A LINGUÍSTICA DE CORPUS NA E PARA A TRADUÇÃO

Stella E. O. Tagnin (USP)

ABSTRACT: This chapter aims at giving an overview of how Corpus Linguistics can be used in and for translation. After a short introduction, Section 2 discusses what a corpus is and presents various types of corpora suitable for translation purposes. The next section describes, in detail, the steps to be followed in order to compile a corpus. Various kinds of annotation of a corpus are the topic of Section 4, and Section 5 addresses computational tools available for investigating it. Section 6 discusses different studies that can be conducted using corpora. Plenty of examples are given throughout the chapter to ensure that all the topics addressed here are clear to a readership of novices in Corpus Linguistics.

KEYWORDS: Corpus Linguistics, Translation, Corpus compilation, Corpus annotation, Corpus tools.

1) Introdução

A Linguística de *Corpus* é uma abordagem empírica para o estudo da língua, em suas diversas dimensões, como, por exemplo, na sintaxe, no léxico e no discurso. Por essa razão, é especialmente útil no estudo da Tradução (OLOHAN, 2004; ZANETTIN, 2012). Essa abordagem parte da observação de uma grande quantidade de textos, reunidos em *corpora* (plural de *corpus*), para, a partir deles, fazer inferências a respeito de como a língua é usada. A observação é feita por meio de ferramentas computacionais, que fornecem dados quantitativos, mas que também permitem análises qualitativas.

Este capítulo constará de uma seção dedicada aos *corpora* (Seção 2), apresentando seus diversos tipos e os objetivos a que se destinam, seguida de uma seção sobre a compilação (Seção 3) e outra sobre a anotação de *corpora* (Seção 4). A Seção 5 tratará das ferramentas computacionais de que se vale a Linguística de *Corpus*. Na Seção 6, será discutido como os diversos tipos de *corpora* podem oferecer uma contribuição relevante às várias áreas da tradução. A seção final é dedicada à conclusão.

2) O que é um *corpus*?

Os *corpora* são bancos de textos de linguagem autêntica, criteriosamente construídos, destinados à pesquisa e legíveis por computador. Existem *corpora* que podem ser consultados *on-line*, e *corpora* que podem ser consultados *off-line*. Muitos desses últimos são, em geral, corpora especializados compilados por pesquisadores conforme seus objetivos (vide Seção 3). Quando for esse o caso os chamaremos de *corpora* personalizados.

Na Linguística de *Corpus*, enfatiza-se sempre a linguagem autêntica, pois a língua é tida como um sistema probabilístico (HALLIDAY, 1961), ou seja, um sistema em que, embora muitas construções sejam possíveis, algumas delas têm probabilidade maior de

ocorrer. Em outras palavras, nem tudo o que é possível ou gramatical numa língua de fato ocorre. Assim, por exemplo, apesar de *amigo próximo* ser uma combinação possível em português brasileiro, a colocação mais provável, mais comum é *amigo íntimo*. Isso pode ser confirmado no *Corpus* do Português[12] em que *amigo próximo* totaliza 2 ocorrências somente no Português Europeu (PE), e *amigo íntimo* resulta em 46 ocorrências ao todo, isto é, no PE e no Português Brasileiro (PB).

Os *corpora* devem ser criteriosamente construídos, de acordo com o objetivo a que se destinam, isto é, deve-se ter a certeza de que os textos compilados são representativos do campo que se deseja pesquisar e são de fonte confiável, caso contrário os resultados podem não ser. Esse tópico será discutido em detalhe na Seção 3.

Finalmente, os textos devem estar em formato digital para poderem ser lidos por ferramentas computacionais.

2.1) Tipos de corpora

Numa visão geral, há *corpora* disponíveis para consulta *on-line* ou *off-line*. *Os corpora* podem ainda ser monolíngues e bi- ou multilíngues, paralelos e comparáveis. Vejamos cada um desses tipos.

2.1.1) Corpora disponíveis *on-line* e *corpora off-line*

Em geral, os *corpora* disponíveis *on-line* podem ser consultados diretamente na Internet e costumam disponibilizar ferramentas embutidas que produzem resultados em vários formatos. Em contraste, os *corpora off-line*, como dissemos acima, são, em geral, compilados por pesquisadores ou tradutores para um objetivo específico e necessitam de ferramentas externas para serem explorados.

2.1.2) Corpora monolíngues

Há vários *corpora* monolíngues *on-line*, a maioria deles para a língua inglesa (vide o capítulo de Viana neste livro). Podem ser compostos pelo que se denomina 'língua geral', ou por textos de 'língua de especialidade'. Os de língua geral contêm textos jornalísticos, acadêmicos, literários, científicos e, por vezes, também transcrições de textos falados. Costumam ter uma interface bastante amigável, muitas vezes com tutoriais para orientar seu uso. Um desses *corpora* é o *Corpus* do Português, com 45 milhões de palavras, cobrindo o

[1] Todos os *corpora* mencionados neste capítulo são detalhados em Viana (2014), neste volume.
[2] *Corpus* do Português: http://www.corpusdoportugues.org.

período do século XIV ao XX, com textos do PB e PE. Foi compilado por Mark Davies da Brigham Young University e Michael J. Ferreira da Universidade de Georgetown. A tela abaixo mostra uma busca pela palavra 'implicar' (Seta 1) apenas em textos dos séculos XIX e XX (Seta 2), o número de ocorrências no corpus – 88 (Seta 3) e os respectivos exemplos (Seta 4).

