A funcionalidade da linguística de corpus no ensino de língua estrangeira: oportunidades e desafios

*Lucas Rezende Almeida (PUC-Rio)**

Resumo

Este trabalho pretende apresentar as potencialidades do corpus no ensino de português como segunda língua. Para isso, demonstrou-se de que forma, utilizando uma plataforma online de corpus brasileiro e português, é possível encontrar dados sobre o uso das combinações do tipo NOME + VERBO e, a partir dos resultados, criar exercícios para alunos não nativos.

1. Introdução

A carência de material didático para o ensino de português como Segunda Língua e o caráter realista dos dados obtidos nas pesquisas com corpus nos servem como principal motivador para propor um trabalho que articule o uso do corpus e o ensino de português como Segunda Língua. Adicionalmente, a ampla utilização de corpus no ensino de inglês como Segunda Língua, associada à qualidade e à quantidade de corpora disponíveis para a língua portuguesa, sobretudo no formato anotado, motivam ainda mais a exploração neste campo ainda pouco explorado no que se refere à língua portuguesa. Dessa forma, esse estudo funciona mais como a apresentação de uma proposta de exploração nova no ensino de PLE do que como resultado de um processo já traçado.

Esta pesquisa é segmentada em uma rápida apresentação histórica da utilização de corpus no Ensino de Português para Estrangeiros no Brasil. Em seguida, através de um serviço disponibilizado pela Linguateca, mostra-se de que forma podemos fazer esse exercício interdisciplinar. Por fim, analisa-se os impactos positivos e negativos dessa associação.

2. Linguística empírica

A utilização de corpus na Linguística permite uma nova maneira de estudar e descrever a língua, criando hipóteses e observando fenômenos raros ou ainda não constatados pelos meios de análise convencionais, que costumam ser a introspecção (Sampson , 2002). *Corpus*, por sua vez, nada mais é do que uma coleção de documentos produzidos naturalmente, com uma dimensão considerável e em formato eletrônico, podendo ter informações linguísticas associadas (*corpora* anotados) ou não (*corpora*

* Não veio nota de rodapé autor

não-anotados). Segundo Sinclair, (2005, s.p.), "a *corpus* is a collection of pieces of language text in eletronic form, selected according to external criteria to represent as far as possible, a language variety as source of data fo linguistic research".

Sinclair, antes de apresentar sua definição, explica a preferência por certas palavras empregadas nessa definição. Segundo ele, os corpora devem ser baseados em uma linguagem produzida de maneira espontânea e devem ser legíveis por máquinas. Os textos que integram o corpus devem, ainda, ser selecionados segundo sua representatividade lingüística e adequação ao propósito da pesquisa, que pode ser variado (Freitas , 2015).

Originalmente, o trabalho com corpus possuía um interesse lexicográfico, ou seja, criavam-se listas de palavras e observavam a frequência com que elas ocorriam com o intuito de procurar por padrões de uso recorrente. Esse tipo de pesquisa informava a criação de dicionários de língua, para aprendizes de língua e para falantes nativos.

Desde então, o trabalho com corpus vem se associando ao ensino de língua estrangeira, sobretudo ao ensino de Inglês. As documentações produzidas em corpus anotados têm se tornado potenciais gramaticais descritivos do inglês. Por exemplo, a gramática "Longman Grammar of Spoken and Writeen English" é inteiramente baseada em um projeto de pesquisa que documenta o corpus "Groundbreaking". Ao longo das opções linguísticas feitas ao etiquetar as palavras em corpus marcados, o linguista registra suas escolhas em documentos. Essas escolhas devem ser estáveis e concretas, já que serão, portanto, lidas por uma ferramenta virtual. Esse processo descritivo gera como produto uma análise morfossintática da língua que funciona, portanto, como uma gramática descritiva.

A Linguística de Corpus em Língua Portuguesa começa na década de 60 em Portugal e na década de 70 é desenvolvida no Brasil por meio do Projeto NURC (Norma Urbana Culta) que descreveu a língua falada em 5 grandes capitais, produzindo diferentes corpus orais. Os corpus anotados, entretanto, começaram nos anos de 1993, através de anotações de erros como suporte e criação de corretores gramaticais em língua portuguesa.

O trabalho com corpus tem trazido para a Linguística em geral uma experiência mais empírica por ser baseado em coletâneas de textos autênticos, ou seja, produzidos em contextos situacionais específicos tanto em modalidade escrita quanto oral. O estudo da linguagem deixa de se basear em intuições dos falantes ideais, conforme defendido por Chomsky (1972), para olhar os dados reais da linguagem.

A essa característica empírica da língua, muito nos interessa quando queremos relacionar a Linguística de/com corpus com o ensino de língua portuguesa brasileira para estrangeiro.

