A ASSOCIAÇÃO DO VOCÁBULO *ACCESS* COM *INTERNET* NO REGISTRO ESCRITO DE DOIS GÊNEROS TEXTUAIS EM INGLÊS: UM ESTUDO À LUZ DA LINGUÍSTICA DE *CORPUS*

Jesiel Soares Silva

RESUMO

Este trabalho é o resultado de uma investigação empírica acerca da força de associação do termo *access* com *internet* entre um corpus de textos escritos formal e um informal. Usamos como medida de associação a *Multual Information*(I = Log_2 O/E).). Como resultado, verificamos uma saliência quando o *access* é colocado com *internet* (5,13 e 31,19, respectivamente). Esta saliência apontou uma força de associação de 4,8 (formal) e 7,2 (informal).

PALAVRAS-CHAVE: linguística de corpus; força de associação; colocados.

1. Considerações iniciais

Postular que a língua está em constante transformação é notadamente um lugar comum nos estudos acerca da linguagem. Ao longo dos anos, várias questões de ordem social, econômica, geográfica, tecnológica têm promovido mudanças significativas na maneira em que a língua opera entre os atores que a realizam.

Um fator determinante nessas mudanças pode ser atribuído ao constante avanço tecnológico nos meios de comunicação, sobretudo após o surgimento da internet. Devido à abrangência e alternativas comunicativas da grande rede, muitas construções linguísticas em diversos idiomas têm sido cunhadas,

adaptadas, relocadas. Na língua portuguesa, por exemplo, vários termos em inglês, advindos do mundo digital, já são recorrentes no uso diário da língua, bem como já estão incorporados nos dicionários, por exemplo: *site*, *link*.

Além de incorporar novos vocábulos à língua, percebemos que as novas tecnologias da informação e comunicação fazem com que termos tradicionais passem a adquirir novos significados e/ou significados mais abrangentes, por exemplo, os termos *baixar* e *conectar*, que, hoje em dia, além de seus significados tradicionais, também se referem a "fazer *download"* e "ter acesso à internet", respectivamente.

Para compreendermos essas transformações de uma forma mais abrangente, é necessária uma investigação empírica da língua em uso, ou seja, um tipo de arcabouço teórico-metodológico capaz de gerar evidências acerca dessas transformações, dos contextos, das evoluções temporais. Uma área dos estudos de linguagem capaz de tal feito é a Linguística de *Corpus*, que trabalha com a compilação de textos orais, escritos ou multimodais em diversos gêneros e, por meio de procedimentos estatísticos, de mensuração, consegue identificar alguns fatores, características, recorrências, padrões da linguagem em uso.

Por intermédio da linguística de *corpus,* somos capazes de acessar e investigar a frequência das palavras, as maiores recorrências de alguns termos, bem como a associação de vocábulos com outros. Podemos ainda empreender um estudo em perspectiva diacrônica e entendermos o comportamento linguístico ao longo de um determinado espaço de tempo.

Pensando nessas questões das transformações da língua influenciadas pelo avanço tecnológico e fazendo uso da linguística de *corpus* como base teórica e metodológica, este artigo pretende fazer uma investigação empírica do vocábulo *access* da língua inglesa e buscar uma relação desse vocábulo com a palavra *internet*. Nossa hipótese é de que, devido às constantes transformações linguísticas subsidiadas pelos avanços tecnológicos, sobretudo a internet, o termo *access,* em gêneros textuais mais informais, tem sido associado fortemente ao termo *internet* de maneira mais contundente do que em gêneros mais formais.

A justificativa para a nossa hipótese reside no fato de que, por causa da entrada da internet em certa medida na vida cotidiana das pessoas em geral, a palavra *access* tem sido fortemente usada para se referir à *internet* mais do que para se referir a outros termos, principalmente em gêneros textuais mais informais, nos quais as mudanças notadamente ocorrem de forma mais rápida

do que em gêneros mais formais.

Para tal, compilamos um *corpus* de textos escritos em língua inglesa com características mais informais e iremos compará-lo a um *corpus* de um gênero mais formal, buscando a força de associação entre os dois termos mencionados. Assim, para os textos informais, construímos um *corpus* de aproximadamente 40000 palavras de *essays* (ensaios) em língua inglesa disponíveis na internet.

Esses ensaios têm características de textos mais informais por se tratarem da opinião de internautas sobre diversos temas (política, religião, educação, entre outros). Compararemos, então, esse *corpus* criado com o *Corpus of Contemporary American English* (*COCA*), na parte do gênero de escrita acadêmica, representando assim o que consideramos uma escrita mais formal.

