The ESPecialist: Descrição, Ensino e Aprendizagem, Vol. 38 No. 1 jan-jul 2017
https://revistas.pucsp.br/esp ©*The ESPecialist*
ISSN 2318-7115

PREPARAÇÃO DE MATERIAL DIDÁTICO PARA ENSINO DE LÍNGUAS COM BASE EM *CORPORA*

Preparing Corpus-Based Language Teaching Materials

Tony Berber SARDINHA
Pontifícia Universidade Católica de São Paulo, (PUC-SP), São Paulo, Brasil

Maria Claudia DELFINO
Faculdade de Tecnologia de São Paulo (FATEC-PG), São Paulo, Brasil

Marianne RAMPASO
Pontifícia Universidade Católica de São Paulo, (PUC-SP), São Paulo, Brasil

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

RESUMO: *O presente artigo tem por objetivo principal discutir a preparação de material didático para ensino de língua materna e estrangeira com base na Linguística de Corpus, com foco em alunos e professores de ensino básico, de idiomas e do ensino universitário. O texto tem como interlocutores principais os professores envolvidos no ensino que desejam desenvolver material de ensino que reflita o uso corrente da língua. A preparação de material baseada em corpus pode ser feita de diversas formas, tanto pelo professor quanto pelos alunos. No artigo, apresentamos sugestões de aplicação de corpora no ensino em torno de conceitos teóricos da Linguística de Corpus, quais sejam: probabilidade, padronização e variação. Espera-se que os conceitos bem como sua aplicação possam ser úteis ao ensino de línguas. O artigo inclui exemplos tanto de corpora online quanto de corpora produzidos pelos próprios professores.*

PALAVRAS-CHAVE: Material didático; Linguística de *Corpus*; Ensino de línguas

ABSTRACT: *The goal of this article is to discuss the preparation of teaching materials for teaching L1 and L2 from a corpus linguistic perspective, with a focus on students and teachers in primary and secondary schools, language institutes and universities. The major goal of corpus-based teaching materials is to have students get in contact with large quantities of evidence of language use. It is argued that such work can lead to greater learner autonomy and collaboration in the classroom. Preparing materials based on corpora can be done in many different ways, by both teachers and students. In this article, the preparation is guided by three major concepts from Corpus Linguistics research, namely probability, patterning and variation. It is argued that that the application of these concepts in the classroom can help instill confidence, independence and an awareness of linguistic patterning that are essential for successful language learning. The article includes ideas for using corpus resource and techniques illustrated with both online and self-compiled corpora.*

KEY WORDS: Teaching materials; Corpus Linguistics; Language teaching

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

1. Introdução

Neste artigo, apresentamos uma proposta para preparação de material didático para ensino de língua estrangeira com base em aportes teórico-metodológicos da Linguística de *Corpus*, tendo em mente alunos e professores da educação básica, universitária e de institutos de idiomas. Assim, a fim de justificar o nosso aporte teórico, apresentamos um breve panorama da Linguística de *Corpus*, uma área dos estudos da linguagem que se ocupa da coleta e análise dos chamados *corpora* (plural da palavra latina *corpus*), que são coletâneas de textos falados e escritos mantidos em arquivo de computador (BERBER SARDINHA, 2004). A Linguística de *Corpus* como a conhecemos hoje deve muito a pioneiros, como John Sinclair, que nos anos 1960 coletaram dados textuais e os transferiram para o computador e, a partir daí, analisaram aspectos como a frequência das palavras e, mais importante, a probabilidade de atração mútua e coocorrência entre palavras nos textos. Do outro lado do Atlântico, Henry Kucera e Nelson Francis dedicavam-se a criar um *corpus* eletrônico pioneiro, o *corpus* Brown (assim chamado porque foi construído na Universidade Brown, em Rhode Island, EUA), que tinha a extensão (inimaginável para a época) de um milhão de palavras. O *corpus* Brown foi utilizado em muitas pesquisas desde então, para desvendar vários aspectos do inglês. Desde então, muitos outros *corpora* foram coletados, de diversas línguas, para diversas finalidades. Ambos casos ilustram como a necessidade de coletar dados linguísticos antecedeu em muito a disponibilidade da tecnologia; ou seja, em épocas em que os computadores eram escassos, caros e limitados a grandes empresas e universidades, linguistas se viram dispostos a enfrentar as dificuldades de acesso e as limitações técnicas a fim de conseguir respostas para perguntas que muitos julgavam já terem sido respondidas por meios outros do que pelos dados, como pela teoria linguística, por instrumentos como dicionários e gramáticas ou pela própria intuição de falante nativo. Hoje sabemos, por meio das pesquisas com base em *corpora*, que esses pioneiros tinham razão em buscar uma descrição empírica da língua em uso, pois os *insights* das pesquisas com *corpora* são ricos e informativos, desvendando a complexidade e beleza da língua em uso.

Hoje em dia, o texto digital, em vez da exceção, é a norma. Cada vez mais, o texto impresso perde espaço para o texto em formato eletrônico, como jornais, revistas e livros. Ao mesmo tempo, qualquer indivíduo munido de um *smartphone* produz textos digitais em formatos que não possuem equivalente no mundo analógico. Alunos e professores vivem, assim, em um mundo digital, cercados de textos escritos, falados e visuais, que por sua vez tornam a existência de *corpora* eletrônicos muito mais palpáveis do que há cinquenta anos. Qualquer pessoa pode coletar seu próprio *corpus* com facilidade, a partir da web, por exemplo, em quantidade e variedade sequer sonhada pelos precursores da Linguística de *Corpus.* Ao mesmo tempo, o que vemos é uma escassez de materiais de ensino de língua materna e estrangeira baseados em *corpora* em todos os níveis.