Corpus em português já tem oferecido documentações extensas do Português Brasileiro como percebemos no projeto LINGUATECA. Entretanto, esse trabalho

não pretenderá relacionar essa documentação com criação de uma gramática para o ensino de PLE, mesmo sabendo do potencial que esse tipo de relação tem a oferecer para o ensino de Português como Segunda Língua.

3. A linguística de corpus e o ensino de segunda língua

A linguística de Corpus permite que a língua seja analisada tanto no eixo paradigmático, por meio das frequências e lemas, como no eixo sintagmático, através das linhas de concordância. Essa interação entre os eixos saussurianos revolucionou a visão da linguagem modular, em que léxico e gramática eram vistos como pares indissociáveis:

> In this respect, it can be said that the corpus revolution has introduced a new theoretical perspective on linguistic structuring: one in bold contrast to the mainstream paradigm of Chomsky (e.g. Chomsky 1965:84-88) whereby grammar and lexicon are two clearly distinct components. It also challenges a tradition long established in language study, whereby grammars and dictionaries provide distinct kinds of information about a language, and are published in separate covers (LEECH, 2010, pg. 12),

Outra vantagem notável da utilização de corpus no ensino de segunda língua é a imparcialidade da intuição do autor. Ao imaginar que os materiais preparados para estrangeiros são feitos em boa parte por nativos esse processo de "desautomatização" da língua pode ser auxiliado por essa ciência. Para Gabrielatos (2003: 2), "a intuição do falante nativo nem sempre é confiável e a condição de falante nativo não nos garante, automaticamente, uma visão consciente, clara e abrangente da língua em todos seus contextos de uso". Junto a essa "naturalização" das evidências linguísticas, tem-se a autenticidade dos textos utilizados, conforme nos aponta Maciel:

> "Nesse contexto, a Linguística de Corpus abre novos caminhos para que o professor e aluno percebam, a partir de realizações textuais autênticas, a complexidade do inter-relacionamento do léxico, da sintaxe e da semântica e possam fazer suas descobertas selecionando elementos lexicais e regras gramaticais de acordo com significado que desejam expressar na comunicação. Em tal integração, torna se possível desenvolver a conscientização linguística e a autonomia do aluno no uso da língua, tão valorizadas no processo pedagógico-didático da comunicação linguística" (2005, pag. 129).

Por outro lado, a utilização de corpus também possui pontos que são criticados em uma perspectiva pedagógica. A utilização das linhas de concordância, como exemplo, é considerada uma análise linguística atomicista e descontextualizada. Entretanto, boa parte das ferramentas que lidam com corpus permitem o prolongamento das linhas de concordância para sentenças maiores, oferecendo o co-texto dessas concordâncias a fim de contextualizá-las, conforme destaca Sardinha :

> "Uma terceira crítica diz respeito à possível incompatibilidade entre o uso de concordâncias e o ensino comunicativo de línguas, já que as concordâncias promoveriam a descontextualização da linguagem por mostrarem apenas pequenos trechos provenientes de várias textos (Aston, 1995). Este problema pode ser evitado fornecendo aos alunos maiores quantidades de texto em cada concordância, ou permitindo aos alunos que tenham acesso a concordanceador que ofereça a visualização dos textos do corpus na íntegra." (1999, s.p..)

Diante dessa discussão, criaram-se dois tipos de abordagem: a top-down approach, inicia-se pelo discurso para após chegar as linhas de concordância, e a bottom-up approach, inicia-se nas linhas para após chegar ao discurso. Ambas as abordagens têm seus objetivos e seus papéis no campo educacional, deve o professor observar qual a adequada para aquele conteúdo, como observado nas palavras de Biber:

> "Functional analysis is primary in top-down approaches; functional distinction are determined on a qualitative basis, to determine the set of relevant discourse types and to identify specific discourse units within texts. In contrast, linguistic analysis is primary in bottom-up approaches; a wide range of linguistic distributional patterns are analysed quantitatively, again being used to determine the set of relevant discourse types and to identify specific discourse units within texts." (Biber et al. 2007b: 241)

A utilização de corpus é vista como um método direto e indutivo de ensino de língua. O aluno normalmente é apresentado a dados e ele deve notar as semelhanças e diferenças linguísticas. O fragmentado abaixo apresenta as vantagens desse tipo de abordagem ditas por Sardinha:

> "The main reason why concordances were adopted as a technique for exploring the corpus with the students was that they provide students with the opportunity to engage in discovery activities. It is argued that concordances have a positive impact on the learners, the teachers and on language learning itself (Johns, 1994). Learners become very effective researchers because concordances provide motivation for inquiry and speculation. In addition, as soon as they start working with the data themselves, students they become active researchers instead of passive recipients of knowledge" (1999, s.p..)

