GIL, C. B. Revisitando Sinclair: o princípio idiomático e o princípio da escolha aberta em um corpus de aprendiz. *ReVEL*, v. 21, n. 40, 2023. [www.revel.inf.br].

REVISITANDO SINCLAIR: O PRINCÍPIO IDIOMÁTICO E O PRINCÍPIO DA ESCOLHA ABERTA EM UM *CORPUS DE* APRENDIZ

Revisiting Sinclair: The Idiom Principle and the Open Choice Principle in Learner Corpora

Cristina Borges Gil[1]

cristinaborgesgil@gmail.com

RESUMO: Este artigo tem como objetivo principal detectar indícios do princípio idiomático e do princípio da escolha aberta na produção escrita de alunos brasileiros em inglês como língua estrangeira. A base teórica desta investigação é a Linguística de *Corpus*, uma área que proporciona a pesquisa, o estudo e a exploração da língua em uso e que se baseia na visão probabilística da linguagem. Sinclair (1991, 2004) considera a linguagem como sistema probabilístico a partir de dois princípios complementares: o idiomático e o da escolha aberta. A metodologia consistiu na coleta de um *corpus* de escrita de aprendizes brasileiros de inglês e do subsequente exame de todas as sequências de palavras de cada um dos textos do *corpus*, comparando-as com um *corpus* de referência representativo do inglês. Esse procedimento, conhecido por rastreamento de colocações, foi introduzido por Berber Sardinha (2014a). A análise dos resultados indicou que os dois princípios coexistem nos textos analisados, como aventado por Sinclair (1991). Além disso, também revelou que há nas redações dos aprendizes nuances nos dois princípios propostos por Sinclair (1991), que denominamos princípio idiomático tipo I e II, e princípio da escolha aberta tipo I e II. A pesquisa pretende dar uma contribuição original à Linguística de *Corpus*, assim como à Linguística de *Corpus* de Aprendiz, na medida em que foi realizada uma investigação descritiva da linguagem do aprendiz baseada em corpora e observado variantes dos princípios nos textos dos aprendizes que não se encontram em textos de falantes nativos letrados da língua.
PALAVRAS-CHAVE: linguística de *corpus; corpus* de aprendiz; princípio idiomático; princípio da escolha aberta.

ABSTRACT: The aim of this research is to find evidence of both the idiom principle and the open-choice principle in the written production of Brazilian students of English as a foreign language. The theoretical basis of this study is Corpus Linguistics, an area which supports the research and the study of language in use, and which is based on the view of language as a probabilistic system. Sinclair (1991, 2004) sees language as a probabilistic system with two complementary principles: the idiom principle and the open-choice principle. The methodology consisted of the collection of a written corpus of Brazilian students of English as a foreign language and the subsequent analysis of all the sequences of

[1] Doutoranda e mestre pelo Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem da Pontifícia Universidade de São Paulo – PUC-SP.

words used in each text of the corpus. This procedure, known as 'collocation tracking', was introduced by Berber Sardinha (2014a). The findings point out that the two principles coexist in the texts as proposed by Sinclair. In addition, they also reveal nuances in the principles described by him in the written production of the learners. We called them the idiom principle I and II, and the open-choice principle I and II. The study presented here intended to have made an original contribution to Corpus Linguistics and to the study of Learner Corpora as it carried out a descriptive investigation of learner language and observed variant forms of the principles which are not found in texts written by educated native speakers.
KEYWORDS: Corpus Linguistics; Learner Corpora; Idiom Principle; Open-Choice Principle.

INTRODUÇÃO

Este artigo é uma versão reduzida da dissertação de mestrado intitulada "A incidência do princípio idiomático e do princípio da escolha aberta na produção escrita de alunos brasileiros de inglês como língua estrangeira" (Gil 2017)[2].

Uma das questões fundamentais para o aprendiz de uma língua estrangeira é conseguir expressar-se de modo natural e fluente, seja na escrita seja na fala. De acordo com a Linguística de *Corpus* (doravante LC), a produção da linguagem não se dá apenas pela aplicação de regras gramaticais, como também através do uso de sequências linguísticas – por exemplo, sequências de palavras típicas de um dado contexto ou de situação de uso da língua (Sinclair 1991; Biber, Conrad, Reppen, 1998; Partington 1998; Hunston 2002; Berber Sardinha 2004). À medida que usa tais sequências, o indivíduo imprime à sua fala ou escrita um ritmo e uma qualidade "natural", pois respeita as expectativas dos seus interlocutores diante daquele contexto ou situação de uso. Ou seja, o uso de sequências pré-fabricadas – que existem no repertório comum dos interlocutores – cria um fluxo de fala ou escrita que respeita as expectativas mútuas desses interlocutores.

A essa sequência de palavras próprias a determinadas circunstâncias de uso deu-se o nome de colocação. Segundo Berber Sardinha (2014a), esse construto é uma das contribuições mais significativas da LC para a nossa compreensão da língua em uso (cf. O'Keefe, McCarthy, Carter 2007). Apesar de a noção básica de colocação como justaposição de palavras já existir há séculos (Barnbrook, Mason, Krishnamurthy 2013), contudo, "[...] foi necessário o surgimento de *corpora* eletrônicos para a colocação se estabelecer como um construto central na teoria

[2] Disponível em: https://repositorio.pucsp.br/jspui/handle/handle/19852. Acesso em fev. 2023.

linguística”[3] (Berber Sardinha 2014a: 9)[4]. A possibilidade de analisar *corpora* de forma sistêmica por meio de programas de computador abriu a chance de explorar a natureza associativa do léxico da língua.

Entretanto, ressalta Berber Sardinha (2014a), essa mudança de perspectiva também significou que perdemos de vista a importância da colocação na construção de textos. Embora *corpora* contenham textos, os limites entre eles são normalmente perdidos nessas coleções, (ou porque não foram levados em consideração durante a compilação ou porque o *software* de análise os ignora); então, com a maioria dos *corpora*, quando descobrimos colocações, ficamos com uma imagem da colocação na língua em vez de nos textos representados nessas coleções[5].

Tal fato leva Berber Sardinha (2014a) a concluir que o que vemos nos *corpora* são abstrações dos usos individuais das colocações em um dado contexto. Por extensão, podemos identificar grandes listas de colocações, embora estejam desconectadas dos textos em que foram originariamente usadas. Nisso há vantagens e desvantagens, salienta Berber Sardinha. Uma vantagem é que “[...] podemos abstrair informação a partir de exemplos individuais para obtermos uma imagem da língua como um todo, ou de uma variedade da língua”[6], na medida em que há em comum entre esses exemplos um padrão de uso linguístico. Por outro lado, uma desvantagem é que podemos perder de vista “[...] a importância que cada exemplo individual tem na constituição do próprio texto em que é encontrado”[7] (2014a: 10). A presença de colocações nos textos é, em suma, uma questão importante tanto para a linguística em geral quanto para a LC em particular, “[...] uma vez que o seu uso no texto é visto

[3] “[...] it took the advent of electronic corpora for collocation to be established as a central construct in linguistic theory. Electronic corpora and the tools needed to explore them revealed the systematic nature of collocation as well as its pervasiveness in language.” (Tradução minha. Doravante todas as traduções são feitas pela autora.)

[4] Tradução minha. Doravante, todas as traduções são feitas pela autora.

[5] “Although corpora contain texts, the textual boundaries are normally lost in these collections (either because they were not taken into account during compilation or because the analysis software ignores them); thus, with most corpora, when we retrieve collocations we end up with a picture of collocation presence in the language rather than in the texts represented in those collections.”

[6] “[...] we can abstract information from the individual instances to achieve a picture of a whole language or a variety, for instance.”

[7] “[...] the importance that each individual instance plays in the constitution of the actual text in which it is found.”

como uma característica definidora de um texto natural e fluente [...]"[8] (Berber Sardinha 2014a: 11).

Partington (1998: 15) relata que o termo colocação, "[...] tal como é conhecido, foi cunhado em seu moderno sentido linguístico pela primeira vez pelo linguista britânico Firth"[9]. Barnbrook, Mason e Krishnamurthy (2013: 36) ressaltam que Firth tornou possível

> [...] considerar a colocação não apenas como um efeito observável da língua em uso, mas como um importante elemento das causas de padrões da língua. Essa mudança de perspectiva proporcionada por sua descrição de colocação foi desenvolvida por vários linguistas na segunda metade do século XX, mais notadamente por John Sinclair.[10]

Ao explicar o papel da coocorrência de palavras na produção de significado no texto, Sinclair (1991, 2004) conceitua dois dos princípios de organização geral da língua em uso: o princípio da escolha aberta (**open-choice principle**) e o princípio idiomático (**idiom principle**), detalhados na seção 2.2. Como tais princípios propostos por Sinclair se articulam no âmbito do ensino e aprendizagem de língua estrangeira é o assunto de que trata este trabalho. Para tal, realizamos uma investigação descritiva da linguagem do aprendiz*,* analisando a ocorrência de colocações em cada um dos textos que compõem o *corpus* de estudo, por meio de um método intensivo de busca de colocações – o rastreamento de colocações –, de modo que se avaliasse a incidência dos princípio idiomático e da escolha aberta nessas produções escritas. Nesse contexto, apresentamos a seguir o objetivo e os questionamentos que nortearam esta pesquisa.

Este trabalho tem como objetivo detectar a existência de indícios do princípio idiomático e do princípio da escolha aberta, formulados por Sinclair (1991, 2004), na produção escrita de alunos brasileiros em inglês como língua estrangeira. A fim de atingir esse intento, foram formuladas as seguintes perguntas:

1. Quais são os indícios de presença do princípio idiomático nos textos produzidos?

[8] "[...] since the use of collocation in text is seen as a defining characteristic of natural, fluent text [...]"
[9] "[...] as is well known, was first coined in its modern linguistic sense by the British linguist J.R.Firth."
[10] "[...] to consider collocation not just as an observable effect of language use, but as an important element of the causes of language patterns. This shift in perspective made possible by his description of collocation was developed by several linguists in the second half of the twentieth century, most notably John Sinclair."

2. Quais são os indícios de presença do princípio da escolha aberta nos textos produzidos?

3. Como são distribuídas as colocações ao longo dos textos, e o que isso mostra sobre a incidência dos princípios idiomático e da escolha aberta na acepção de Sinclair (1991, 2004)?

Além da introdução, neste artigo fazemos na seção 1 uma breve resenha bibliográfica. Na seção 2, apresentamos a fundamentação teórica; na seção 3, expomos a metodologia; na seção 4, apresentamos os resultados; na seção 5, esses resultados são discutidos. Por fim, há as considerações finais.

1. RESENHA BIBLIOGRÁFICA

Nesta seção fazemos uma breve resenha bibliográfica. Para tal, escolhemos pesquisas que trabalham com *corpora* de aprendiz e que analisam as combinações de palavras usadas por aprendizes.

