

Evidence of semantic association in technology-related vocabulary: a corpus study of collocations

*Evidência de associação semântica em vocabulário relacionado
com a tecnologia: um estudo de corpus sobre colocados*

Jesiel Soares Silva¹

Luiz Henrique Mendes Brandão²

Brenda Lorraine Grillo Silva³

Abstract: This paper presents a semantic study of corpora-based words which meanings are related to technology, exploring how such words have been semantically coined over time regarding their most frequent collocations. In this study, we counted on three corpora (DAVIES, 2010): The Corpus of Contemporary American English (COCA) and the Corpus of Historical American English (COHA) were used as source of data about the English language. The Portuguese language data was assessed through the *corpus do português* NOW (News on the Web). The screening procedures were respectively: a) Selection (words that probably have changed semantically over the years); b) Comparison (check the collocates and their frequency); collection: a file was created with all the comparisons regarding frequency and strength of attachment). In order to verify the existence of statistically different groups, we applied T-test and considered the percentage of appearance for the target-collocations as well as the MI score. Results indicated meaningful changes to words related to internet milieu when time is considered and regarding collocation. Similar results have been found for both Portuguese and English. Moreover, data revealed that the three most frequent collocates for the target-words when present time was considered not only demonstrated a clear change towards internet-related meaning, but also showed a higher strength of attachment.

¹ Professor adjunto de língua inglesa na Universidade Federal dos Vales do Jequitinhonha e Mucuri. E-mail: jesielsoaressilva@gmail.com

² Mestrando em linguística teórica e descritiva pela Universidade Federal de Minas Gerais. E-mail: luizhenriquemendesbrandao@gmail.com

³ Graduada de Letras na Universidade Federal dos Vales do Jequitinhonha e Mucuri. E-mail: brendagrillo@live.com

Keywords: collocates; word association; word frequency.

Resumo: Este artigo apresenta um estudo semântico de palavras baseadas em corpora cujos significados estão relacionados à tecnologia, explorando como tais palavras foram semanticamente cunhadas ao longo do tempo em seus colocados mais frequentes. Neste estudo, contamos com três corpora distintos (DAVIES, 2010): The Corpus of Contemporary American English (COCA) e o Corpus of Historical American English (COHA) foram utilizados como fonte de dados sobre a língua inglesa. Os dados da língua portuguesa foram coletados do corpus do português NOW (Notícias na Web). Os procedimentos de triagem foram, respectivamente: a) Seleção (palavras que provavelmente mudaram semanticamente ao longo dos anos); b) Comparação (verificação das colocações e sua frequência); coleta: foi criado um arquivo com todas as comparações quanto à frequência e força de fixação). Para verificar a existência de grupos estatisticamente diferentes, aplicou-se o teste T e considerou-se a porcentagem de aparição para os colocados-alvo e também o MI score. Os resultados indicaram mudanças significativas nas palavras relacionadas ao meio da internet quando o tempo e a posição do colocado foram considerados. Resultados semelhantes também foram encontrados tanto para o português quanto para o inglês. Além disso, os dados revelaram que os três colocados mais frequentes para as palavras-alvo, quando considerado o tempo presente, não só demonstraram uma mudança clara no significado relacionado à tecnologia, mas também apresentaram uma maior força de vinculação.

Palavras-chave: colocados; associação de palavras; frequência de palavras.

Introduction

This paper presents a semantic study of corpora-based words which are culled deriving out of technology-related meaning, following by how such words have been semantically chiseled over time regarding their most frequent collocations. The choice of using a corpus for data collection is due to the fact that corpora present “a large sample of how people have used language” (STUBBS, 2001, p. 23). Stubbs (op. cit.) draws out the pertinence in acknowledging word collocates in pursuance of labeling semantic patterns; particularly, for this study, such semantic patterns imply meaningful changes in these words anent meaning.

As it is asserted by Gablasova, Brezina and McEnery (2017), formulaic expressions (group where collocates are inserted) have been object of linguistic and psycholinguistic studies and have long been at the forefront of interest in corpus linguistics (CL, henceforth). The techniques to identify patterns of co-occurrence of linguistic units plus the description of collocates language have been a fruitful niche of collocations studies. An example is the research conducted by Gablasova, Brezina and McEnery (2017) which analyzed the nature of collocates as found and identified in corpora, also the implications of these language items towards the language learning research. Another study having collocations as object is Zaabalawi (2019). Zaabalawi worked with the differences between English and Arabic collocations and reached the conclusion that the main reasons for the failure of Arabic language users to master the

use of English language collocations is that these units of formulaic language are not used as frequently in Arabic than they are in English, and the Arabic ones are not patterned as rigidly as those in English.