Porém, alunos possuem diferentes estilos cognitivos (FLOWERDEW, 2008). Estudantes que são dependentes do campo, são considerados cooperativos e gostam de discussão. Esses alunos são beneficiados por essa metodologia. Alunos que são independentes do campo, preferem instruções e regras a seguirem, o método dedutivo é mais aconselhável.

Dessa forma, Carter e MCCarthy (1995) criaram três estágios para a aplicação desses métodos. O primeiro estágio refere-se à "ilustração", aquele em que os alunos olham os dados. O segundo, a "interação", aquele em que as observações são discutidas

e compartilhadas. O terceiro, a "indução", cria uma regra para as regularidades encontradas naquela amostra. Um quarto estágio, chamado de "intervenção", poderia ser acrescido caso os alunos tenham dificuldade com as etapas anteriores. O professor auxiliaria os alunos a obterem suas próprias conclusões por meio do esclarecimento de elementos linguísticos não compreensíveis.

4. Linguística de corpus e o ensino de português para estrangeiros.

A utilização de Corpus auxiliando no Português para Estrangeiro tem se dado prioritariamente em Portugal e no Brasil (BACELAR DO NASCIMENTO, 1997).

Em Portugal, a Faculdade de Lisboa usa os corpus tanto como materiais para a aplicação do método comunicativo, por meio de um contato direto e com usos reais e diversificados da língua alvo, como também para refletir sobre a língua, através da documentação que registra as opções de anotação desses corpus.

Dois grandes projetos relacionam a Linguística de Corpus com o PLE. Um deles chama-se "Português Falado, Variedades Geográficas e Sociais". Trata-se da produção de material didático para alunos do nível B1 a C2 com o intuito de oferecer áudios com transcrições de situações de fala reais e divulgar os resultados das análises lexicais, sintáticas e discursivas com base nesse próprio sub-corpus. O outro projeto chama-se "Dicionário de Combinatórias do Português" e procura identificar grupo de palavras com diferentes graus de cristalização e de proximidade entre elementos. (BACELAR DO NASCIMENTO, 1997).

No Brasil, os estudos que envolvem Linguística de Corpus estão ligados ao ensino de Português na Universidade Federal do Fluminense e na Pontifícia Universidade Católica do Rio de Janeiro e de São Paulo. O artigo de Berber Sardinha (1999) é considerado como o primeiro trabalho de corpus em PLE. Sua pesquisa mostra o resultado da exploração de um corpus coletado da internet para o ensino de português na Grã-Bretanha. Além de Sardinha (1999), temos a dissertação de mestrado de Cavalcante (2006) que analisou a linguagem usada em um livro didático de PLE, com o objetivo de entender se a frequência de tempos e modos verbais encontradas no livro correspondia aquela da linguagem falada e escrita. Em 2007, Carvalho analisou como as imagens da identidade estrangeira eram construídas nos livros didáticos de PLE por meio de uma pesquisa quantitativa e qualitativa do vocabulário encontrado em oito livros didáticos com relação a identidade social. O trabalho de Mestrado de Ferreira (2010) identificou o grau de autenticidade em livros didáticos por meio da comparação de trigamas e pacotes lexicais com outros corpus de referência (Banco do Brasil –BP). Ainda há outras pesquisas como a de Dell'sola (2002) e Alencar (2004). Essas pesquisas supracitadas envolvem mais a análise de livros didáticos do que o processo de criação para um material, apresentando ferramentas e discussões na produção de um original.

Em comum, Brasil e Portugal possuem um grande corpus que é mantido virtualmente para a utilização de pesquisadores ou professores através do site da Linguateca (SANTOS,2011)

A Linguateca surgiu com os materiais contrastantes obtidos no protótipo pedagógico chamado PoNTE criado para o desenvolvimento de tradução entre o Português e o Norueguês. A partir dessa nova ênfase, ferramentas educacionais como o "Ensinador" (Simões e Santos, 2011) foi desenvolvido para fornecer material didático em formato de teste.

5. Construções lexicais

Entendendo a língua não mais como como unidades segmentadas e modularizadas de léxico e gramática, mas como um compêndio léxico-gramatical, este trabalho irá apresentar o processo de criação de um material didático sobre construções lexicais do tipo: VERBO + NOME.

Segundo Sinclair (1919) uma característica da linguagem autêntica é a sua idiomaticidade, ou seja, o conjunto de combinações lexicogramaticais que é considerado natural e inerente a cada língua. Dessa forma, a medição da idiomaticidade é verificada por meio dos "pacotes lexicais" que, segundo Sardinha (2007), são sequência de palavras fixas de extensão variável e que ocorrem com certas frequências em determinados tipo de textos. Expressões de cumprimento são exemplos de "pacotes lexicais". Essas estruturas possuem, assim, um certo grau de padronização como definem Hunston e Francis (2000, p.37):

> The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it (HUNSTON e FRANCIS, 2000, p. 37).