Ellis et al. (2015) ao analisar o que eles denominam linguagem formulaica (**formulaic language**) em *corpora* de aprendiz, um indício do que Sinclair (1991) denominou de princípio idiomático, abordam duas questões: quão proficientes seriam aprendizes de uma segunda língua em usar sequências pré-fabricadas, e se essas sequências desempenhariam algum papel na aquisição de uma segunda língua. Os autores afirmam que há definições robustas, mas divergentes, do que seria formulaicidade (**formulaicity**), embora o estudo não tenha considerado que o uso ou a ausência do uso de sequências de multipalavras (**multi-word sequences**) poderia ser um indício dos princípios idiomático e da escolha aberta.

Ebeling e Hasselgard (2015) ressaltam ser geralmente aceita em linguística a visão de que, em grande medida, a língua baseia-se em combinações de palavras que habitualmente coocorrem. Tais combinações, prosseguem as autoras, são consideradas a fraseologia, ou o "**phrasicon**", de uma língua. Podem ser identificadas de duas formas: por um método descendente (**top down**) ou ascendente (**bottom up**). O primeiro seria mais ligado à tradição da Europa Oriental, no qual o grau idiomático das unidades de multipalavras (**multi-word units**) é visto como mais relevante. O segundo, por outro lado, leva menos em

consideração o aspecto composicional e "adota uma definição mais ampla para o seu objeto de estudo [...] que não necessariamente constitui uma unidade semântica"[11]. As autoras dão exemplos de pesquisa de fraseologia com o método ascendente: pacotes ou feixes lexicais (**lexical bundles**) (Biber, Conrad, Cortes 2004), n-gramas (**n-grams**) (Stubbs 2009), colocações (Sinclair 1991), e "coloconstruções" (**colloconstructions**) (Stefanowitsch, Gries 2003). Ao desenvolver um pouco mais o construto da colocação, porém, as autoras não fazem referência à colocação como uma evidência do princípio idiomático, como apontado por Sinclair (1991).

Ao fazer uma síntese das pesquisas de fraseologia em *corpora* de aprendiz, Paquot e Granger (2012) salientam que "o tipo de sequência formulaica frequentemente mais estudada em pesquisa de *corpus* de aprendiz são as colocações verbo-substantivo"[12]. Em outro artigo, Granger e Bestgen (2014) discutem o uso de colocações em textos de alunos não-nativos intermediários versus avançados, analisando o uso de bigramas (pares de palavras diretamente adjacentes[13]). Elas observam nesse estudo que o uso de colocações de aprendizes intermediários ou avançados é caracterizado por uma mistura de colocações de alta e baixa frequência, sendo que a primeira é mais característica de alunos intermediários e a segunda de alunos avançados.

As pesquisas acima reconhecem a importância do uso de linguagem formulaica, sendo vista como um indicativo de quanto um aprendiz adquiriu uma língua. Também foi ressaltado que no estudo da fraseologia há basicamente duas abordagens – uma ascendente (**bottom up**) e outra descendente (**top down**). Entretanto, a colocação, que foi um dos exemplos dados de abordagem ascendente, e que é o tipo de frequência formulaica mais estudada em pesquisas de *corpus* de aprendiz, não foi apresentada como um indício do princípio idiomático (Sinclair, 1991). Diferentemente dos textos acima, o presente estudo tem por objetivo analisar se a combinação de palavras (verbos, substantivos, adjetivos e advérbios) usadas pelos aprendizes em seus textos podem ou não ser colocações, e possam, portanto, ser consideradas indícios da incidência do princípio idiomático ou do princípio da

[11] "[...] adopts a broader definition for its object of study that [...] do not necessarily constitute a semantic unit."
[12] "[...] the most frequently studied type of formulaic sequence in learner corpus research is verb-noun collocations [...]"
[13] "[...] directly adjacent word pairs."

escolha aberta, revisitando dessa forma os princípios desenvolvidos por Sinclair (1991) em um *corpus* de aprendiz.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção faremos uma breve introdução à linguística de *corpus* de aprendiz, discutiremos os dois princípios que nortearam esta pesquisa: o princípio idiomático e o princípio da escolha aberta, e por último, examinaremos o conceito de colocação.

2.1 LINGUÍSTICA DE *CORPUS* DE APRENDIZ

Segundo Granger (1998, 2002), só em fins da década de 1980 e início da década de 1990 acadêmicos e editoras começaram a coletar *corpora* de inglês não nativo, denominados *corpora* de aprendiz (**learner corpora**), reconhecendo o seu potencial prático e teórico. A LC de aprendizes, ou CLC, sigla de **computer learner corpora**, usa os principais princípios, ferramentas e métodos da LC e tem por objetivo proporcionar melhores descrições da língua de aprendizes (Granger 2002: 1). De acordo com a mesma autora, essa área de investigação linguística criou uma importante ligação entre a LC e a pesquisa de Aquisição de Segunda Língua, que tenta compreender quais os mecanismos de aquisição de uma segunda língua, e a pesquisa do Ensino de Língua Estrangeira, que tem por objetivo aperfeiçoar o ensino-aprendizagem de línguas estrangeiras (Granger 2002: 2).

Granger (2002), ao definir o que seja um *corpus* de aprendiz, afirma que apesar de poder ser definido como uma coleção de dados produzidos por aprendizes, esse tipo de definição deve ser evitado porque pode fazer com que o termo seja usado para tipos de dados que de fato não são *corpora*. Ela sugere (2002: 4) adotar uma definição baseada na definição de *corpora* de Sinclair:

> *Corpora* de aprendiz computadorizados são coleções eletrônicas de dados textuais autênticos de Língua Estrangeira ou Segunda Língua reunidos de acordo com critérios de desenho explícitos para um determinado propósito para Aquisição de Língua Estrangeira/Ensino de Língua Estrangeira. São

> codificados de forma padrão e homogênea assim como são documentados suas origens e procedências.[14]

A partir dessa definição podemos depreender os critérios a serem usados na compilação de um *corpus* de aprendiz. Abaixo se encontra uma tabela resumindo esses critérios.

Língua	**Aprendiz**
Meio (escrito ou falado)	Idade
Variedade textual	Sexo
Tópico	Língua materna
Condições da atividade	Região
	Outra língua estrangeira
	Nível de proficiência
	Contexto de aprendizagem
	Experiência prática

Tabela 1: Critérios para a compilação de um *corpus* de aprendiz. Fonte: Granger (1998: 8).

O contexto de aprendizagem distingue os alunos que estão aprendendo inglês num país de língua inglesa (Inglês como Segunda Língua – ISL) ou não (Inglês como Língua Estrangeira – ILE). A experiência prática refere-se ao tempo que o aprendiz estuda inglês, ao material que usa, à quantidade de horas-aula por semana e a se o aluno já esteve em um país de língua inglesa (Granger 1998: 6).

2.2. O PRINCÍPIO DA ESCOLHA ABERTA E O PRINCÍPIO IDIOMÁTICO

Sinclair (1991), ao explicar de que forma o significado se manifesta no texto, desenvolve dois princípios de organização geral da língua em uso. Nesta seção serão apresentados esses dois princípios: o princípio da escolha aberta e o princípio idiomático.

[14] "Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardized and homogeneous way and documented as to their origin and provenance."

2.2.1 O PRINCÍPIO DA ESCOLHA ABERTA

O princípio da escolha aberta (**open-choice principle**) é um modo de ver a língua como o resultado de um grande número de escolhas complexas. Segundo Sinclair (1991: 109), "a cada ponto em que uma unidade é completada – uma palavra, uma frase, uma oração – uma grande amplitude de escolhas se abre, e a única restrição é a gramaticalidade"[15].

Para Sinclair (1991), esse é provavelmente o modo habitual pelo qual livros didáticos e materiais de referência, tais como os livros de gramática, veem a língua em uso, e também como a descrevem e a ensinam. É frequentemente chamado de modelo de abertura-e-enchimento (**slot-and-filler model**). Esse modelo considera os textos como uma série de aberturas que têm de ser preenchidas por um léxico que satisfaça as restrições daquela abertura – e, em cada uma delas, ocorrer virtualmente qualquer palavra. Sinclair ressalta que, similarmente, todas as gramáticas são construídas segundo o princípio da escolha aberta (1991: 110).

2.2.2. O PRINCÍPIO IDIOMÁTICO

Para Sinclair, as palavras não se sucedem aleatoriamente num texto, assim como o princípio da escolha aberta não é suficiente para explicar todas as ocorrências de palavras. Segundo esse autor, não produziríamos um texto normal simplesmente operando com o princípio da escolha aberta (1991: 110):

> [o] princípio idiomático consiste em o usuário da língua dispor de um vasto número de frases semiconstruídas que constituem escolhas únicas, ainda que possam ser analisadas em segmentos. De certa forma, isso pode refletir a recorrência de situações similares em assuntos humanos, ou pode ilustrar uma tendência natural para a economia de esforços, ou pode ainda ser motivada em parte pelas exigências da comunicação em tempo real. Seja como for, foi relegado a uma posição inferior pela maioria dos linguistas, porque não se encaixa no modelo da escolha aberta.[16]

[15] "At each point where a unit is completed (a word, a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness."
[16] "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 analyzable into segments. To some extent, this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies

Na sua forma mais simples, explica Sinclair (1991), o princípio idiomático pode ser visto na escolha aparentemente simultânea de duas palavras, por exemplo, **of course** (claro). Essa frase, prossegue o autor, opera efetivamente como se fosse uma única palavra. O **of** em **of course** não é a preposição encontrada em livros de gramática. A preposição **of** é normalmente encontrada depois do substantivo, que é o núcleo de um grupo nominal, ou depois de um quantificador, como em **a pint of**... (um quartilho de...). No modelo de escolha aberta, a preposição **of** pode ser seguida de qualquer grupo nominal. Similarmente, **course** em **of course** não é o substantivo explicado nos dicionários; seu significado não é propriedade da palavra, mas da frase. Se fosse um substantivo enumerável no singular, em inglês, para ser gramatical, ele deveria ser precedido por um determinante (**determiner**) – então, conclui Sinclair, claramente não o é (1991: 111). "O mesmo tratamento poderia ser dado a centenas de frases similares – qualquer ocasião na qual uma decisão acarreta o uso de mais de uma palavra no texto [...], por exemplo, [...] expressões idiomáticas, provérbios, termos técnicos, verbos frasais [...]"[17].

Segundo Sinclair (1991), a esmagadora natureza dessa evidência faz com que o princípio idiomático deixe de ser uma característica de menor importância, e que seja tão importante quanto a gramática na explicação de como o significado se manifesta no texto.

Sinclair (1991) esclarece que os limites entre os trechos elaborados segundo esses dois princípios não serão normalmente bem definidos, assim como nem todos os trechos possuem evidências tão claras de não terem sido elaborados segundo regras gramaticais como **of course**. Também afirma que deveria ser reconhecido que "[...] os dois modelos de língua que estão em uso são incompatíveis um com o outro. Não há uma transformação gradativa de um para outro; a mudança de um modelo

of real-time conversation. However it arises, it has been relegated to an inferior position in most current linguistics, because it does not fit the open-choice model."