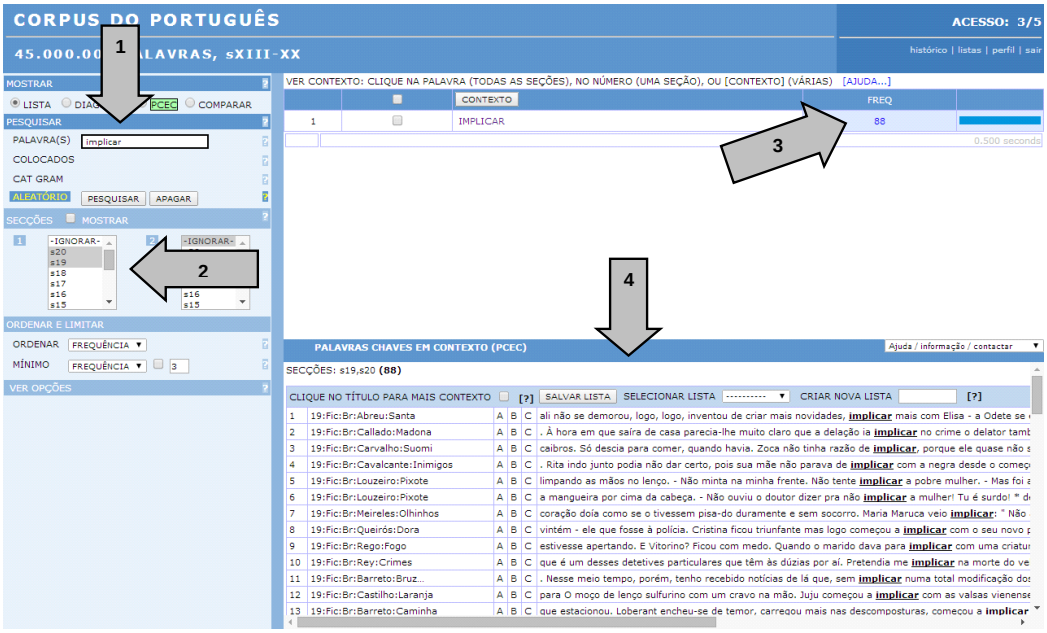


Figura 1: Tela do *Corpus* do Português

Dentre os *corpora* de língua geral disponíveis *on-line* para o português brasileiro, temos ainda o Lácio-Ref, dentro do projeto Lácio-Web[3] (ALUÍSIO et alii, 2003), o *Corpus* Brasileiro,[4] além dos *corpora* disponibilizados pela Linguateca, como o NILC/São Carlos[5] e o ECI-EBR.[6] Como cada *corpus* tem uma interface distinta, cabe ao usuário familiarizar-se com cada uma delas.

Dentre os *corpora* de língua de especialidade, ou seja, *corpora* especializados, podemos citar os *corpora* que compõem o CorTec, tanto em inglês quanto em português.

[3] Lácio-Web: http://www.nilc.icmc.usp.br/lacioweb/.
[4] Corpus Brasileiro: http://corpusbrasileiro.pucsp.br.
[5] NILC/São Carlos: http://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS.
[6] ECI-EBR: http://www.linguateca.pt/acesso/corpus.php?corpus=ECI-EBR.

2.1.3) Corpora bi- ou multilíngues

São *corpora* em duas ou mais línguas, como o próprio nome já diz. São esses os mais usados nos estudos relativos à Tradução, pois permitem a comparação entre línguas. Podem ser de dois tipos: paralelos e comparáveis.

2.1.3.1) Corpora paralelos

Corpora paralelos são compostos por textos originais em determinada língua e suas respectivas traduções em uma ou mais línguas. Podem ser unidirecionais, ou seja, originais numa língua A e traduções na língua B; ou bidirecionais, com originais tanto na língua A quanto na B e, portanto, respectivas traduções também nas duas línguas. Há três *corpora* paralelos disponíveis *on-line* que lidam exclusivamente com o par inglês-português, todos bi-direcionais: o COMPARA[7], o CorTrad[8] e o COPA-TRAD[9]. O COMPARA é composto por extratos de obras literárias de quatro variantes do português (Angola, Brasil, Moçambique e Portugal) e três do *inglês* (África do Sul, Estados Unidos e Reino Unido). Permite uma grande variedade de pesquisas, de simples a ultra-avançadas. A tela abaixo mostra uma pesquisa simples, na direção português-inglês para a palavra 'feijão'.

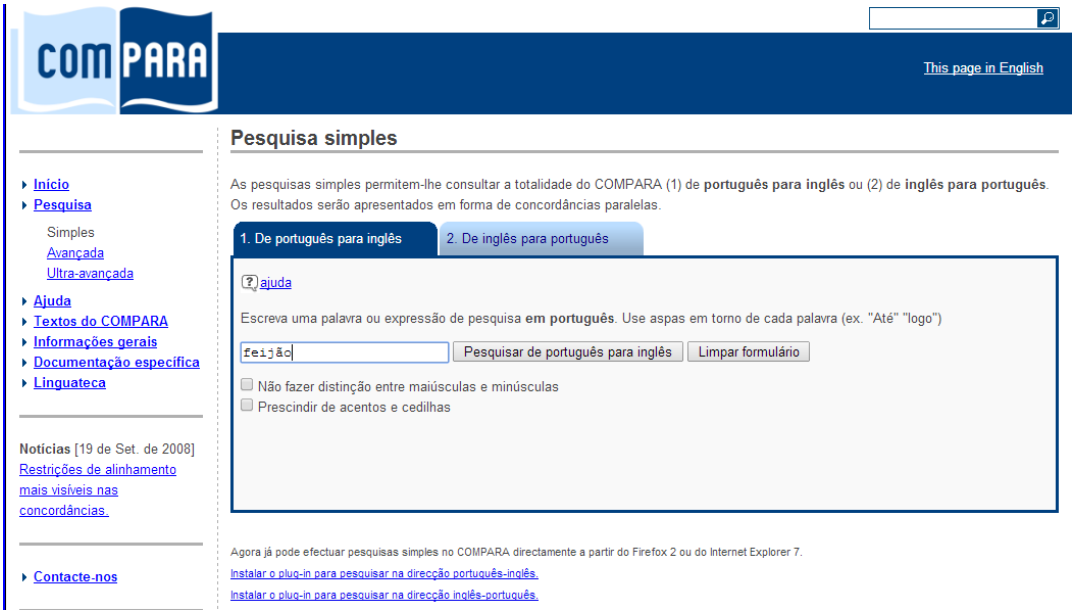


Figura 2: Tela do COMPARA

[7] COMPARA: http://www.linguateca.pt/COMPARA.
[8] CorTrad: http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html.
[9] COPA-TRAD: http://copa-trad.ufsc.br/

O COPA-TRAD, sediado na Universidade Federal de Santa Catarina, é composto por cinco *corpora* distintos: (1) COPA-LIJ (*Corpus* Paralelo de Literatura Infantil e Juvenil), (2) COPA-MDT (*Corpus* Paralelo de Meta-Discurso em Tradução, (3) COPA-RAC (*Corpus* Paralelo de Resumos Acadêmicos), (4) COPA-TEJ (*Corpus* Paralelo de Textos Jurídicos, e (5) COPA-TEL (*Corpus* Paralelo de Textos Literários).