Além dessas expressões idiomáticas, há os casos considerados "collocations" cunhado pelo linguística J. R. Firth (1966). Segundo o autor, "collocations" são palavras que andam juntas sem necessariamente terem uma explicação para essa ligação. A principal diferenciação dessas estruturas para aquelas anteriores é que as últimas são associações estatísticas que tendem a ocorrerem em conjuntos específicos em vez de expressões relativamente fixas.

A escolha por prepararmos um material que vislumbrasse esse tópico linguístico encontra-se justamente pela dificuldade de consegui-lo explicar ao estrangeiro. Ao ouvido do falante nativo, essas expressões são consideras ora mais naturais ora mais artificiais. A explicação dessa "naturalidade" ou "estranhamento" está na ocorrência contínua ou não nos discursos sociais advindos daquele grupo

que julga a sua recorrência . Com o uso de corpus, essas recorrências podem ser mensuráveis por meio das frequências obtidas em uma pesquisa quantitativa. Este é um exemplo de uma perfeita ligação entre corpus – por meio das suas ferramentas estatísticas – e conteúdo – através da temática lexicogramatical.

6. Proposta pedagógica: padrões lexicais do tipo: v+n

6.1 investigação teórica

Nossa pesquisa foi realizada por meio de uma ferramenta disponível online com corpus retirados do Brasil e de Portugal chamada "Linguateca". Os dados extraídos do site http://www.linguateca.pt/ correspondem ao projeto AC/DC e foram retirados no dia 01 de junho de 2014.

Com intuito de ensinar os padrões lexicais do tipo binômio "VERBO + NOME", inicialmente procuramos pelos verbos mais recorrentes em todos os corpus, sem distinguir os brasileiros dos portugueses. O resultado está representado na tabela abaixo:

Dados retirados do corpus: todos juntos

Distribuicão por lema: [pos="V"]

ser	15081291
ter	3925349
estar	2542667
poder	2015553
fazer	1937321
haver	1162678
ir	1160961
dever	1100593
dar	888209
dizer	841581
apresentar	753299
ver	738507
passar	604701
ficar	594601
realizar	535173
vir	527223
querer	510031
partir	505982
encontrar	435865
utilizar	435246

chegar	432504
levar	404991
receber	387922
saber	380132

Após sabermos quais os verbos mais recorrentes na amostra, selecionamos os dezesseis primeiros verbos e analisamos quais eram as construções possíveis efetuadas entre um verbo seguido de um nome, procurando, neste casos, por padrões lexicais.

Dados retirados do corpus: todos juntos

Distribuição das colocações por lema: [lema="____" & pos="V"] @[pos="N"]

Verbo	**Nome**
Ser	Verdade, objeto, criado, alvo, vítima
Ter	acesso, direito, condição, certeza
Estar	associado, preso, interessado, presente, sujeito
Poder	auxiliar, politico, local, central...
Fazer	parte, uso, questão, referencia, sentido
Haver	diferença, necessidade, dúvida, aumento.
Ir	direto, campeão, auxiliar, professor, aluno
Deber	dinheiro, obediência, explicação, satisfação, respeito, favor.
Dar	conta, origem, inicio, lugar, continuidade, aula, resposta
Dizer	coisa, estudo, pesquisa, palavra, jornal, ministro
apresentar	coisa, diferença, resultado, problema, característica
Ver	quadro, texto, caixa, televisão, problema, filme.
passar	fome, férias, hora, informação, necessidade...
Ficar	preso, encarregado, detido, desempregado, subordinado, mudo, prisioneiro
realizar	estudo, atividade, pesquisa, análise, teste.
Vir	gente, pessoa, dia, auxiliar, juntas, chuva

Com os resultados coletados, percebemos que a maior parte dos nomes mais recorrentes que acompanhavam os verbos não poderiam ser classificados como padrões lexicais efetivos. Eles apareciam de forma recorrente provavelmente pelas temáticas em comum que alguns corpus maiores possuíam em consideração com outros. Alguns casos de binômios lexicais também não foram compreensíveis, boa parte porque o corpus possuía tanto o português na variante brasileira quanto portuguesa.

Entretanto, uma das observações que fazemos como falantes da língua é que boa parte dos resultados apresentam construções consideradas como do registro escrito, formal da linguagem. A própria ocorrência de verbos como "realizar" entre

os mais usados e expressões como "ficar encarregado", "realizar análise", "apresentar análise" nos parece mais como recorrentes no discurso escrito do que no oral.

Entendendo que o intuito dessa pesquisa é servir pedagogicamente para alunos estrangeiros no ensino de português, achamos necessário investigar mais especificadamente um corpus do discurso oral e que fosse apenas da variante brasileira para que, assim, pudéssemosfazer uma análise contrastiva com os resultados obtidos. Para isso, utilizamos o corpus "C-ORAL BRASIL" com a mesma metodologia que no trabalho anterior com todos os corpus: procuramos saber quais seriam os verbos mais recorrentes.

