FREEDOM OF COMBINATION AND HETEROGENEITY: A CORPUS LINGUIST'S LOOK AT TWO SAUSSUREAN INSIGHTS

Tony Berber Sardinha[1]
(São Paulo Catholic University)

ABSTRACT

This article offers a reexamination of two of Saussure's insights from the point of view of corpus linguistics—namely, freedom of combination and heterogeneity in language in use. Regarding the first insight, an analysis of word combinations in a corpus of newspaper texts written in Brazilian Portuguese was carried out to determine how many of these combinations were actual collocations—that is, were used frequently enough in a very large reference corpus (the Brazilian corpus) to warrant statistical significance. The results suggested that most word combinations are not free; rather, they follow previously established preferences among speakers. Regarding the second notion, that of heterogeneity, the collocations in the newspaper texts were tracked as they were deployed one after the other along each text, and this flow was visually depicted. The inspection of the charts revealed unique patterns of the distribution of collocation, thereby suggesting that the evidence supports the view of heterogeneity. A cluster analysis was later conducted on the amount of collocations in each text, revealing three basic collocation bands onto which all the texts can be fitted. This was interpreted as suggesting that heterogeneity, despite being present and noticeable, is constrained rather than limitless. The article concludes that the methods and techniques afforded by present-day corpus linguistics can shed light onto Saussure's many valuable insights.

KEYWORDS: Saussure, corpus linguistics, collocation, freedom of combination, heterogeneity

PALAVRAS-CHAVE: Saussure, linguística de corpus, colocação, liberdade de combinação, heterogeneidade

> If corpus linguistics is there to stay we have to present it as a novel way to look at language, to discuss language as a social phenomenon, and to establish it as the kind of *parole*-linguistics that Saussure failed to deliver. (TEUBERT, 2009, p.23)

1. Introduction

Saussure's work has inspired linguists for nearly a century and will continue to do so for years to come. His ideas are constantly being put to the test in the face of different theories being put forth and developed over the years and, more recently, compared to evidence drawn from electronic corpora, which are large collections of language use stored in electronic form. In this article, the main goal is to shed light on two of Saussure's observations from a corpus linguistics perspective (BERBER SARDINHA, 2004; BIBER, CONRAD, REPPEN, 1998; MCENERY, HARDIE, 2012), through evidence drawn from an electronic corpora of Brazilian Portuguese. The first statement is this:

> The characteristic of speech is a freedom of combination... (SAUSSURE, 1916/1986, p. 112 [172])

This assertion was made in the context of a discussion on syntagmas (in Harris's 1986 translation, whereas in Baskin's translation they are called syntagms). Saussure sees syntagmas as essentially fixed word sequences and, as such, they do not belong in speech because, in his view, speech is fundamentally defined by a freedom of combination:

> Words as used in discourse, strung together one after another, enter into relations based on the linear character of languages (cf. p. [103]). Linearity precludes the possibility of uttering two words simultaneously. They must be arranged consecutively in spoken sequence. Combinations based on sequentiality may be called syntagmas. (SAUSSURE, 1916/1986, p.121 [170])

Similarly, sentences are also part of speech because they vary extensively from one another:

> If we think of all the sentences which could be uttered, what strikes us most forcibly is the lack of resemblance between them. (SAUSSURE, 1916/1986, p.104 [148])

Following this reasoning, we would predict that texts are free combinations of words and that little resemblance exists among texts. The notion that linguistic units join together freely in language use has been challenged in corpus linguistics, mainly by John Sinclair and his followers (BAKER, FRANCIS, TOGNINI-BONELLI, 1993; BARNBROOK, KRISHNAMURTHY, MASON, 2013; CLEAR, BAKER, FRANCIS, TOGNINI-BONELLI, 1993; CLEAR, FOX, FRANCIS, KRISHNAMURTHY, MOON, 1996; HERBST, FAULHABER, UHRIG, 2011; HUNSTON, 2002; JONES, SINCLAIR, 1973; MACKIN, 1978; MOON, 1998; RENOUF, SINCLAIR, 1991; SINCLAIR, 1966; 1987; 1991; SINCLAIR, JONES, DALEY, 1970/2004; STUBBS, 1996); (STUBBS, 2011; TOGNINI-BONELLI, 2001), according to whom language users base their language production to a large extent on prefabricated units. Ample evidence exists of the recurrence of word combinations in corpus-based studies, which has been taken to suggest that a language-organizing principle is in fact at play and responsible for the ubiquity of recurring word sequences, which came to be known as the phraseological view of language, in which is couched the idiom principle:

> the principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments (SINCLAIR, 1991, p. 110)

Similarly, Bolinger (quoted in WRAY, 2002, p.8) stated that:

our language does not expect us to build everything starting with lumber, nails, and blueprint, but provides us with an incredibly large number of prefabs. (BOLINGER, 1976, p.1)

Recurrent word combinations have received many different labels in corpus linguistics, such as collocations, multi-word units, and lexical bundles. In this study, I use collocation as the unit for investigating the notion of freedom in language in use. In Sinclairean corpus linguistics, collocations are regarded as the building blocks of language use, as units of meaning that reflect the famous Firthian quote of "you shall judge a word by the company it keeps" (FIRTH, 1957/1968, p.11), which ultimately states that the juxtapositions of words in text are not mere tokens of stylistic preference, but rather

key units of meaning. As meanings are shared by individuals in society, these meanings are patterned into chunks of words that are also shared by language users in particular contexts. Chunks become primed in the users' heads with repeated encounters (HOEY, 2005), thereby improving their chance of coming up more often in one's speech or writing. Corpora, being samples of language in use, incorporate many of these instances, and when analysts peruse the corpus, they come across lots of instances of the pattern, thereby noting and cataloguing collocation. As can be seen, collocation is discovered by the linguist only after it has been accepted in society as a token of discourse and its existence captured in a corpus. Firth might have been the one who envisaged the concept of collocation, but as Hoey (2009, p.34) defines it, Sinclair was the discoverer of collocation and the one who saw it in action in corpora.