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

2. Metodologia

Atualmente há diversos *corpora* prontos que podem ser acessados online pelos usuários. Destacamos aqui dois portais que concentram uma gama de *corpora* de línguas diferentes: os *corpora* BYU (da Universidade Brigham Young, EUA; corpora.byu.edy) e os *corpora* do portal SketchEngine (sketchengine.co.uk). A vantagem desse tipo de *corpus* é que o usuário não precisa ocupar seu computador com os textos, nem instalar programas de computador para analisar os dados. Devido a seu tamanho, o acesso local a esses *corpora* seria bastante lento; na versão online, o acesso é geralmente muito rápido.

Dentre os *corpora* da BYU, destacamos o COCA (*Corpus of Contemporary American English*), cujo acesso é disponibilizado online gratuitamente. Este *corpus* é composto por mais de 520 milhões de palavras, distribuídas em 220.225 textos, e dividido em cinco seções: língua falada, ficção, revistas, jornais e escrita acadêmica. Já no portal SketchEngine é possível encontrar mais de 200 *corpora* de dezenas de línguas. Aqui destacamos dois *corpora* desse portal, o enTenTen, de inglês, e o ptTenTen, de português. O primeiro é composto por aproximadamente 23 bilhões de palavras, e o segundo, por 3,2 bilhões de palavras. O SketchEngine também apresenta uma ferramenta online gratuita, o SkELL (*SketchEngine for English Language Learning*; skell.sketchengine.co.uk), composto por mais de um bilhão de palavras, de fácil acesso e manuseio, destinada a professores e alunos de inglês como língua estrangeira.

Uma segunda alternativa é obter um *corpus* coletado por terceiros e salvá-lo no seu próprio computador. Alguns *corpora* acima mencionados, como o *Brown Corpus* e o COCA podem ser adquiridos e usados pelos interessados. Além desses, o *Corpus* Brasileiro (um bilhão de palavras de português brasileiro; CEPRIL, LAEL, PUCSP, Fapesp, CNPq) pode ser também obtido para uso próprio. A fim de ilustração, citamos o *corpus Business English Corpus* (BEC), que foi compilado por Nelson (2000) e é composto por 1076 textos falados e escritos de linguagem de Inglês para os Negócios produzidos por falantes nativos, contabilizando 1.017.050 palavras e 28 registros, como contratos, entrevistas de emprego, atas de reunião, etc. O *corpus* BEC é uma alternativa para a produção de material didático para alunos de Inglês para os negócios, já que apresenta os principais registros usados por esse público-alvo.

Além disso, o trabalho com *corpora* em sala de aula possibilita ao aprendiz o trabalho com material autêntico, extraído de interações reais na língua e não baseado em linguagem elaborada para fins pedagógicos, como muitas vezes ocorre nos livros didáticos convencionais. Rampaso (2016) utilizou o *corpus* BEC para a preparação de unidades didáticas para ensino de inglês para os negócios em um contexto de aulas particulares, tentando assim aproximar o contexto profissional do aprendiz ao trabalho em sala de aula, contando com dados de falantes reais nas situações de negócios.

Conforme mencionamos, além dos *corpora* online os professores e alunos podem criar seus próprios *corpora*, a fim de satisfazer as necessidades próprias dos seus

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

contextos de ensino-aprendizagem. A fim de ilustração, citamos o *corpus Corpus of English Lyrics* (CoEL), que foi compilado por Delfino (2016), a partir de pedidos de alunos de uma faculdade de tecnologia que queriam aprender inglês através de letras de música. *Corpus* este composto por 585 letras de música das bandas Beatles, Bon Jovi e Maroon 5 e do cantor Bruno Mars, contabilizando 154.487 palavras. O *corpus* CoEL é uma alternativa para produção de material didático de Inglês Geral, onde a música torna-se o elemento central da sala de aula, levando tanto o aluno como o professor a trabalharem com a língua de forma prazerosa, utilizando material autêntico.

3. Conceitos Norteadores para Propostas de material didático

Propomos neste trabalho que a preparação de material de ensino de língua materna e estrangeira seja baseado em pesquisa com *corpora*, a fim de que o conteúdo ensinado tenha suporte em evidências de uso efetivo da língua (BERBER SARDINHA, 2013b; 2016). Esperamos, neste artigo, ilustrar algumas maneiras de concretizar essa proposta. Ao mesmo tempo, não é nossa intenção apresentar uma proposta pronta, finalizada, mas sim sugestões de aplicação de *corpora* no ensino em torno de conceitos teóricos da Linguística de *Corpus*.

Abaixo discutimos e ilustramos como e por que a probabilidade (HALLIDAY, 1993), a padronização (BERBER SARDINHA, 2013a; SINCLAIR, 1966; SINCLAIR, 1991) e a variação (BERBER SARDINHA, 2000a; BERBER SARDINHA, VEIRANO, 2014; BIBER, 1988) podem ser usadas no ensino de língua materna e estrangeira por meio da Linguística de *Corpus* (BERBER SARDINHA, 2000b, 2011). Embora os exemplos sejam de ensino de inglês como língua estrangeira, os conceitos podem ser aplicados a outras situações de ensino. Além disso, propomos o uso de *corpora* online, pois estão prontamente disponíveis aos usuários. Contudo, isso não impede que os professores e alunos criem seus próprios *corpora*, como também ilustramos aqui.