O objetivo maior é investigar, por meio de um método estatístico, se o nível de associação do termo *access* com o colocado *internet* é mais significativo no *corpus* de textos mais informais do que no *corpus* de textos formais. Dessa forma, poderíamos fazer inferências acerca do processo de abrangência de significado do termo devido ao avanço da tecnologia, sobretudo da internet.

Este artigo está assim dividido: além das considerações iniciais, faremos uma breve revisão teórica acerca da linguística de *corpus*, bem como seu lugar epistemológico nas teorias linguísticas. Em seguida, apresentaremos a metodologia do estudo, o percurso da criação do *corpus*, assim como as medidas estatísticas utilizadas. Por conseguinte apresentaremos a análise e discussão dos dados. Por fim apresentaremos as considerações finais acerca do trabalho.

2. Fundamentação teórica: bases epistemológicas, históricas e conceituais da Linguística de *corpus*

A busca pela compreensão de algumas transformações, alcances, abrangência, contextos da linguagem tem sido objeto de diversos estudos, correntes teóricas e escolas linguísticas ao longo das décadas. A complexidade dos fenômenos da linguagem vem sendo investigada sobre diversos aspectos, tanto por um ponto de vista mais estrutural e modelos teóricos, quanto por abordagens mais pragmáticas, que consideram a língua em uso. Seja na filosofia, na filosofia de linguagem ou nos estudos linguísticos, uma dualidade sempre recorrente e teoricamente antagônica é entre a visão racionalista e a visão empírica da linguagem.

As raízes filosóficas para essa dualidade entre o racionalismo e o empirismo residem no século XVII na Europa, sobretudo entre René Descartes e seus seguidores, com o racionalismo, e alguns filósofos da escola inglesa, entre eles John Locke, com o empirismo. Para Descartes, o conhecimento provinha de ideias inatas e inerentes ao homem ao nascer. Reside neste pensamento o fato de que as ideias são universais e se diferenciam unicamente na adequação ao contexto. Um exemplo deste pensamento, para Descartes, seriam as ideias matemáticas (MORA, 1982)[1].

Contrários a isso, alguns filósofos ingleses como Francis Bacon, David Hume e John Locke postulavam que só é possível conceber o conhecimento baseado nas impressões sensitivas, ou seja, nas bases da sensibilidade dos nossos sentidos. Locke (1979) definiu a mente como um quadro em branco sobre o qual é gravado o conhecimento por intermédio das sensações. Consequentemente, a sensação, a experiência, as tentativas e os erros culminam no conhecimento. Ao admitir que o conhecimento era proveniente das sensações, Locke (1979)[2] refutou tanto a ideia de um conhecimento *a priori* quanto as ideias baseadas na abstração de conceitos não formulados pela sensibilidade e a experiência.

Nos estudos acerca da linguagem, a visão do inatismo cartesiano culminou na concepção racionalista de linguagem, cujo principal represente é Chomsky. Por outro lado, o empirismo dos pensadores ingleses foi a base epistemológica para a visão empírica da linguagem, cujo principal expoente é Halliday (BERBER SARDINHA, 2004)[3].

Embasado no modelo gerativo-transformacional de concepção da linguagem, Chomsky (1957)[4] postulou que a língua deve ser compreendida no conjunto de *possibilidades* de sua realização vinculadas à competência dos falantes. Em outras palavras, na língua há possibilidades finitas de realização das estruturas e essas podem ser definidas e estabelecidas *a priori*.

Por outro lado, Halliday (1992)[5] assevera que a linguagem é um siste-

---

1 MORA, J. F. *A filosofia analítica*: mudança de sentido em filosofia. Trad. Fernando Leorne. Rés: Porto, Portugal, 1982.

2 LOCKE, J. *An essay concerning human understanding*. New York: Oxford University, 1979.

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

4 CHOMSKY, N. *Syntactic Structures*. The Hague: Mouton & Co., 1957.

5 HALLIDAY, M. A. K. *Language as system and language as instance:* the *corpus* as a theoreti-

ma de *probabilidades*, portanto, alguns traços linguísticos são mais recorrentes que outros, a depender de determinados aspectos. Ou seja, mesmo que vários traços linguísticos sejam possíveis de ocorrer, eles não vão acontecer em frequências regulares ou equivalentes. Para Halliday (1992) a língua não deve ser analisada pelas *possibilidades* universais, mas pelas *probabilidades* contextuais, pois há uma correlação entre características linguísticas e situacionais (BIBER, 1988)[6].