Thus, we have two competing accounts of language in use: one in which freedom of combination predominates (Saussure), and the other in which freedom of combination is limited (Sinclair). Spoiler alert: As is typical of bottom-up corpus linguistic research, the answer will certainly not be "all or nothing" (i.e., complete freedom or an absence of freedom), but rather somewhere in between. The real question is therefore not whether freedom in language production exists, which of course it does (I can say or write anything I want if I don't mind the consequences, which people normally do because they live in society and language is a social phenomenon), but rather what the degree of freedom is. To investigate this issue empirically, I will make use of corpus linguistics techniques, including two corpora and a host of computer tools to extract collocations from one corpus and compare these with the other.

2. Method

The method consists of comparing all the collocations in individual texts with the collocations in a language-representative reference corpus (that does not contain the individual texts) to determine to what extent the authors of the individual texts used word combinations that match the collocations present in the reference corpus. The reasoning is that each word sequence in the individual text that matches a collocation in the comparison corpus signals a lack of freedom of combination because the author chose (consciously or not) a lexical sequence that others had already chosen multiple

times. Therefore, the author's choice, although in principle free of constraints of word combination,[2] was actually constrained by the repertory of pre-existing lexical combinations. In theory, the author could have chosen any word combination that he/she desired, but due to the many constraints acting on his/her choice, the preference fell on a particular collocation that is typical of the language.[3]

On the other hand, if the author used a particular word combination that has no counterpart in the comparison corpus, meaning that word sequence is not frequent enough to be considered a collocation, it was interpreted as a token of the freedom of combination. In this case, the author had a number of choices with which to express a particular proposition and chose a lexical formulation that does not occur frequently enough in the language in use to qualify as a collocation. There could be several reasons for this, including creativity on the part of the author (the combination is so unique that it has not been used by anyone else in the reference corpus) and specialization (the combination is so technical or specific to a particular context that its uses are not common for most speakers). A clarification is needed: The comparison was made with the attested collocations in the reference corpus, not with any word combinations, meaning that only those word sequences that met the statistical threshold of co-occurrence (measured by logDice) were valid as reference corpus units; hence, a word combination in the text might well be found in the reference corpus, but its frequency is not high enough for it to qualify as a collocation. Such cases are also considered to be a freedom of choice because the parameter is collocations, not mere co-occurrence, as argued below.

The first corpus is the newspaper register subcorpus of the Corpus Brasileiro de Variação de Registro (Brazilian Register Variation Corpus; CBVR), and the second is the Corpus Brasileiro (Brazilian Corpus). The newspaper corpus contains 20 newspaper stories, published in Portuguese in Brazil in national newspapers, totaling 11,467 words; the Brazilian Corpus is a register-diversified corpus with 1.1 billion tokens (see Table 1) of Brazilian Portuguese from different sources. The newspaper corpus is not included in the Brazilian corpus. The Brazilian Corpus is available on both Sketch Engine (sketchengine.co.uk, see KILGARRIFF, JAKUBÍÈEK, POMIKALEK, BERBER SARDINHA, WHITELOCK, 2014) and Linguateca (linguateca.pt, see SANTOS, 2014).

Subcorpus*	Tokens	%
Theses and dissertations	310,972,387	28.58%
Articles	258,585,002	23.76%
Newspapers	253,732,527	23.32%
Education, various	89,398,389	8.22%
SESSIONS OF CONGRESS	77,139,578	7.09%
Wikipedia	45,910,768	4.22%
Reports and manuals	13,742,224	1.26%
Legislation, various	9,097,447	.84%
Literature, various	8,659,955	.80%
Conference proceedings	6,947,244	.64%
INTERVIEWS	4,003,975	.37%
STATE SENATE PROCEEDINGS	3,977,450	.36%
PRESIDENTIAL SPEECHES	1,803,404	.17%
Religion, various	914,786	.08%
Bible	859,004	.08%
Manuals	708,239	.07%
Biographies	534,965	.05%
Magazines	494,974	.05%
SCREENPLAYS	289,389	.03%
Essays (crônicas)	160,525	.01%
Drug labels	113,228	.01%
SOCCER BROADCASTS	86,323	.01%
Short stories	60,777	.01%
TELEVISED PRESIDENTIAL DEBATES	22,033	<.01%
Horoscopes	4,319	<.01%
Total	1,088,218,912	100%

*Subcorpora in UPPER CASE are spoken.

Table 1: Composition of the Brazilian Corpus

It is crucial to distinguish between simple co-occurrence and collocation; I will do so here based on Hoey (2005, p. 3), who stated that "[w]henever I need to refer to the occurrence of two or more words within a short space of each other, I shall talk of 'lexical co-occurrence.'" Thus, the mere presence of words near others in a text (e.g., two words standing three words apart from each other) is a lexical co-occurrence, but if their frequency of occurrence is such that it matches or exceeds a critical value (measured by the logDice statistic offered by the SketchEngine tool at sketchengine.co.uk), then a collocation is formed. In this paper, collocation is defined as "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey 1991, pp. 6-7).

The word sequences were identified in the newspaper texts as follows. Each sentence of a text (a unit of speech, according to Saussure) was broken down into sequences of up to 11 running words, maintaining the same order as they appear in the text. For each window, a single word is selected as the node word (the focus word for the window), and the other words occurring in the window are seen as collocate candidates. They are called collocate candidates because the status of collocation depends on the statistical association with a node, which was determined in the reference corpus.