Uma das formas de apresentação dos resultados é denominada concordância, que apresenta, em contexto, todas as ocorrências da palavra de busca no *corpus.* No caso de um *corpus* paralelo como o COMPARA, a concordância é exibida de forma alinhada, ou seja, os exemplos aparecem lado a lado nas duas línguas (vide Figura 3).

PBOL1(506):	Pelas manhãs, depois que o marido, negociante de **feijão** e milho, vai para o trabalho, se debruça à janela pintada de verde.	In the morning, after her husband, a cereal merchant, goes to work, she leans out of the green window.
PBOL1(847):	O negociante de milho e de **feijão**, felizmente, não quis acompanhá-la em seu lamentado começo de viagem; meus pais não me deixariam andar na rua, sem vigilância, depois das oito horas.	Luckily the cereal merchant refused to accompany her in the lamented beginning of her voyage; my parents would not let me go out in the streets, unsupervised, after eight.
PPJS1(1603):	Tu tens **feijão** em casa, que chegue para nós?	have we got enough beans in the house for all of us?´
PPJS1(1628):	Estava preparando afanosamente o almoço, e, com o mais juvenil dos sorrisos, chamou-me a que olhasse o interior de uma panela que destapou: era **feijão**.	She was painstakingly preparing lunch and, with a girlish smile directed me to peer inside a pan from which she removed the lid: beans.

Figura 3: Algumas linhas de concordância para 'feijão' no COMPARA.

Além de concordâncias, o COMPARA, na opção ultra-avançada, também apresenta resultados por categoria gramatical (substantivo, adjetivo, etc.), campo semântico (cor, roupa, etc.), tempo verbal (presente, pretérito, etc.), fonte do texto, autor, entre outras.

O CorTrad é um *corpus* multiversão, pois apresenta, sempre que possível, o original e várias versões da tradução desse texto. É composto por um sub*corpus* Jornalístico, um Literário e um Técnico-científico.

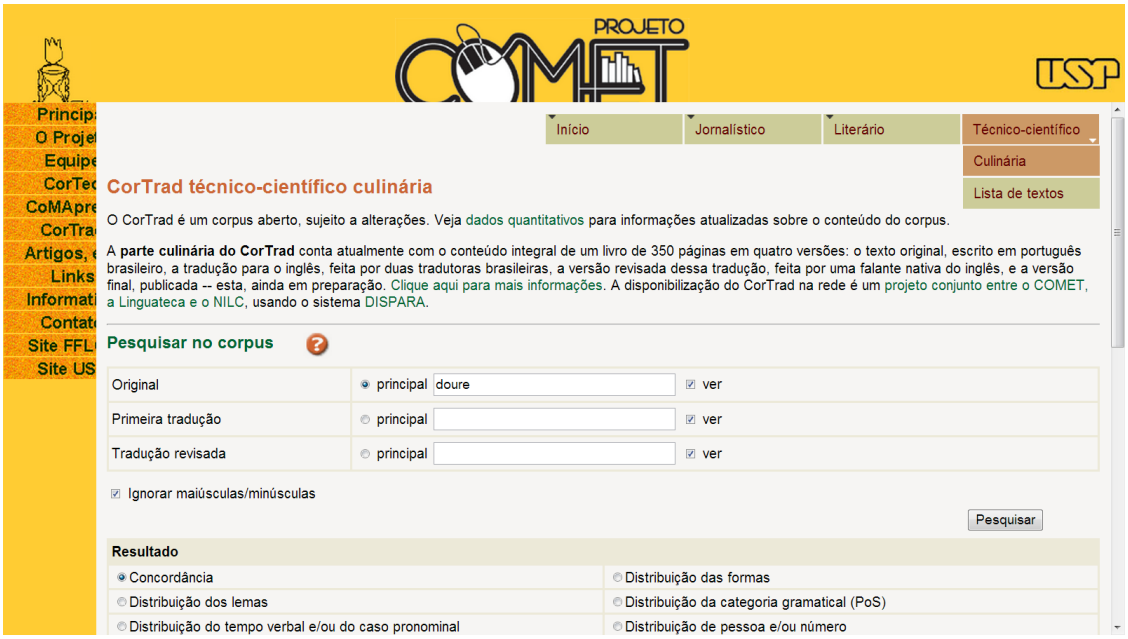


Figura 4: Tela inicial do CorTrad mostrando uma busca pela palavra ‘doure’

Uma busca pela palavra ‘doure’, no sub*corpus* Técnico-científico, composto, por ora, por um livro de culinária, originalmente escrito em português, traz 94 ocorrências, como indicado na Figura 5.