É importante ressaltarmos que a variação ocorrida na frequência dessas ocorrências não é aleatória, pois está condicionada a determinados aspectos como contexto de realização, associação entre termos, prosódia semântica, entre outros. A língua tem padrões de realização e a compreensão desses padrões é fundamental quando o que se busca é fazer inferências acerca dela (BERBER SARDINHA, 2004).

Baseados em Berber Sardinha (2004), podemos afirmar que, enquanto na visão racionalista o foco das investigações está na competência linguística e nos universais linguísticos, a linguística de *corpus*, amparada pela visão empirista, contempla o desempenho linguístico e a descrição linguística. Poderíamos sintetizar que o empreendimento da linguística de *corpus* consiste na investigação da linguagem por intermédio das evidências empíricas extraídas dela através da compilação de dados linguísticos (BERBER SARDINHA, 2004).

Desde que Sinclair (1966)[7] publicou *Beginning the study of lexis,* a Linguística de *Corpus* tem se ocupado em investigar a linguagem humana em uma perspectiva empírica. A partir de então, o campo de estudos com *corpora* tem se desenvolvido em vários países e o aparato tecnológico utilizado para os estudos tem obtido avanços consideráveis. Isso torna mais eficaz a compilação de *corpora*, bem como promove uma maior sofisticação no armazenamento, formatação e análise dos dados. Além do trabalho pioneiro de Sinclair (1966), vários outros autores posteriormente fomentaram as pesquisas e teorizações

---

cal construct. In: SVARTIK, J. (org.) Directions in *corpus* linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlim/Nova York, De Gruyter, 1992. p. 61-78.

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

7 SINCLAIR, J. McH. *Beginning the study of lexis.* In: BAZELL, C. E. In memory of R. Firth. Londres: Longman, 1966, p. 410-430.

acerca dos estudos com *corpora* linguísticos (c.f BIBER, 1988; FRANCIS e KUCERA, 1982)[8].

Berber Sardinha (2004, p. 18) considerou a definição de Sanchez (1995)[9] interessante, pois nela está contida as bases e princípios da linguística de *corpus*: a origem, o propósito, a composição, a formatação, a representatividade, a extensão:

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para descrição e análise. (SANCHEZ, 1995, p. 8-9)

Uma característica importante da linguística de *corpus* é sua vinculação com o processamento computacional dos dados, pois a análise de *corpora* está sujeita à tecnologia dos computadores eletrônicos. Entretanto, para Berber Sardinha (2000)[10], antes mesmo da existência do computador já havia alguns trabalhos com característica de compilação de *corpus* – por exemplo, na Grécia antiga, quando Alexandre, o grande compilou o *corpus* helenístico, ou ainda na Idade Média com os grandes armazenamentos de citações bíblicas.

Já na era dos computadores, o marco principal foi o *Taggit,* primeiro etiquetador morfossintático para computador, lançado em 1970 (BERBER SARDINHA, 2004) e desde então a instrumentação computacional tem se desenvolvido significativa e constantemente, facilitando o aprimoramento no tratamento com os dados de *corpora*.

---

[8] FRANCIS, W. N.; KUCERA, H. *Frequency analysis of English usage*: lexicon and grammar. Boston: Houghton Mifflin, 1982.

[9] SANCHEZ, A. *Definición e historia de los corpos.* In: SANCHEZ, A. et al. (orgs.). CUMBRE: *corpus* linguístico de español contemporaneo. Madri: SGEL, 1995, p. 7-24.

[10] BERBER SARDINHA, T. Linguística de *corpus*: histórico e problemática. D.E.L.T.A. Vol. 16, N.º 2, 2000, p. 323-367.

Na Linguística de *Corpus*, uma das formas de se investigar a recorrência de termos é através de seus colocados. Podemos analisar um termo (*node* ou nódulo) e o padrão que se forma em relação a seus colocados mais próximos. Assim, é possível, por intermédio de medidas estatísticas, mensurar o nível da força de associação de dois termos e delimitar se essa associação é aleatória ou se há alguma saliência estatisticamente significativa.