[17] "The same treatment could be given to hundreds of similar phrases – any occasion where one decision leads to more than one word in a text. Idioms, proverbs, technical terms, phrasal verbs [...]"

para outro é abrupta. Os modelos são diametralmente opostos"[18] (1991: 114). E acrescenta que

> para textos normais, podemos propor que o primeiro modo a ser aplicado é o princípio idiomático, uma vez que a maior parte do texto será interpretável por esse princípio. Sempre que houver uma boa razão, o processo interpretativo muda para o princípio da escolha aberta, e rapidamente de volta para o primeiro. Escolhas lexicais que são inesperadas em seu ambiente irão presumivelmente ocasionar uma mudança; escolhas que, se gramaticalmente interpretadas, seriam pouco usuais são uma afirmação do princípio idiomático em ação.[19] (Sinclair 1991: 114)

Sinclair (1991) ressalta que alguns textos podem ser compostos em uma tradição que faça mais uso do princípio da escolha aberta, como por exemplo, declarações legais e poesia, mas esses são textos especializados que requerem uma prática adicional para a sua compreensão.

A partir dessas evidências, Sinclair conclui que um modelo de língua (**language**) que separe gramática e léxico, e use a gramática para proporcionar uma série de pontos de escolha lexical, é um modelo secundário. Não pode ser abandonado, porque um texto tem muitos pontos nos quais o princípio da escolha aberta entra em ação:

> [O princípio da escolha aberta] tem uma relevância abstrata, no sentido de que muito do que se observa em um texto apresenta o potencial de ser analisado como o resultado de escolhas abertas, mas é o outro princípio, o princípio idiomático, que domina. A análise de escolhas abertas pode ser imaginada como um processo analítico que a princípio acontece o tempo todo, mas cujos resultados são apenas intermitentemente requeridos[20] (Sinclair 1991: 114).

[18] "[...] the two models of language that are in use are incompatible with each other. There is no shading of one into another; the switch from one model to the other will be sharp. The models are diametrically opposed."
[19] "[...] for normal texts, we can put forward the proposal that the first mode to be applied is the idiom principle, since most of the text will be interpretable by this principle. Whenever there is good reason, the interpretative process switches to the open-choice principle, and quickly back again. Lexical choices which are unexpected in their environment will presumably occasion a switch; choices which, if grammatically interpreted, would be unusual are an affirmation of the operation of the idiom principle."
[20] "[...] has an abstract relevance, in the sense that much of the text shows a potential for being analysed as the result of open-choices, but the other principle, the idiom principle dominates. The open-choice analysis could be imagined as an analytical process which goes on in principle all the time, but whose results are only intermittently called for."

Em algumas ocasiões, "[...] as palavras parecem ser escolhidas em pares ou grupos, e não estão necessariamente adjacentes"[21] (Sinclair 1991: 115).

A partir do desenvolvimento desses dois princípios explicadores de como o significado emerge do texto, Sinclair considera o papel da colocação – conceito que ilustra o princípio idiomático.

2.3. O CONCEITO DE COLOCAÇÃO

Berber Sardinha (2004) salienta que, no estudo de *corpus*, o fenômeno da colocação é o mais tradicionalmente analisado. Foi cunhado em seu significado linguístico moderno por Firth, em 1957, juntamente com a frase "you shall judge a word by the company it keeps"[22]. Partington (1998) apresenta três definições de colocações: a textual, a psicológica ou associativa, e a estatística. Para caracterizar colocação textual, Partington cita Sinclair (1991: 170): "Colocação é a ocorrência de duas ou mais palavras num texto em um curto espaço uma da outra"[23]. Uma palavra é colocada de outra se a palavra está em algum lugar próximo do nódulo, em um texto (Partington 1998: 15).

Para a segunda definição, colocação psicológica ou associativa, Partington faz referência a Leech (1974: 20): "O sentido colocacional consiste nas combinações adquiridas por uma palavra por causa dos significados das palavras que tendem a ocorrer no seu ambiente."[24] Partington (1998: 16) esclarece que essa definição de colocação faz parte da competência comunicativa de um falante nativo, que sabe quais as colocações são esperadas e quais não o são em determinados textos.

A terceira e última definição é a da colocação estatística. Esse outro aspecto do fenômeno foi salientado por Hoey (1991 *apud* Partington 1998: 16): "Colocação tem sido o nome dado à relação que um item lexical tem com os itens que aparecem com probabilidade maior do que aleatória no seu contexto (textual)"[25]. Partington ressalta

[21] "[...] words seem to be chosen in pairs or groups and these are not necessarily adjacent."
[22] Julgue a palavra por sua companhia.
[23] "Collocation is the occurrence of two or more words within a short space of each other in a text."
[24] "Collocative meaning consists of the associations a word acquires on account of the meanings of words which tend to occur in its environment."
[25] "Collocation has long been the name given to the relationship a lexical item has with items that appear with greater than random probability in its (textual) context."

que é uma boa definição para aqueles que estudam LC, na qual grandes quantidades de texto podem ser analisadas por computador.

Stubbs (1995: 1) define colocação como "a relação habitual de coocorrência entre as palavras (lemas ou formas lexicais)"[26] e salienta que "a força de atração entre elas pode ser medida quantitativamente"[27]. Esse autor ressalta que, embora nativos da língua possam com frequência dar exemplos de colocados de uma palavra, eles certamente não conseguem documentar colocações em profundidade, assim como também não conseguem dar estimativas precisas da sua frequência ou da distribuição de diferentes colocações.

O *script* rastreador de colocação desenvolvido por Berber Sardinha (2014a), e também usado nesta pesquisa, usa três medidas estatísticas – IM (Informação Mútua), Escore T e logDice. Essas medidas estatísticas de associação de palavras usam as frequências observadas em um *corpus* de referência, o **enTenTen12** (Jakubíček et al. 2013), para estimar o nível de coocorrência entre um nódulo e um colocado. Elas fornecem dados quantitativos para que se possa dizer com mais firmeza se as combinações de palavras observadas podem ou não ser chamadas de colocação, e, portanto, ser consideradas um indício da incidência do princípio idiomático ou do princípio da escolha aberta.

3. METODOLOGIA

A metodologia deste estudo está dividida em três etapas, descritas a seguir: 1) a compilação do *corpus* de estudo, um *corpus* de aprendiz; 2) o rastreamento de colocações; e 3) os procedimentos realizados para a obtenção dos resultados.

3.1. O *CORPUS* DE ESTUDO – UM *CORPUS* DE APRENDIZ

A compilação do *corpus* de estudo foi feita segundo os critérios indicados por Granger (1998; 2002): (a) autenticidade – no caso de aprendizes, textos produzidos em uma genuína atividade de aula; (b) controle de variáveis, tais como as

[26] "[...] a relationship of habitual co-occurrence between words (lemma or word forms)."
[27] "[...] – the strength of association between words can be measured in quantitative terms [...]"

relacionadas a tarefas – por exemplo, o tempo disponível para a sua execução – e o nível de proficiência dos aprendizes, e (c) adequação aos objetivos da pesquisa.

O *corpus* de estudo é composto por 39 textos, cada um de 120 a 180 palavras, escritos por aprendizes brasileiros da língua inglesa, com um total de 6.650 palavras (**tokens**). Todos são a respeito do mesmo tema: uma redação argumentativa a respeito do conto *Through the Tunnel,* da escritora Doris Lessing, na versão original, não simplificada, lido em sala de aula. As redações foram escritas nos computadores da escola em formato de documento do Word em uma aula de 75 minutos. Os alunos podiam usar o corretor ortográfico e acessar a Internet, por exemplo, para pesquisar a respeito da autora do conto, consultar dicionários ou sítios de tradução. Ao terminarem o trabalho, enviaram o texto para a professora pelo portal da escola, que adota a plataforma Moodle[28]. Essa atividade de produção escrita não era uma prova, mas era uma atividade avaliada. As redações foram baixadas do portal e arquivadas no formato .txt, exigido pelo *script collocation tracking*[29], usado nesta pesquisa. As redações foram incluídas no *corpus* da maneira como foram escritas nessa ocasião pelos alunos, sem nenhuma intervenção do professor.

Os autores dos textos eram alunos de uma escola particular da zona oeste da cidade de São Paulo e tinham entre 14 e 17 anos. Estavam divididos em três turmas, duas do primeiro e uma do segundo ano, por níveis de proficiência de acordo com a pontuação em um teste classificatório que faziam antes de iniciar o Ensino Médio. Os níveis, assim como o material didático usado, têm como referência a *Common European Framework of Reference for Languages*[30] (doravante CEFRL). O nível dos estudantes desta pesquisa era B2. Segundo a CEFRL (p. 24), um aprendiz B2

> consegue entender as ideias principais de textos complexos de temas tanto concretos quanto abstratos, incluindo discussões técnicas na sua especialidade, interagir com um grau de fluência e espontaneidade que faz com que a interação habitual com falantes nativos seja possível sem esforço para ambas as partes, e produzir textos claros e detalhados a respeito de temas variados, além de ser capaz de explicar o seu ponto de vista a respeito

[28] MOODLE é o acrônimo de *Modular Object-Oriented Dynamic Learning Environment* (Ambiente de Aprendizagem Dinâmico e Modular Orientado a Objetos). Foi criado por Martin Dougiamas. Cf. https://moodle.org/mod/forum/discuss.php?d=332821.
[29] Rastreador de colocação.
[30] Quadro Europeu Comum de Referência para Línguas.

> de um tema da atualidade apresentando as vantagens e desvantagens de várias opções[31].

O *corpus* de estudo, com 39 textos produzidos por cada um dos 39 alunos, é 100% representativo das redações argumentativas das três turmas que foram a fonte da análise. A razão para empregar um número limitado de textos se justifica pela opção feita por realizar uma pesquisa intensiva com o rastreador de colocações. Dado que o *script* extrai todas as combinações de palavras usadas pelos alunos a fim de detectar a presença do princípio idiomático e o da escolha aberta, o número de textos não pode ser muito alto, pois tanto o processamento computacional quanto a interpretação do resultado do processamento são laboriosos (cf. Seção 3.2).

Todas as 39 produções escritas dos aprendizes foram rastreadas. Neste artigo, por uma limitação de espaço, serão analisados dois textos representativos do *corpus* (textos 17 e 01) para a discussão dos princípios idiomático e da escolha aberta.

3.2. O RASTREAMENTO DE COLOCAÇÕES E OS *CORPORA* DE REFERÊNCIA

O rastreador de colocações (**collocation tracking**) é um *script* desenvolvido por Berber Sardinha (2014a) que visa detectar a presença de colocações em cada frase dos textos de um *corpus*. Esse método consiste em comparar todos os pares de palavras de cada frase com um *corpus* de referência extenso da língua para determinar em que medida os candidatos à colocação correspondem a colocações atestadas no *corpus* de referência.

Para fazer esse rastreamento, cada sentença do texto é subdividida em sequências de até 11 palavras consecutivas – o nódulo e as cinco palavras que estão antes e depois dele. Tais segmentos formam, na mesma ordem em que aparecem no texto – as chamadas "janelas"[32]. Dentro de cada janela, uma palavra é selecionada para ser o nódulo – ou seja, a palavra que se encontra na posição central da janela, a

[31] "Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options."