It is broadly known that language evolves as time goes by and, as the culture molds itself towards the globalization, newfangled aspects of language arise. Departing from this perspective, this study demonstrates the volatility of words meaning towards sociocultural changes, and some evidence of it being occurred over time. As the world changes constantly, new technologies come out altering the manner people interact and how they convey meaning. Such changes modify not only the way we see the world, but also the way we represent it linguistically. The semantic representations we associate with certain words in a given language seem to amend synchronically with the world's contemporaneity. Some words, regarding frequency of appearance, have their meanings increased, contrarily some other meanings may even cease to exist.

Brandão & Silva (2020) have shown recently that with the constant growth of the internet, some words had their meanings changed in order to adapt. These words had been used before mostly with a given meaning (e.g.: *web*, to express mainly the structure sewn by a spider), however nowadays, considering frequency of appearance in corpora, this word is most related to the World Wide Web. Considering this evolving scenario, this study aims at presenting evidence of semantic changing of five words in Portuguese and five words in English according to their most frequent collocations. Analysis for frequency and collocates were applied to sustain the hypothesis that such words have been changing across the years.

1 Corpora and crosslinguistic studies

In order to get acquainted with the premises endowed in this research endeavor, one should delve into the context of the corpus and relevant studies. Corpora are a compilation of written texts and transcribed speeches serving as common-data for linguistic studies. The corpora explored in this study were withdrawn from the Corpus of Contemporary and Historical English (COHA) (MARK, 2001). Both samples plus an English sample from COCA presented the cross-language we relied on the development of the study on word-collocates evolution.

Regarding the corpus, it provides a minimum of 45 million words (Portuguese sample – Genre/Historical) collected from subjects' speeches of real-context language on diverse content. Each word contains specifications such as: frequency, context informing the date, bearer, and fragments from which the content comes. The corpus also reveals the collocates pre or postposition of the chosen word or compare one to another or yet show the patterns in which a word occurs. Corpora such as this one, give the contexts in which speakers use one variant or another, and lead to insights about extra-linguistic factors that favor one variant (ROBINSON & DUNCAN, 2019).

The choice of using an online corpus was to find the most adequate approach in the recent studies regarding the language and its uses. The corpora offer to the linguistic community an empirical approach to the description of language with a much larger data than any other sources. A recent study about employing technology to analyze language use (MCENERY, 2019) has shown how corpus data from computational technology can provide a large data in linguistic studies. McEnery's study specifically aimed at analyzing the language use linked to technology in second language acquisition speakers. The author highlighted the word “network” and displayed its associations, demonstrating the important role that the word plays in the discourse:

The analysis reveals that the more general evaluative adjective good is more frequent in L2 speech (4,812 occurrences in almost 2 million words) than the adjective great (590 occurrences) and has more collocates in the corpus examined (84 vs. 9) (MCENERY, et al. 2019, p. 75).

Moreover, not only McEnery but also Brandão & Silva (2020) have presented studies in which we see ourselves into today, the evolution of the language since the technology molded our lives. In their study they presented evidence that target-words turned into, over the years, being more associated to collocates connected to technology. With its research, the word site, which presents a word connected with technology use, was linked with collocations such as web and browse, but previously in the 90s, that same word was followed by terms such as construction or with the adjective displayed near. This information provides us more an empirical association with the study around collocation words that could change once associated to technology-related words.

1.1 Collocates as evidence for language change

According to Poznanski (2002), a collocation is a group of words that are closely related and each of these words is called collocate. Some collocates are sequential, condition in which the collocates are consecutive in the sentence, others are non-sequential. Berber Sardinha (2004) displayed a widely knowing idea of collocations under the sentence: "You shall know a word by the company it keeps". It expresses an idea in lexical semantics that a word can have its semantic representation derived through investigation of patterns of co-occurrences with another lexical item (collocation). Conversely, a collocate can be, according to Jia-ying and Yuan-yuan (2014), taken as a broader term, meaning a co-occurrence that reaches a statistically significant level, or as a narrow definition, collocation is strictly influenced by grammar and grammatical structures.

In an embracing analysis of the role played by collocates in the various language research fields, Gablasova, Brezina and McEnery (2017) assert that formulaic language (a group of words used together very often) has long been of prominent interest within corpus linguistics. Moreover, the idea that corpora represent a vast repository, and from this repository data related to the regularity, frequency and distribution of formulaic patterns in language can be explored. According to the authors, collocations, as units of formulaic language, have received considerable attention in the last 10 years in the corpus-based studies. Gablasova, Brezina and McEnery (2017) also point that there are traditionally a twofold taxonomy of collocability when it comes to language learning: absolute frequency and strength of association between word combinations.