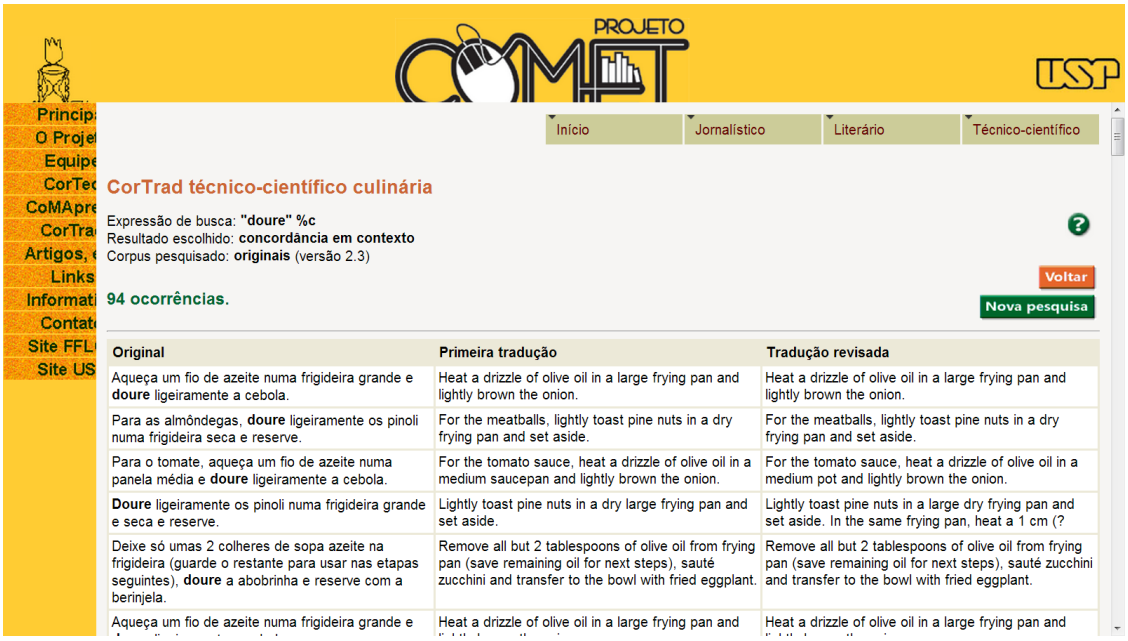


Figura 5: Concordância parcial para ‘doure’ no *subcorpus* Técnico-científico do CorTrad

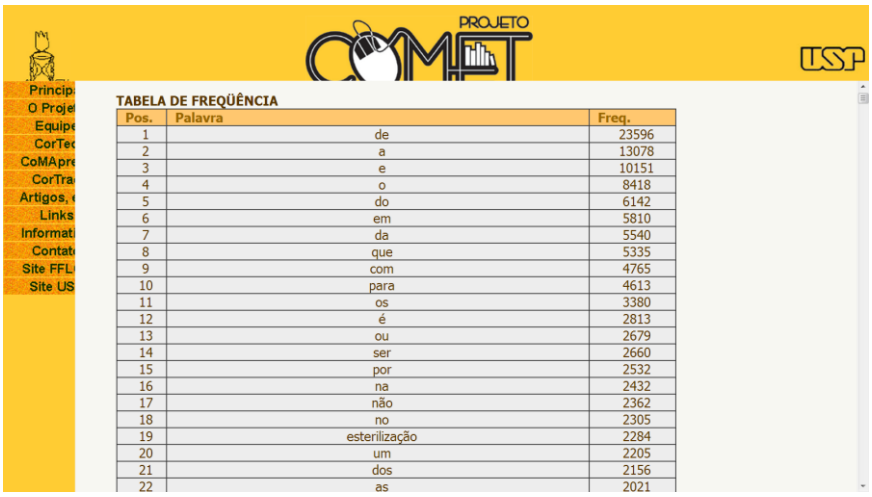
Observe-se que a palavra de busca, ‘doure’, aparece em negrito na primeira coluna. As outras duas colunas não apresentam a tradução dessa palavra em negrito, mas exibem as sentenças nas versões traduzidas correspondentes às sentenças no texto original. Assim, cabe ao pesquisador identificar essas traduções. Por exemplo, na primeira ocorrência, pode-se afirmar que “doure ligeiramente a cebola” foi traduzido por “lightly brown the onion” nas duas versões da tradução.

Outro tipo de *corpus* bi- ou multilíngue é o *corpus* comparável, que veremos a seguir.

2.1.3.2) *Corpora* comparáveis

Na realidade, essa terminologia pode gerar certa confusão por ser empregada para vários tipos de *corpora*. Em primeiro lugar, pode ser empregada, num sentido bastante amplo, para quaisquer *corpora* que se queira comparar. Neste volume, no entanto, *corpora* comparáveis devem ser entendidos como aqueles com textos originais em duas (ou mais) línguas, numa determinada área de domínio. Para efeitos de comparação, os *corpora* devem ser constituídos por textos de mesma tipologia, de temática semelhante e cobrindo períodos similares. A esse processo denominamos balanceamento do *corpus*. Por serem textos autênticos nas duas línguas, são comumente usados para a extração de terminologia especializada ou mesmo para certificar-se do termo correto empregado em determinado contexto (BOWKER & PEARSON, 2002). O CorTec[10] é um *corpus* técnico desse tipo, em inglês e português, com cerca de 20 sub*corpora* de distintas áreas de especialidade, dentre elas: Autoclaves, Culinária, Futebol, Hipertensão, Informática, Insuficiência Renal, Medidores Eletromagnéticos de Vazão, e Prostodontia. O CorTec como um todo é um *corpus* dinâmico, de modo que novos *subcorpora* podem ser incorporados sempre que disponíveis. Possui ferramentas embutidas, dentre as quais um concordanciador, um gerador de n-gramas (sequência de palavras) e um gerador de lista de palavras, que elenca todas as palavras contidas no *corpus*, conforme se vê, parcialmente, na Figura 6.

[10] CorTec: http://www.fflch.usp.br/dlm/comet/consulta_cortec.html.



Pos.	Palavra	Freq.
1	de	23596
2	a	13078
3	e	10151
4	o	8418
5	do	6142
6	em	5810
7	da	5540
8	que	5335
9	com	4765
10	para	4613
11	os	3380
12	é	2813
13	ou	2679
14	ser	2660
15	por	2532
16	na	2432
17	não	2362
18	no	2305
19	esterilização	2284
20	um	2205
21	dos	2156
22	as	2021

Figura 6: As 22 palavras mais frequentes no sub*corpus* de Autoclaves do CorTec

É interessante notar, na Figura 6, que as primeiras palavras são palavras gramaticais, como preposições e artigos. Aliás, isso ocorrerá em qualquer *corpus*, pois são sempre as palavras mais frequentes A primeira palavra de conteúdo, "esterilização", aparece na posição 19 e já dá alguma indicação do teor desse *subcorpus*.

As várias ferramentas empregadas na análise de *corpora* serão apresentadas na Seção 5. A seguir discutimos os passos para a compilação de um *corpus*.