[32] Optou-se por usar na ferramenta o mesmo alcance padrão das janelas do Sketch Engine – 5:5, que aparece quando se clica em Search, e depois em Context.

partir do qual são definidos os candidatos à colocação. Para cada par formado pelo nódulo mais uma palavra da janela, que é o candidato a colocação, é feita uma busca no *corpus* de referência do *script*. Caso esse candidato exista no *corpus* de referência, ele passa a ser considerado uma colocação propriamente dita – um indício, para a pesquisadora, da aplicação do princípio idiomático. Em caso contrário, ou seja, quando o candidato a colocação não é localizado no *corpus* de referência, essa instância foi considerada como um indício do princípio da escolha aberta, na medida em que se trata de uma combinação de palavras pouco comum, rara ou atípica da língua inglesa – como, por exemplo, **precise** + **fashion**.

Berber Sardinha (2014b: 184) explica que a comparação é feita com colocações atestadas no *corpus* de referência. Isso significa que somente as sequências de palavras que tenham um valor igual ou maior que o limite estatístico de coocorrência são válidas[33]. Dito de outra forma, não basta a combinação de palavras usada no texto do aprendiz ser encontrada no *corpus* de referência. Sua frequência no *corpus* de referência tem de ser suficientemente notável para que ela se qualifique como colocação. O *corpus* de referência escolhido para ser usado no rastreamento de colocações foi o enTenTen12[34], por ter sido o maior *corpus* da língua inglesa no momento da pesquisa, com mais de 11 bilhões de palavras. As estatísticas de associação usadas no presente artigo – Informação Mútua, Escore T e logDice – foram calculadas pelo Sketch Engine.

Nesta pesquisa foram usados mais dois *corpora* de referência, o Corpus of Contemporary American English (doravante COCA) e o Sketch Engine for Language Learners (doravante SkELL). O COCA é um *corpus* de inglês americano então composto de mais de 520 milhões de palavras, igualmente divididas em blocos de tamanho quase idêntico na forma de textos de ficção, revistas, jornais, periódicos acadêmicos, e transcrição de conversas improvisadas em programas de rádio e TV. Foi escolhido por estar disponível gratuitamente na internet e permitir pesquisar se as escolhas lexicais dos alunos são mais frequentes em linguagem oral ou escrita.

[33] "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-occorrence [...] were valid as reference corpus units."

[34] O enTenTen12 está disponível no sítio Sketch Engine (https://www.sketchengine.co.uk), mas seu acesso não é gratuito.

O SkELL (Baisa e Suchomel 2014) é a interface de busca no *corpus* com mais de um bilhão e meio de palavras no momento da pesquisa, coletadas especialmente para o ensino da língua inglesa, pelo portal do Sketch Engine. É composto de artigos em inglês da Wikipedia, de textos do Projeto Gutenberg, de um subconjunto do enTenTen14 com documentos de domínios da rede catalogados na **dmoz.org**[35], do British National Corpus, entre outros. Também está disponível gratuitamente na internet[36]. Além disso, a concordância que ele mostra é mais compreensível do que a do COCA, pois mostra cada padrão em frases.

3.3. OS PROCEDIMENTOS PARA A OBTENÇÃO DOS RESULTADOS

Primeiramente foi feita uma análise quantitativa dos dados gerados pela ferramenta. Em função do resultado dessa análise, selecionaram-se os textos para a análise qualitativa.

3.3.1. A ANÁLISE QUANTITATIVA

Para a análise quantitativa, três aspectos foram levados em consideração: as medidas estatísticas usadas no rastreador de colocações, a média da porcentagem compartilhada das colocações e a porcentagem de janelas válidas com pelo menos uma colocação.

O rastreador de colocações usa três medidas de associação estatística: a Informação Mútua (IM), o Escore T e o logDice. Cada uma delas tem um valor mínimo abaixo do qual a associação de palavras não é considerada uma colocação. O valor mínimo para o IM é 3; para o Escore T, é 2; e para o logDice, é 1[37] (Stubbs 1995; Hunston 2002; Berber Sardinha 2004; Rychlý 2008).

Stubbs (1995), em seu estudo a respeito de colocações e perfil semântico, emprega três medidas. Segundo esse autor, os grupos de palavras que sobrevivem aos filtros de mais de uma medida de associação estatística indicam combinações lexicais,

[35] *DMOZ* é o mais amplo e abrangente diretório da Web editado por humanos.
[36] Pode-se acessar o SkELL pelo endereço: https://www.sketchengine.co.uk/skell/.
[37] Rychlý (2008) usa zero, mas esse valor significa que não há associação. Nesta pesquisa optou-se por usar 1 como valor mínimo, para considerar que a associação tem significância estatística.

baseados numa sólida evidência quantitativa, para ulterior investigação humana. Esses são os casos em que podemos estar confiantes de que há uma forte associação entre o nódulo e o colocado.[38] Igualmente, Biber (2012) questiona o uso de apenas uma única medida de associação lexical. Optou-se, então, por rejeitar as colocações que não tiveram o valor mínimo em pelo menos duas das medidas de associação estatística.

Essa filtragem com duas medidas influencia outros aspectos da análise quantitativa, como a porcentagem compartilhada e as janelas com pelo menos uma colocação, na medida em que a porcentagem compartilhada registra o número de candidatos a colocado que formam uma colocação com o nódulo. A título de exemplificação, escolheu-se uma janela com três palavras consecutivas, na sentença 1 do texto 01. Nela, o nódulo **NOT** tem dois candidatos a colocado.

- **NOT** (nódulo) + **CAN** (candidato a colocado)

- **NOT** (nódulo) + **THINK** (candidato a colocado)

CAN tem -0,783 de logDice, 4,524 de IM, e 26,543 de Escore T, ultrapassando o valor mínimo na IM e no Escore T. **THINK** também registra valores maiores que o mínimo em duas das medidas, 1,923 no logDice e 3,422 na IM. Isso significa que são colocações. Na janela 3, da sentença 1 do texto 01, temos, com o nódulo em negrito:

"[...] *I ca**n't** think of* [...]"

Nessa janela, os dois candidatos à colocação são atestados no **corpus** de referência, ou seja, essa janela tem 100% de porcentagem compartilhada.

Em contraposição, na janela 91 do mesmo texto 01, o nódulo é o substantivo **POINT** e nenhum dos seus quatro candidatos a colocado formou uma colocação com

[38] "The important thing is that we have a replicable procedure for filtering out cases which might be entirely due to chance. The cases which survive the filters provide a set of words, based on solid quantitative evidence, for further human interpretation. These are the cases where we can be confident that there is a strong association between node and collocate." (STUBBS, 1995, p. 40)

o nódulo – **MOTHER**, **ALWAYS**, **BE** e **INSECURE**. A porcentagem compartilhada é, portanto, 0%. Abaixo, a janela 91, com o nódulo em negrito:

"[...] *like his mother in the* ***point*** *of always be* [sic] *insecure by* [...]"

Neste *corpus* de estudo, a média aritmética da porcentagem compartilhada de todas as redações do *corpus* alcançou 43%. Alguns textos se destacaram por terem as maiores porcentagens compartilhadas em relação à média geral de 43% observada no *corpus* de estudo, indicando que possivelmente seu número superior de colocações os posicionam mais próximos do princípio idiomático. Outros textos, por terem as menores porcentagens compartilhadas, e, portanto, um número inferior de colocações em relação à porcentagem média observada no *corpus* de estudo, sinalizam que possivelmente estão mais próximos do princípio da escolha aberta. Dois deles foram selecionados para serem aqui discutidos.

3.3.2. A ANÁLISE QUALITATIVA

A análise qualitativa foi realizada em três etapas. Na primeira, releram-se os textos dos aprendizes, selecionados a partir da análise quantitativa, buscando indícios do princípio idiomático e do princípio da escolha aberta. Esses indícios foram atestados nos dois outros *corpora* de referência usados nesta pesquisa, o COCA e o SkELL. A seguir, buscou-se evidência desses princípios a partir dos dados gerados pelo rastreador de colocações: a porcentagem compartilhada das janelas, a porcentagem de janelas com pelo menos uma colocação, e os valores das medidas estatísticas de associação. Por último, comparou-se o texto dos alunos com a versão de uma revisora norte-americana, sendo essas versões também rastreadas pelo *script*.

4. RESULTADOS

Nesta seção são apresentados e analisados os resultados obtidos em resposta às perguntas de pesquisa que nortearam o trabalho.

1. Quais são os indícios de presença do princípio idiomático nos textos

produzidos?

2. Quais são os indícios de presença do princípio da escolha aberta nos textos produzidos?

3. Como são distribuídas as colocações ao longo dos textos e o que isso mostra sobre a incidência dos princípios idiomático e da escolha aberta?

A apresentação dar-se-á por meio de dois textos produzidos pelos alunos, o 17 e o 01.

4.1. QUESTÃO 1: QUAIS SÃO OS INDÍCIOS DE PRESENÇA DO PRINCÍPIO IDIOMÁTICO NOS TEXTOS PRODUZIDOS PELOS APRENDIZES BRASILEIROS DA LÍNGUA INGLESA?

O texto de número 17 tem a porcentagem compartilhada mais alta dos textos rastreados, 57%. Ou seja, de cada 10 combinações de palavras, aproximadamente seis são colocações do inglês. A sua porcentagem de janelas com pelo menos uma colocação é de 76%. Esses dados indicam que o aluno tem um bom repertório de colocações. Nessa redação, podemos destacar o seguinte trecho (Exemplo 1):

> (1) Because of the certain limit of pages and words, short stories tend to be simple and quick to read. But at the same time, very hard to write because you have to create a whole story, start, middle and end, in a short period of time.

Nessa passagem, observa-se que o aluno usou a colocação **whole** + **story**, bastante comum no inglês, com 1,803 de ocorrências COCA, e também **short** + **period** + **of** + **time**, que tem 993 de ocorrências no mesmo *corpus* – duas evidências empíricas do princípio idiomático, tal qual caracterizado por Sinclair (1991, 2004). O rastreador de colocação também dá mais exemplos da incidência desse princípio, no mesmo trecho. Para fazer esse rastreamento, o *script* leva em consideração o nódulo e os candidatos a colocado de cada janela, como explicitado na metodologia. Por exemplo, na janela 28, "**be simple and quick to read***"*, o nódulo **quick** tem três candidatos a colocado:

- **QUICK**[39] (nódulo) +[40] **BE** (candidato a colocado),

- **QUICK** (nódulo) + **SIMPLE** (candidato a colocado),

- **QUICK** (nódulo) + **READ** (candidato a colocado).