Dados retirados do corpus: C-ORAL BRASIL
Distribuicão por lema: [pos="V"]

ser	9644
estar	3149
ter	2880
ir	2645
fazer	1478
falar	1283
ficar	935
ver	902
dar	872
saber	870
poder	721
querer	658
pegar	454
passar	412
vir	389
olhar	321
chegar	303
sair	301
pôr	262

Comparando esta tabela com a que temos de todos os corpus, já percebemos alguns elementos diferentes que poderiam ser topicalizados da seguinte forma:

¥

O verbo "ser", "estar", "ter" e "ir", os primeiros verbos ensinados majoritariamente no ensino de PLE, são os mais recorrentes no discurso oral, diferente do discurso escrito.

¥

Alguns verbos encontrados na lista no discurso escrito não aparecem no discurso oral como "utilizar, realizar, encontrar", enquanto que quase todos os verbos do discurso oral estão presentes também no discurso escrito com posições de frequência diferentes.

Em seguida, analisamos os correspondentes binômios dos quatorze primeiros verbos encontrados na tabela acima:

Dados retirados do corpus: C-ORAL BRASIL

Distribuicão por lema: [lema="_____" & pos="V"] @[pos="N"]

Verbo	Nomes
Ser	hhh, verdade, tipo, coisa, gente
Estar	joia, doida, puto, preso
Ter	jeito, gente, problema, certeza, dinheiro
Ir	tar, pa, pó, prum, pedra
Fazer	sentido, parte, força, favor...
Falar	pa, c, p, cós, coisa, p, besteira
Ficar	hhh, pai-avô, preso, gente, amigo, miudinho, sozim
Ver	todos os casos de palavras ocorreram apenas uma vez, exceto o "hhh".
Dar	aula, conta, licença, sinuca, tempo
Saber	todos os casos de palavras ocorreram apenas uma vez, exceto o "hhh".
Poder	tar, n
querer	peito, ajuda, frango, arroz-doce, pinga...
Pegar	fogo, galinha,
passar	Quarta, roupa

Nesta tabela, temos os casos mais recorrentes dos nomes encontrados depois de verbos, entretanto, alguns verbos em especifico chamaram a nossa atenção durante a pesquisa, fazendo com que, ao observássemos, fossemos direto as linhas de concordância:

Verbos	Linhas de concordância
Ir	- Inda bem que cê **vai tar** com+o seu cabelo curto, pa n fazer aquela touca pra cima / / 268 porque Deus me livre . - Mas eu n **vou tar** aqui mais quando esse quarto tiver pronto hhh . - Mas o Rodrigo **vai tar** lá . - E ' já tinha tomado banho, pronto pa **ir pa** escola - **Vai pa** Lagoa Santa, de Lagoa Santa, vai pa Belo Horizonte, né . - Essa aqui **vai po** Rio de Janeiro Que o cara **vai po** paraíso, né - Ai, eu acho que eu queria **ir prum** Spa, que ninguém ficasse mandando eu levantar, eu queria ir prum Spa que eu ficasse dormindo assim, e alguém fazendo massagem em mim
Falar	- Tô **falando** c ' ocê que tava caprichado . - Tá vendo, eu **falo p** ' ocê que esse negócio aqui é horrível . - O povo já **falou pa** caramba . - Aí eu **falei cos** menino
Poder	- Cê **pode tar** até desanimada né, Gigi, agora, falou numa construção . - Muitas vezes o carro **pode tar** até a oitenta quilômetro por hora, e ela desaparece .

Através destes três verbos, destacamos a importância do professor em recorrer a observação das linhas de concordância antes de preparar seu material pedagógico. Nestes casos, a plataforma de busca reconheceu expressões como "c, p, tar, pa, cum" como nomes, entretanto, eles são abreviações recorrentes de preposições e de verbos na fala e portanto, não cabem ao nosso estudo dos padrões lexicais binários tipo "VERBO+NOME".

Após os dados coletados, retornamos as tabelas acima, reconhecendo quais eram os possíveis padrões lexicais binários relacionados a cada verbo encontrado em ambas as investigações por meio de uma análise das linhas de concordância. Chegamos ao seguinte resultado final, considerando que o corpus "todos juntos" privilegia o discurso escrito, enquanto que o corpus "c-oral Brasil" destaca o discurso oral:

Verbos	Discurso oral Corpus: C-ORAL BRASIL NOMES	Discurso escrito Corpus: TODOS JUNTOS NOMES
SER	Verdade, tipo	Verdade, alvo, vitima
ESTAR	Joia	sujeito
TER	Jeito, certeza, problema	direito, certeza, acesso
IR	"tar"	direto
FAZER	Sentido, parte, favor, força	
FICAR		Mudo, encarregado