3) Compilação de um *corpus*

O primeiro passo para a compilação de um *corpus* é definir seus objetivos, isto é, para quê servirá o *corpus*: para consulta do tradutor? Para extração de terminologia de determinada área? Para estudar estratégias de tradução? Para comparar diversas traduções de um mesmo original? Uma vez definido o objetivo, devem ser estabelecidos os parâmetros de sua composição. Por exemplo:

a. será um *corpus* estático (i.e. fechado) ou dinâmico (i.e. com atualizações)?

b. incluirá apenas textos escritos ou também língua falada?

c. será monolíngue ou multilíngue?

d. se for multilíngue, será paralelo ou comparável?

e. que tipos de textos serão incluídos: jornalísticos, manuais, artigos acadêmicos, teses e dissertações, etc.?

f. qual será a proporção entre os diversos tipos?

g. quais domínios abrangerá? Por exemplo, natação, sustentabilidade, segurança bancária, obras de determinado autor, obras de determinada época, etc.

h. qual será a fonte dos textos a serem compilados: Internet ou material impresso? No caso de material impresso, é preciso lembrar que esse deverá ser escaneado ou, na pior das hipóteses, digitado.

i. os textos serão completos ou apenas fragmentos? Por exemplo, os textos que compõem o *Brown Corpus* são constituídos de trechos de 2.000 palavras, enquanto o *Bank of English* abriga textos completos. Já o COMPARA é constituído, no geral, de 30% das obras contempladas.

j. como será estabelecido o tamanho do *corpus*, pelo número de textos ou pelo número de palavras? Essa decisão muitas vezes terá de ser modificada durante o processo de compilação devido à especificidade dos textos. Esse tópico é discutido abaixo.

Embora esses itens devam ser estabelecidos antes de se começar a coleta dos textos, alguns podem sofrer ajustes ao longo do processo de compilação. Por exemplo, no caso de um *corpus* de contratos em inglês e português (CARVALHO, 2007), em que inicialmente se determinou o número de 500.000 palavras para cada língua, logo se observou que isso produziria dois *corpora* com conteúdos bastante diversos uma vez que esses documentos tinham extensões distintas nas duas línguas. Enquanto os contratos em português eram sucintos por fazerem apenas referência à legislação vigente, os contratos em inglês eram muito mais longos por explicitarem a legislação pertinente. Assim, esse critério foi alterado para incluir 28 tipos contratuais com 5 exemplares de cada.

Outro problema que pode surgir refere-se à disponibilidade de textos nas várias línguas. Como o inglês é a língua mais representada na Internet, a coleta de textos nessa língua costuma ser bem mais fácil do que nas outras, o que acaba demandando, por vezes, um tempo maior para a realização da tarefa. Em determinados casos pode ser necessário recorrer a textos impressos para se obter um corpus balanceado, isto é, um *corpus* bilíngue com conteúdo equiparável nas duas línguas.

Um fator frequentemente discutido na literatura (e.g. LEECH 1991, BIBER 1993, McENERY et al., 2006) diz respeito à representatividade de um *corpus*, ou seja, o fato de que um *corpus* deve ser representativo daquilo que se pretende investigar, contendo uma amostragem significativa do universo a ser estudado. Quando isso é alcançado, os resultados obtidos podem ser generalizados para todo o universo em estudo (LEECH, 1991). No entanto, o que seria uma amostragem significativa? Segundo Biber (1993, p. 243), a representatividade diz respeito ao quanto uma amostra inclui toda a gama de variabilidade da produção linguística que se pretende investigar. Dessa forma, um *corpus* geral, por exemplo, deve incluir o maior número de gêneros e tipos textuais para assegurar sua representatividade.

Apesar disso, Leech (1991, p. 27) adverte que toda representatividade "deve ser encarada como um ato de fé",[11] uma vez que ainda não há métodos científicos que mensurem esse aspecto.

Uma vez coletados os textos, eles devem ser salvos no formato .txt para poderem ser lidos pelas ferramentas computacionais. Para tanto, devem antes ser limpos de todos os elementos que não sejam de interesse para a pesquisa linguística como, por exemplo, tabelas, URLs, etc. Em seguida, cada arquivo deve ser nomeado de forma a deixar claro seu conteúdo. O nome pode incluir a área (por exemplo, "derm" para dermatologia, "inf" para informática, "cul" para culinária), o tipo de texto ("jor" para jornalístico, "aca" para acadêmico, "rec" para receita), a língua ("pt" para português, "in" para inglês), e quaisquer outras informações que sejam consideradas relevantes. Assim, por exemplo, um arquivo denominado: culrecmaspt01.txt deixa claro tratar-se de um texto de culinária ("cul"), mais especificamente de uma receita ("rec") de massa ("mas") em português ("pt"). O número indica ser a primeira receita de várias outras de massa.

Os arquivos devidamente nomeados devem ser armazenados em pastas. No exemplo acima, pode-se ter uma pasta denominada "Culinária" e uma subpasta indicando a língua, no caso "Português".

Feito isso, os textos estão basicamente prontos para serem explorados por ferramentas computacionais. Porém, eles ainda podem ser enriquecidos com vários tipos de anotação.

4) Anotação de *corpora*

A anotação de *corpora* mais comum é a *morfossintática* (*POS* em inglês, de *part-of-speech*), em que cada palavra recebe uma etiqueta correspondente a sua categoria gramatical naquele contexto. Uma anotação *lexical* permite identificar, por exemplo, estrangeirismos, entidades nomeadas (nomes próprios), termos técnicos, etc. Campos *semânticos* como cor, roupa, sentimentos, modalidade (e.g. obrigação e possibilidade) também podem etiquetados. Por exemplo, no COMPARA, as cores são etiquetadas semanticamente. Assim, uma busca pela palavra 'branco' retorna 200 ocorrências em que essa palavra de fato denota uma cor, 96 em que se refere a raça, 9 em que é usada de forma metafórica ou idiomática, e 8 em que se refere a vinho. Finalmente, temos etiquetas *discursivas*, que indicam as várias partes de um texto, por exemplo, num texto acadêmico: resumo, introdução, materiais e métodos; ou numa

[11] No original: "At present, an assumption of representativeness must be regarded largely as an act of faith" (tradução da autora).

receita: ingredientes, modo de fazer, etc. Na realidade, cada pesquisador pode criar as etiquetas que forem mais relevantes à sua pesquisa.

