACE. This process is important to support the assumption that the data obtained in the academic corpus are significant – an attempt to indicate that the use of phraseological units have a much higher frequency in the academic speech than in the general language speech (non-specialized).

Among the non-parametric tests that can be used to compare samples, we used the Wilcoxon rank sum test (to compare two different independent samples) in order to find out whether there is difference between the samples. We elaborated two hypotheses, the null hypothesis (H_0) and the alternative hypothesis (H_1).

For the use frequency of these expressions, we have:

H_0 = there is no difference between the use frequency of the expressions in the ACE and the BNC;
H_1 = there is difference between the use frequency of the expressions in the ACE and the BNC;

A standard level of significance of 5% ($\alpha = 0.05$) was chosen. We identify that there is statistical significance when the p-value is lower than the significance level adopted ($\alpha = 0.05$). Whenever a statistical significance is found, the alternative hypothesis is accepted.

The result of the tests show that the value obtained of $p = 0.000023$. As the value of the test is inferior to the value of alpha (0.05), the null hypothesis can be rejected and the alternative hypothesis is then accepted. Consequently, the data of frequency featured in the ACE differs significantly from those in the BNC – attesting that the ACE does bring highly functional typical expressions of the context in which they occur. From a statistical standpoint, the analysis of the phraseological units bring up a highly significant use in the academic texts when compared to the level of use of the general language, what allows to state that these units have an important role in academic written communication.

Furthermore, the information from the ACE can be used in pedagogical activities. Individuals who need to practice reading, for example, can resort to the context that the phraseological units occur to memorize their meaning. English language instructors of Language Arts or Translation have at their disposal a series of real uses of phraseological units with which they can elaborate exercises based on gap filling, multiple choice, and assignments of version/translation of specific units. The excerpts below belong to the corpus 'Linguistic, Literature and Arts' and bring three examples of the uses of three phraseological units: 'a wide range of', 'with respect to the' and 'the extent to which':

a wide range of

*That leads to **a wide range of** specialized knowledge that will serve as a base for our research.*

*As we have seen in the previous section, two-year-old speakers of **a wide range of** languages use rather similar syntactic and morphological simplifications.*

*The conversations, which cover **a wide range of** subjects, present a romanticized and whimsical view of Scotland.*

with respect to the

*… to evaluate the current educational politics adopted in recent years in the United States **with respect to the** education of Spanish speakers and other minority bilingual groups.*

*… and activities from various method sources and can be regarded as innovative only **with respect to the** purposes for which they are recommended and the ways they are used.*

*Tasks also displayed some differences **with respect to the** characteristics of LREs.*

the extent to which

The extent to which the participants had received instruction in grammar, either of their native language or some other target language, is described in only the most general terms.

Hence, the following presentation of qualitative data will explore **the extent to which** these results may be attributed to the divergent approaches.

He has also doubted **the extent to which** translation is possible between languages.

The possibility of selection of the knowledge area along with the phraseological unit can provide a perspective of personalized learning. Considering the relevance of Phraseology, the conscious treatment of units can offer valuable pedagogical resources to the individuals that need to deal with academic texts in English.

Finally, assuming that the results stemming from the exploration of ACE are trustworthy, we have conditions to compare our list with Chen and Baker's list (2010), a quantitative approach based on corpus to identify the most common phraseology (constituted of four elements) in academic essays written in the English language. Two corpora were explored: the Freiburg-Lancaster-Oslo/Bergen (FLOB) in which only the academic essays section was used, called FLOB-J, created by several specialists who are native English speakers. The FLOB-J is constituted of texts extracted from academic journals and scientific books; and also, the British Academic Written English (BAWE), with essays of college students proficient in English language, having been selected two nationalities for sample: Chinese (BAWE-C) and British (BAWE-E).

We verified the presence/absence of each of the 100 most functional phraseological units of ACE in the list presented by FLOB-J, BAWE-E and BAWE-C. From the most common 100 phraseological units in the English language of the academic communication, 46% of them were also the ones most used in FLOB-J. In BAWE-E, 39%. In BAWE-C, 31%.

In the educational context, both for translators and for language teachers, these results emphasize the importance of drawing attention to the problem of academic phraseology. The divergences of use between the texts FLOB-J and BAWE-C may serve to foster further studies on teaching-learning and/or writing/reading academic texts.

Final considerations

With the framework of Phraseology and Corpus Linguistics, we identified and analyzed the academic-scientific phraseological units in English, shared by eight major areas of knowledge (Exact and Earth Sciences, Biological Sciences, Engineering, Health Sciences; Agrarian Sciences, Applied Social Sciences, Linguistic, Literature and Arts). As we sought to demonstrate, the simple absolute frequency of the phraseological units would not suffice to investigate the most common ones. The mere observation of the normalized frequencies in the different corpora would not be enough to state that a phraseological unit has a uniform occurrence among different areas. A phraseological unit with a high frequency could be identified in one or a few corpora, not serving to represent the type of phraseological unit, which is focused in the present study. On the other hand, another phraseological unit could have an equitable distribution among all corpora, but a low frequency. This would

undermine the scope of the research. Our intention, therefore, was to select the expressions that, besides being frequent, were well distributed in the eight research corpora.

Concerning the research problems previously formulated (which are the 100 most frequent expressions from an interdisciplinary point of view?), we identify that the phraseological units, for example, 'in the case of', 'as a result of', 'at the end of', 'the end of the', and 'at the same time' have a high frequency and a high dispersion index. This means that they are used in all areas of knowledge analyzed in this work, with a considerable level of occurrence.