Todas as três colocações foram atestadas no enEnTenTen12, o *corpus* de referência. O logDice de cada uma delas, respectivamente, é 1,866, 1,012 e -0,394; a IM, de 3,440, 6,324 e 4,918, nessa ordem; e o Escore T de cada uma é de 23,294, 12,175 e 8,261, por essa ordem. Com exceção do logDice de **QUICK** e **READ**, que está abaixo do valor mínimo para essa medida nesta pesquisa, todas as outras medidas indicam atração entre as palavras. Esses indícios de ocorrência do princípio idiomático, tal qual definido por Sinclair (1991, 2004), observados nos textos produzidos pelos aprendizes de língua inglesa, são denominados neste artigo de princípio idiomático do tipo I.

Nesse mesmo excerto do texto escrito pelo aprendiz (Exemplo 1), a frase "**because of the certain limit of**" (janela 12) tem 100% de porcentagem compartilhada[41]. Nessa janela, o nódulo é **CERTAIN** e o único colocado é **LIMIT**. Essa colocação é atestada no *corpus* de referência com um logDice de 0,031, IM de 5,360 e Escore T de 4,363, indicando uma razoável atração entre as palavras, como registrado pela IM e o Escore T. Contudo, apesar de ser uma colocação possível, *certain* não é a palavra mais adequada, como se pode perceber quando lemos o restante da frase: "**because of the certain limit of pages and words, short stories**", ou quando verificamos o significado do adjetivo **certain**. De acordo com o Cambridge Dictionary Online[42], **certain** quer dizer **without doubt** (sem dúvida). Pode ter também o significado de **limited** (limitado), quando usado em combinação com **degree** (grau) ou **extent** (alcance) na sequência **to** + **a** + **certain** + **degree/extent**, segundo o mesmo dicionário. Igualmente no COCA, a sequência

[39] Letras maiúsculas indicam lemas.
[40] Esse sinal indica que as palavras não são adjacentes. No caso do *script* rastreador de colocações, há cinco palavras para direita e cinco para a esquerda.
[41] Conforme explicado na metodologia, ter 100% de porcentagem compartilhada significa que todos os candidatos a colocados na janela em questão foram atestados no *corpus* de referência.
[42] Disponível em: http://dictionary.cambridge.org/dictionary/english/certain Acesso em: 3/7/2016

certain + **limit**, com 30 ocorrências, está sempre precedida pelo artigo indefinido *a*, assim como não há registro para a sequência **the** + **certain** + **limit** (vide Figura 1)[43]. No SkELL[44], **certain limit** geralmente aparece no plural (**certain limits**) e é precedido por palavras como **within**, **beyond** e **up to** (vide Figura 2).

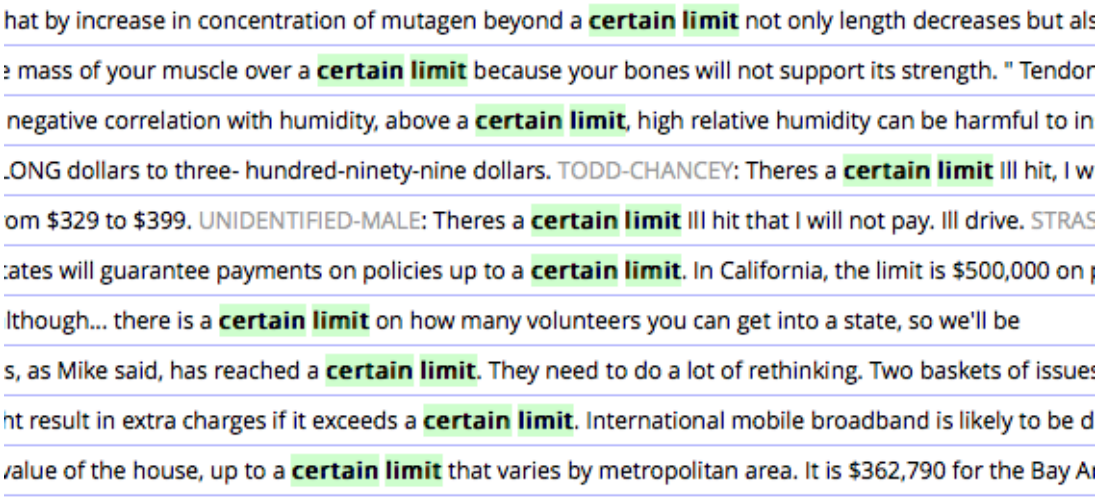


Figura 1: Extrato de concordância de **certain limit** no portal COCA

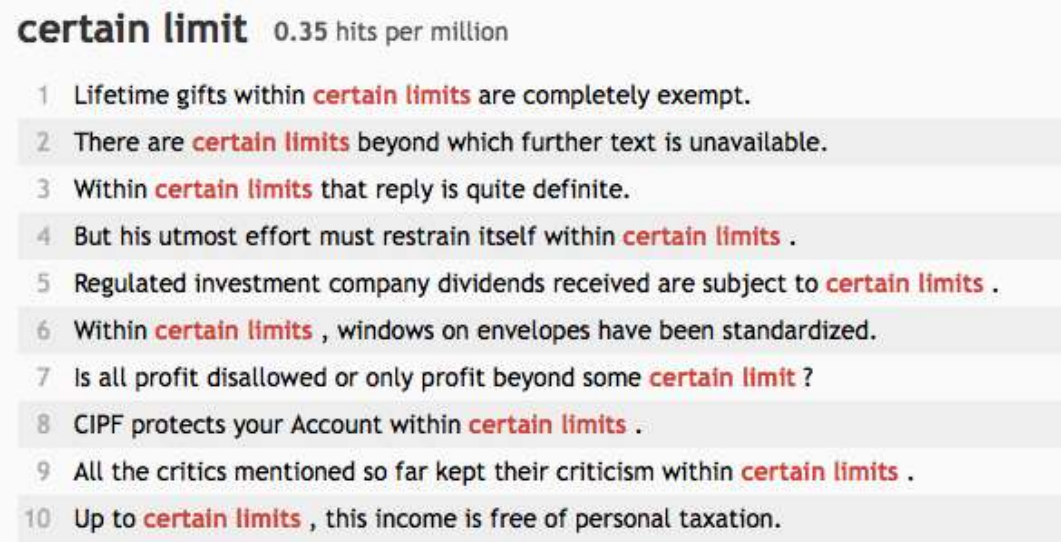


Figura 2: Extrato de concordância de **CERTAIN LIMIT** no portal SkELL

[43] A busca da sequência **certain limit** foi feita na caixa **List** do portal COCA, assim como todas as outras buscas, caso não seja mencionada outra forma de investigação.
[44] O SkELL usa lemas como padrão, mesmo que a palavra buscada não seja digitada em letras maiúsculas - a busca independe das letras serem maiúsculas ou minúsculas.

Nesse exemplo específico, portanto, percebe-se que a palavra **certain** não está adequada à frase na qual está inserida, ainda que, a julgar pelos dados dos *corpora* de referência, seja um indício da idiomaticidade do texto. O que o aluno quis dizer é algo como "devido ao limite de tamanho", referente ao tamanho máximo permitido para o texto, o que em inglês pode ser expresso por colocações como **size restrictions**, **limits on the number of** (**words** etc.), **limited word counts**, ou **limited size** (vide Figura 3).

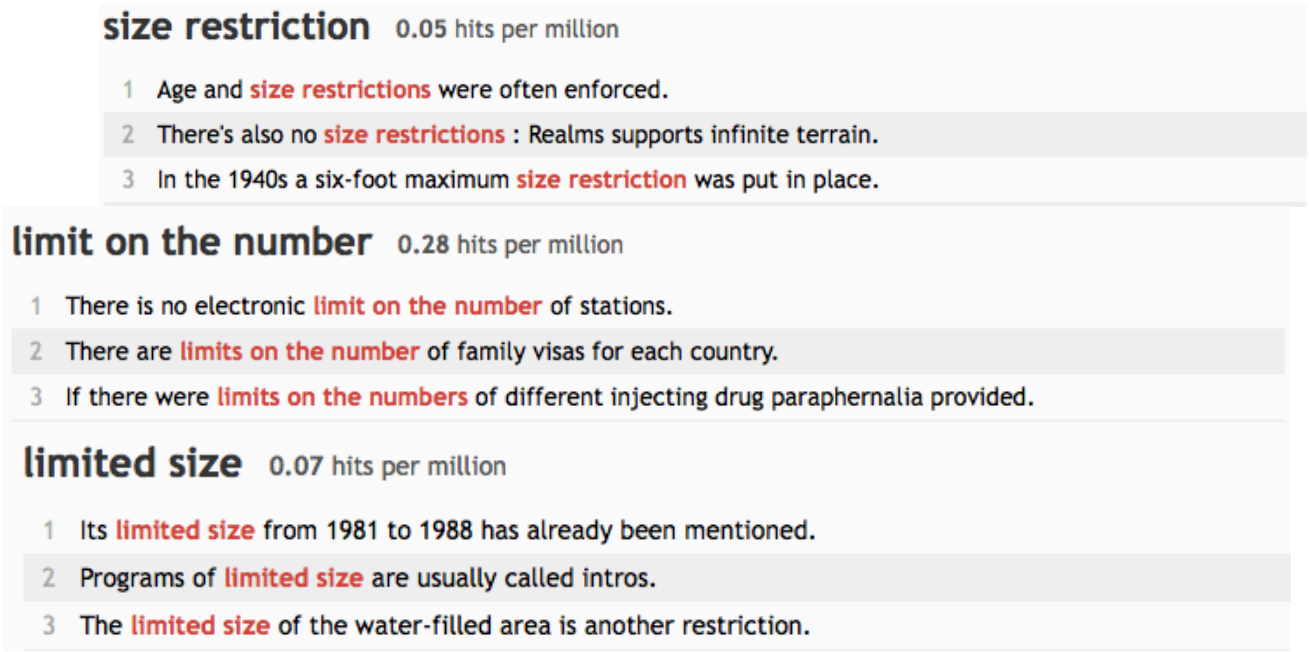


Figura 3: Extrato de concordâncias de alternativas para **certain limit** no portal SkELL

Para esse tipo de evidência, no qual a combinação de palavras usadas existe no inglês, mas não expressa o significado desejado no texto, deu-se o nome de princípio idiomático do tipo II. Ou seja, o aluno usou uma colocação existente com um sentido diferente daquele que a colocação tem normalmente.

Na versão modificada, a revisora norte-americana opta por substituir **certain** por **limited**, conforme mostrado abaixo no Exemplo 2, apesar de a colocação **limited words** ser muito rara no inglês: no SkELL, apenas quatro ocorrências existem (vide Figura 4). Nesse caso, a estratégia da revisora foi a de retirar **certain** e

usar **limited**, apesar de haver outras colocações mais idiomáticas do inglês, conforme mostrado acima. Em outras palavras, do ponto de vista da revisão de texto, essa estratégia parece conveniente, pois implica uma menor intervenção no texto alheio, embora não tenha sido escolhida melhor colocação para o sentido pretendido.

> (2) Because of their **limited**[45] words and pages, short stories tend to be simple and quick to read, **but very hard to write** because you have to create a whole story **— start, middle, and end —** in a short period of time.