Essas anotações possibilitam pesquisas bastante elaboradas. O CorTrad, um *corpus* de originais e respectivas traduções, tem anotação para cor e roupa. Assim, pode-se investigar, por exemplo, como são traduzidas as diversas cores em qualquer um dos *subcorpora* que o compõem (i.e. jornalístico, literário, técnico-científico). Santos et al. (2012), por exemplo, estudaram o uso de cores nesses *subcorpora* e suas respectivas traduções. Concluíram que, contrariando expectativas, as cores são muito empregadas na terminologia técnica, mas variam de uma língua para outra, aumentando, assim, as possibilidades de erros tradutórios.

A seguir veremos as principais ferramentas que permitem a exploração de um *corpus*.

5) Ferramentas de investigação

Os *corpora* que não possuem ferramentas embutidas necessitam de ferramentas externas para serem investigados. Os programas mais conhecidos que contêm essas ferramentas são o *AntConc* (ANTHONY, 2004, 2012[12]) e o *WordSmith Tools* (doravante *WST*) (SCOTT, 1996[13]) . O primeiro, desenvolvido por Laurence Anthony da Universidade de Waseda, no Japão, pode ser baixado gratuitamente; o segundo, desenvolvido por Mike Scott, é um programa comercial e pode ser adquirido *on-line*. Ambos apresentam as três ferramentas básicas para a análise de um *corpus*: lista de palavras, lista de palavras-chave e concordâncias.

5.1) Lista de palavras

Já vimos, na Figura 6, um exemplo de lista de palavras do CorTec. A seguir apresentamos uma lista de palavras gerada pelo *AntConc* e outra pelo *WSTWST* – ambas para um *corpus* de Culinária.[14]

[12] AntConc: http://www.antlab.sci.waseda.ac.jp/antconc_index.html.
[13] WordSmith Tools: http://www.lexically.net/wordsmith/index.html.
[14] Esse *corpus* está disponibilizado no CorTec e foi compilado por Teixeira (2008).

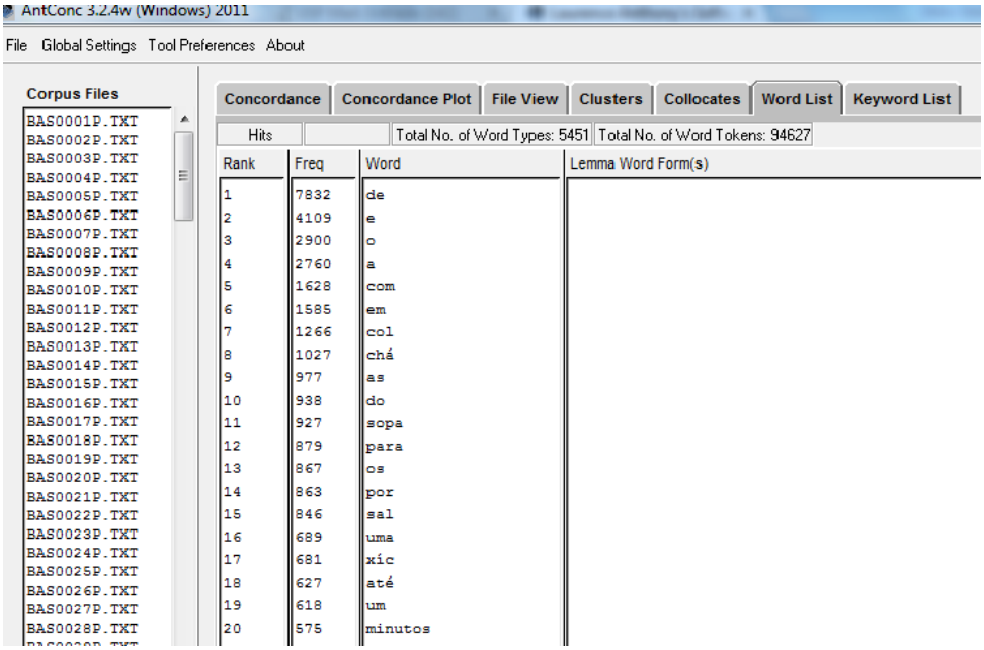
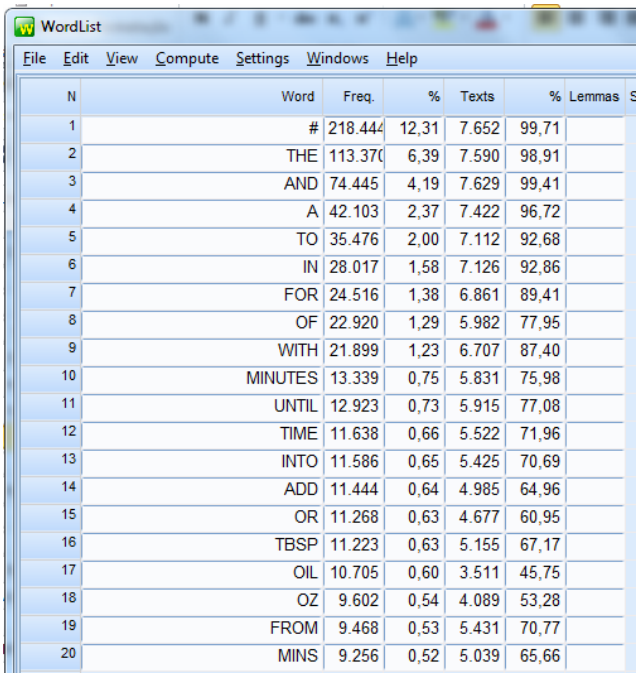


Figura 7: Primeiras vinte palavras mais frequentes para um *corpus* de Culinária em português listadas pelo *AntConc*