Several corpus-based research endeavors have relied on collocations to investigate linguistic changes. Making use of the Corpus of Contemporary American English (COCA) as source of their inquiry, Brandão & Silva (2020) found that after the popularization of the access to the World Wide Web, some words began to be used together with specific collocates highlighting the semantic change of those words towards a strongly relation to internet-related milieu. For instance, the word *cloud*, was then strongly related to natural phenomena and accompanied by collocates like *dust*.

Differently, nowadays it is strongly related to web, indeed, this very term is now the first collocate related to the given word. Brandão & Silva (2020) also reported the growth in the frequency of use of those words; the word *cloud*, for example, appear accompanied by its most prominent collocate in the beginning of the 1990's 24 times in the COCA, while it appears in the year 2017 with its number one collocate 5896 times.

Another study that used collocates as a tool of its inquiry was the one of Laufer and Waldman (2011). For this study, the authors selected LOCNESS (which is a corpus of young adult native English speakers) and compared it to another corpus composed by essays written by Hebrew native speakers who are learners of English. Results suggested that the successful use of collocates in a language is related to the proficiency of the user of that language. In this case, the native users of English far surpassed the native users of Hebrew in the use of the collocates in the English language.

2 Methodology

In order to conduct a study of language change over a certain period of time, we need to rely on a trustworthy data source. The online corpora are crucial tools for such endeavor. The corpus linguistics is a fertile field of studies that has been the basis for several discoveries in the language research worldwide, since it was established as such by Sinclair (1966).

In this study, we counted on three different corpora, all of them conceived by Davies (2010). The Corpus of Contemporary American English (COCA) and the Corpus of Historical American English (COHA) were used as source of data about the English language. The COCA is, according to Davies (2010), the first big and non-static corpus ever created. It contains more than 1 billion words from 1990-2019. COCA is balanced by genre, being probably the most widely used corpus of English. The COHA is, according to Davies (2010), the largest structured corpus of historical English and contains more than 400 million words of texts from the 1810s-2000s. The COHA is also balanced by genre.

The Portuguese language data was assessed with the help of the *corpus do português NOW* (News on the Web), that has a storage of around 1.4 billion words from 2012 to present time from newspapers and magazines, as stated by Davies (2010). The historical counterpart was the *corpus do português genre/historical*, a tool that comprises about 45 million words from the 1300s to 1900s and is corpus based, also according to Davies (2010).

Much research has been conducted based on online corpora by researchers from different areas of knowledge. Viola (2018) conducted research based on several Italian corpora to investigate the hypothesis that, over time, the words *grazie* and *ringraziare*

(thanks and to thank respectively in English) have been increasingly used in combination with *per* (“for”, in English) due to an influence from the English thanks/to thank for.

2.1 Words and collocates selection procedure

Crystal (2013) postulates how the rise of the internet has generated language patterns that characterized the way we speak English. For instance, FacebookTM has shaped feed which has a food-related meaning into a place where we can feed our minds with several pieces of online information, mostly about our circle of friends and relatives. With Facebook appropriation of semantics, the popularization of social media and new technologies, news terms came along shaping the way that we use the language, not only in English but also in other languages.

The website Datareportal⁴ (2019) developed a research in which it is argued that as of the third quarter of 2019, there are over 4.33 billion active internet users and based on a population of 7.74 billion people, it equates to an internet penetration rate of approximately 56%. Technology through internet has been shaping the world which we live in today, although people are not only talking about perception but also about the language, the mother that holds our communicative system.

Based on the previous studies of this research, each word was searched in the corpora and highlighted in its results. The word *Perfil* in Portuguese also translated to profile, has several meanings being the first to appear (and most common one): “a set of face trace from someone who is being looked from the side position”⁵ (PERFIL, 2020). Following it, in 1800-1900 the word *Perfil* was mostly found with collocates such as *usuário* (user), *linhas* (lines) and *adequado* (adequate). However, when we look for the same word in the first semester of 2019, we are able to see that its collocates have changed drastically. As so, we adjusted as procedures: selection of words, comparison and data collection.

- a) Selection: in this part sought to find in our daily routines words that don't have the same semantic role as before, since the new technology era.