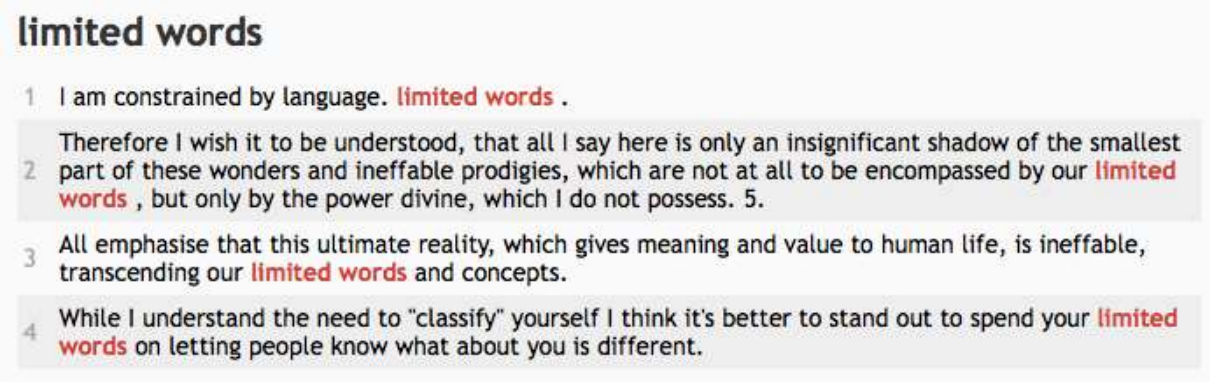


Figura 4: Extrato de concordâncias de **LIMITED WORD** no portal SkELL

Em conclusão, o texto 17, ilustrado pelo Exemplo 1, com 57% de porcentagem compartilhada e 76% de janelas com pelo menos uma colocação, apresenta evidências empíricas do princípio idiomático. Presumivelmente, a despeito de suas imprecisões, a redação satisfaria as expectativas lexicais de leitores da língua inglesa. Contudo, observou-se que, além do princípio descrito por Sinclair (1991, 2004), há também o que se pode denominar uma pequena variação desse princípio no texto escrito pelo aprendiz. A essa variante demos o nome de princípio idiomático tipo II.

[45] As mudanças feitas pela revisora norte-americana estão destacadas em negrito nesse e em todos os outros exemplos.

4.2. QUESTÃO 2: QUAIS SÃO OS INDÍCIOS DE PRESENÇA DO PRINCÍPIO DA ESCOLHA ABERTA NOS TEXTOS PRODUZIDOS PELOS APRENDIZES BRASILEIROS DA LÍNGUA INGLESA?

O texto 01 tem 50% de porcentagem compartilhada, e 80% de janelas com pelo menos uma colocação. Nessa redação, podemos destacar o seguinte trecho (vide Exemplo 7):

> (7) I can't think of anyone in the world who never overcome [sic; overcomes] a limit or wish [sic; wishes] to do it.

Na frase "(...) **in the world who never overcome** [sic; **overcomes**] a **limit or wish** [sic; **wishes**] **to** (...)", janela 9, cujo nódulo é **overcome**, há quatro candidatos a colocado:

- **OVERCOME** (nódulo) + **WORLD** (candidato a colocado),
- **OVERCOME** (nódulo) + **NEVER** (candidato a colocado),
- **OVERCOME** (nódulo) + **LIMIT** (candidato a colocado),
- **OVERCOME** (nódulo) + **WISH** (candidato a colocado).

Três colocações são atestadas no *corpus* de referência: **OVERCOME** + **WORLD**, **OVERCOME** + **NEVER**, e **OVERCOME** + **LIMIT**, com logDice de 1,619, 1,658, 0,179, IM de 3,690, 3,696, e 5,421, e Escore T de 7,773, 1,598 e 3,660, respectivamente. Contudo, não há registros correspondentes no COCA para a sequência **overcome** + **a** + **limit** usada pelo aprendiz. No COCA, há apenas 9 ocorrências para **OVERCOME** + **LIMIT**[46], como podemos ver na Figura 5. A sequência **overcome a limit** não foi encontrada no SkELL. O verbo **overcome** aparece usado em combinação com **limitations**, como se pode ver na Figura 6.

[46] A busca foi feita em **Collocates** no COCA. **OVERCOME** *entrou na caixa* **Word/frase** e **LIMIT** na caixa **Collocates**, com um alcance de 4 palavras à direita e 4 à esquerda.

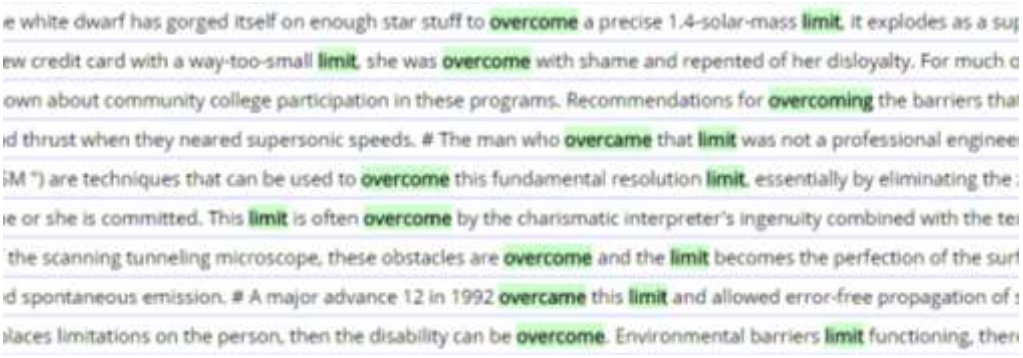


Figura 5: Extrato de concordância de *OVERCOME* + *LIMIT* no portal COCA

Figura 6: Extrato de concordância de *OVERCOME* + *LIMITATION* no portal SkELL

A revisora, ao modificar o texto, mudou o tempo verbal do presente simples para o presente perfeito (**present perfect**) e trocou *it* por *so,* mas manteve a mesma sequência, como pode se ver abaixo, no Exemplo 8:

(8) I can't think of anyone in the world who **has never overcome a limit** or **wished** to do **so**.

Em outras palavras, embora a sequência **overcome** + **a** + **limit** usada pelo aprendiz não seja encontrada no COCA nem no SkELL, é uma sequência possível. Ela respeita a norma culta da língua, um verbo transitivo seguido de um substantivo no singular, precedido pelo artigo indefinido, e faz sentido – um exemplo do princípio da escolha aberta tipo I – o princípio descrito por Sinclair (1991, 2004).

Noutro trecho (Exemplo 9), há um exemplo do princípio da escolha aberta que denominamos tipo II. O aluno usa a sequência **insist** + **in** + **his** + **goals**. Na janela,

cujo nódulo é **INSIST**, e **BOY**, **CONTINUE** e **GOAL** são candidatos a colocado, a colocação **INSIST** + **GOAL** não foi atestada no *corpus* de referência. Também no COCA não foi encontrado registro correspondente para a sequência **insist in his goal**, embora haja 33 ocorrências para **INSIST** + **GOAL**[47]. Gramaticalmente, no entanto, essa sequência é possível: um verbo (**insist**) seguido de preposição (**in**), pronome possessivo (**his**) e, por fim, o substantivo (**goals**). Ou seja, nesse exemplo específico, o aprendiz segue as regras da gramática normativa e da sintaxe da língua, mas usa uma palavra cujo significado não expressa o sentido desejado. O verbo **insist** não é adequado ao texto, como podemos verificar pelo seu significado em inglês[48]: dizer ou demandar alguma coisa com firmeza, principalmente quando alguém discorda ou se opõe ao que você diz. Também a preposição escolhida não está adequada. O modelo de abertura-e-enchimento (**slot-and-filler**), construído por regras gramaticais, e que tem como limite único a gramaticalidade (**gramaticalness**), como sintetizado por Sinclair (1991), foi seguido pelo aprendiz, mas ele não consegue se expressar a contento na língua inglesa, produzindo uma combinação textual que não faz sentido nesse idioma.

> (9) On the other hand, there were some friends that liked the fact that the boy **continue** [sic; **continued**] and **insists** [sic; **insisted**] **in** [sic; **on**] **his goals** until the end, and stayed really connected to the story.

Na versão alterada, a revisora opta por modificar a pontuação do aluno, apesar de correta, e acrescenta o sujeito que estava faltando na frase seguinte. Ela não usa **there were**, o que deixa o texto mais conciso, e substitui **insist**, que não tem o sentido apropriado, por **pursue**, na colocação **PURSUE** + **GOAL**, como mostra a passagem abaixo (Exemplo 10). A colocação **PURSUE** + **GOAL** é bastante idiomática no inglês; exemplos são oferecidos na Figura 7.

[47] Essa busca foi feita em **Collocates**, com **INSIST** na caixa **Word/phrase** e **GOAL** na caixa **Collocates**, com um alcance de 4 palavras à esquerda e 4 à direita.
[48] "to say firmly or demand forcefully, especially when othersdisagree with or oppose what you say" Disponível em: <http://dictionary.cambridge.org/dictionary/english/insist> Acesso em: 7/1/2017.

(10) On the other hand, some friends liked the fact that the boy continued to **pursue his goals** until the end**. These friends** stayed really connected to the story.

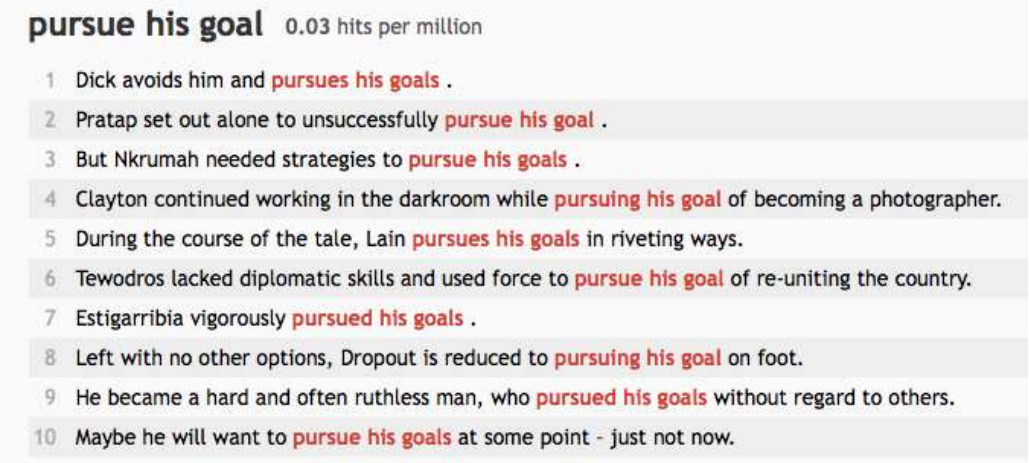


Figura 7: Extrato de concordância de **PURSUE +HIS +GOAL** no portal SkELL

Em suma, no texto 01 há indícios do princípio da escolha aberta. Isso significa que o aprendiz conhece e usa satisfatoriamente as regras da gramática normativa, o que não garante que sempre consiga expressar perfeitamente as suas ideias. Para tal, também é necessário conhecer o significado das palavras escolhidas, e avaliar se estão adequadas ao sentido desejado para o texto no qual estão inseridas.