4. Probabilidade

Probabilidade é aqui entendida como a chance de um ou mais elementos linguísticos ocorrerem em uma determinada situação. A probabilidade é um conceito central do funcionamento da linguagem, pois aponta quais características linguísticas são mais pertinentes para o funcionamento da língua, variedade textual ou texto. Na sala de aula, o conceito de probabilidade ajuda, por exemplo, a selecionar quais palavras, expressões ou pontos gramaticais são mais relevantes e, por conseguinte, deveriam ser ensinados primeiro. A probabilidade de uso aparece de modo mais simples na frequência de ocorrência, que vem a ser quantas vezes uma determinada palavra aparece num determinado *corpus*, ou seja, podemos dizer que embora muitos traços linguísticos sejam possíveis teoricamente, eles não ocorrem com a mesma frequência. Berber Sardinha (2000a) nos dá um exemplo disso no nível morfossintático, onde a frequência de

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

substantivos (no inglês e, com certeza, no português) é maior do que qualquer outra categoria; cerca de 25% das palavras são substantivos (KENNEDY, 1998, p.103). Desse modo, a probabilidade de um traço ser um substantivo é maior do que outra classe gramatical. E em segundo lugar, embora seja teoricamente possível se aninhar orações relativas *ad infinitum* (o gato que está no tapete é meu, o gato que está no tapete que é meu é pardo, o gato que está no tapete que é meu que é pardo está dormindo, etc), à primeira vista a frequência de ocorrência de frases com mais de uma oração relativa é muito maior do que com sucessivas orações. Em resumo, as possibilidades da estrutura não se realizam todas com a mesma frequência.

O mais importante da diferença de frequências entre os traços é o fato de essas diferenças não serem aleatórias. Se o fossem, então o fato das possibilidades estruturais se realizarem com frequências diferentes não seria significativo, isto é, não acrescentaria informação a respeito da própria estrutura. Entretanto, pelo contrário, há um mapeamento regular entre a frequência maior ou menor de um traço e um contexto de ocorrência. Segundo Biber (1988, 1995), há uma correlação entre características linguísticas e situacionais (os contextos de uso). Sinclair (1991, p.67) argumenta que não há escolhas aleatórias na língua, ao mostrar que a linguagem é padronizada, ou seja, é formada por porções lexicais ou *chunks*, cuja maior ou menor frequência é atestada pelo uso real na língua pelos falantes.

Na atividade proposta abaixo, mostramos a frequência das palavras de um *corpus* de letras de música americanas e britânicas (CoEL; DELFINO, 2016). A tabela abaixo indica que a frequência das palavras na música pop é bem diferente daquela encontrada na linguagem em geral. Na sala de aula, os alunos podem comparar listagens de frequência de palavras de registros/gêneros diferentes a fim de perceberem como a frequência do conjunto de palavras reflete a situação de uso da língua. O professor pode preparar listas de frequência com software como WordSmith e AntConc

Tabela 1: Palavras mais frequentes nos *corpora* COEL (*Corpus of English Lyrics*) e COCA (*Corpus of Contemporary American English*)[1]

Palavra	*Ranking* no CoEL / Frequência	Textos / %	*Ranking* no COCA / Frequência (por milhão de palavras)
YOU	1 / 5.975	535 / 91,45	14 / 0,75
I	2 / 5.720	525 / 89,74	10 / 4,60
THE	3 / 4.747	549 / 93,85	01 / 23,83
TO	4 / 3.333	540 / 92,31	04 / 11,14
ME	5 / 2.788	452 / 77,26	62 / 791,46

[1] Por questões de espaço, relatamos apenas as 10 primeiras palavras mais frequentes no *corpus* CoEL e seu comparativo no *corpus* COCA.

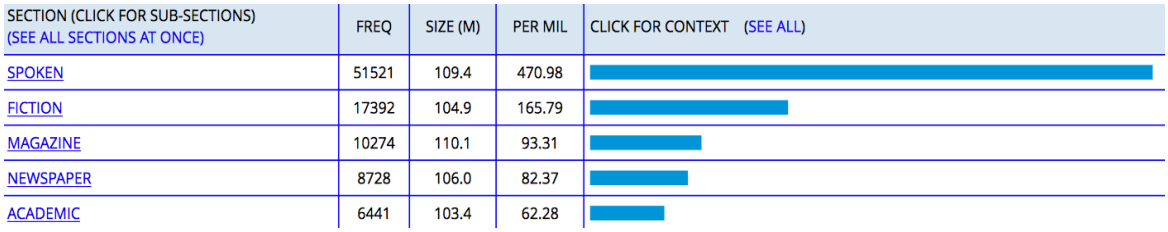
DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

AND	6 / 2.698	524 / 89,57	2 / 11,802
A	7 / 2.691	500 / 85,47	5 / 10,057
IT	8 / 1.971	402 / 68,72	11 / 4,235
MY	9 / 1.804	404 / 69,06	55 / 1,012
IN	10 / 1.711	445 / 76,07	7 / 7,996

Os alunos podem ainda discutir quais registros poderiam ter frequências parecidas com as de música pop. Nesse caso, conforme mostraram Delfino (2016) e Bertoli-Dutra (2014), o registro de música pop possui afinidades com o de conversação, visto que entre as palavras mais frequentes de ambos encontramos palavras das categorias gramaticais pronomes de primeira e segunda pessoa (*I, you*), além de verbos no presente e contrações, que são características linguísticas marcantes de um diálogo.

