A INFLUÊNCIA DO LEXICÓGRAFO NO *CORPUS*

PHILIPPE HUMBLÉ

Universidade Federal de Santa Catarina

A lingüística de *corpus* está em evidência pelo menos desde que o Cobuild[1] começou a publicar dicionários usando exemplos autênticos tirados do *Bank of English* (Fox, 1997). Ninguém questiona a utilidade de um grande número de textos para nos ajudar a ter uma melhor idéia do que a linguagem é na realidade. Entretanto, a questão se a língua real deve ser usada diretamente num dicionário sem nenhuma modificação é objeto de controvérsia.

Entre aqueles que eloqüentemente defendem o ponto de vista de que os exemplos devem ser autênticos está John Sinclair (1987; 1988; 1991). Segundo ele, as frases inventadas dão uma idéia errada da língua, e as amostras tiradas de um *corpus* contradizem a intuição dos lexicógrafos.

Sem querer colocar em questão a utilidade de *córpora* para a compilação de dicionários, fiquei interessado em descobrir até que ponto exemplos selecionados de um *corpus* ainda mantêm as características deste mesmo *corpus*. Será que a personalidade do lexicógrafo intervém no processo de seleção e, em caso positivo, de que maneira? Para responder a essas perguntas, montei um pequeno experimento na Universidade Federal de Santa Catarina com a equipe que trabalha no *Dicionário de Uso Português-Espanhol*. Quando conduzi este experimento em 1995, o grupo de lexicógrafos consistia de cinco estudantes de lingüística, do sexo feminino, entre 19 e 32 anos de idade. Como naquela época o Cobuild era um projeto

relativamente recente e eu estava muito impressionado, insisti para que todos os exemplos do dicionário fossem autênticos, e retirados do *corpus* compilado pelo projeto VARSUL[2].

Na época do experimento, o número de palavras no *corpus* era de apenas 500.000, mas já que o objetivo do *Dicionário de Uso Português-Espanhol* é ajudar estudantes brasileiros a se expressarem oralmente em Espanhol, o fato de o *corpus* do VARSUL consistir exclusivamente de entrevistas (orais) se adequava perfeitamente ao nosso propósito.

Contudo, devido ao pequeno tamanho do *corpus*, percebemos que não havia exemplos suficientes para certos sentidos de algumas palavras muito comuns e não tivemos outra saída a não ser inventá-los. É importante notar que as estudantes-lexicógrafas nunca se impressionaram muito com os resultados de suas pesquisas no *corpus* e sempre se opuseram ao uso exclusivo de exemplos autênticos, que elas achavam confusos. No final, concordei que alguns exemplos seriam inventados e outros autênticos.

O experimento que organizei visava comparar primeiro as características dos exemplos escolhidos do *corpus* pelas lexicógrafas com as características do próprio *corpus*. Em segundo lugar eu queria comparar o conjunto de exemplos inventados pela minha equipe com o *corpus* do VARSUL.

A primeira hipótese era de que os exemplos autênticos teriam as mesmas características do *corpus* de onde tinham sido tirados. A segunda hipótese era de que os exemplos inventados teriam características bem diferentes dos autênticos, principalmente em termos de pronomes.

Foi essa última hipótese que se revelou ser parcialmente verdadeira. Os exemplos inventados mostraram diferentes padrões em termos de uso de pronomes e também na relação *type/token*. Já no que se refere à primeira hipótese, os resultados, surpreendentemente, mostraram que os exemplos autênticos possuíam características bem diferentes do *corpus* do qual tinham sido tirados e exibiam, por outro lado, muitas características em comum com os exemplos inventados.

Os dados

Eu trabalhei com três tipos de *córpora*. O primeiro foi o *corpus* do VARSUL. Utilizei três entrevistas deste *corpus* para fazer um '*corpus monitor*' de 25.297 palavras. Decidi manter este *corpus* relativamente pequeno para que a relação *type-token*, sobre a qual vou falar mais tarde, fizesse algum sentido.

O segundo *corpus* incluía todos os exemplos inventados pelas lexicógrafas, perfazendo um total de 3.199 palavras, ou 430 frases.