The basic algorithm is as follows:

-for each text
-for each sentence
-for each window
(i.e., a sentence segment up to 11 words long)
-grab the node word
(i.e., the word at the center of the window)
-grab the collocation candidates
(i.e., the remaining words in the window)
-for each valid node word
(i.e., content word nodes, except proper nouns)
-run a concordance for it in the Brazilian Corpus
(i.e., of up to 10,000 lines)
-extract its collocates
(i.e., the top 2,000 collocates, sorted by logDice)
-compare these collocates to the collocate candidates
-count number of matches

The algorithm was implemented using scripts the author wrote in Unix Shell and Python. The searches in the Brazilian Corpus were carried out using the SketchEngine API. For more details on the computational processing of the data, see Berber Sardinha (2014). In the first part of the processing, sentences are broken down into "windows" (strings of words) of at most 11 tokens. As each window "moves along" each sentence (or as the sentence "slides through" the window), the center word is singled out as the node and the remaining words are considered potential collocates. The node word is paired with each of these potential collocates, and a match for this pair is sought in the reference corpus.

To illustrate, let's take sentence 12 from text 5 (Example 1):

(1) Chirac afirmou ainda não ver necessidade para nova resolução da ONU, conforme os EUA defendem, que autorize claramente o uso da força contra Bagdá – mas ele não quis responder se Paris usaria seu poder de veto contra ela.'

(Chirac said further that he sees no need for a new UN resolution, like the US is pushing for, which will clearly authorize the use of force against Baghdad – but he would not say if Paris would use its veto power against it.)

The first window begins with the first word of this sentence (Chirac) and moves one word to the right, where it stops; the first window has two words: Chirac afirmou (Chirac said). This word combination is then looked for in the reference corpus list of collocations. If found, a match is recorded; if not, a no-match is tallied. The window is open further incrementally by one token at a time, and a comparison with the reference corpus is carried out for each new window, until 11 tokens are entered in the window, at which time a new window begins (this time with the word "afirmou" (said)). Notice that preposition-plus-article contractions in the original text (da, by the) were split up (de a) by the PALAVRAS tagger (which was used to tag the CBVR corpus, of which the newspaper subcorpus is part) and therefore count as two window tokens (Example 2):

(2) Chirac afirmou ainda não ver necessidade para nova resolução de a

(Chirac said he sees no need for the new resolution by the)

When the window reaches a sentence boundary, its right side stops moving, while at the same time the left-hand side continues to push forward, which in effect reduces the size of the window one word at a time. Examples (4) and (5) are the last two windows on that sentence:

(3) veto contra ela.

(4) contra ela.

Collocate candidates are lemmatized (transformed to their base form) and assigned a part of speech by the tagger. Only candidates having a valid part of speech (PoS) category (nouns, verbs, adjectives, adverbs, numerals) were considered as nodes or collocates. For instance, here is a window and its collocates (Example 6):

(5) Window #160, Sentence #12, text #5:

Chirac afirmou ainda não ver necessidade para nova resolução de a

Node: necessidade (need)

Collocate candidates: afirmou (said), ainda (still), Chirac, nova (new), não (not), resolução (resolution), ver (see)

The comparison with the reference corpus is processed as follows. For each window node in each newspaper text, a concordance is run in the Brazilian Corpus for its node lemma, extracting at most 10,000 lines (for tokens with a frequency lower than that, all occurrences are included; for node lemmas with more than 10,000 occurrences, a random sample is taken). A collocate table is generated for the concordance in SketchEngine, including all lemmas occurring three times or more, and the LogDice statistic is calculated. Therefore, going back to the previous example, the collocates of the lemma necessidade (the noun "need") in that window are matched to the collocates of the lemma necessidade in the Brazilian Corpus: necessidade + afirmou, necessidade + ainda, etc. The matching words are stored in a file, and the count of matches is recorded and transformed into a percentage (e.g., if four matches are found in a 10-word window, the value of 40% is recorded for that window).

A script was used to calculate the average percentage of collocates shared between the text and the reference corpus by averaging out the percentage of shared collocates of all windows in a particular text. This percentage represents the average incidence of collocations in the text. If it is higher than zero, it indicates a lack of freedom of combination; if it is equal to zero, it indicates an absence of collocation (freedom of combination).

3. Freedom of combination

Collocation can be quantified in the texts in two different ways. The first way involves computing the average percentage of collocation across all windows. In this case, the average percentage of collocations is 60.8% (see Table 2).

Text	Windows	Collocations	Std. Deviation
1	388	57.0%	33.6
2	438	73.0%	30.2
3	238	53.6%	32.2
4	254	54.9%	34.0
5	260	62.4%	34.6
6	228	60.7%	32.8
7	305	67.6%	30.8
8	213	62.0%	31.0
9	255	58.6%	32.1
10	295	55.1%	33.9
11	363	55.0%	30.2
12	329	65.5%	30.8
13	264	62.3%	31.6
14	279	49.6%	30.6
15	244	56.5%	32.8
16	293	67.1%	33.1
17	382	59.9%	32.7
18	240	60.5%	30.4
19	246	66.0%	30.6
20	291	61.6%	31.9
Total	5805	60.8%	32.5

This is an estimate of the richness of collocation: Approximately 6 out of every 10-word combinations encountered in the texts are collocations in Brazilian Portuguese. The following example shows a window with a collocation density comparable to the corpus average of 61% (window #327 from text 7; the portion in brackets is outside the window, but quoted here for completeness).

(7) ... [um balneário que] faz festa boa, mas também sabe planejar, sabe executar obras.

(a resort that can throw a good party, and also knows how to plan ahead and renovate)

The node sabe (knows) has eight candidates in its neighborhood, five of which are attested collocates in the corpus (boa, good; faz, do/make; festa, party; planejar, to plan; também, also), resulting in 62.5% density.

This figure is slightly higher than the finding in Erman and Warren's (2000, p.37) study, which reported that 55% of the words were part of a prefab, in different English texts. Important differences exist between their study and this one (apart from the language of the texts). First, their study used the concept of prefab, rather than collocation:

> a prefab is a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization. (ERMAN, WARREN, 2000, p.31)

The prefab candidates were then subjected to the criterion of restricted exchangeability:

> By restricted exchangeability is meant that at least one member of the prefab cannot be replaced by a synonymous item without causing a change of meaning or function and/or idiomaticity. (ERMAN, WARREN, 2000, p.32)

Thus, unlike this study, not all word combinations were taken into account; only those that met the criterion of restricted exchangeability were considered. It is not possible to ascertain what effect this might have had on the results if all word combinations were taken into consideration. It is certain, though, that the sample of prefabs was restricted by this criterion.