Os estudos sobre colocados têm sido foco de várias investigações da linguística de *corpus* e, para Berber Sardinha (2004), é o tipo de estudo com *corpora* mais recorrente na atualidade no meio científico. Podemos considerar que os estudos sobre colocados foram introduzidos por Firth, quando ele usou a célebre frase: ""Uma palavra deve ser julgada por sua companhia"[11] (BERBER SARDINHA, 2004, p. 41).

Outros dois grandes pioneiros do trabalho com colocados foram Benson e colaboradores (1986)[12], que propuseram um dicionário de combinações, e Kjellmer (1994)[13], que elaborou o primeiro dicionário de colocações baseado em *corpus* (Dicionário BBI de colocações). Este trabalho foi conduzido pela observação de padrões recorrentes que foram identificados e medidos estatisticamente.

Levando em conta essa concepções e seguindo as definições de Partington (1998, p. 16-17)[14], nosso trabalho se encaixa no *estudo colocacional de ordem estatística*, que averigua a probabilidade significativa de ocorrência de alguns colocados em relação a um *node.*

3. Metodologia: A criação do *corpus*, o *corpus* de referência e a medida de associação

Como mencionado na parte introdutória, pretendemos analisar o nível de associação do termo *access* com o colocado *internet* em dois contextos de

---

[11] *You shall judge a word by the company it keeps.*

[12] BENSON, M., BENSON, E., ILSON, R. (orgs.) *The BBI dictionary of english word combinations.* Amsterdã/Filadélfia: John Benjamins, 1986.

[13] KJELLMER, G. A. *A dictionary of English collocations:* based on the Brown *Corpus*, v. 3. Oxford: Oxford University Press, 1994.

[14] PARTINGTON, A. *Patterns and meanings:* using *corpora* for english language research and teaching. Amsterdã/Filadélfia: John Benjamnins, 1998.

registro escrito da língua inglesa: um mais formal (escrita acadêmica) e o outro mais informal (ensaios livres).

Como registro formal, utilizaremos os dados do *corpus* de referência *COCA,* na parte de escrita acadêmica. Como registro informal, construímos um *corpus* pequeno de aproximadamente 40000 palavras contendo textos de caráter informal. Para fazer a comparação entre os dois *corpora*, utilizaremos a medida estatística de *MI* (*Multual Information*), além da normalização dos dados para que haja a devida proporcionalidade.

Nesta parte metodológica, primeiro apresentaremos uma breve definição do *COCA*; em seguida, trataremos dos procedimentos de compilação do nosso *corpus*; por fim, apresentaremos a delimitação das medidas estatísticas nas quais nos apoiamos para a averiguação da força de associação entre os termos investigados.

3.1 O COCA e o Free Essays

O *COCA* (*Corpus of Contemporary American English*) é um *corpus* online[15] de referência que tem armazenado textos desde 1990 até 2012 e continua crescendo. O *COCA* conta hoje com aproximadamente 450 milhões de palavras e está vinculado à *Brigham Young University.* De acordo com o site do *corpus*, o *COCA* é o maior e mais balanceado *corpus* online com acesso livre.

Criado por Mark Davies, o *COCA* tem seus textos divididos em *spoken, fiction, popular magazines, newspapers*, e *academictexts.* Pelo *COCA,* é possível fazer comparações entre gêneros textuais diversos (*spoken, academic)*, bem como comparar as ocorrências com base na evolução do tempo (desde 1990).

Para fazer a comparação com os dados do gênero escrita acadêmica do *COCA*, fizemos a compilação de um pequeno *corpus* de aproximadamente 40000 palavras. Essas palavras provêm de textos de caráter informal, em forma de ensaio, dispostos no site *www.123helpme.com.*

Chamaremos esse *corpus* de *Free Essays*, visto que no site os internautas podem escrever textos sobre quaisquer temas e postar para que outros tenham acesso. Os textos avaliados, que seguem algum rigor e crivo, exigem que o internauta pague para ter acesso.

[15] http://*corpus*.byu.edu/coca/

Entretanto, há um conjunto de textos que são de livre acesso e que não têm um processo de avaliação, por isso os consideramos mais informais. A partir destes textos gratuitos e informais, compilamos nosso *corpus*, o qual convertemos para o formato *.txt*, e utilizamos o software *AntConc* para efetuarmos o armazenamento e as análises dos dados.

No processo de compilação do *Free Essays,* seguimos as orientações de Berber Sardinha (2004, p. 19). Dessa forma, os textos são autênticos em linguagem natural, e o conteúdo do *corpus* foi escolhido criteriosamente levando em consideração o gênero escrito e o caráter informal.