⁴ <https://datareportal.com/global-digital-overview>

⁵ Translated by the authors from: *Conjunto dos traços do rosto de uma pessoa visto de lado*

- b) Comparison: We searched each and every word that was thought and looked for relevance in each corpus with each language.
- c) Data collection: A file was created with the comparison of the words and then, the most relevant ones were chosen with the intention to analyze if the theory was correct.

3 Data Analysis

First, we applied a two-tailed t-test considering the frequency values of the target-words in Portuguese and English. For Portuguese, the two groups of word frequency were composed by the target-words from 1900 (time- α , henceforth) (M=402.2, SD=408) and the target-words from 2019 (time- β , henceforth) (M=8470, SD=7019). The result, in $t(4)=2.51$, $p<.05$, revealed statistical significance between the groups. Likewise, for English, the two groups of word frequency were composed by the target-words from time- α (M=13041, SD=7445) and the target-words from time- β (M=1595, SD=1105). The result, in $t(4)=3.51$, $p<.05$ also showed statistical significance between the groups. Thus, the target words frequency factor in corpora, regarding the difference between both periods, is a compelling account for our assumption.

As a second step, we explored the data set of Portuguese, considering the MI score⁶. First, we analyzed the group of collocates for the target-words. Regarding the target-word *perfil* related to time- α , the first three collocates were consecutively *usuário* (freq=23, MI=7.31), *linhas* (freq=14, MI=4.73) and *adequado* (freq=11, MI=6.89). Relying on intuition, one could affirm that only *usuário* has a possible semantic relationship with the idea of internet-related concepts. When time- β data are taken into account, the first three collocates for *perfil* are consecutively *instagram* (freq=1080, MI=6.81), *facebook* (freq=659, MI=5.91) and *twitter* (freq=574, MI=5.76). Differently, in time- β the three first collocates have clear relationship with the web-related words. Thus, the results concerning the three most frequent collocates for *nuvem*, considering two different time periods, comply with our hypothesis.

⁶ The Mutual Information (MI) score expresses the extent to which words co-occur compared to the number of times they appear separately. MI Score is affected strongly by the frequency, low-frequency words tend to reach a high MI score which may be misleading. (<https://www.sketchengine.eu/>)

After that, we checked the target-word *nuvem* considering time- α . The primary collocates were consecutively *negra* (freq=45, MI=6.4), *poeira* (freq=42, MI=7.86) and *céu* (freq=34, MI=4.55). In all cases, there is no clear relationship with internet-related words with the three most frequent collocates for *nuvem*. Differently, when time- β were considered, the first three collocates for *nuvem* are *computação* (freq=145, MI=10.23), *serviços* (freq=117, MI=4.54) and *armazenamento* (freq=60, MI=7.73). The three most frequent collocates, especially the first (*computação*) and the third (*armazenamento*) present an affinity with web-related meaning, which also adhere to our hypothesis.

After, we considered the target-word *conexão* and its collocates in time- α . The first three collocates were sequentially *entre* (freq=36, MI=4.32), *íntima* (freq=4, MI=7.17) and *permitir* (freq=4, MI=6.78). For this case, the first and third collocate remain blurred regarding the semantic connection with web-related concepts, while the second cannot be immediately related to the topic. In time- β , the primary collocates for *conexão* were *política* (freq=193, MI=3.83), *internet* (freq=159, MI=5.41) and *portal* (freq=126, MI=5.27). In this case, the first collocate has clearly no connection with web-related semantic ideas. On the other hand, *portal* and *internet* suggests a connection with the web semantic milieu. This result mostly accedes to our hypothesis. The fourth target-word to be analyzed was *navegador*. From time- α , *navegador* had the three first collocates: *português* (freq=35, MI=7.33), *século* (freq=13, MI=5.27) and *inglês* (freq=10, MI=6.40).

There is no clear suggestion that those collocates refer to the context of web-related words. In time- β , the three first collocates were: *através* (freq=1503, MI=6.76), *internet* (freq=923, MI=7.14) and *browser* (freq=914, MI=11.27). With the exception of *através*, the other collocates suggest a semantic context of web-related concepts. This result also mostly confirms our hypothesis.

The last target word for Portuguese was *postar*. Regarding time- α , the collocates were: *abandonou* (freq=2, MI=9.31), *distância* (freq=2, MI=9.31) and *janela* (freq=2, MI=7.19). Both first and second collocates suggest no connection with web-related words and the third one could be linked to meaning close to web-related words. Differently, in time- β , the collocates were: *foto* (freq=200, MI=5.19), *vídeo* (freq=86, MI=5.72) and *redes* (freq=85, MI=5.18). As it is suggested, all collocates for *postar* in time- β resemble a web-related milieu, which confirms our hypothesis. Data analysis from Portuguese target-words are summarized in Table 1.