Secondly, Erman and Warren's study did not compare each candidate with a reference corpus, as done here; rather, they judged

the status of each candidate by hand, based on their interpretation of the restricted exchangeability criterion. Again, it is hard to predict the impact this might have had on the findings. Yet the perils of trusting one's intuitions about collocability have been exposed in the literature (SAMPSON, 1997; SINCLAIR, 1991). Finally, the corpus used in their study differs was not register-specific, unlike ours. It contained 19 samples of texts: seven spoken ones from the London-Lund Corpus of Spoken English, ten from the Lancaster-Oslo-Bergen corpus, and two samples from a children's story. Yet the total word count for the corpus was 10,246 words, which is similar to our corpus. Given these crucial differences between the two studies, it is surprising that the two percentages are so close at all. Nevertheless, more research is needed before a better understanding of the presence of collocations and prefabs is reached and generalizations can be made about the expected rates of collocation in texts.

The second way in which collocation was quantified was by counting how many windows have at least one collocation in them. The result in this case is much higher: 89% of the windows include one collocation or more (see Table 3).

Text	Windows with at least one collocation	Std. Deviation
1	85.1%	35.70
2	92.9%	25.67
3	83.6%	37.09
4	84.7%	36.12
5	86.2%	34.61
6	87.3%	33.39
7	92.5%	26.45
8	92.0%	27.16
9	87.1%	33.63
10	84.4%	36.34
11	90.4%	29.56
12	92.1%	27.02
13	90.2%	29.85
14	86.4%	34.36
15	86.1%	34.70

16	89.8%	30.37
17	89.3%	30.99
18	93.8%	24.26
19	92.3%	26.75
20	90.0%	30.01
Total	88.9%	31.41

Table 3: Windows with at least one collocation

The two figures differ because the former is an average (lower percentages lower the average, higher percentages increase it) whereas the latter is a count (every window counts so long as it has more than 0% collocation). The latter figure suggests that, if we were to read the newspaper texts, we would likely find a collocation in nearly 9 out of every 10 windows. Table 4 illustrates one such case (sentence #27 for text 2) with 100% of the collocation opportunities fulfilled.

Node	Collocates
O	*
deputado	federal (federal)
federal	anunciou (announced), deputado (Congressman)
Raul Jungmann	*
(PMDB-PE)	*
anunciou	federal (federal), hoje (today), iniciar (start), irá (is going to)
que	*
irá	anunciou (announced), coleta (gathering), iniciar (start)
iniciar	anunciou (announded), assinaturas (signatures), coleta (gathering), hoje (today), irá (is going to)
hoje	anunciou (announced), assinaturas (signatures), iniciar (start), irá (is going to)
a	*
coleta	assinaturas (signatures), iniciar (start), irá (is going to)
de	*
assinaturas	coleta (gathering), Congresso (Congress), hoje (today)

no	*
Congresso	abertura (opening), assinaturas (signatures)
para	*
a	*
abertura	Congresso (Congress), investigar (investigate)
de	*
uma	*
CPI	*
para	*
investigar	abertura (opening), caso (case), grampos (wiretap)
o	*
caso	grampos (wiretap), investigar (investigate)
dos	*
grampos	caso (case)
na	*
Bahia.	*

* Function word or proper noun and, therefore, not a valid node.

Translation of the sentence: "Congressman Raul Jungmann (PMDB-PE) announced that he will start gathering signatures in Congress today to open a hearing into the Bahia wiretap case."

Table 4: A sentence with 100% collocation presence

The results challenge the notion of the freedom of combination: More often than not, the word combinations in the texts are conventional expressions that have already been uttered frequently, reflecting the fact that words tend to form collocations rather than combine freely with one another. If there were freedom of combination, most combinations would be novel, not attested in a reference corpus.

4. Heterogeneity within and across texts

The second observations by Saussure that I want to address are the following:

... language in general is heterogeneous. (SAUSSURE, 1916/1986, p.14 [32])

> Language in its totality is unknowable, for it lacks homogeneity. (SAUSSURE, 1916/1986, p.20 [38])

The results reported thus far cannot be used to examine this assertion because the figures refer to the average incidence of collocation across the whole corpus and not on a text-by-text basis, which would permit looking at heterogeneity intertextually. A further analysis is needed that addresses the occurrence of collocation along each text (window per window) and compares this distribution across the texts.

This aspect of language use has been addressed in corpus linguistics generally as variation, not heterogeneity, but arguably the two notions are akin in the sense in which they are considered here. Variation has been at the heart of corpus linguistics in general and at the core of the strand of corpus linguistics pursued by Biber (BIBER, 1988 et seq.) in particular. Biber has repeatedly argued for the centrality of register (text varieties defined by situational/contextual rather than linguistic characteristics) in linguistic description:

> ... it is still the norm that most studies of collocation and lexico-grammatical associations to disregard the possible influence of register differences. [...] we should instead treat this possible influence as a likelihood. [...] Thus, the practice advocated here is to begin a research study with the hypothesis that such register differences exist [...]. (BIBER, 2012, p.34)

In this study, I took heed of Biber's advice and restricted the analysis to a particular register instead of using a corpus of many different undistinguished registers (as often happens in corpus-based studies that aim to describe the language in general). On this general point of the need for considering variation/heterogeneity as central to language use, Saussure and modern-day corpus linguists of the Biberean persuasion go hand in hand. Another kind of variation, that which happens within texts, is also important as it presupposes that the flow of the units of linguistic use are not distributed uniformly in the sample. Rather, the flow has its ups and downs, peaks and valleys, as is normal with any complex system. This view of language use is reminiscent of Pike's wave view of language (PIKE, 1972). Saussure's point about the inherent heterogeneity in language use would not be inconsistent with this flow view of language or with the presupposition that units of language use are not distributed uniformly in a language sample, I would argue.