To answer the second problem (Do the phraseological units of the sample occur at the same frequency of the general non-specialized language?), we used a statistical calculation known as the Wilcoxon rank sum test and compared the frequency of use of the phraseological units in the ACE and in the BNC. The obtained results allow to state that there is a statistically significant difference in the comparison between the two groups.

Our third problem (is there any difference between the phraseological units use in the English language among natives and non-natives?) can be answered by the comparison between the most used phraseological units in the academic environment, identified here, and those highlighted in the study by Chen and Baker (2010). Although the authors have identified similarities in the identification of the phraseology commonly found in written English academic communication of natives and non-natives, when comparing phraseological units found in our research to the results of these authors, we find that natives have a relatively higher use when compared to university students - both Chinese and British.

Among the 100 most used phraseological units in academic communication, we noticed that 46% of them were used by specialists (FLOB-J). British university students (BAWE-E), used 39% and the Chinese (BAWE-C), 31%. We noticed a difference of 15 percentage points between specialists whose native language is English and Chinese university students who do not speak English as their native language.

In pedagogical terms, after a strict selection, the formulaic expressions found in the academic writing of native specialists can be of great value to non-native learners, in terms of a more natural academic style, and should therefore be integrated into the teaching curriculum of English as a second language, as a foreign language or for specific purposes.

It is also worth highlighting the important role that the contributions of the phraseological studies and Corpus Linguistics have to educators and to modern society, because, through their advances, they allow more accessibility and mastery of the knowledge related to phraseological units and the uses of human language. As seen, the establishment of the ACE undertook an attempt to cover as many knowledge sub-areas as possible to cover a wide range of disciplines. Since the phraseological units appear several times in the scientific text, the processing performed by the WST software with the specialty corpora provided relevant data.

The criteria outlined in this work show that labeling an expression as functional implies taking into account two fundamental details: frequency and dispersion (translated by U use). If only frequency or only distribution is highlighted, the analysis will become skewed, prone to error, and inaccurate judgments. We believe the phraseological units with which we work can be called functional because they are not only the most used in academic written communication in English, but also have a specific pragmatic function in the textual genre in which they occur.

The use of statistical calculations in a data analysis allow the discovery of linguistic facts, deserving attention because it gives objectivity in the quantitative exploration in the study of the language and allows new paradigms and new questions - both made feasible thanks to the interdisciplinary approach.

The idea underlying this research is to contribute to the study of phraseological units. In general, researchers, translators and teachers can benefit from the discussions presented here regarding phraseological expressions, for two main reasons: ubiquity of phraseological units in academic-scientific communicative situations and direct applications in professional performance, especially in contact with the English language in written contexts.

References

Babini, M., & Silva, E. B. (2012). A terminologia acadêmica nos textos científicos em língua inglesa: uma abordagem baseada em *corpus*. In A. N. Isquerdo, & M. C. T. C. Seabra (Orgs.), *As ciências do léxico: lexicologia, lexicografia, terminologia* (p. 415-427). Campo Grande, MS: UFMS.

Berber Sardinha, T. (2004). *Linguística de Corpus.* Barueri, SP: Manole.

Bevilacqua, C. R. (2004-2005) Fraseologia: perspectiva da língua comum e da língua especializada. *Revista Língua e Literatura, 6*-7(10-11), 73-86.

Biber, D., Conrad, S., & Reppen, R. (2004). Register variation and English for specific purposes. In D. Biber, S. Conrad, & R. Reppen. *Corpus Linguistics: investigating language structure and use* (p. 135-171). Cambridge, MA: Cambridge University Press.

Biderman, M. T. (2001). *Teoria linguística: teoria lexical e linguística computacional*. São Paulo, SP: Martins Fontes.

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49.

Coxhead, A. (2008). Phraseology and English for academic purposes: challenges and opportunities. In F. Meunier, & S. Granger (Ed.), *Phraseology in foreign language learning and teaching* (p. 149-162). Philadelphia, PA: John Benjamins Publishing Company.

Ellis, N. (2008). Phraseology: the periphery and the heart of language. In F. Meunier, & S. Granger (Ed.), *Phraseology in foreign language learning and teaching* (p. 1-13). Philadelphia, PA: John Benjamins Publishing Company.

Hunston, S. (2010). Corpora and language teaching: issues of language. In S. Hunston. *Corpora in applied linguistics* (p. 137-169). Cambridge, MA: Cambridge University Press.

Larson, R., & Farber, B. (2012). *Estatística aplicada* (4a ed.). São Paulo, SP: Pearson Prentice Hall.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh, UK: Edinburgh University Press.

Orenha-Ottaiano, A. (2009). *Unidades fraseológicas especializadas: colocações e colocações estendidas em contratos sociais e estatutos sociais traduzidos no modo juramentado e não juramentado* (Tese de Doutorado). Universidade Estadual Paulista, São José do Rio Preto.

Paquot, M. (2008). Exemplification in learner writing. In F. Meunier, & S. Granger (Ed.), *Phraseology in foreign language learning and teaching* (p. 101-119). Philadelphia, PA: John Benjamins Publishing Company.

Scott, M. (2011). *Word smith tools.* Version 5. Liverpool, UK: Lexical Analysis Software.

Sinclair, J. (2004a). How to build a corpus. In M. Wynne (Ed.), *Developing linguistic corpora*: *a guide to good practice* (p. 79-83). Oxford, UK: Oxbow Books.

Sinclair, J. (2004b). *Trust the text: language, corpus and discourse.* New York, NY: Routledge.

Received on May 3, 2016.
Accepted on November 23, 2016.