Tal como com o princípio idiomático, observamos que, além do princípio descrito por Sinclair (1991, 2004), há também uma pequena variação desse princípio nos textos escritos pelos aprendizes. A essa variante intitulamos o nome de princípio da escolha aberta do tipo II.

4.3. COMO SÃO DISTRIBUÍDAS AS COLOCAÇÕES AO LONGO DOS TEXTOS E O QUE ISSO MOSTRA SOBRE A INCIDÊNCIA DOS PRINCÍPIOS IDIOMÁTICO E DE ESCOLHA ABERTA?

Ao explicar como o significado se manifesta no texto, Sinclair (1991, 2004) desenvolve dois princípios de interpretação – o princípio idiomático e o princípio da escolha aberta. Ele propõe que os dois princípios coexistem nos textos, são

necessários para o seu entendimento, e que a mudança de um modelo para o outro é abrupta.

O Anexo 1 apresenta dois dos gráficos produzidos pelo rastreador de colocações, representando a distribuição das colocações nos textos 1 e 17 dos aprendizes. Esses gráficos representam visualmente a distribuição colocacional e, no eixo vertical, mostram a porcentagem de colocações por janela. Assim, quanto mais alta a linha, maior é a incidência de colocações, e, portanto, sobressai o princípio idiomático no texto; por outro lado, quanto mais baixa a linha, menor é a incidência de colocações, e, por conseguinte, prevalece no texto o princípio da escolha aberta[49]. Verificamos que não há uma distribuição homogênea na incidência dos princípios, de modo que uma parte do texto fosse toda idiomática e a outra seguisse o modelo abertura-e-enchimento, ou que um texto fosse todo idiomático, ou outro apenas tivesse exemplos de combinações de palavras que sigam exclusivamente o modelo abertura-e-enchimento. O texto com mais janelas acima da linha 50 indica que os alunos usaram mais de 50% de colocações da língua inglesa por trecho selecionado pela janela, e que, consequentemente, é mais idiomático e com prevalência do princípio idiomático. O texto cujo gráfico está majoritariamente abaixo da linha 50, por sua vez, assinala que a porcentagem de colocações por janela está abaixo de 50%, e que os textos têm maior incidência do princípio da escolha aberta. Em outras palavras, os dois princípios coexistem alternadamente nos textos dos aprendizes.

Em comparação, no Anexo 2 observamos os gráficos dos mesmos textos do Anexo 1, porém após a revisão por falante nativo. Esses gráficos são mais idiomáticos do que os originais escritos pelos alunos – todavia, verificamos que essa alternância entre os dois princípios, observada anteriormente, se manteve. Alguns textos têm uma maior quantidade de linhas altas, indicando uma alta porcentagem de colocações por janela, logo, maior incidência de colocações – evidências do princípio idiomático nas redações. Outros têm maior quantidade de linhas mais baixas, abaixo da linha 50, apontando uma menor incidência de colocações – evidências do princípio da escolha aberta na produção escrita dos alunos, mesmo após tendo sido revisadas por uma qualificada e competente corretora.

[49] Não há distinção aqui entre os tipos I e II do princípio idiomático e da escolha aberta.

Esse perfil também ocorre com os textos produzidos por jornalistas nativos da língua portuguesa (Berber Sardinha 2014a, 2014b), cujos gráficos com o mapeamento das colocações estão no Anexo 3. Observa-se nesses gráficos que a maior parte das colocações usadas está acima da linha 50, o que significa que os textos têm acima de 50% de porcentagem compartilhada com o *corpus* de referência usado no estudo. Em outras palavras, são textos idiomáticos. Mas também se observa que há várias colocações abaixo da linha 50, com vários picos na linha zero, o que significa que têm 0% de porcentagem compartilhada com o *corpus* de referência usado, indicando mudanças abruptas do princípio idiomático para o princípio da escolha aberta. Dito de outra forma, são passagens nos textos com 50% ou menos de colocações, ou seja, são evidências da incidência do princípio da escolha aberta em textos produzidos por conhecedores da língua.

O que se constata, em conclusão, ao compararmos os gráficos dos textos produzidos pelos aprendizes de língua inglesa (Anexo 1), os dos mesmos textos revisados pela corretora americana (Anexo 2), assim como os dos textos escritos por conhecedores da língua portuguesa (Anexo 3), é que os dois princípios – o princípio idiomático e o da escolha aberta – se alternam, confirmando o que foi proposto por Sinclair (1991, 2004).

5. DISCUSSÃO DOS RESULTADOS

Ao desenvolver o conceito do princípio idiomático e da escolha aberta, Sinclair (1991, 2004) usou textos de falantes nativos da língua inglesa. De modo semelhante, Berber Sardinha (2014a), ao investigar colocações no português do Brasil, utilizou textos de jornalistas brasileiros falantes nativos. Diferentemente desses dois estudos, o *corpus* de estudo aqui analisado é constituído por redações de aprendizes da língua inglesa, o que nos exigiu repensar questões levantadas por esses trabalhos em relação à linguagem de aprendizes.

Ao se avaliar a incidência do princípio idiomático nos textos dos alunos, observou-se que, por vezes, eles usam colocações do inglês que não condizem com o significado desejado ou esperado naquele contexto. A esse tipo de incidência se deu o nome de princípio idiomático tipo II. À incidência do princípio idiomático nos textos,

tal qual definido por Sinclair (1991, 2004), deu-se o nome de princípio idiomático tipo I.

Na análise da incidência do princípio da escolha aberta nos textos dos aprendizes brasileiros de língua inglesa, também se observou que, por vezes, os alunos usam combinações de palavras que, embora a princípio gramaticalmente possíveis, não fazem sentido na língua inglesa. Nesses casos, os aprendizes levaram às últimas instâncias o modelo de abertura-e-enchimento (**slot-and-filler**), construído por regras gramaticais, e que tem por único limite a gramaticalidade (**gramaticalness**) (Sinclair 1991). A esse tipo de incidência se deu o nome de princípio da escolha aberta tipo II. Denominamos o princípio da escolha aberta, tal qual caracterizado por Sinclair (1991, 2004), de princípio da escolha aberta tipo I.

Mas o que significa um texto estar mais próximo do princípio da escolha aberta ou do princípio idiomático?

Um texto estar mais próximo do princípio idiomático significa que o autor aprendiz empregou um bom repertório de frases semiconstruídas, e que o texto é idiomático. Contudo, diferentemente do que acontece num texto de um falante letrado que tem o português como língua materna, como no caso do estudo de Berber Sardinha (2014a), ou o inglês, no caso de Sinclair (1991, 2004), esse dado por si só não garante que o aluno tenha usado colocações adequadas ao seu texto. Pode ser que parte de sua escolha lexical, em vez de fazer com que seu texto soe natural e atenda às expectativas de um leitor de língua inglesa, comprometa o próprio sentido da ideia que o aluno deseja expressar.

Um texto de aprendiz mais próximo do princípio da escolha aberta também apresenta nuances que não são existentes em textos produzidos por falantes letrados da língua. Pode indicar que o aluno usou combinações pouco comuns, talvez relacionadas à própria especificidade do assunto sobre o qual está discorrendo, mas que fazem sentido e expressam com acuidade o seu ponto de vista. Mas podem também sinalizar que, além de não ser uma combinação possível, não traduzem o sentido adequado ao texto no qual está inserido. Contudo, tanto no primeiro caso quanto no outro, o texto é pouco idiomático.

CONSIDERAÇÕES FINAIS

Conforme mencionado na Introdução, uma das dificuldades do aprendiz da língua inglesa é conseguir se expressar de modo natural e fluente. Muitas vezes produzem textos, que embora gramaticalmente adequados, soam truncados e artificiais. Uma característica definidora da naturalidade e fluência de um texto é o uso de colocações (Berber Sardinha 2014a; O'Keefe, McCarthy, Carter 2007). O uso de colocações, por sua vez, ilustra a incidência do princípio idiomático, assim como a sua ausência, indica a incidência do princípio da escolha aberta.

Nesta pesquisa, analisamos a ocorrência de colocações em cada um dos textos de aprendiz que compõe o *corpus* de estudo, de modo a avaliar a incidência do princípio idiomático e da escolha aberta na produção escrita de cada um dos alunos. Estabelecemos que uma porcentagem compartilhada igual ou abaixo de 43% indica que o texto tem uma maior incidência do princípio da escolha aberta.

Essas observações a respeito da incidência do princípio idiomático ou o da escolha aberta nas redações dos aprendizes de inglês chamaram a nossa atenção para nuances que não tinham sido pensadas para os textos de falantes letrados nativos, possivelmente por não ter se investigado a incidência desses princípios em *corpora* de aprendizes.

Numa futura pesquisa, pode ser verificado se 43% é um valor limite indicativo do princípio da escolha aberta que pode ser aplicado a outros *corpora* indiscriminadamente, assim como analisar outras variedades textuais. Outra possibilidade seria examinar até que ponto a língua materna influencia o uso de colocações, e consequentemente, a incidência do princípio idiomático, ao comparamos a produção escrita de alunos de diversas nacionalidades.

REFERÊNCIAS

BAISA, Vít; SUCHOMEL,Vít. SkELL: Web Interface for English Language Learning. In: *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Tribun EU, 2014. ISSN 2336-4289.

BARNBROOK, Geoff; MASON, Oliver; KRISHNAMURTHY, Ramesh. *Collocation: Applications and Implications*. Basingstoke: Palgrave Macmillan, 2013.

BERBER SARDINHA, Tony. *Linguística de* corpus. São Paulo: Manole, 2004.

BERBER SARDINHA, Tony. *Pesquisa em linguística de* corpus *com Wordsmith Tools*. Campinas: Mercado das Letras, 2009.

BERBER SARDINHA, Tony. Looking at Collocations in Brazilian Portuguese through the Brazilian *corpus*. In: BERBER SARDINHA, Tony; FERREIRA, Telma de Lourdes São Bento (Eds.). *Working with Portuguese Corpora*. London: Bloomsburry, 2014a.

BERBER SARDINHA, Tony. Freedom of combination and heterogeneity: a corpus linguist's look at two Saussurean insights. In: *Matraga*. Rio de Janeiro, v. 21, n. 34, jan./jun. 2014b.

BIBER, Douglas. Register as a Predictor of Language Variation. In: *Corpus Linguistics and Linguistics Theory*. De Gruyter, n. 8, v. 1, 2012. p. 9-37.

BIBER, Douglas; CONRAD, Susan; CORTES, Viviana. 'If you look at...': Lexical bundles in university teaching and textbooks. *Applied Linguistics*, n. 25, v. 3, 2004, p. 371-405.

BIBER, Douglas; CONRAD, Susan; REPPEN, Randi. *Corpus Linguistics*: *Investigating Language Structure and Use*. Cambridge: CUP, 1998.

Common European Framework of Reference for Languages: learning, teaching, assessment. Disponível em:

<http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf>. Acesso em: 4 jul. 2016.