To investigate both of these points, the percentage of collocation use in each window was plotted in charts, each for an individual text, and the charts were examined. The charts appear in Figures 1 through

4 (in the Appendix; the numbers on the left-hand side reflect the collocation percentages). A visual inspection identified ample variation within each text and that the norm is for texts to have many peaks and valleys in collocation use. No two texts are alike in the distribution of collocation within them. In addition, an ANOVA was conducted, with the percentage of collocations per window being the dependent variable and text the independent variable. The results suggest significant cross-text variation (F=10.073, df=19, p=.000). This might be interpreted as evidence of (one kind of) heterogeneity, relative to the distribution of units in the flow of usage, which makes each text different from the other. Put another way, texts from the same register are heterogeneous with respect to the distribution of collocation units within them.

Having established that collocation in texts is distributed heterogeneously, the question that arises is whether groups of texts sharing similar incidences of collocation exist. A hierarchical cluster analysis was run on the collocation percentages, which generated a dendrogram showing the clustering process of the data (see Figure 5). The length of the line joining the texts reflects how similar or different their means are, with shorter lines signaling texts with comparable means and longer lines indicating disparate texts. The dendrogram pointed to the presence of three different clusters (Table 3).

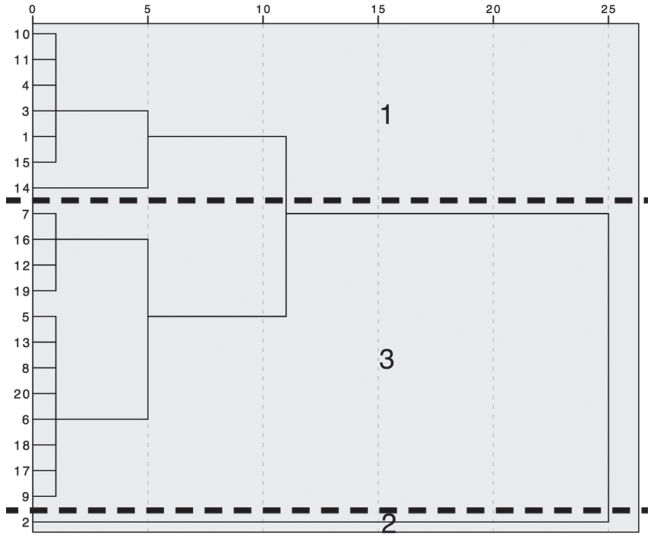

Figure 5: Cluster analysis dendrogram for the percentage of collocation in the texts

Table 5 shows that cluster 1 has seven texts, cluster 2 has one text, and cluster 3 has 12 texts, with means that are statistically different (F=29.954, df=2, p=.000): a low band (cluster 1, 54.5% collocation), a mid band (cluster 3, 62.8%), and a high band (cluster 2, 73%). Cluster 3 has the most texts (the mid band), with 60% of the texts. Importantly, these three clusters account for 77.9% of the variation in the texts (R^2=.779), indicating that these three clusters (or collocation bands) capture most of the heterogeneity across the texts and that the heterogeneity can be collapsed into just three groups.

Cluster	Collocation average	Std. Deviation	N
1 (low)	54.5%	2.46	7 (35%)
2 (high)	73.0%	—-	1 (5%)
3 (mid)	62.8%	2.97	12 (60%)
Total	60.4%	5.64	20

Table 5: Bands of collocation density across texts

4. Closing remarks

This paper examined two observations put forth by Saussure from a corpus linguistic perspective, the first of which claims a freedom of combination in speech (*parole*). Using word combinations as the unit of analysis, the results indicated that language users are constrained by the expected patterns of lexical combination in their choice of wording and generally tend to use conventional collocations attested in previous texts rather than creating unique combinations. This is interpreted as evidence that points toward a rejection of the notion of the freedom of combination as far as word sequences are concerned, thereby confirming previous corpus linguistic analyses (BARNBROOK, KRISHNAMURTHY, MASON, 2013; HERBST, FAULHABER, UHRIG, 2011; SINCLAIR, 1966) that underscored the prevalence of the idiom principle and of the phraseological view of language. Its competing formulation that would support the freedom of combination, the slot-and-filler principle (SINCLAIR, 1991), is much

less common (in texts like newspaper stories, at least) in actual language production. The results confirm that, although language users have in principle the full freedom to choose how they word their utterances, what actually happens in practice is a much more constrained environment, where typical choices are used over and over again.

In connection with the issue of the freedom of combination, a second claim was further addressed—namely, that of the heterogeneity in language use. The distribution of collocation within the texts was depicted visually and then analyzed statistically, showing that the texts were strikingly different from each other with respect to how the collocations follow each other within the sentences and from sentence to sentence. This was in turn interpreted as evidence that seems to confirm heterogeneity as an inherent property of language use, as predicted by Saussure. At the same time, further statistical analysis indicated the presence of three groups of texts, each with a different level of collocation use, thereby suggesting that the heterogeneity is not uninhibited, but rather constrained.

As previously mentioned, many other studies have also provided evidence that suggests a lack of freedom in word combinations (SINCLAIR, BOLINGER, MOON, ERMAN, PHILLIPS 1989, etc.). Where this study innovates is with respect to the method, because entire texts were taken into account rather than individual target words, and the judgment of the status of collocation was passed based on evidence taken from a large reference corpus, instead of by intuition. This enabled the verification of the rate at which whole texts employ typical word combinations and how these collocations (or the lack thereof) are distributed along the texts.

Linguistic observations made nearly a century ago did not enjoy the luxury of ample language data collections or the computer power such as we have today. When they are confirmed using today's techniques, these observations are even more remarkable. When they are not, they are important nonetheless because they will have spearheaded new investigations that might lead to new findings or new methods. Computers and the modern-day corpora are game changers in linguistics, but they can only find or attest what someone is looking for. The Course in General Linguistics is a treasure trove of keen insights that should be revisited by corpus linguists more often.