Em relação ao tamanho do *corpus*, compilamos o maior número de palavras que conseguimos, pois a representatividade do *corpus* está diretamente ligada à sua extensão. Ainda seguindo as orientações de Berber Sardinha (2004), em relação à tipologia, o *Free Essays* é um *corpus* escrito, contemporâneo, de amostragem, especializado e com a finalidade de estudo.

3.2 Medidas de associação: o MI Score

Para Berber Sardinha (2004), há três modelos estatísticos para se calcular a força de associação entre termos lexicais. Primeiramente, a *razão O/E* (observado/esperado), medida que exprime a probabilidade de vezes em que dois vocábulos podem ocorrer juntos dentro de um agrupamento limitado de palavras. Outra medida é o teste T (*T score),* que considera a posição dos itens analisados, ou seja, a palavra e seu colocado. Por fim, há uma medida de associação chamada de Informação Mútua (*MI score*), que utilizaremos neste trabalho.

O cálculo[16] do *MI score* também leva em consideração a razão logarítmica entre o que foi observado e o que é esperado (BERBER SARDINHA,

[16] Matematicamente, podemos dizer que *Multual Information* é a mensuração da dependência inerente que pode ser explicitada em uma distribuição conjunta de *x* e *y* por si só em função da distribuição conjunta de *x* e *y* sob o pressuposto de independência. Partindo desse modelo matemático distributivo, *Multual Information* mensura o nível de dependência da seguinte maneira: I (*x,y*) = 0 se, e somente se, *x* e *y* forem variáveis aleatórias independentes. Esse modelo pode ser expresso na equação:

$$\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) = \log 1 = 0.$$

2004) e pode ser entendido na fórmula: $I = Log_2 O/E$. O objetivo principal em um exercício de medida de *MI score* é verificar o quanto uma determinada frequência de um colocado em relação a um *node* é *não-aleatória.* Em outras palavras, quanto maior for o nível de não-randomização, teremos uma saliência estatisticamente importante no *corpus* em relação à linearidade esperada.

Posto isso, faremos a análise do termo *access* com o colocado *internet* levando em consideração quatro colocados à direita e quatro colocados à esquerda. Optamos por esse número, uma vez que gostaríamos de analisar uma possibilidade maior de associação, posto que o colocado pode se referir ao *node*, mesmo estando a certa "distância" dele.

Para não incorrermos no risco de o colocado não fazer referência ao *node*, analisamos todas as ocorrência de *access+internet (4e, 4d)* no *Free Essays,* bem como analisamos algumas ocorrências do mesmo caso no *COCA.* Assim pudemos verificar que, em um espaço de quatro *gaps* à esquerda e quatro à direita, o colocado *internet* é referente ao *node access* (confira no *Anexo*).

4. Análise e discussão dos dados

Inicialmente, para estabelecer um aspecto geral da nossa investigação acerca da força de associação dos respectivos termos, fizemos uma análise da frequência de *access* no geral *(access*),* nos dois *corpus* analisados, bem como a frequência do termo dentro do gênero de escrita acadêmica do *COCA,* uma vez que esse gênero será nossa base de comparação com o *Free Essays*. O intuito maior foi termos uma ampla visão da ocorrência *access,* independente do colocado, em ambos os *corpus*:

Tabela 1 – Frequência bruta e normalizada de *access* no *FreeEssays* e no *COCA*

Vocábulo	***corpus***	**gênero**	**frequência**	***per mi***
*access**	*FreeEssays*	------	51	138.48
*access**	*COCA*	Geral	65168	140.44
*access**	*COCA*	*academic*	25259	277.37

Como podemos observar na *Tabela 1,* tanto no aspecto geral quanto na parte de escrita acadêmica (sobretudo na parte acadêmica) e, proporcionalmente considerando a normalização dos dados (*per mi),* a ocorrência de *access*

é mais frequente no *COCA* do que no *Free Essays.* Uma primeira inferência que os dados nos apontam em relação ao *esperado* é a ideia de que, se há maior ocorrência do termo *access* no tanto no geral quando na parte acadêmica do *COCA,* ao se analisarmos o mesmo termo com qualquer colocado, esperamos que essa análise aponte uma maior frequência de ocorrência do termo com o respectivo colocado no *COCA*, ao mesmo tempo, esperamos que ocorra uma menor frequência de ocorrência no *Free Essays*.