Table 1: Summary of the target-words, collocates and MI score for Portuguese

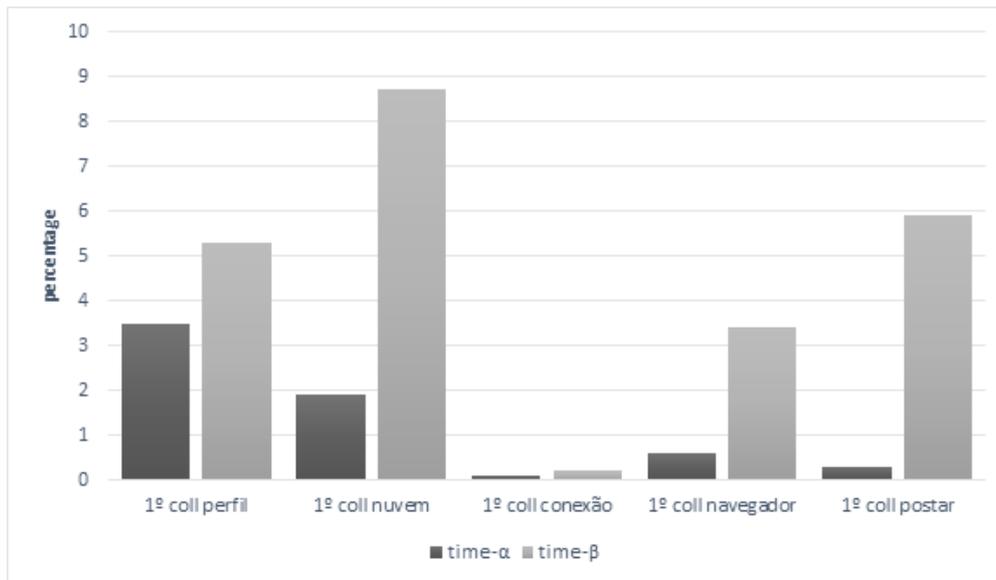
<i>target words</i>	<i>collocates</i>	<i>MI</i>	<i>target words</i>	<i>collocates</i>	<i>MI</i>		
perfil	time- α	usuário	7.3	navegador	time- α	português	7.3
		linhas	4.7			século	5.2
		adequado	6.8			inglês	6.4
	time- β	instagram	6.1		time- β	através	6.7
		facebook	5.9			internet	7.1
		twitter	5.7		browser	1.1	
nuvem	time- α	negra	6.4	postar	time- α	abandonou	9.3
		poeira	7.8			distância	9.3
		ceu	4.5			janela	7.1
	time- β	computação	10		time- β	foto	5.1
		serviços	4.5			vídeo	5.7
		armazenamento	7.7		redes	5.1	
conexão	time- α	entre	4.3	conexão	time- α	entre	4.3
		intima	7.1			intima	7.1
		permitir	6.7			permitir	6.7
	time- β	política	3.8		time- β	política	3.8
		internet	5.4			internet	5.4
		portal	5.2		portal	5.2	

Source: Authors' elaboration

Following, we set the MI score of the 3 most frequent collocates for the 5 target words in time- α in comparison with the 3 most frequent collocates for the 5 target-words in time- β regarding Portuguese. Our goal was to examine whether both groups are statistically different. A paired-sample t-test revealed the difference between time- α (M=6.6, SD=1,5) and time- β (M=6,4, SD=2,0) concerning the MI scores revealed $t(14)=0,40, p> .05$, which suggest there is no effect concerning both groups.

This previous result is informative to the fact that collocates attached to our target-word in time- α have shown a similar strength to those in time- β . However, as time went by, this word bond did not remain, since the position of these collocates (which are first, second and third positions) changed enormously in time- β . Figure 1 is illustrative of the percentage of occurrence for the first position of the 3 most frequent collocates to exemplify:

Figure 1: Percentage of the first position for the 3 most frequent collocates in Portuguese



Source: Authors' elaboration

Subsequently, we analyzed the group of collocates for the target-words in English. Regarding the target-word *search* related to time- α , the first three collocates were consecutively *truth* (freq=298, MI=3.10), *warrant* (freq=219, MI=8.23) and *object* (freq=210, MI=3.45). Intuitively, we may affirm that the three most frequent collocates for *search* are not closely related to the concept of internet. When time- β data are considered, the first three collocates for *search* are consecutively *warrant* (freq=91, MI=8.23), *engine* (freq=88, MI=7.70) and *result* (freq=49, MI=3.76). Different from our hypothesis, in time- β the three first collocates do not show clear links to the web-related words.