DAR	Aula, conta, licença, tempo	Conta, inicio, lugar
PEGAR	Fogo	
PASSAR	roupa	Fome, ferias hora, informação, necessidade
DEVER		Satisfação, favor
APRESENTAR		Resultados
REALIZAR		Estudo, pesquisa
VIR		chuva

Instrumentalizado com os dados acima, são diversas as formas pedagógicas que o professor de PLE pode abordar com esse tipo de combinação lexical nas salas de aula, apresentamos algumas propostas que achamos interessante sem, tentar por meio delas, engessar as inúmeras possibilidades que podem ser aproveitadas dessa amostra:

- O professor pode trabalhar com o uso funcional das expressões gramaticalizadas que correspondem, como no caso a seguir, a concordância e assentimento: "SER+VERDADE" e "ESTAR JOIA";
- ele também pode tratar de expressões recorrentes do discurso oral como o caso das estruturas: "VAI + TAR" e "VAI + CUM, P" que não propriamente se classificam como padrões lexicais, mas são um fenômeno importante a ser destacado;
- ou tratar de padrões lexicais que normalmente aparecem próximos em determinados gêneros como o caso do verbo "apresentar" e "realizar". É mais recorrente "apresentar" resultados do que "estudo" e "pesquisa" que são, por sua vez, "realizados" e assim vice e versa.

1.2. Investigação prática

Referimos a investigação prática como a capacidade de criarmos exercícios por meio do mesmo portal virtual que utilizamos para a investigação teórica acima através de uma ferramenta chamada "Ensinador". Sobre o "Ensinador", Simões e Santos dizem:

> The kind of exercise that Ensinador produces is a cloze test, that is, the student is given sentences where one or more words have been removed, and s/he should fill the blanks, therefore restoring the original text (2011, pg.303).

Apresentaremos a seguir duas propostas de exercícios baseadas nos corpus já apresentados acima de acordo com a tabela conclusiva que chegamos na parte anterior, reunindo os padrões lexicais de maior ocorrência. Selecionamos os verbos "dar", "passar", "fazer" e "ter" por ser aqueles com maior número de colocações dentre os demais.

O primeiro exercício pretende testar a habilidade do estrangeiro em detectar se as expressões tratam de padrões lexicais com sentido específico ou se são expressões usuais da língua. Para isso, utilizamos o verbo "ter" com o seguinte comando no Ensinador: [lema= "ter" & pos="V"] [pos="N"]. Como resultado, apareceram diversas linhas de concordância com e sem padrões lexicais. Diante da tabela final da sessão anterior

buscamos pelas expressões mais recorrentes nas linhas de concordância e selecionamos algumas que não tratavam de padrões lexicais. O Ensinador apresentou as sentenças sem as expressões, já que o programa foi feito para preparar exercícios do tipo "complete a lacuna". Porém, queríamos as sequências completas, e para isso clicamos no botão "Ver solução" e exportamos as concordâncias para o word. Este foi o resultado final:

A. O verbo "ter" é muito recorrente na língua brasileira, ele apresenta, portanto, várias expressões do tipo V+N com sentidos próprios e outros, usuais. Procure identificar nas frases abaixo quais das expressões possuem um sentido específico e quais estão no seu sentido literal.

1. *E001-PT-873*: Não **ter dinheiro** para pagar aos funcionários é o pior que pode acontecer.
2. *E016-PT-111*: É por isto que eu utilizo o balneário, porque não **tenho** á**gua** quente em casa.
3. *E020-PT-649*: É o tal ditado: «Nós devemos **ter pena** destes desgraçados que **têm vergonha** de pedir e não saem à rua.
4. *E109-BR-19*: Regressaram a terra para trabalhar porque a pesca não dá ou porque não **têm jeito** para o mar.
5. *E160-BR-620*: Na verdade, a SOS Mata Atlântica, ela começou, eu acho que todo movimento desse tipo na época, hoje nem tanto, mas na época embora talvez hoje isso seja válido, não **tenho certeza** .
6. *E173-BR-1291*: Como a gente **tem problema** de calendário escolar, geralmente a gente tem que espera um ano para o outro.

As sentenças desse exercício foram retiradas do corpus "Museu da Pessoa". Esse corpus foi o que apresentou o menor número de padrões lexicais na lista dos verbos mais recorrentes, conforme pode ser observado na tabela abaixo:

Verbo	Nomes
Ser	Criança, verdade,
Ter	Condição, idéia, tempo,
Fazer	Parte, mal, falta,
Estar	A vontade,
Ir	Direto,
trabalhar	Sem ocorrência de colocações.
querer	Sem ocorrência de colocações.
Dizer	Respeito.
Dar	aula, conta, trabalho, condição, espaço, tempo
Falar	Besteira, bobagem.
Ver	Sem ocorrência de colocações.
Poder	Sem ocorrência de colocações.
Ficar	Sem ocorrência de colocações.
Vir	Sem ocorrência de colocações.