Appendix

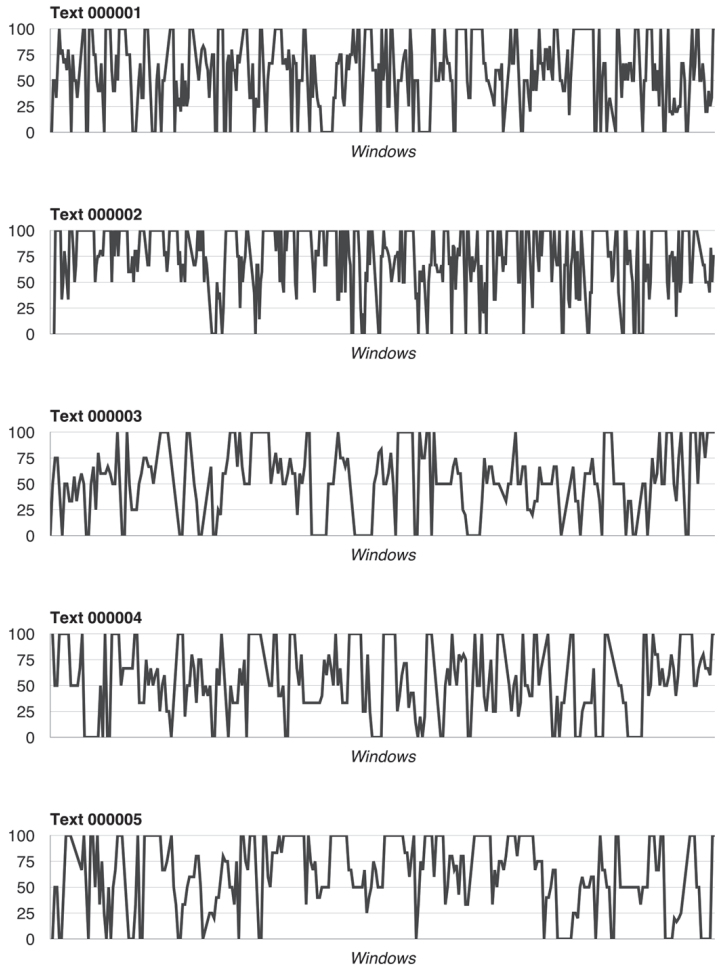

Figure 1: Distribution of collocation in texts 1 through 5

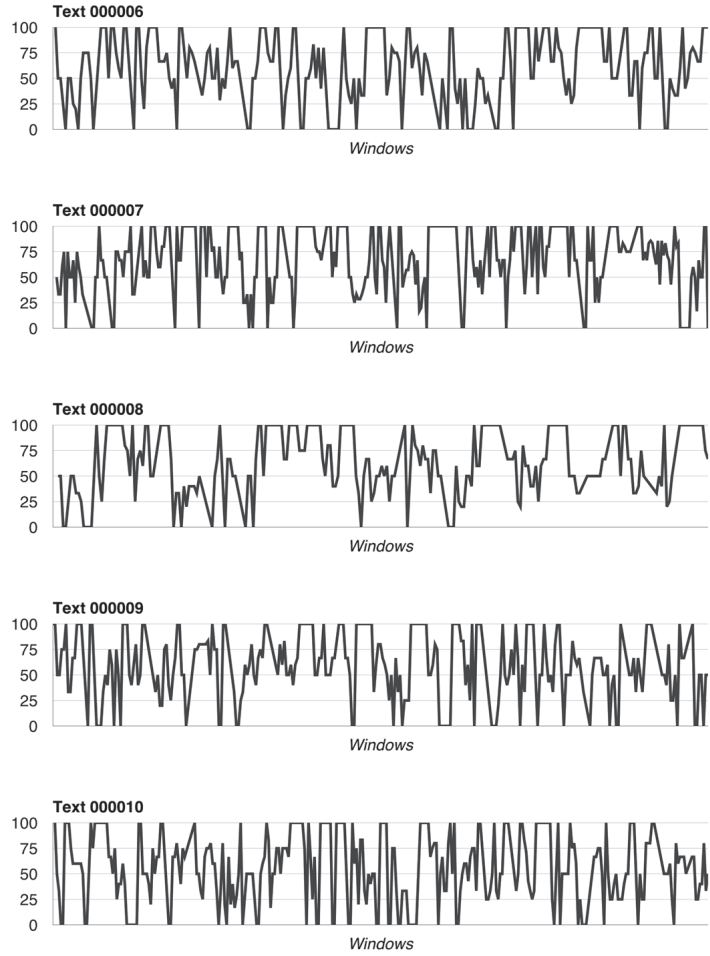

Figure 2: Distribution of collocation in texts 6 through 10

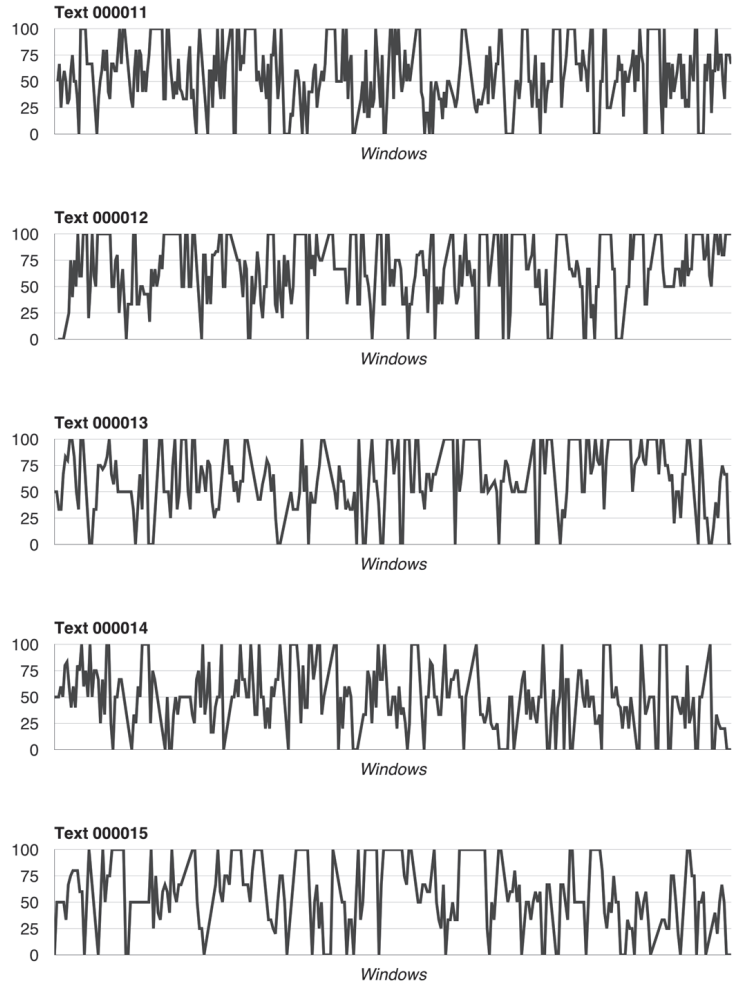

Figure 3: Distribution of collocation in texts 11 through 15

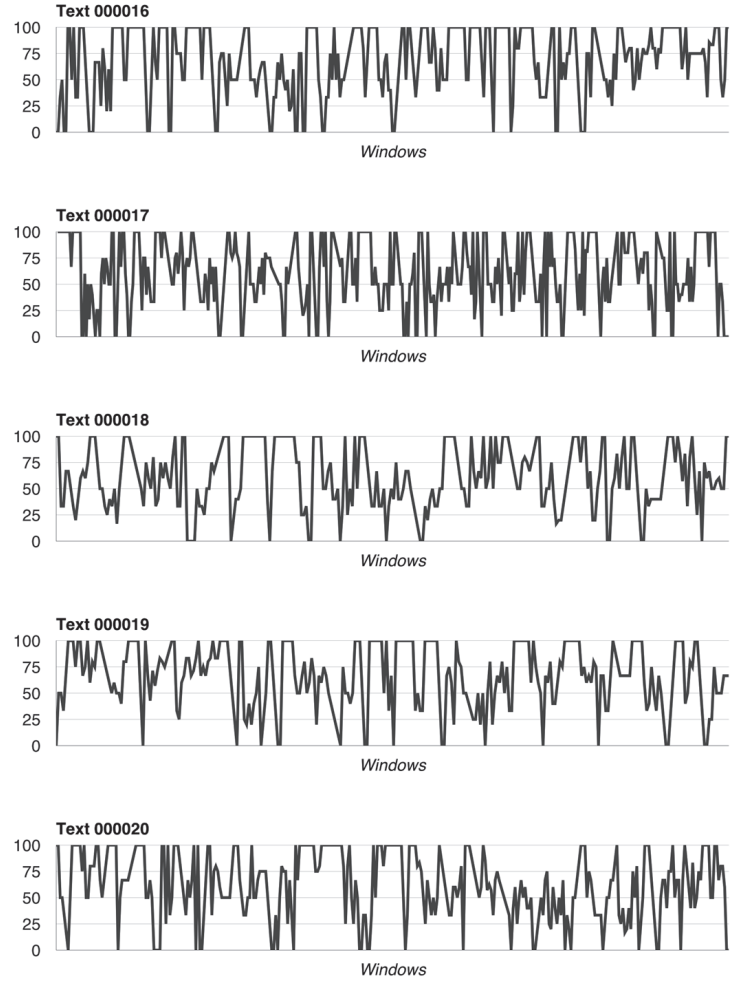

Figure 4: Distribution of collocation in texts 16 through 20

RESUMO

O artigo reexamina dois dos insights de Saussure a partir da perspectiva da linguística de corpus, a saber a liberdade de combinação e a heterogeneidade no uso da língua. Com relação ao primeiro, foi feita uma análise de combinações de palavras em corpus de textos de jornais para determinar quantas eram realmente colocações, isto é, quantas eram usadas com frequência suficiente num corpus de referência (o Corpus Brasileiro). Os resultados sugerem que a maioria das combinações de palavras não são livres, mas seguem preferências previamente estabelecidas pelos sujeitos falantes. Com relação à segunda noção - heterogeneidade - as colocações dos textos de jornal foram acompanhadas conforme eram empregadas uma após a outra ao longo dos textos, sendo esse fluxo capturado de forma visual. A inspeção dos diagramas revelou