Following we checked the target-word *access* considering time- α . The first collocates were consecutively *easy* (freq=399, MI=5.29), *free* (freq=382, MI=4.48) and *gain* (freq=250, MI=6.03). In all cases, there is no immediate relationship with internet-related words with the three most frequent collocates for *access*. Differently, when time- β were considered, the first three collocates for *access* are *digital* (freq=439, MI=7.80), *open* (freq=107, MI=3.81) and *internet* (freq=67, MI=5.25). Within the three most frequent collocates for *access*, both *digital* and *internet* present closeness with web-related meaning.

After, we considered the target-word *feed* and its collocates in time- α . The first three collocates were sequentially *clothes* (freq=260, MI=9.82), *cattle* (freq=183, MI=5.75) and *hungry* (freq=159, MI=5.58). In this scenario, none of the three

collocates present a semantic connection with web-related words. In time-β, the primary collocates for *feed* were *Facebook* (freq=133, MI=6.95), *posted* (freq=1219, MI=7.86) and *animal* (freq=38, MI=6.047). In this case, only the third collocate has no clear connection with web-related semantic ideas.

The fourth target-word to be analyzed was *streaming*. From time-α, *streaming* had the three first collocates: *tears* (freq=482, MI=7.56), *eyes* (freq=333, MI=4.27) and *face* (freq=305, MI=4.28). There is no clear indication that those collocates refer to the context of web-related words. In time-β, the three first collocates were: *service* (freq=128, MI=7.30), *video* (freq=53, MI=6.10) and *Netflix* (freq=35, MI=7.82). The three collocates suggest a semantic context of web-related concepts. Even service, when put together with *streaming* relates to the internet.

The last target word for English was *cloud*. Regarding time-α, collocates were: *dust* (freq=901, MI=7.07), *smoke* (freq=6.88, MI=6.42) and *black* (freq=504, MI=3.74). None of the collocates suggest connection with web-related words. Differently, in time-β the collocates were: *red* (freq=93, MI=6.28), *computing* (freq=29, MI=8.70) and *services* (freq=25, MI=4.77). As it is suggested, with the exception of *red*, the collocates for *cloud* in time-β resemble a web-related milieu. Data analysis from English target words are summarized in Table 2.

Table 2: Summary of the target-words, collocates and MI score for Portuguese⁷

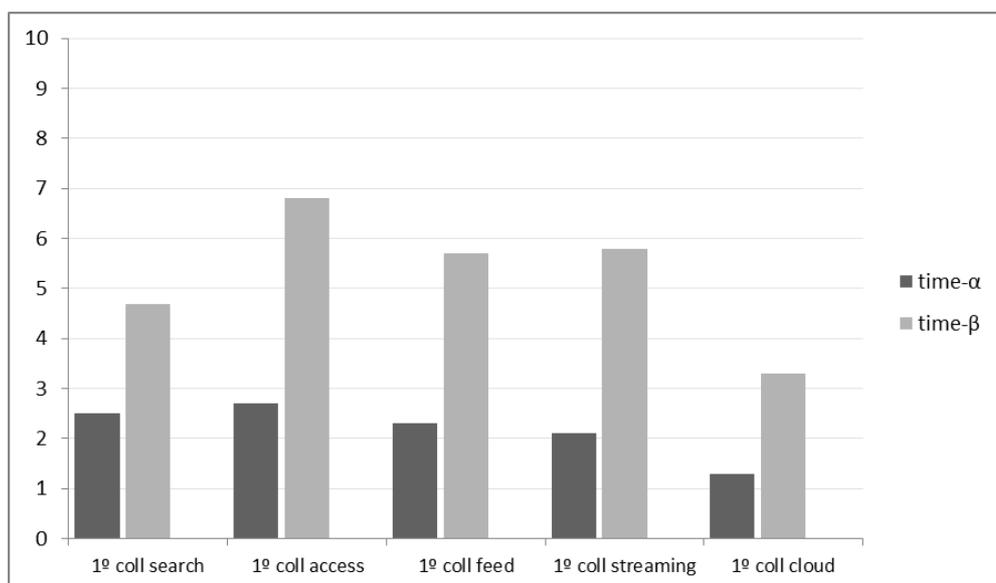
<i>target words</i>		<i>collocates</i>	<i>MI</i>	<i>target words</i>		<i>collocates</i>	<i>MI</i>
search	time-α	truth	3.1	streaming	time-α	tears	7.5
		warrant	8.2			eyes	4.2
		object	3.4			face	4.2
	time-β	warrant	8.2		time-β	servisse	7.3
		engine	7.7			video	6.1
		results	3.7			Netflix	7.8
access	time-α	easy	5.2	cloud	time-α	dust	7.0
		free	4.4			smoke	6.4
		gain	6.0			black	3.7
	time-β	digital	7.8		time-β	red	6.2
		open	3.8			computing	8.7
		internet	5.2			services	4.7
feed	time-α	clothes	9.8	feed	time-α	cattle	5.7
		hungry	5.5			hungry	5.5
		Facebook	6.9			Facebook	6.9
	time-β	posted	7.8		time-β	posted	7.8
		animal	6.0			animal	6.0