Além de trabalhar com frequências das palavras em um registro e de comparar registros diferentes, os alunos podem ainda usar listas de palavras-chave, que são aquelas cujas frequências são estatisticamente diferentes no *corpus* de estudo em relação ao de referência (BERBER SARDINHA, 2004, 2009). Para trabalhar com palavras-chave nas atividades de ensino, o professor pode se valer de diversos instrumentos de visualização, como os gráficos de frequência; a Figura 1 mostra um gráfico que ilustra as frequências por milhão de palavras da palavra *sort* (substantivo (‘tipo’) ou verbo (‘ordenar’)) no COCA (corpus.byu.edu/coca; Rampaso, 2016). Para produzir um gráfico desse tipo, o aluno ou professor faz uma busca no COCA por meio da opção LIST. Ao clicar nas opções da coluna SECTION, são apresentados exemplos de uso da palavra.

Figura 1: Frequência de *sort* nas seções do COCA (*Corpus of Contemporary American English*)



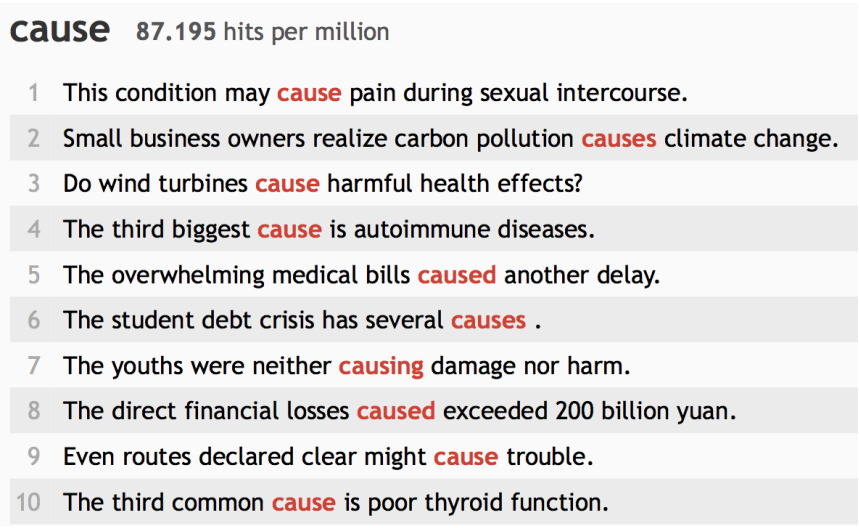
4.2 Padronização

Padronização é concebida na Linguística de *Corpus* como a tendência de as características linguísticas se aglutinarem em grupos e confluírem para formas repetidas de uso. Os principais padrões de linguagem são a colocação, a coligação, a preferência semântica e a prosódia semântica, que são formas de entender a coocorrência lexical. Por colocação, entendemos ‘a associação entre itens lexicais, ou entre o léxico e o campo semântico’ (BERBER SARDINHA, 2004; 2013a). Um exemplo de colocação seriam as combinações *career plan, career ladder*, etc. Já a coligação, conforme este mesmo autor, é a ‘associação entre itens lexicais e gramaticais ou itens gramaticais entre si’. Um

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

exemplo bem conhecido de coligação são as estruturas *start studying e begin to study*. A preferência semântica, em seu turno, enfoca a relação entre conjuntos lexicais e conjuntos semânticos (STUBBS, 2001). A título de exemplo, temos a palavra *jam*, que se associa a outros itens linguísticos relacionados à comida, como *doughnuts* e *tarts*. Por fim, a prosódia semântica (SINCLAIR, 1991; LOUW, 1993) é um fenômeno colocacional em que ocorre a associação entre itens lexicais e instância avaliativa (p.ex. positiva, negativa, etc.). Um caso clássico de prosódia semântica do inglês é a palavra *cause* (verbo ou substantivo), que é normalmente associada a conotação negativa, como *diseases, problems*, etc. A opção EXAMPLES do site SKELL (http://skell.sketchengine.co.uk) permite que o aprendiz obtenha linhas de concordância da palavra de busca, dando-lhe a oportunidade de visualizar o contexto em que a palavra aparece e as colocações mais comuns associadas a ela. Na Figura 2 aparece um exemplo de busca de *cause* com a opção EXAMPLES no SKELL.