Assim, se no *Free Essays* a ocorrência de *access* com um determinado colocado for maior ou igual ao número de ocorrência no *COCA*, teríamos uma saliência passível de ser investigada. Para averiguar melhor esta questão, precisamos entender a frequência do termo *access* distribuída dentro dos gêneros do *COCA:*

Tabela 2 – Distribuição da frequência de *access* nos diversos gêneros do COCA

	GÊNEROS					
	spoken	***fiction***	***magazine***	***newspaper***	***academic***	***All***
Freq.	7515	3391	16442	12561	25259	65168
Per mi	78.63	37.50	172.06	136.96	277.37	140.44

Como podemos notar na *Tabela 2*, a frequência da ocorrência de *access* na parte de escrita acadêmica do *COCA é* significativamente maior do que nos outros gêneros, sobretudo nos gêneros que consideramos mais informais, como o *spoken* e o *magazine*. A exceção apresentada pelos dados é em relação ao gênero *fiction*, que tem um número baixo de ocorrência comparado com os outros registros, sobretudo com o *academic*.

Ao compararmos os dois *corpora* de maneira proporcional (dados normalizados), podemos ver nas tabelas 1 e 2 que a frequência de *access* no *academic* do *COCA* (277.37) é significantemente mais expressiva que a frequência do mesmo termo no *Free Essays* (138.48).

Essa superioridade de frequência do respectivo termo no *COCA* (*academic*) em relação ao *Free Essays* nos faria pressupor que, na maioria dos casos, se *access* for exposto a algum colocado no *COCA,* a frequência de ocorrência com este colocado seria maior do que no *Free Essays.* Se acontecer o contrário, é porque há uma saliência importante que pode ser investigada; em outras palavras, haveria uma ocorrência que vai além do *esperado*.

Primeiramente, se analisarmos os principais colocados de *access* (4e, 4d), o principal colocado de *access* no academic do COCA é *information;* enquanto que no *Free Essays* é *internet*. Temos assim o primeiro indício de que *access* tem uma força de associação maior com *internet* no *Free Essays* do que no *academic* do *COCA*. Entretanto, essa ainda não é uma medida estatística. Posto isso, a seguir apresentaremos a frequência de *access* com o colocado *internet* distribuída nos gêneros do *COCA*:

Tabela 3 - Distribuição da frequência de *access*com *internet* nos diversos gêneros do *COCA*

	GÊNEROS					
	spoken	***fiction***	***magazine***	***newspaper***	***academic***	***all***
Freq.	219	39	653	583	467	1961
Per mi	2.29	0.43	6.83	6.36	5.13	4.23

Se compararmos as tabela 2 e 3 perceberemos algo importante. Quando se trata frequência de *access* no geral, com os mais diversos tipos de colocados, o *academic* do COCA aparece em destaque com um número significativamente maior do que os outros gêneros. Entretanto, quando analisamos o vocábulo *access* com o colocado *internet*, percebemos que a frequência no *academic* é menor do que em gêneros mais informais como *magazine.* Dessa forma, podemos inferir que no *academic* do *COCA* há muita frequência de *access* com diversos colocados, mas, quando se trata do colocado *internet*, a frequência é menor do que nos gêneros mais informais.

Essa inferência nos leva ao ponto chave desta pesquisa: a comparação entre a força de associação de *access* com *internet* nos dois *corpora* para, através da media estatística do *MI score*, determinarmos se a diferença entre os dois casos tem alguma significância estatística ou se apenas é uma aleatoriedade:

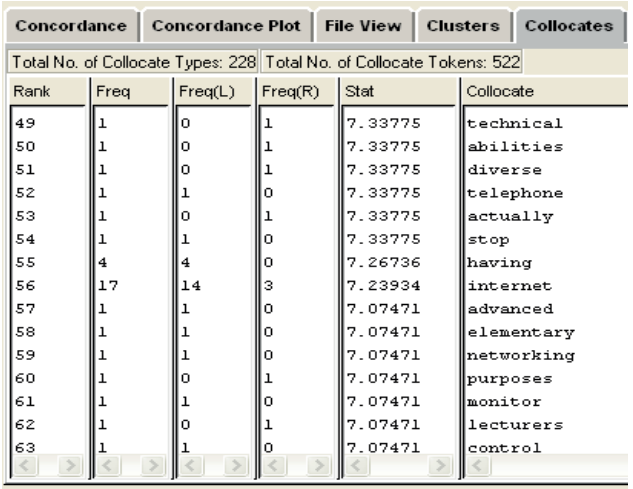
Tabela 4 – comparação entre a associação de *access* com *internet* no *FreeEssays* e no *COCA*