O terceiro *corpus* continha todos os exemplos que as lexicógrafas haviam selecionado do *corpus* do VARSUL. Este terceiro *corpus* totalizava 6.564 palavras, ou 820 frases. Usei também os dados disponíveis em Sinclair (1991), que se referem à freqüência de palavras no léxico inglês.

Depois de comparar a lista de freqüência de cada um desses *córpora*, escolhi apenas aqueles itens que considerei especialmente discrepantes. Desses itens discutirei os pronomes pessoais, a referência, a relação *type/ token* e algumas particularidades referentes a gênero. Como o material analisado era um pouco reduzido (apesar de que, há apenas uma década, este material teria sido considerado enorme), os resultados desta pesquisa não devem ser tomados em termos absolutos, mas somente como indicativo de uma tendência.

Pronomes pessoais

Enquanto estava trabalhando no dicionário, sempre tive a impressão de que a presença de pronomes pessoais era diferente nos três *córpora*. Isso foi confirmado em uma análise mais apurada. A tabela abaixo mostra a classificação de freqüência relativa de pronomes pessoais: primeiro no *corpus* original do VARSUL; segundo, no *corpus* dos exemplos selecionados; terceiro, no *corpus* de exemplos inventados; finalmente, no *corpus* do Cobuild e de acordo com os dados disponíveis em Sinclair (1991). Na tabela abaixo, classifiquei os pronomes pessoais de acordo com a freqüência de cada um nos vários *córpora*. A comparação da coluna Cobuild mostra a freqüência de pronomes pessoais em inglês de acordo com *The Bank of English.*

VARSUL	exemplos selecionados	exemplos inventados	COBUILD
eu	eu	ele	I
você/tu	ele	eu	he
ele	você/tu	ela-você/tu	you
nós	eles	A gente	they
a gente	a gente	nós	we
ela	ela	Eles	she
eles	nós	vocês	
vocês	elas	Elas	
elas	vocês		

Tabela 1. Pronomes pessoais. Classificação vertical. Mais freqüentes no topo.

A segunda tabela mostra como os diferentes pronomes se relacionam entre si. Em cada linha (horizontal) atribui valor 1 para cada item no *corpus* com mais ocorrências de um pronome em particular. Quando, como no caso de *eu*, *eu* ocorre mais entre os exemplos selecionados, dei a esses o valor 1 e reduzi o número de outras ocorrências a este valor.

Item	*Corpus* do VARSUL	Exemplos selecionados	Exemplos inventados
eu	0,82	1	0,3
você	1	0,42	0,25
tu	1	0,51	0,17
ela	0,84	0,78	1
ele	0,51	0,47	1
a gente	0,7	1	0,46
nós	1	0,54	0,34
vocês	1	0,5	1
elas	0,29	1	0,66
eles	0,46	1	0,19

Tabela 2. Comparação entre pronomes pessoais[3].

Dessas tabelas pode-se deduzir o seguinte: em todos, menos no *corpus* dos exemplos inventados, *eu* é o pronome mais freqüente. As pessoas parecem falar preferencialmente sobre o que elas mesmas fazem e pensam. Não há diferença entre brasileiros e falantes de inglês. Entretanto, da tabela 2 pode-se deduzir que há maior freqüência do pronome *eu* nos exemplos selecionados do que no próprio *corpus* de onde foram tirados, apesar de haver freqüência bem menor nos exemplos inventados pela mesma equipe que selecionou os exemplos do *corpus*.

Mais surpreendente ainda é o caso de *ele*. Esse pronome sobe da terceira posição no *corpus* original para a segunda no *corpus* de exemplos selecionados pelas lexicógrafas, terminando em primeiro nos exemplos inventados. Isto significa que as lexicógrafas consideram a terceira pessoa masculina singular mais adequada do que qualquer outra para figurar nos exemplos de um dicionário. Este fato confirma, aliás, as tendências gerais em inglês e em português como ficam exemplificadas pelos *córpora*. *Ela* fica muito atrás de *ele* e até em último lugar na classificação dos pronomes segundo a lista apresentada em Sinclair (1991).