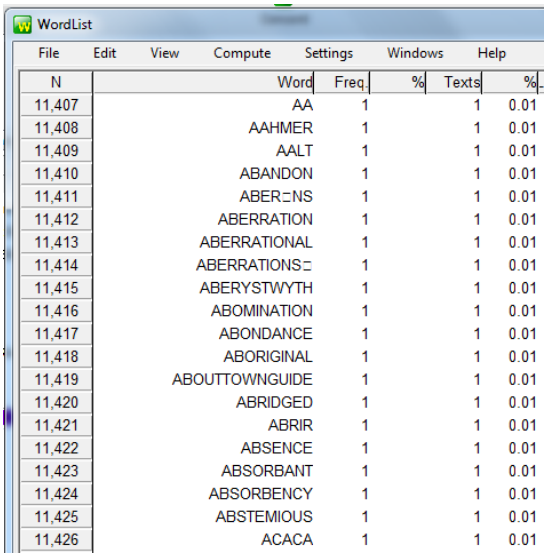


N	Word	Freq.	%	Texts	%	Lemmas
1	#	218.444	12,31	7.652	99,71	
2	THE	113.370	6,39	7.590	98,91	
3	AND	74.445	4,19	7.629	99,41	
4	A	42.103	2,37	7.422	96,72	
5	TO	35.476	2,00	7.112	92,68	
6	IN	28.017	1,58	7.126	92,86	
7	FOR	24.516	1,38	6.861	89,41	
8	OF	22.920	1,29	5.982	77,95	
9	WITH	21.899	1,23	6.707	87,40	
10	MINUTES	13.339	0,75	5.831	75,98	
11	UNTIL	12.923	0,73	5.915	77,08	
12	TIME	11.638	0,66	5.522	71,96	
13	INTO	11.586	0,65	5.425	70,69	
14	ADD	11.444	0,64	4.985	64,96	
15	OR	11.268	0,63	4.677	60,95	
16	TBSP	11.223	0,63	5.155	67,17	
17	OIL	10.705	0,60	3.511	45,75	
18	OZ	9.602	0,54	4.089	53,28	
19	FROM	9.468	0,53	5.431	70,77	
20	MINS	9.256	0,52	5.039	65,66	

Figura 8: Primeiras vinte palavras mais frequentes para um *corpus* de Culinária em inglês listadas pelo *WST*

Como vemos, a lista de palavras exibe todas as palavras do *corpus* (ou texto) que está sendo investigado por ordem de frequência, como nos exemplos acima, ou alfabética. No *AntConc*, os totais aparecem no topo da tela, subdivididos em *types* (palavras distintas ou

formas) – 5.451 – e *tokens* (número total de palavras ou ocorrências) – 94.627. O *WST* exibe o número total de formas no final da tela – no caso, 11.141. Pela alta frequência das palavras gramaticais em qualquer tipo de texto, são elas que vão encabeçar a lista. As primeiras palavras de conteúdo geralmente indicam o tema do *corpus*. Pelas listas acima, fica fácil identificar ambos os *corpora* como sendo da área da culinária, pois em português sobressaem *col*, *chá*, *sopa*, *sal*, *xíc*, *minutos*, enquanto em inglês temos *minutes*, *time*, *add*, *tbsp*, *oil*, *oz*, *mins*. Ao final da lista, constam as palavras com uma única ocorrência, denominadas *hapax legomena* (vide Figura 9).


WordList

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%
11,407	AA	1		1	0.01
11,408	AAHMER	1		1	0.01
11,409	AALT	1		1	0.01
11,410	ABANDON	1		1	0.01
11,411	ABER□NS	1		1	0.01
11,412	ABERRATION	1		1	0.01
11,413	ABERRATIONAL	1		1	0.01
11,414	ABERRATIONS□	1		1	0.01
11,415	ABERYSTWYTH	1		1	0.01
11,416	ABOMINATION	1		1	0.01
11,417	ABONDANCE	1		1	0.01
11,418	ABORIGINAL	1		1	0.01
11,419	ABOUTTOWNGUIDE	1		1	0.01
11,420	ABRIDGED	1		1	0.01
11,421	ABRIR	1		1	0.01
11,422	ABSENCE	1		1	0.01
11,423	ABSORBANT	1		1	0.01
11,424	ABSORBENCY	1		1	0.01
11,425	ABSTEMIOUS	1		1	0.01
11,426	ACACA	1		1	0.01

Figura 9: Algumas das palavras com uma única ocorrência no *corpus* de Culinária em inglês listadas pelo *WST*

As palavras com uma única ocorrência podem indicar tanto termos de uma área especializada quanto vocábulos usados criativamente.

5.2 Lista de palavras-chave

O *AntConc* e o *WST* também podem produzir listas de palavras-chave (em inglês, *keyword list*). As palavras-chave resultam da comparação de duas listas de palavras, uma do *corpus* de estudo e outra de um *corpus* que serve de comparação, geralmente denominado *corpus* de referência, mas pode também ser chamado de *corpus* de contraste ou mesmo de comparação.

Para fazer essa comparação, os programas recorrem a testes estatísticos: *log-likelihood* (teste estatístico de máxima verossimilhança), *T-score* (escore-T) ou *mutual information*

(informação mútua). Nesse processo, as palavras que tem frequências estatisticamente similares desaparecem, restando apenas as que são peculiares ao *corpus* de estudo. Assim, uma lista de palavras-chave para um *corpus* de culinária em português[15] apresentará os itens indicados na Figura 10.