	Ocorrênciade *access** no geral		Ocorrênciade *access** com *internet*		
COCA (*academic*)	frequência	*per mi*	frequência	*per mi*	*MI score*
	7236 (0.06%)	**18.8**	**467 (6.43%)**	**5.13**	**4.8**
	frequência	*permi*	frequência	*per mi*	*MI score*
FreeEssays	**51 (0.12%)**	**12.6**	**17 (33.3%)**	**31.19**	**7.2**

Pela *Tabela 4*, notamos que o termo *access* tem uma maior força de associação com *internet* nas *Free Essays* (7.2) do que no acadêmico do *COCA* (4.8). Sendo assim, temos evidências estatísticas para afirmar que, nos casos estudados, a força de associação de *access* com *internet* é maior em textos informais do que formais. Para confirmar essa força de associação apresentada nos dados, no quadro a seguir são apresentados os resultados obtidos no AntConc acerca dos colocados de *access* ordenados pela medida de MI Score.

Podemos perceber neste *ranking* que vários colocados têm um valor de *MI score* próximo do valor obtido com *internet.* Entretanto, se olharmos a frequência de ocorrência dos ermos, veremos que o número de *internet* é consideravelmente maior.

Figura 1 – Ranking dos colocados de *access* considerando o MI



Vemos, por exemplo, que o colocado *advanced* tem uma força de associação de 7.07, mas houve apenas uma ocorrência no *corpus*; em outras palavras, *advanced* ocorreu apenas uma vez no *corpus* e quando ocorreu foi como colocado de *access.* Por outro lado, no caso de *internet*, há um maior número de ocorrências e ainda assim o *MI score* é 7.2, um valor significativo que corrobora com a nossa hipótese inicial.

5. Considerações finais

Retomando a ideia inicial, este trabalho mostrou um pequeno indício de que a evolução tecnológica contribui para algumas mudanças na língua, seja na criação de novos termos, seja na abrangência de significação de outros. Além disso, pudemos notar que, no gênero escrito, essas mudanças ocorrem primeiramente em textos notadamente mais informais.

Como limitação deste trabalho, temos o fato de que o *corpus* construído foi pequeno; mesmo que tenhamos normalizado os dados, tivemos apenas uma nuance da representação que pretendíamos. Sugerimos, para estudos futuros que abordem a mesma questão, que sejam feitos levantamentos históricos em *corpus* que tenham essa característica para assim se conhecer a evolução dos termos e seus colocados ao longo dos anos. Sobretudo, neste aspecto do termo *access,* seria válido fazer um levantamento de seus colocados antes e depois do surgimento da internet.

O constante avanço dos aparatos tecnológicos favorece a Linguística de *corpus* nesse aspecto, por isso acreditamos que investigar a língua em seu caráter probabilístico, buscando padrões e fazendo inferências, é uma agenda importante nos estudos acerca da linguagem humana.

THE ASSOCIATION OF THE WORD 'ACCESS' WITH 'INTERNET' IN THE TWO WRITTEN TEXT GENRES IN ENGLISH: A STUDY ON OF CORPUS LINGUISTICS

ABSTRACT:
This paper is the result of an empirical research on the association between the word 'access' with 'internet' in two corpora of written text: formal and informal. As

an association measure we made use of *Multual Information Score* ($I = Log_2 O/E$).). As a result, we verified a highlight when 'access' is displayed with 'internet'(5,13 e 31,19, respectively). This fact has guided us to an association of 4,8 (formal) and 7,2 (informal)

KEYWORDS: Corpus Linguistics; Association; Collocates

Recebido em: 21/01/2013
Aprovado em: 12/08/2013

ANEXO

Ocorrência do node *access* colocado com *internet* (4d, 4e) no *corpus Free Essays*