A presença de *tu* em segundo lugar no *corpus* do VARSUL e em terceiro no *corpus* do Cobuild se deve provavelmente ao fato de que o *corpus* do VARSUL seja de língua falada e o *corpus* do Cobuild seja predominantemente escrito.

O que pode ser concluído desses resultados é que, quando confrontados com os dados originais, a equipe de lexicógrafas tentou compatibilizar os exemplos com a opinião que tinham sobre o verbete em pauta. No caso de *a gente/nós*, por exemplo, as frases selecionadas não refletem o uso real, e esse é o caso da maioria das lexicógrafas que selecionam exemplos de um *corpus* autêntico com frases. Provavelmente em muitos casos, e isso me foi confirmado verbalmente por mais de um lexicógrafo do Cobuild, os lexicógrafos têm um exemplo já pronto na mente antes de consultarem o *corpus*. E no *corpus* procuram algo que se

assemelhe. É claro que a opinião do lexicógrafo que trabalha com um *corpus* pode ser mudada pelos dados que ele vê com tanta profusão na tela, enquanto que, no caso de exemplos inventados, existe somente o lexicógrafo e sua própria mente.

Por outro lado, esses resultados, que parecem sugerir cautela no uso de um *corpus,* não invalidam a asserção segundo a qual os exemplos autênticos, considerados individualmente, refletem um uso particular autêntico da língua. Somente quando juntamos todos os exemplos num *corpus* específico é que eles não refletem mais o uso real como um todo, e isso é de interesse mais sociológico do que lingüístico. A escolha de exemplos, portanto, é grandemente influenciada pela pessoa que os escolhe e o item *referência* também sugere isso.

Referência

De acordo com Halliday e Hassan (1993), a *referência* é um dos principais elementos de coesão. Ela une a frase a um contexto extra- ou intra-textual. A presença de itens de referência se torna especialmente visível quando não há esse contexto, como no caso de exemplos inventados. Em uma frase inventada, como *Aquela loja só vende roupa de marca,* a palavra *aquela* tem claramente a intenção de unir *loja* a um mundo lá fora para dar ao leitor a sensação de que a frase, sim, é autêntica. "Autenticidade", no caso de exemplos, implica, por definição, estar ligado a alguma outra coisa.

Em seguida, analisarei a ocorrência de alguns itens de referência.

Pronomes Demonstrativos

Uma das hipóteses iniciais desta pesquisa era de que os pronomes demonstrativos eram usados com mais freqüência em sentenças inventadas do que em sentenças autênticas porque os lexicógrafos tentariam corrigir a falta de naturalidade de exemplos inventados pelo uso exagerado de elementos referenciais. Uma característica essencial de sentenças autênticas é que elas são proferidas num determinado contexto, material ou escrito. Pressupõe-se que o leitor ou interlocutor usará esse contexto para deduzir o significado de uma frase. Por outro lado, um lexicógrafo quando inventa exemplos pretende que eles sejam auto-suficientes e a referência tende a ser supérflua.

Na tabela 3 estão relacionadas as ocorrências de pronomes demonstrativos. Para tornar possível uma comparação, usei o *corpus* com mais ocorrências como base para comparar os outros *córpora.*

	VARSUL	Exemplos selecionados	Exemplos inventados
esta	0,41	0,48	1
essa/s	0,35	0,48	1
esse/s	0,36	0,38	1
isso	1	0,78	0,84
aquele	0,62	0,42	1
aquela/s	0,34	0,92	1
aquilo	0,27	1	0,54

Tabela 3. Pronomes demonstrativos[4]