Figura 2: Exemplos de concordância com a palavra *cause*



Esse tipo de busca pode ser trabalhado na sala de aula com atividades deste tipo:

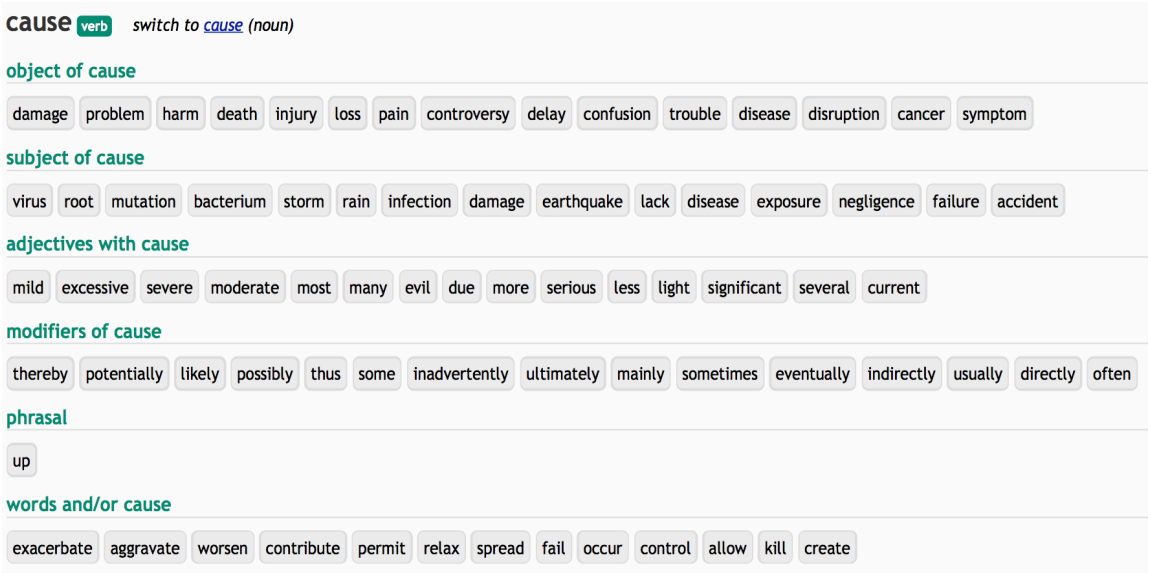
(1) Identifique as palavras que ocorrem próximas a *cause*.
(2) Marque, para cada uma, o teor da avaliação que a frase exprime: positivo, negativo ou neutro. Justifique sua resposta.
(3) Sabendo que *cause* possui essa conotação, quais outras palavras você acha que seriam atraídas por ela normalmente?

Para ajudar a responder o item 3 acima, os alunos podem usar a opção WORK SKETCH do SkELL. Essa opção mostra os itens linguísticos mais comuns que acompanham a palavra, tais como verbos, preposições, adjetivos, etc. Como se percebe

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

na figura 3 abaixo, palavras com maior probabilidade de coocorrer com *cause* são *damage, problem, harm, virus,* e *mutation*, entre outras.