padrões únicos de distribuição de colocações, evidenciando dessa forma a visão da heterogeneidade. Uma análise de agrupamento foi feita sobre as colocações em cada texto, revelando três níveis de colocabilidade. Esses níveis indicaram que a heterogeneidade, apesar de presente e aparente, sofre coerções e tem limites. O artigo conclui que os métodos e técnicas da linguística de corpus atual podem iluminar muitos dos *insights* valiosos propostos por Saussure.

PALAVRAS-CHAVE: Saussure, linguística de corpus, colocação, liberdade de combinação, heterogeneidade

REFERÊNCIAS

BAKER, M.; FRANCIS, G.; TOGNINI-BONELLI, E. *Text and technology - In honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins, 1993.

BARNBROOK, G.; KRISHNAMURTHY, R.; MASON, O. M. A. *Collocation: Applications and Implications*. Basingstoke: Palgrave Macmillan, 2013.

BERBER SARDINHA, T. *Linguística de Corpus*. São Paulo: Manole, 2004.

_____. Looking at collocations in Brazilian Portuguese through the Brazilian Corpus. In: BERBER SARDINHA, T.; SÃO BENTO FERREIRA, T. (Ed.). *Working*

with Portuguese Corpora. London / New York: Bloomsbury/Continuum, 2014. p. 9-32.

BIBER, D. *Variation across speech and writing*. Cambridge: Cambridge University Press, 1988.