EBELING, Signe Oksefjell; HASSELGARD, Hilde. Learner Corpora and Phraseology. In: GRANGER, Sylviane; GILQUIN, Gaëtanelle; MEUNIER, Fanny (Eds.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP, 2015.

ELLIS, Nick et al. Learner Corpora and Formulaic Language in Second Language Acquisition Research. In: GRANGER, Sylviane; GILQUIN, Gaëtanelle; MEUNIER, Fanny (Eds.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP, 2015.

GIL, Cristina Borges. *A incidência do princípio idiomático e do princípio da escolha aberta na produção escrita de alunos brasileiros de inglês como língua estrangeira*. Dissertação (Mestrado em Linguística Aplicada). LAEL, PUC, São Paulo, 2017.

GRANGER, Sylviane. *Learner English on Computer*. London; New York: Addison Wesley Longman, 1998.

GRANGER, Sylviane. A Bird's Eye View of Learner *Corpus* Research. In: GRANGER, Sylviane; HUNG, Joseph; PETCH-TYSON, Stephanie (Eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins, 2002.

GRANGER, Sylviane; BESTGEN, Yves. The Use of Collocations by Intermediate vs. Advanced Non-Native Writers: A Bigram-Based Study. In: De Gruyter, *IRAL,* n. 52, v. 3, 2014.

HOEY, Michael. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

HOEY, Michael. A Word Beyond Collocation: New Perspectives on Vocabulary Teaching. In: LEWIS, Michael (Ed.). *Teaching Collocation:* Further Developments in the Lexical Approach. Boston: Thomson Heinle, 2000.

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

JAKUBÍČEK, Miloš; KILGARRIFF, Adam; KOVÁŘ, Vojtěch, RYCHLÝ, Pavel; SUCHOMEL, Vít. The TenTen corpus family. In: *7th International Corpus Linguistics Conference CL*, July, 2013. Anais...

LEECH, Geoffrey Neil. *Semantics: The study of meaning*. Harmondsworth: Penguin, 1974.

O'KEEFE, Anne; McCARTHY, Michael; CARTER, Ronald. *From Corpus to Classroom:* Language Use and Language Teaching. Cambridge: CUP, 2007.

PAQUOT, Magali; GRANGER, Sylviane. Formulaic language in learner corpora. In: *Annual Review of Applied Linguistics*, 32, March 2012, DOI: https://doi.org/10.1017/S0267190512000098

PARTINGTON, Alan. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins, 1998.

RYCHLÝ, Pavel. A Lexicographer-Friendly Association Score. In: SOJKA, Petr; HORÁK, Aleš. (Eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008*. Brno: Masaryk University, 2008.

SINCLAIR, John. *Corpus, Concordance, Collocation*. Oxford: OUP, 1991.

SINCLAIR, John. *Trust the Text*: *Language, Corpus and Discourse*. London: Routledge, 2004.

STEFANOWITSCH, Anatol; GRIES, Stephan Th. Collostructions: Investigating the interaction of words and constructions. In: *International Journal of Corpus Linguistics*, n. 8, v. 2, 2003, p. 209-243.

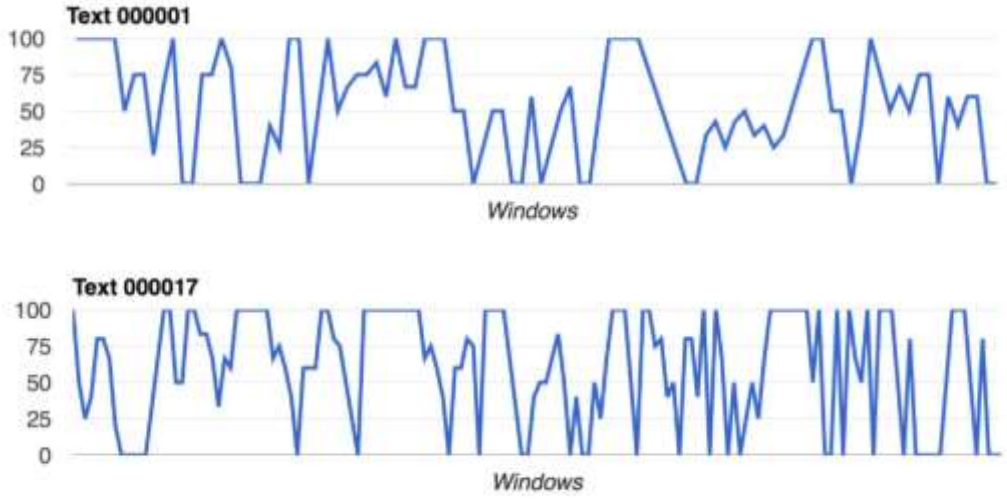
STUBBS, Michael. The search for units of meaning: Sinclair on Empirical Semantics. In: *Applied Linguistics*, n. 30, 2009, p. 115-137.

STUBBS, Michael. Collocations and Semantic Profiles: on the Cause of the Trouble with Quantitative Studies. In: *Functions of language*, n. 2, v. 2, dez. 1995.

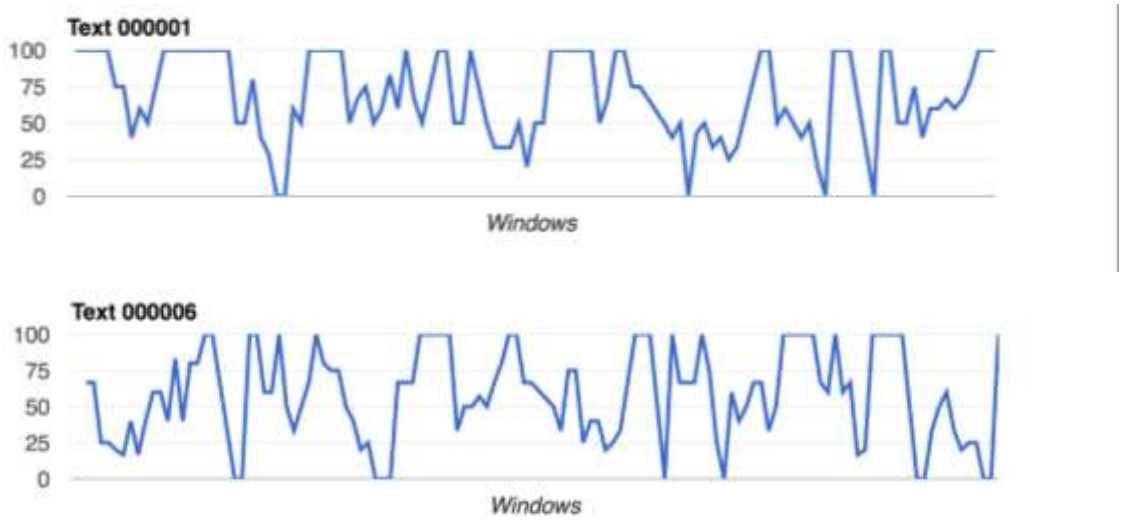
Artigo recebido em 05 de dezembro de 2022.
Artigo aceito para publicação em 04 de março de 2023.

Anexos

Anexo 1: Gráficos com as distribuições de colocações nos textos dos alunos



Anexo 2: Gráficos com as distribuições de colocações nos textos revisados selecionados para a análise qualitativa



Texto revisado	Texto original do aprendiz
000001	1
000006	17

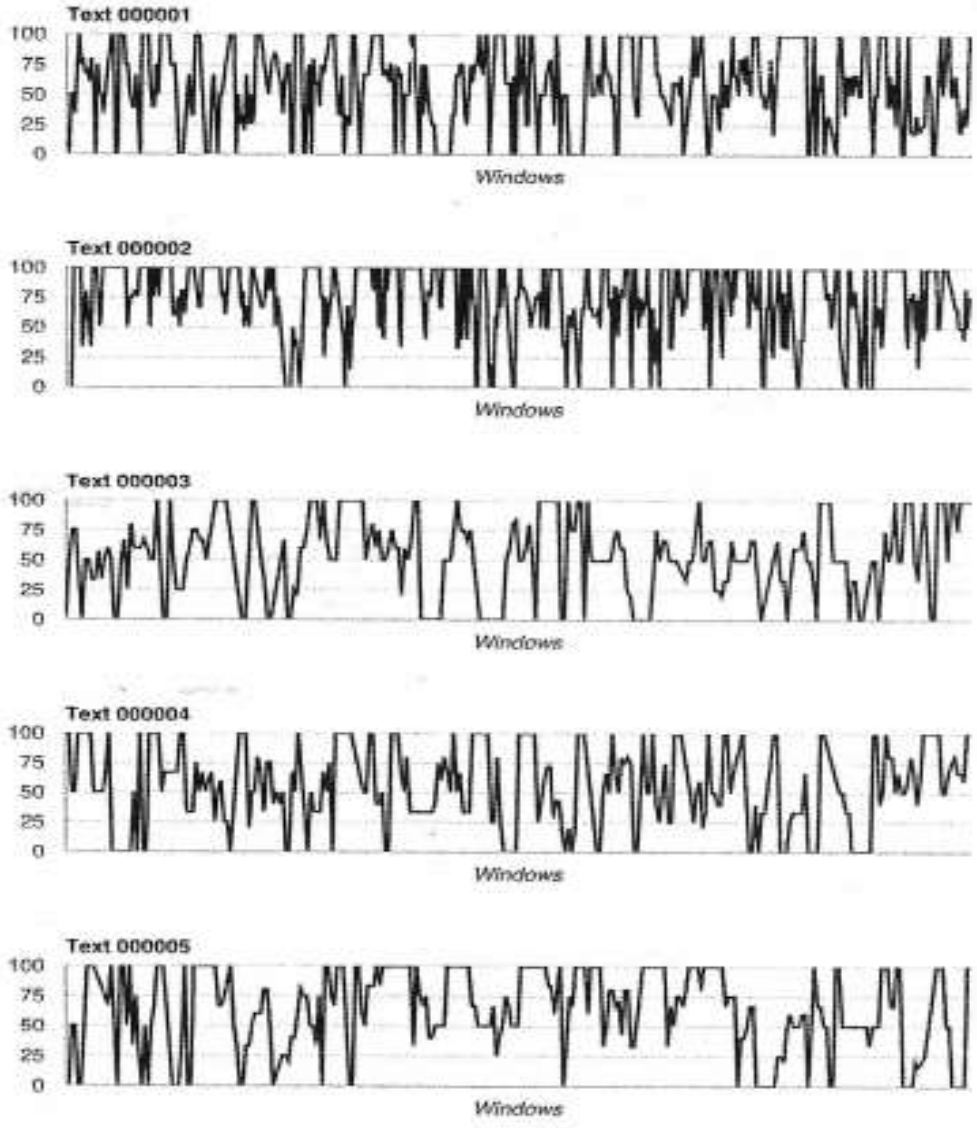


Tabela Anexo 2: Numeração dos textos revisados e dos textos dos aprendizes

Anexo 3: Gráficos com as distribuições de colocações por janela em textos de jornalistas brasileiros (Berber Sardinha 2004b: 198)