Figura 3: Funções e classes gramaticais dos colocados da palavra *cause*

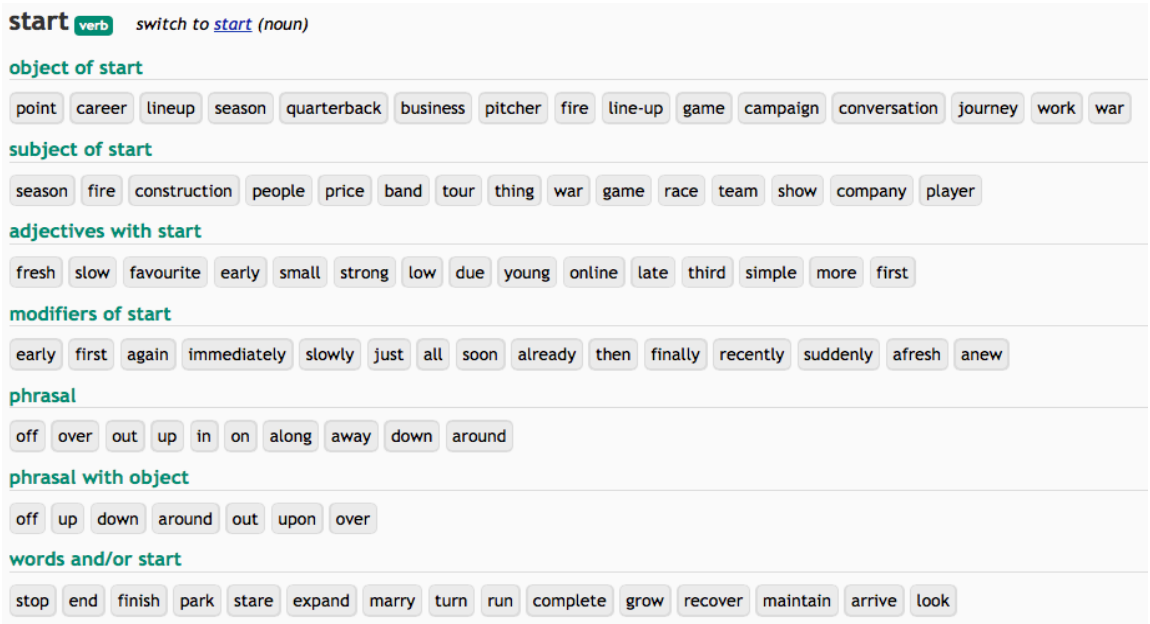


Na sala de aula, o conceito de padronização auxilia na difícil tarefa de ensinar os aprendizes a empregar formas padronizadas e esperadas de uso. Entram aqui fraseologias, combinatórias, expressões e outras formas estereotipadas, como dizer *get out of here* em vez de **leave from here* para mandar alguém ir embora. Contudo, a padronização vai muito além dessas formas fixas ou semifixas, abrangendo principalmente a expectativa de uso de uma determinada palavra devido à presença de outras ao seu redor. É essa atração ou expectativa mútua que cria o fluxo harmonioso da fala e da escrita que associamos com a fluência e boa expressão, que nós professores tanto enfatizamos em aula. É importante que fique claro que tal harmonia no fluxo lexical tem pouco a ver com a construção gramatical, pois a formação de frases sintaticamente aceitas pela norma culta não leva necessariamente à criação de palavreados (*wordings*) fluentes, naturais. Esse descompasso entre a gramática e a fluência fica talvez mais claro no ensino-aprendizado de língua estrangeira, durante o qual os aprendizes, muitas vezes, produzem frases gramaticalmente corretas mas lexicalmente canhestras, levando a reações como 'o que você escreveu não tem erro, mas soa estranho' ou 'não é assim que se diz em inglês/espanhol/etc.'. A padronização está ligada à probabilidade, uma vez que resulta do grau maior do que o normal de coocorrência entre formas lexicais. A padronização ajuda a ensinar, por exemplo, diferenças de uso entre palavras que parecem sinônimas, como *begin* e *start.* As Figuras 4 e 5 mostram os colocados mais prováveis de cada palavra. Como se nota, há diferenças e semelhanças na padronização das duas palavras: colocados como *work* e *story* são mais comuns com *begin*, enquanto *fire* e *journey*, mais típicos de *start*. Ao mesmo tempo, há semelhanças: ambas palavras são igualmente prováveis ao

lado de *career* e *season.* Os alunos podem estudar essas colocações de diversos modos, como por exemplo, por meio de *flashcards*, escrevendo o colocado de um lado e o verbo de outro. Uma outra possibilidade é o professor fazer uma tabela e pedir aos alunos para preencher com os colocados mais típicos de cada verbo.

Figura 4: Colocados de *begin*

Figura 5: Colocados de *start*



DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

4.3 Variação

O último conceito norteador da preparação de materiais didáticos que apresentamos é o de variação, e para isso nos baseamos no trabalho de Biber (1988 *et seq.*), mais especificamente na Análise Multidimensional (AMD), metodologia desenvolvida por ele para mapear a variação linguística entre variedades textuais (registros, gêneros) a partir da coocorrência de uma gama de traços linguísticos (BIBER, 1988; BERBER SARDINHA, 2000). Para tanto, busca identificar os padrões de ocorrência estabelecidos pelas relações de correlação entre os traços linguísticos dos diferentes textos do *corpus,* para em seguida comparar os registros aos quais estes textos pertencem. Nesse trabalho ela foi utilizada para que o perfil lexicogramatical pudesse ser traçado a partir das cinco dimensões de variação identificadas por Biber em seu estudo seminal de 1988. O primeiro passo da análise de um *corpus*, classificando-o a partir da análise multidimensional, é o procedimento de etiquetagem por meio do programa computacional *Biber Tagger*; na sequência, faz-se a checagem manual do *corpus* para, se necessário, corrigir problemas de interpretação de grafia das palavras pelo programa. O terceiro passo é o cálculo automático dos valores das variáveis com o auxílio do programa computacional *Biber Tag Count*, que apresenta os escores de dimensão de cada texto. As dimensões de variação de registro da língua inglesa aparecem no Quadro 2. Cada dimensão é composta por diversas características linguísticas que coocorrem nos textos. Devido ao escopo deste artigo, não é possível descrever as características linguísticas de cada dimensão; o leitor pode se referir a Biber (1988), Delfino (2016) e Rampaso (2016).

QUADRO 2 – DIMENSÕES DE VARIAÇÃO DA LÍNGUA INGLESA

Dimensão 1	Produção com Interação *versus* Informacional
Dimensão 2	Discursos Narrativos *versus* Não Narrativos
Dimensão 3	Referência Elaborada *versus* Dependentes do Contexto
Dimensão 4	Argumentação Explícita
Dimensão 5	Estilo Abstrato *versus* Estilo Não Abstrato

Fonte: Delfino (2016), adaptado de Biber (2009).

Com base no seu modelo multidimensional de variação de registro, Biber (1988, 1995) mostra como características linguísticas variam sistematicamente com relação a textos típicos de contextos comunicativos específicos. O autor observou que há variação de características linguísticas entre diferentes textos ou registros, de tal modo que se compararmos uma redação de aluno e um relatório de negócios, poderemos verificar que são formados por características linguísticas que ocorrem com frequência diferente em um e em outro, ou seja, apresentam variação sistemática, não aleatória, que podemos chamar de padronizada (*patterned*). Antes do estabelecimento da Análise

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

Multidimensional, era comum investigar-se a variação entre variedades textuais com base em poucas características linguísticas e/ou poucos dados (BERBER SARDINHA, VEIRANO PINTO, 2014). Para Biber, a variação entre as variedades linguísticas deve levar em conta uma grande quantidade de características linguísticas, uma vez que nenhuma característica por si só é suficiente para uma descrição adequada.