Source: Authors'elaboration

⁷ Correction: instead of *servisse*, read *service*

Following, we checked the MI score of the 3 most frequent collocates for the 5 target words in time- α in comparison with the 3 most frequent collocates for the 5 target-words in time- β in English. The objective was to verify whether both groups are statistically different. A paired-sample t-test revealed the difference between time- α (M=5.6, SD=1.2) and time- β (M=5.4, SD=1.8) concerning the MI scores revealed $t(14) = .38, p > .05$, which suggest there is no effect between groups. This result revealed that collocates attached to our target-word in time- α have shown a similar strength to those in time- β . However, as a result of time passing, this word companionship did not remain, since the position of these collocates changed considerably in time- β . Figure 2 is illustrative of the percentage of occurrence for the first position of the 3 most frequent collocates to exemplify:

Figure 2: Percentage of the first position for the 3 most frequent collocates in English



Source: Authors' elaboration

As we note in Figure 2, the percentage of occurrence for the first most frequent collocates for English considering our five target words remain different from time- α to time- β . In all cases investigated, the percentage of occurrence in time- β is superior to time- α which is an evidence that the attachment of these collocates seem to be stronger as time goes by.

Conclusion

Data from this study have been suggesting meaningful changes to words related to internet milieu when time is considered and regarding collocation. Similar results

have been found for both Portuguese and English. In Portuguese, target words such as *perfil* has revealed that in the past time, words such as *usuário* or *linhas* used to be highly frequent as collocates. Both words were not related to technology, as well as the subsequent collocates. However, when it comes to present time, words such as *Instagram* or *Twitter* appears as the most frequent. One could argue that such change is due to the sources corpora are formed from (websites/chats, etc.) and that social media has become popular recently. However, not just collocates have changed enormously but the old collocates decreased significantly regarding frequency of appearance.

Such inversion of frequency for these collocates that ‘lost’ position with the target words agrees to how the language has been used. Those collocations therefore exposed reflects in which contexts the words have been used for, although these words are still being spoken by speakers of the language, they are not as relevant as they used to be and this reason justifies that its use is even thought in different and several ways and aspects, change within the habitat of the language. In the XXI century, technology enfolds everything around, social media has become the center of relations and daily life of people, language follow its relation naturally.

When we compare target-words easily translated between languages, such as the pair ‘*nuvem/cloud*’ (Portuguese/English), we noticed that the words, when taken with their collocates, behaved somehow similar. For *nuvem* the three first collocates in the past time were not related to technology/internet (*negra, poeira and céu*). Similarly, in English the most frequent collocates for *cloud* in the past were not related to internet (dust, smoke, black). Moreover, we can note some similarity between the collocates in both languages (*negra/black, poeira/dust*). When the present time was taken into account, both *cloud* and *nuvem* presented internet-related collocates (*computação, armazenamento, computing, services*). Such evidence demonstrate that the advance of technology seems to interfere in the most common use of some words regardless the language.

These changes in frequency noticed through the analysis of collocates associated with words that are now related to the internet was brought by Brandão & Silva (2020) who pointed out that some words, even some of them that have always been considerably frequent in the English language, have grown to be much more frequently used after the popularization of the internet access in the first half of the 1990’s and some of the collocates before strongly related to these words have lost their positions

and are now, in most cases, much less frequent in this association (some aren't even displayed anymore in the corpora list of collocation relation with the target word).

As data revealed, the three most frequent collocates for the target words when present time was considered not only demonstrated a clear change towards internet-related meaning, but also showed that the strength of attachment of these words nowadays are considerably higher, what confirms our hypothesis. This perception of semantic change towards collocation is also raised by Gablasova, Brezina and McEnery (2017) when they point that there are traditionally a twofold taxonomy of collocability when it comes to language learning: absolute frequency and strength of association between word combinations. For those criteria the target words define how the language with function in the time exposed.