Analisando os dados da tabela acima, vemos que em 5 de 7 casos, a equipe de lexicógrafas fez mais uso de pronomes demonstrativos[5] do que o *corpus* original, e, algumas vezes de uma maneira muito explícita. Mas, mesmo quando estavam meramente selecionando os exemplos, as lexicógrafas mostraram uma preferência por sentenças que tinham um pronome demonstrativo. *Esta* é levemente mais freqüente nos exemplos selecionados do que no *corpus* original, mas é muito mais freqüente nos exemplos inventados. O caso de *essa(s)* e *esse(s)* é semelhante. A tendência é levemente diferente no caso de *aquele*, mas mesmo assim o pronome é ainda mais freqüente no *corpus* inventado. Somente no caso de *isso* e *aquilo* os exemplos inventados não atingem a freqüência mais alta. Não tenho explicação para isso. Pode ser que estes pronomes sejam mais freqüentes no *corpus* original do VARSUL, porque *isso* também funciona como um marcador de conversação, indicando confirmação. Finalmente, deve-se ressaltar que o sistema de pronomes demonstrativos do português brasileiro tem uma tendência a confundir a forma *este* com a forma *esse* sem que isso seja fácil de diferenciar na audição. Na transcrição das fitas do VARSUL, *essa* deve ter sido transcrita como *esta* muitas vezes. Isso, entretanto, não afeta o assunto em questão.

Ao todo, as tabelas para pronomes demonstrativos mostram a mesma tendência como as referentes aos pronomes pessoais: as frases selecionadas tendem a ser intermediárias entre autênticas e inventadas. Em nove entre dezenove casos, há uma clara linha ascendente do *corpus* do VARSUL para o *corpus* inventado.

Advérbios Demonstrativos

Um dos resultados mais surpreendentes desta pesquisa é o uso muito específico de advérbios demonstrativos em exemplos inventados e exemplos selecionados. Para todos eles, o uso varia grandemente. No *corpus* do VARSUL, os advérbios demonstrativos estão entre os itens mais freqüentes: *aqui* 0.6%; *lá* 0.7%; *ali* 0.6% de todos os itens do *corpus*. Não é o que acontece com os exemplos selecionados nem com os exemplos inventados.

Novamente, ao *corpus* que tinha o maior número foi dado valor 1 e aos outros foram dados valores reduzidos de acordo com este critério.

	VARSUL	Exemplos selecionados	Exemplos inventados
Cá	8 occ.	0 occ.	0 occ.
Aqui	1	0,45	0,3
Lá	1	0,38	0,9
Ali	1	0,35	0,2

Tabela 4 Advérbios demonstrativos (números horizontalmente comparáveis).

Pode-se deduzir, a partir da tabela 4, que os advérbios demonstrativos, ao contrário dos adjetivos demonstrativos, não são vistos como caracterizando a fala autêntica. Caso contrário, as lexicógrafas, instruídas por mim a imitar esse tipo de frase, os teriam usado mais. Embora os advérbios demonstrativos também relacionem a sentença a um contexto, não é tão obviamente seu papel como no caso dos adjetivos demonstrativos. *Cá*, por exemplo, não foi usado uma única vez, mas isso é compreensivo já que *cá* é quase que exclusivamente usado nas expressões *vem cá* e *pra cá*. Outra explicação para a diferença no uso entre advérbios e pronomes demonstrativos é que pronomes como *essa*, *esse*, *aquela* e outros são usados igualmente na língua escrita e falada, enquanto os advérbios demonstrativos são típicos da fala. As lexicógrafas devem tê-los achado estranhos quando vistos no papel. Novamente, neste caso, a seleção de sentenças autênticas não refletiu as características do *corpus* original.

Relação *type/token*

A relação *type/token* nos fala sobre a variedade de palavras usadas em um texto. Quanto mais alta a relação *type/token*, menos uma palavra é repetida. Textos escritos repetem muito menos itens do que textos falados. Eles são lexicalmente mais densos e atingem um índice maior. Na tabela seguinte, comparei a relação *type/token* dos três *córpora*.

VARSUL	13.1%
Sentenças selecionadas	32.6%
Sentenças inventadas	29.4%

Tabela 5. Relação type/token.