Biber partiu da perspectiva de que os falantes de uma língua possuem, além do conhecimento lexical, conhecimento sobre a estrutura e o uso da língua (BIBER, 1988, p. 8), para usá-la de acordo com as exigências funcionais e situacionais. Com essa abordagem, o autor investigou a variação na língua inglesa confrontando textos dos mais variados registros (desde conversas telefônicas a resenhas jornalísticas, de livros didáticos e artigos acadêmicos a programas de rádio e TV), para delinear as características linguísticas típicas de cada registro, obtendo uma descrição abrangente do inglês e esclarecendo como os vários registros variam por dimensões, que podem ser entendidas como espaços de coocorrência de características linguísticas. Assim, essas dimensões são como um contínuo em que os textos se distribuem de acordo com suas características linguísticas. De acordo com Biber (1988, p.13), uma "dimensão linguística é determinada a partir de uma correlação consistente de padrões entre as características. Ou seja, quando um grupo de características ocorre com frequência em textos, essas características definem uma dimensão linguística"[2].

A importância do trabalho com as dimensões de variação no ensino advém do fato de que o professor pode trabalhar com os itens linguísticos mais frequentes identificados em registros específicos. Na perspectiva da Análise Multidimensional, cada registro tem sua gramática própria, e a variação é entendida como uma propriedade de qualquer língua em uso. Para ilustrar o conceito de dimensão de variação, tomemos o caso do *corpus* de letras de música CoEL (Delfino, 2016), formado por canções americanas e britânicas de estilo pop. Esse *corpus* possui o seguinte perfil multidimensional em relação às dimensões de Biber (1988): envolvido (Dimensão 1), não narrativo (Dimensão 2), dependente do contexto (Dimensão 3), persuasivo (Dimensão 4) e não abstrato (Dimensão 5). Com base nesse perfil, o professor pode criar materiais de ensino de inglês baseados em música para ensinar pontos lexicais e gramaticais específicos. Por exemplo, sabendo que as canções são 'envolvidas', em vez de 'informacionais', segundo a dimensão 1, pode-se depreender que características linguísticas como pronomes pessoais de primeira pessoa, verbos no presente e contrações são muito comuns nesse tipo de música. Um perfil dimensional distinto foi obtido com a análise do *corpus* BEC (*Business English Corpus;* Rampaso 2016). Segundo esse *corpus*, *emails*, por exemplo, possuem o seguinte perfil: moderadamente informacional (Dimensão 1), não narrativo (Dimensão 2), dependente de contexto (Dimensão 3), persuasivo (Dimensão 4) e não abstrato

[2] No original: "A linguistic dimension is determined on the basis of a consistent co-occurrence pattern among features. That is, when a group of features consistently co-occurs in texts, those features define a linguistic dimension".

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

(Dimensão 5). Assim como as músicas, o registro *email* pode ser trabalhado em sala de aula com base nas dimensões de variação. Sabendo que *emails* são moderadamente informacionais, o professor pode mostrar aos alunos ocorrências de adjetivos atributivos (*tough meeting*), preposições e substantivos, por exemplo, por meio de concordâncias.

5. Considerações Finais

A preparação de material didático a partir da análise de dados à luz da Linguística de *Corpus* tem o potencial de trazer para a sala-de-aula uma grande variedade de padrões de linguagem. Uma seleção de padrões encontrada pelo professor no *corpus* investigado pode ser utilizada no desenvolvimento de materiais de ensino, como os ilustrados no presente artigo ou como uma "inspiração" para que cada professor crie suas estratégias de acordo com seu contexto de ensino e com as ferramentas disponíveis para o seu trabalho.

De modo geral, a Linguística de *Corpus* pode possibilitar um redirecionamento importante e inovador na preparação de material didático, ao colocar à disposição do professor evidências de uso da língua que podem ser aproveitadas na sala de aula a fim que o ensino seja baseado na língua efetivamente usada para comunicação, em vez da linguagem artificial e distanciada dos reais contextos de uso que aparece em alguns materiais didáticos. Da mesma forma, os aprendizes podem ter a oportunidade de desenvolver autonomia na aprendizagem, fazendo buscas nos *corpora*, discutindo seus achados com o grupo e com o professor, agindo como protagonistas do processo de construção do conhecimento sobre o idioma que aprendem. Diferentemente do que muitas vezes ocorre em aulas de língua estrangeira, a aplicação pedagógica do *corpus* requer uma abordagem autônoma (Viana, 2011). Freire (1970) criticou o modelo de educação que chamou de bancária, em que o educador é o sujeito que conduz os educandos à memorização mecânica do conteúdo narrado. Foi além, quando afirmou que a narração os transforma em 'depositários', em recipientes a serem enchidos pelo educador. A proposta do trabalho com *corpus* em sala de aula se contrapõe a essa educação bancária, uma vez que o uso de corpus no ensino traz à tona, de um lado, a busca pela autonomia na aprendizagem, e de outro, a colaboração entre alunos e professores. Conforme colocou Johns (1991: 1), a "tarefa do aprendiz é descobrir a língua estrangeira, enquanto a do professor é fornecer um contexto no qual o aprendiz possa desenvolver estratégias para essa descoberta – estratégias que o possibilite aprender a aprender"[3].

[3] No original: the task of the learner is to 'discover' the foreign language, and the task of the language teacher is to provide a context in which the learner can develop strategies for discovery – strategies through which he or she can 'learn how to learn'.

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12

Referências bibliográficas

BERBER SARDINHA, T. , 2000a. Análise Multidimensional. *Delta, v.* 16, n. 1, p. 99-127.

_____, 2000b. Computador, corpus e concordância no ensino de léxico-gramática de língua estrangeira. In V. Leffa (Ed.), *As palavras e sua Companhia - O Léxico na Aprendizagem* (pp. 45-72). Pelotas, RS: EDUCAT / ALAB.