BIBER, D. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, v. 8, n. 1, p. 9-37, 2012.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus Linguistics - Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

BOLINGER, D. Meaning and memory. *Forum Linguisticum*, v. 1, n. 1, p. 1-14, 1976.

CLEAR, J.; BAKER, M.; FRANCIS, G.; TOGNINI-BONELLI, E. From Firth principles - Computational tools for the study of collocation*Text and technology - In honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins, 1993. p. 271-292.

CLEAR, J.; FOX, G.; FRANCIS, G.; KRISHNAMURTHY, R.; MOON, R. Cobuild: the state of the art. *International Journal of Corpus Linguistics*, v. 1, n. 2, p. 303-314, 1996.

ERMAN, B.; WARREN, B. The idiom principle and the open choice principle. *Text*, v. 20, n. 1, p. 29-62, 2000.

FIRTH, J. R. A synopsis of linguistic theory, 1930-55. In: PALMER, F. R. (Ed.). *Selected Papers of J. R. Firth 1952–59*. London: Longmans, 1957/1968. p. 168-205.

HERBST, T.; FAULHABER, S.; UHRIG, P. *The phraseological view of language: a tribute to John Sinclair*. Berlin ; Boston: De Gruyter Mouton, 2011.

HOEY, M. *Lexical Priming: A New Theory of Words and Language*. London, New York: Routledge, 2005.

_____. Corpus-driven approaches to grammar: The search for common ground. In: SCHULZE, R.; RÖMER, U. (Ed.). *Exploring the Lexis-Grammar Interface*. Amsterdam: John Benjamins, 2009. p. 34-48.

HUNSTON, S. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.

JONES, S.; SINCLAIR, J. M. English lexical collocations - A study in computational linguistics. *Cahiers de Lexicologie*, v. 23, n. 2, p. 15-61, 1973.

KILGARRIFF, A.; JAKUBÍÈEK, M.; POMIKALEK, J.; BERBER SARDINHA, T.; WHITELOCK, P. PtTenTen: a Corpus for Portuguese Lexicography. In: BERBER SARDINHA, T.; SÃO BENTO FERREIRA, T. (Ed.). *Working with Portuguese Corpora*. London: Bloomsbury, 2014. p. 111-130.

MACKIN, R. On collocations: Words shall be known by the company they

keep. In: STREVENS, P. (Ed.). *In honour of A.S. Hornby*. Oxford: Oxford University Press, 1978. Cap.149-166.

MCENERY, T.; HARDIE, A. *Corpus linguistics : method, theory and practice*. Cambridge: Cambridge University Press, 2012.

MOON, R. *Fixed Expressions and Idioms in English - A Corpus-Based Approach*. Oxford: Clarendon Press, 1998.

PIKE, K. L. Language as particle, wave, and field. In: BREND, R. M. (Ed.). *Kenneth L Pike - Selected writings*. Hague: Mouton, 1972. p. 129-143.

RENOUF, A.; SINCLAIR, J. M. Collocational frameworks in English. In: AIJMER, K.; ALTENBERG, B. (Ed.). *English Corpus Linguisitcs - Studies in honour of Jan Svartvik*. London: Longman, 1991. p. 128-144.

SAMPSON, G. *Educating Eve: The "language instinct" debate*. London ; Washington, DC: Cassell, 1997. (Open linguistics series).

SANTOS, D. Corpora at Linguateca: Vision and roads taken. In: BERBER SARDINHA, T.; SÃO BENTO FERREIRA, T. (Ed.). *Working with Portuguese Corpora*. London: Bloomsbury, 2014. p. 219-236.

SAUSSURE, F. D. *Course in general linguistics*. Tradução de HARRIS, R. Chicago and La Salle, IL: Open Court, 1916/1986.

SINCLAIR, J. M. Beginning the study of lexis. In: BAZELL, C. E. (Ed.). *In Memory of J R Firth*. London: Longman, 1966. p. 410-430.

_____. Collocation: a progress report. In: STEELE, R.; THREADGOLD, T. (Ed.). *Language Topics - Essays in Honour of Michael Halliday*. Amsterdam/ Philadelphia: John Benjamins, 1987. p. 319-332.

_____. *Corpus, Concordance, Collocation*. Oxford, New York: Oxford University Press, 1991. (Describing English language).

SINCLAIR, J. M.; JONES, S.; DALEY, R. *English Lexical Studies: The OSTI Report*. London/New York: Continuum, 1970/2004. (English Collocation Studies: The OSTI Report).

STUBBS, M. *Text and Corpus Analysis — Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell, 1996.

STUBBS, M. A tribute to John McHardy Sinclair. In: HERBST, T. *et al* (Ed.). *The Phraseological View of Language: A Tribute to John Sinclair*. Berlin: De Gruyter Mouton, 2011. p. 2-16.

TEUBERT, W. Aproximación a la Lingüística de Corpus y su contribución en la elaboración de diccionarios (Interviewed by Cristina Martin Herrero). *Cuadernos del Instituto Historia de la Lengua*, v. 3, p. 11-24, 2009.

TOGNINI-BONELLI, E. *Corpus linguistics at work*. Amsterdam ; Philadelphia: J. Benjamins, 2001. (Studies in corpus linguistics,, 6).

WRAY, A. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press, 2002.

NOTAS

[1] The author wishes to thank CNPq (Brasília, DF) and Fapesp for supporting this research study. Many thanks to Tania Shepherd for her support. An earlier version of the results included in this paper was presented elsewhere (BERBER SARDINHA, 2014).

[2] This is not to say that no constraints are acting on the choices, because they are, including syntactic, pragmatic, topical, and many others.

[3] The same collocation can be used in different ways, with different purposes, and therefore it is not meant here that by using an existing collocation, a language user is reproducing the same intentions, ideologies, and discourse characteristics or sharing the presuppositions or contexts of other users who employed the same collocation in different texts. The meanings emanating from collocations can be quite subtle and complex, but it is not the goal of this paper to examine these issues.

Recebido em 8 de maio

Aprovado em 22 de maio