Escolher esse corpus para que o aluno fizesse a diferenciação foi mais viável, já que encontraríamos um grande número de estruturas que não se caracterizariam como padrões lexicais com sentidos específicos. Durante a seleção das sentenças foi preciso uma atenção do professor para que elas fossem compreensíveis no co-texto oferecido pelas linhas de concordância.

O segundo exercício pretende que os alunos completem as sentenças com as expressões oferecidas de forma a torná-las compreensíveis. Neste exercício, utilizamos os verbos "dar", "passar" e "fazer". Para isso, usamos os seguintes comandos sucessivamente por meio do botão "ADICIONAR": [lema= "dar" & pos="V"] [lema="conta|lugar|licença|tempo"] e [lema= "passar" & pos="V"] [lema="roupa|fome"] e [lema= "fazer" & pos="V"] [lema="sentido|parte"]. Selecionou-se das frases apresentadas aquelas que eram compreensíveis no próprio co-texto oferecido pelas linhas de concordância.

A. Complete as frases abaixo com as expressões disponibilizadas na caixa a seguir:

DAR CONTA DAR LUGAR DAR LICENÇA DAR TEMPO PASSAR ROUPA PASSAR FOME FAZER SENTIDO FAZER PARTE

1. Brincadeira tem hora; você vai me _____ , mas vou subir a serra.
2. : "... Se _____ adiantasse, todo gordo seria magro, e o magro seria obeso... "
3. Quando _____ dos tiros e da velocidade de uns e de outros, numa descida quase que morremos.
4. O filme corre com muito bom humor, e confesso que analizar interpretações em rotoscopia não cabe a mim, especialmente em se tratando de Keanu Reeves, que sempre me lembra de uma tábua de _____ , e Winona Ryder, a menina chorona do bairro.
5. E, embora seja cedo para tirar conclusões, pode acontecer que a experimentação formal, que dominou as artes do espectáculo dos anos 80, tenha de _____ ao renascimento da palavra falada, à exposição de um sentido e de um conteúdo.
6. Quero muito jogar e espero _____ do recado.
7. Não queremos é construir um estádio para ficar às moscas, não queremos esbanjar o dinheiro do município, onerando o orçamento da Câmara Municipal, hipotecando-a exclusivamente durante três anos a construir um campo de futebol novo para alimentar ambições mesquinhas que não _____ .
8. A venda judicial da sede da Casa do Douro, marcada para ontem, foi suspensa por 30 dias para _____ à conclusão de todos os trâmites legais relativos ao acordo para pagamento da dívida à Cofipsa.
9. Ontém sai com um grupo, do qual _____ uma senhora de cabelos brancos.

Na elaboração desse exercício observamos sentenças mais complexas que são justificadas pelo nível de proficiência elevado de alunos que aprendem padrões lexicais em uma língua estrangeira. As linhas de concordância foram retiradas do corpus: "todos juntos", já que o intuito era reunir uma maior diversidade de ocorrências das

expressões em diversos tipos de discursos: sejam eles orais ou escritos. Este exercício também poderia ser aproveitado para a criação de um terceiro com o seguinte enunciado:

B. Após completar as frases com as expressões, reescreva-as, tentando substituir os padrões lexicais por outras estruturas com o mesmo sentido.

Essa proposta permite que o profesor verifique se os alunos realmente entenderam o sentido das expressões naqueles contextos apresentados.

7. Considerações finais

Por meio deste trabalho, percebemos o crescente interesse na união da Linguística de/com Corpus e o ensino de português como segunda língua. As ferramentas e os corpus já disponibilizados para o português nos permitem explorar o uso didático dessa área da ciência da linguagem assim como tem sido feito para outras línguas.

Conclui-se, por meio da descrição da preparação de um conteúdo didático, as vantagens e as dificuldades na aplicação de corpus no ensino de língua estrangeira. A forma de acesso a corpus e a atenção do professor para os resultados coletados, como também, para as concordâncias selecionadas na preparação de exercícios aparecem como os principais problemas. A neutralidade da intuição do falante frente aos dados quantitativos obtidos pelo corpus nos permite acreditar em uma fidelidade maior com a realidade linguística, destituindo-a dos nossos juízos de valores como nativos. Essa sim, sem sombra de dúvida, é a maior qualidade da utilização de corpus.

O ensino de português como segunda língua tem crescido em todo o mundo e é dever dos professores assumirem a responsabilidade sobre o que é apresentado como pertencente ao nosso país. O cuidado com a seleção de dados ocorre por meio da prática educativa e pela leitura das referências que constrõem abordagens e métodos de ensino. Esse trabalho é uma proposta de mediar a relação tríade entre aluno, língua e professor, apresentando ferramentas que estão à disposição para agirem no processo linguístico-cultural desse aprendiz estrangeiro.