From an applied linguistics perspective and considering English (as L2) teaching, these data have shed light on the content teachers approach when teaching vocabulary. It was shown the importance of taking into account some aspects of semantic change across the years in the teaching practice. By delving into the changes some words present due to social changes, teacher will be able to update these meanings and discuss these changes.

References

- CRYSTAL, D. **The Effect of New Technology on English**. 2013. Disponível em: <<https://eltbookshelf.wordpress.com/2015/05/24/david-crystal-the-effect-of-new-technologies-on-english/>>. Acesso em: 29 maio.2020.
- DAVIES, M. The Corpus of Contemporary American English as the first reliable monitor corpus of English, *Literary and Linguistic Computing*, Volume 25, n. 4, p. 447–464, 2010. Disponível em <<https://academic.oup.com/dsh/article-abstract/25/4/447/997323>>. Acesso em: 29 maio.2020.
- PERFIL In: DICIO, Dicionário Online de Português. Porto: 7Graus, 2020. Disponível em: <<https://www.dicio.com.br/perfil/>>. Acesso em: 29 maio.2020.
- GABLASOVA, D.; BREZINA, V.; MCENERY, T. Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. **Language learning**, v. 67, n. 1 p. 155-179, 2017. Disponível em: <<https://onlinelibrary.wiley.com/doi/10.1111/lang.12225>>. Acesso em: 20 set.2020.
- JIA-YING, H.; YUAN-YUAN, S. A Corpus-based Study on Collocation and Colligation of "Soil" in Agricultural English. **Journal of Northeast Agricultural University (English Edition)**, v. 21, n. 4, p. 86-92, 2014. Disponível em: <<https://www.sciencedirect.com/science/article/abs/pii/S1006810415300246>>. Acesso em: 20 set.2020.

LAUFER, B.; WALDMAN, T. Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. **Language Learning**, 2011, v. 61, n. 2, p. 647-672, 2011. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9922.2010.00621.x>>. Acesso em: 20 set.2020.

MCENERY, T. et al. Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. **Annual Review of Applied Linguistics**, 2019, v. 39, p. 74-92, 2019. Disponível em: <<https://www.cambridge.org/core/journals/annual-review-of-applied-linguistics/article/corpus-linguistics-learner-corpora-and-sla-employing-technology-to-analyze-language-use/9043EF90E605FCE276CEF6C375A4C9C8>>. Acesso em: 20 set.2020.

MCENERY, T. et al. Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use, **Annual Review of Applied Linguistics**, 39, p. 74-92, 2019. Disponível em: <https://www.researchgate.net/publication/334632975_Corpus_Linguistics_Learner_Corpora_and_SLA_Employing_Technology_to_Analyze_Language_Use>. Acesso em: 20 set.2020.

ROBINSON, M., Duncan, D. Holistic Approaches to Syntactic Variation: Wh-all Questions in English, **University of Pennsylvania Working Papers in Linguistics**, v. 25, n. 1, 2019. Disponível em: <<https://repository.upenn.edu/pwpl/vol25/iss1/23>>. Acesso em: 29 maio.2020.

STUBBS, M. **Words and Phrases: Corpus Studies of Lexical Semantics**. Blackwell Publishers, 2001, p. 1-267.

POZNANSKI, V. et al. An apparatus for, and a method of, identifying collocates in order to distinguish readily between different collocations. **PATENTE** número 2002/0007266A1, registrada em 03 de setembro de 1998. Disponível em: <<https://www.freepatentsonline.com/y2002/0007266.html>>. Acesso em: 20 set.2020.

VIOLA, L. A corpus-based investigation of language change in Italian: The case of grazie/ringraziare di and grazie/ringraziare per. **Journal of Historical Linguistics**. v. 7, n. 3, p. 372-388, 2017. Disponível em: <https://www.researchgate.net/publication/322727266_A_corpus-based_investigation_of_language_change_in_Italian_The_case_of_grazieringraziare_di_and_grazieringraziare_per>. Acesso em: 20 set.2020.

BRANDÃO, L. M; SILVA, J. S. Transformações léxico-semânticas correlatas à influência da internet. **Trama**, v.16, n. 37, p. 4-17, 2020. Disponível em: <<http://e-revista.unioeste.br/index.php/trama/article/view/23604>>. Acesso em: 20 set.2020.

ZAABALAWI, R. S. (2020) English Collocations versus Arabic Collocations: A Pedagogical Dimension. **The Reading matrix: An international online journal**, v. 19, n. 1, p. 167-180, 2019. Disponível: <<http://www.readingmatrix.com/archive/19/1>>. Acesso em: 20 set.2020.