_____, 2004. *Linguística de Corpus*. São Paulo: Manole

_____, 2011. Como usar a Linguística de Corpus no ensino de língua estrangeira. Ou: Por uma linguística de corpus educacional brasileira. In: TAGNIN, S.; VIANA, V. (Orgs.). *Corpora no Ensino de Línguas Estrangeiras*. São Paulo: Hub. p. 301-356.

_____, 2013a. Lexicogrammar. In C. Chapelle (Org.), *The Encyclopedia of Applied Linguistics*. Hoboken, NJ: Wiley, p. 3365-3370.

_____, 2013b. Teaching Grammar and Corpora. In C. Chapelle (Org.), *The Encyclopedia of Applied Linguistics*. Malden, MA: Wiley-Blackwell, p. 5578–5584.

_____, 2016. Corpus-based teaching in LSP. In: MARTIN-MONJE, E. *et al* (Orgs.). *Technology-enhanced language learning for specialized domains*. Oxon: Routledge, p. 203-315.

BERBER SARDINHA, T.; VEIRANO PINTO, M. (Orgs.), 2014. *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*. Amsterdam/Philadelphia, PA: John Benjamins.

BERTOLI-DUTRA, P. , 2014. Multi-dimensional analysis of pop songs. In: BERBER SARDINHA, T.; VEIRANO PINTO, M. (Orgs.). *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*. Amsterdam/Philadelphia, PA: John Benjamins, p. 149-175.

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

_____, 1995. *Dimensions of Register Variation - A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

DELFINO, M. C. N. , 2016. *Uso de Música para o Ensino de Inglês como Língua Estrangeira em um Ambiente Baseado em Corpus*. (Dissertação de Mestrado) - LAEL, PUCSP.

FREIRE, P. 1970. *Pedagogia do Oprimido.* Rio de Janeiro: Editora Paz e Terra, 17ª edição.

HALLIDAY, M. A. K. 1993. Quantitative studies and probabilities in grammar. In M. Hoey (Org.), *Data Description Discourse -- Papers on the English Language in Honour of John McH Sinclair on his Sixtieth Birthday*. London: HarperCollins, p. 1-25.

KENNEDY, G. , 1998. *An Introduction to Corpus Linguistics*. New York: Longman

LOUW, B., 1993. Irony in the text or insincerity in the writer: the diagnostic potential of semantic prosodies. In: BAKER, M. *et al* (Orgs.). *Text and technology - Essays in honor of John McH Sinclair*. Philadelphia/Amsterdam: John Benjamins. p. 157-176.

NELSON, M. *A*, 2000. *Corpus-Based Study of Business English and Business English Teaching Materials*. University of Manchester, Manchester

RAMPASO, M., 2016. *Elaboração de material didático voltado aos alunos de inglês para os negócios com base na linguística de corpus*. (Dissertação de Mestrado) - LAEL, PUCSP

SINCLAIR, J. McH. (1966). Beginning the study of lexis. In BAZELL, C. E. (Org.), *In Memory of J R Firth*. London: Longman, p. 410-430.
_____, 1991. *Corpus, Concordance, Collocation*. Oxford, New York: Oxford University Press.
STUBBS, M., 2001. *Words and Phrases - Corpus-based studies of lexical semantics*. Oxford: Routledge
VIANA, V. 2011. *The politics of Corpus Linguistics.* In: VIANA, V., ZYNGIER, S., BARNBROOK, G. (Orgs.), *Perspectives on Corpus Linguistics.* Amsterdam / Philadelphia: John Benjamins, p. 229-246.

Tony Berber Sardinha *is an associate professor of Applied Linguistics at São Paulo Catholic University, Brazil. He holds a BA in English and an MA in Applied Linguistics, both from Sao Paulo Catholic University (PUCSP), and a PhD from the University of Liverpool (UK). His current research interests include corpus-based approaches to the study of the relationship between culture, history and language use, the analysis of metaphor in everyday language, and the application of multidimensional methods to the analysis of national identities, lexis, and register. E-mail: tonycorpuslg@gmail.com*

Maria Claudia Nunes Delfino *has a master's degree in Applied Linguistics and Language Studies from São Paulo Catholic University (PUCSP), where she focused on Corpus Linguistics and Language Teaching, under Tony Berber Sardinha's supervision. She has a specialist degree in Linguistics and Language Teaching from Oswaldo Cruz School (COC-SP) and serves as faculty in São Paulo Technical College (FATEC-PG). Her main academic interest is in the application of Corpus Linguistics to foreign language teaching and learning. E-mail: mariac.delfino@yahoo.com.br*

Marianne Rampaso *has a major in Languages, Translation and Interpretation English-Portuguese granted by Ibero-American University (Unibero, SP, Brazil) in 2002, in addition to a specialist degree in English Language from São Judas Tadeu University. She was awarded a master's degree in Applied Linguistics and Language Studies from São Paulo Catholic University (PUCSP) in 2016, where she looked at the uses of Corpus Linguistics in language teaching, under Tony Berber Sardinha's supervision. She teaches private English lessons for both general and specific purposes, and develops materials for Distance Learning. Her main research interest is in the application of Corpus Linguistics to foreign language teaching and learning. E-mail: ladymariann@ig.com.br*

DOI http://dx.doi.org/10.23925/2318-7115.2017v38i1p12