Referências

BACELAR DO NASCIMENTO, Maria Fernanda. 1997. *A exploração de corpora linguisticos no ensino/aprendizagem do português*. In Seminário internacional do Português como Língua Estrangeira, Macau, 21-24, Maio.

BERBER SARDIHA, T. Computador, corpus e concordância no ensino do léxico-gramática da língua estrangeira. In: LEFFA, V. (ED.). *As palavras e sua companhia* – o léxico na aprendizagem. Pelotas: ALAB/EDUCAT, 2000, p. 45-72.

BERBER SARDIHA, T. *Beginning Portuguese Corpus Linguistics:* exploring a corpus to teach Portuguese as a Foreign Language. D.E.L.T.A., v. 15, n. 2, p.289-299, 1999.

BIBER, D., Connor, U. & Upton, T. 2007b. *Conclusion:* Comparing the analytical approaches In D. Biber, U. Connor & T. Upton (Eds.), Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure. Amsterdam/Philadelphia: John Benjamins, 239–259.=

FERREIRA, Telma de Lurdes São Bento. *Linguistica de corpus e Autenticidade de Livros Didáticos:* o caso do português como Língua Estrangeira (PLE). Dissertação de Mestrado: PUC-SP, 2010.

FLOWERDEW, Lynne. *Applying corpus Linguistics to pedagogy*. 14:3 (2009) 393-417.

HUNSTON, S., & FRANCIS, G. (2000). *Pattern Grammar* – A corpus-driven approach to the lexical grammar of English. Amsterdam/ Philadelphia: John Benjamins.

LEECH, Geoffrey N. 2011. *Frequency, corpora and language learning*". In A Taste for Corpora: In honour of Sylviane Granger, Meunier, Fanny, Sylvie De Cock, Gaëtanelle Gilquin and Magali Paquot (eds.), 7 ff.

MACIEL, Anna Maria Becker. *Novos Horizontes para o ensino do léxico*. PPG Letras-UFRGS: Revista Língua e Literatura: v. 6 e 7, n 10/11, 2004/2005, p 123- 130.

REPPEN, R. (2009). *English language teaching and corpus linguistics:* Lessons from the American National Corpus. In P. Baker (Ed.).Contemporary Approaches to Corpus Linguistics. London: Continuum Press. pp. 206 – 215

SANTOS, Diana. *Corpora at Linguateca:* Vision and roads taken, in Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), Working with Portuguese Corpora, Bloomsbury, 2014, pp. 219-236.

SINCLAIR, J. 2005. *"Corpus and Text -Basic Principles"* in Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne. Oxford: Oxbow Books: 1-16. Available online from http://ota.ahds.ac.uk/documents/creating/dlc/chapter1.htm

SIMÕES, Alberto & Diana Santos. *Ensinador:* corpus-based Portuguese grammar exercises", Procesamiento del Lenguaje Natural 47, septiembre de 2011, pp. 301-309

Anexo

Linhas de concordância do exercício B:

1. *par=7683*: -- Brincadeira tem hora; você vai me **dar licença** , mas vou subir a serra.
2. : "... Se **passar fome** adiantasse, todo gordo seria magro, e o magro seria obeso... "
3. : Quando **dei conta** dos tiros e da velocidade de uns e de outros, numa descida e nós até à rasca, quase que chocávamos.
4. <p>: O filme corre com muito bom humor, e confesso que analizar interpretações em rotoscopia não cabe a mim, especialmente em se tratando de Keanu Reeves, que sempre me lembra de uma tábua de **passar roupa** , e Winona Ryder, a menina chorona do bairro.
5. *par=171*: E, embora seja cedo para tirar conclusões, pode acontecer que a experimentação formal, que dominou as artes do espectáculo dos anos 80, tenha de **dar lugar** ao renascimento da palavra falada, à exposição de um sentido e de um conteúdo.
6. *par=fut68688*: «Quero muito jogar e espero **dar conta** do recado.
7. *par=DC-N1685-1*: Não queremos é construir um estádio para ficar às moscas, não queremos esbanjar o dinheiro do município, onerando o orçamento da Câmara Municipal, hipotecando-a exclusivamente durante três anos a construir um campo de futebol novo para alimentar ambições mesquinhas que não **fazem sentido** .
8. *par=54217*: A venda judicial da sede da Casa do Douro, marcada para ontem, foi suspensa por 30 dias para **dar tempo** à conclusão de todos os trâmites legais relativos ao acordo para pagamento da dívida à Cofipsa, continuando também agendada para 26 de Fevereiro uma hasta pública sobre outros bens da instituição.
9. par=2934: Ontém sai com um grupo, do qual **fazia parte** uma senhora de cabelos brancos.