Como pode ser concluído a partir dessa tabela, a densidade lexical acaba sendo quase três vezes mais alta no *corpus* de exemplos selecionados do que no *corpus* original, sendo ainda mais visível do que nas sentenças inventadas. Novamente há uma fácil interpretação psicológica. Quando estavam selecionando frases de um *corpus* falado que constariam em um livro, as lexicógrafas tiveram a tendência de escolher aquelas sentenças que soavam mais como 'língua escrita'. Ao inventar sentenças, esta tendência

era moderada pelo fato de que a elas foi solicitada a produção de uma língua falada (ainda que não muito). É surpreendente, no entanto, que sentenças vindas de um *corpus* com uma relação *type/token* específica possam fornecer um conjunto de frases no qual essa relação seja tão diferente.

Conclusão

O número relativamente pequeno de dados que analisei não me permite ser muito categórico em minhas conclusões e o objetivo deste artigo é apenas indicar tendências que devem ser confirmadas por pesquisas futuras. Ademais, nada foi dito em relação à freqüência nesses três *córpora*, no que se refere a itens como metáforas, ordem das palavras, tipo de sujeito gramatical e outros tópicos que poderiam ser igualmente interessantes. Entretanto, algumas conclusões podem ser tiradas.

De todos os diferentes itens analisados, um fato surge claramente: a partir do momento em que há uma seleção de um *corpus*, não importa quão autêntico ele seja, essa seleção não mais reflete este *corpus* fielmente. Se a equipe do Cobuild e Sinclair em particular enfatiza que a intuição do lexicógrafo não oferece uma boa orientação, esta intuição interfere inevitavelmente. Quando as lexicógrafas têm à sua disposição centenas de exemplos, a possibilidade de escolher um exemplo que se pareça com um que inventariam é alta.

Isso não é uma crítica ao uso de um *corpus* mas à crença de que a intuição e as convicções dos lexicógrafos são eliminadas tão logo se use um *corpus* autêntico. A análise de um item é, por outro lado, grandemente facilitada pelo uso desse *corpus* e deve influenciar, por sua vez, as convicções do lexicógrafo.

Exemplos selecionados de um *corpus* autêntico podem ter características diferentes. Sua seleção reflete o que o lexicógrafo pensa que melhor representa a língua autêntica. Mesmo assim, o resultado pode ser melhor do que se os exemplos fossem todos inventados.

Os lexicógrafos não podem fazer uma abstração do objetivo de seu trabalho. Eles têm consciência do perfil do seu público e de suas necessidades e eles não podem evitar o fato de terem uma opinião sobre como essas necessidades devem ser satisfeitas. No caso de dicionários elaborados especificamente para ajudar na aprendizagem de uma língua estrangeira, eles têm em mente uma teoria sobre essa aprendizagem e é também de acordo com essa teoria que eles vão selecionar exemplos do *corpus*.

NOTAS

1 Cobuild refere-se a Collins Birminghan University.

2 O objetivo do projeto VARSUL é elaborar um atlas lingüístico abrangendo doze cidades diferentes nos três estados do sul do Brasil: Paraná, Santa Catarina e Rio Grande do Sul.

3 Estas tabelas são relativas e devem ser comparadas horizontalmente, não verticalmente. Utilizei o *corpus* que tinha mais ocorrências de uma característica em particular como base, dei o valor 1, e reduzi o número de ocorrências nos outros dois *córpora* para esta base.

4 Os valores devem ser lidos horizontalmente e são comparáveis desta maneira. Eles são relativos e não representam o número real de ocorrências dos itens no *córpora*.

5 Neste caso: *esta*, *estas*, *essa*, *esse*, *dessa*, *disso*, *desse*, *nessa*, *aquelas*, *aquele*, *naquele*.

BIBLIOGRAFIA

Fox G. 1987. The case for examples. Sinclair J. (Ed.) *Looking Up*, London and Glasgow: Collins ELT.

Sinclair J. M. (Ed.) 1987. *Looking Up*, Collins, London and Glasgow.

Sinclair J. M. 1988. Naturalness in Language. McCarthy, M. (Ed.) 1988. *Naturalness in Language*. English Language Research Journal Vol. 2. University of Birmingham.

Sinclair J. M. 1991. *Corpus*, *Concordance*, *Collocation*, Oxford University Press, Oxford.