Essentially, this expansion along with Internet	*Access*	created a new way to access and send information
Georgia's K12 school districts have Internet	*Access*	and use a variety of on-line educational software
Internet in its simplest form enables one to	*Access*	emails; this too is a source of information
a cell phone or Blackberry is by having	*Access*	to the internet which allows students
solutions to stop students from gaining	*Access*	to the internet have been found to
money and we can still monitor your Internet	*Access*	Also for disciplinary purposes if you fail
century learning work for us. Internet	*Access*	and more constructivist teaching practices
objective was to see if well-supported Internet	*Access*	changes practice in constructivist directions
was observed had a level of Internet	*Access*	technical support and staff development
Thus, the impact of classroom Internet	*Access*	could be examined in an environment where the typical
In 1996 schools brought Internet	*Access*	into the classroom, allowing students and teachers
Yet another benefit having computers and internet	*Access*	in the classroom is the ability of students to be able
It can be quite a liability to have internet	*Access*	in the classroom; teachers must be sure they protect
In 2001, almost all public schools with internet	*Access*	(96 percent) used various technologies or procedures
technologies or procedures to control internet	*Access*	to inappropriate material on the internet
the prohibitively high cost of telephone and Internet	*Access*	charges. Educators also face the challenge of designing
to an infinite amount of information with	*Access*	to the Internet. With so much information available

Exemplos de Ocorrência do node *access* colocado com *internet* (4d, 4e) no *corpus COCA*

of the file system of the CRIS UNS server. The folder is not directly **accessible** through the **Internet**, but digital contents can be downloaded using a Java servlet.
school oersonnel can use the computers for grading, writing papers or assignments, and **accessing** information on the **Internet**. # The implications for parents are also impo
provide schools with a very convenient bad-weather backup -- as long as students' **Internet access** spreads as quickly. # Source: Chinese Web media company Sina, May 1
). As having a family decreases in importance in the next 30 years, **access** to the **Internet** will increase, suggesting that the impulse for human contact will take
and (d) share updates on action plan items. Smart phones with **Internet access** also are convenient ways to transmit information and stay connected. Because all participan
general. Shelburne noted problems with Digital Rights Management (DRM) and with **Internet access** along with other technical difficulties such as the need for special reader
, the e-books discussed are aggregations of titles purchased from vendors in package deals and **accessible** over the **Internet**. # What few articles do discuss e-readers in a
, responsible citizen participation. contributed articles challenge. ICTs, including cellphones and **Internet access**, serve as technological mediation between political administr
the world and said, " Here, fly into the world of information, **access** the **Internet** and the Worldwide Web, " they would reply, " I'm
to in the first place. But if a government were to shut down **Internet access** or ban cell phones, it would risk radicalizing otherwise pro-regime citizens or harming the
Most notably, Clinton announced funding for the development of tools designed to reopen **access** to the **Internet** in countries that restrict it. This " instrumental " approach
of its operations under the federal E-Rate program, which helps fund telephone and **Internet access** for schools and libraries. The provider was SBC Communications Inc., a
of websites by lawyers. " A lawyer website can provide to anyone with **Internet access** a wide array of information about the law, legal institutions and the value of
the stakes are even higher. # GENERAL TRENDS Despite the fact that their **Internet access** is free of outside manipulation, most Palestinian activists do not reveal their nam
of Statistics estimates that in 2009, 28.5 per cent of Palestinian households had **Internet access2** though these statistics do not account for the widespread use of hundreds
the majority of the Arab world as Israel provides the Palestinian territories with unfettered **Internet access.4** # FDD instructed ConStrat not to provide percentages for the s
hybrid learning will find a way to help bridge the gap for learners without **Internet access**. # THE LIBRARY IN THE CENTER Our blended online and face-to-face program thr
transgender people often comes through the media. Today, the **internet** is providing greater **access** to information and people are gaining this awareness at a younger age.
. 2004; 94(12):2050. # 64. Gilmour JA. Reducing disparities in the **access** and use of **Internet** health information. a discussion paper. Int J Nurs Stud.
methods to recruit panelists. To ensure participation of low-income individuals and those without **Internet access**, KN provides access to the **Internet** and hardware if need
. To ensure participation of low-income individuals and those without **Internet** access, KN provides **access** to the **Internet** and hardware if needed. 39 40 # We invited a ran
as heightened concern for the environment (Bhattarai and Hammig 2004). Increased **Internet access** (Godfray 2007) and more published field guides (Pearson and Shetterl
. # It was 1994. Newly devised graphical Web interfaces made the **Internet** widely **accessible**. Scanning technology was just becoming commercially available. In a bold mo
their growing presence include concerns about the environmental and other consequences of neoliberal globalization, **access** to the **Internet**, and the emergence of relativel
hampers the freedoms of the print and broadcast media, Malaysians enjoy widespread and uncensored **access** to the **Internet**, in large part because the government is eage