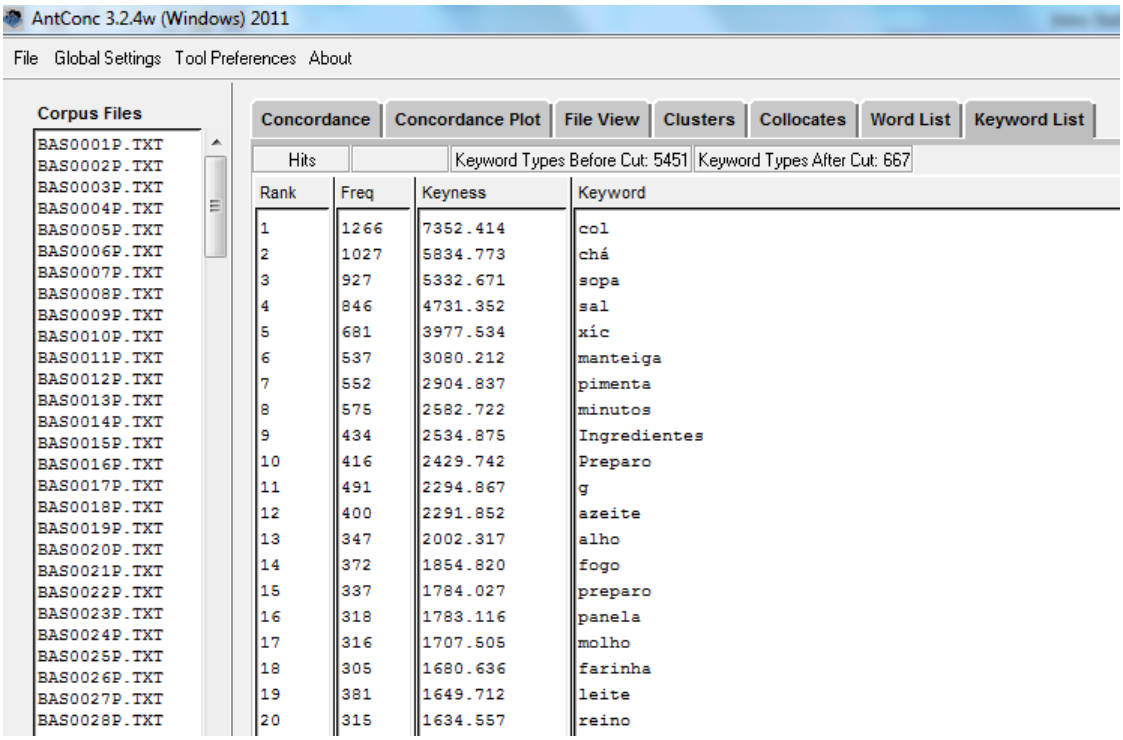
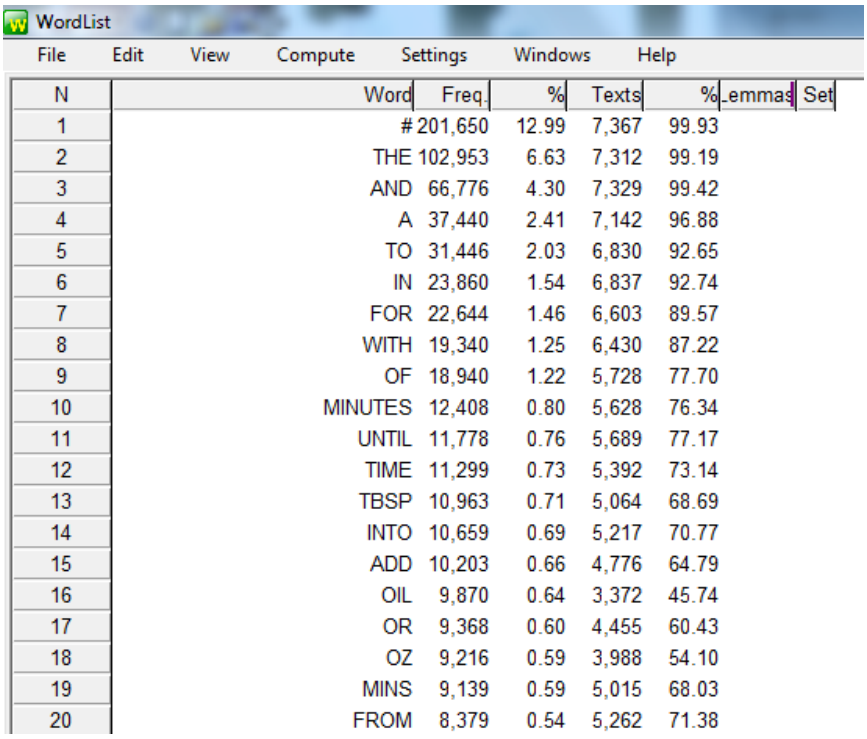


Figura 10: Primeiras vinte palavras-chave para um *corpus* de culinária em português geradas pelo *AntConc*

Como se pode observar, a lista gerada pelo *AntConc* contém apenas as palavras peculiares ao *corpus* de culinária com suas respectivas frequências brutas (Freq). Além disso, indica também a chavicidade (*Keyness*), uma medida que indica o quão típica é a palavra naquele *corpus*.

A lista apresentada pelo *WST* (vide Figura 11) apresenta, além dessas informações, também a frequência no *corpus* de referência (RC. Freq.), além das frequências relativas, em termos percentuais, tanto do *corpus* que está sendo investigado (%) quanto do *corpus* de referência (RC. %).

[15] Esse corpus está disponibilizado no CorTec e foi compilado por Teixeira (2008).



N	Word	Freq.	%	Texts	%	Lemmas	Set
1	#	201,650	12.99	7,367	99.93		
2	THE	102,953	6.63	7,312	99.19		
3	AND	66,776	4.30	7,329	99.42		
4	A	37,440	2.41	7,142	96.88		
5	TO	31,446	2.03	6,830	92.65		
6	IN	23,860	1.54	6,837	92.74		
7	FOR	22,644	1.46	6,603	89.57		
8	WITH	19,340	1.25	6,430	87.22		
9	OF	18,940	1.22	5,728	77.70		
10	MINUTES	12,408	0.80	5,628	76.34		
11	UNTIL	11,778	0.76	5,689	77.17		
12	TIME	11,299	0.73	5,392	73.14		
13	TBSP	10,963	0.71	5,064	68.69		
14	INTO	10,659	0.69	5,217	70.77		
15	ADD	10,203	0.66	4,776	64.79		
16	OIL	9,870	0.64	3,372	45.74		
17	OR	9,368	0.60	4,455	60.43		
18	OZ	9,216	0.59	3,988	54.10		
19	MINS	9,139	0.59	5,015	68.03		
20	FROM	8,379	0.54	5,262	71.38		

Figura 11: Primeiras vinte palavras-chave para um *corpus* de culinária em inglês geradas pelo *WST*

A ferramenta *Keywords* do *WST* apresenta primeiramente as palavras-chave positivas, que acabamos de ver (vide Figura 11). Mas, apresenta também – em vermelho –, ao final, as palavras-chave negativas, ou seja, aquelas que são mais frequentes no *corpus* de referência do que no *corpus* de estudo. Se contrastarmos um *corpus* de receitas brasileiras (como *corpus* de estudo) com um *corpus* de receitas portuguesas (como *corpus* de referência) (TAGNIN & TEIXEIRA 2004), por exemplo, teremos como palavras-chave positivas as do *corpus* brasileiro e como negativas as do *corpus* português (vide Figura 12).

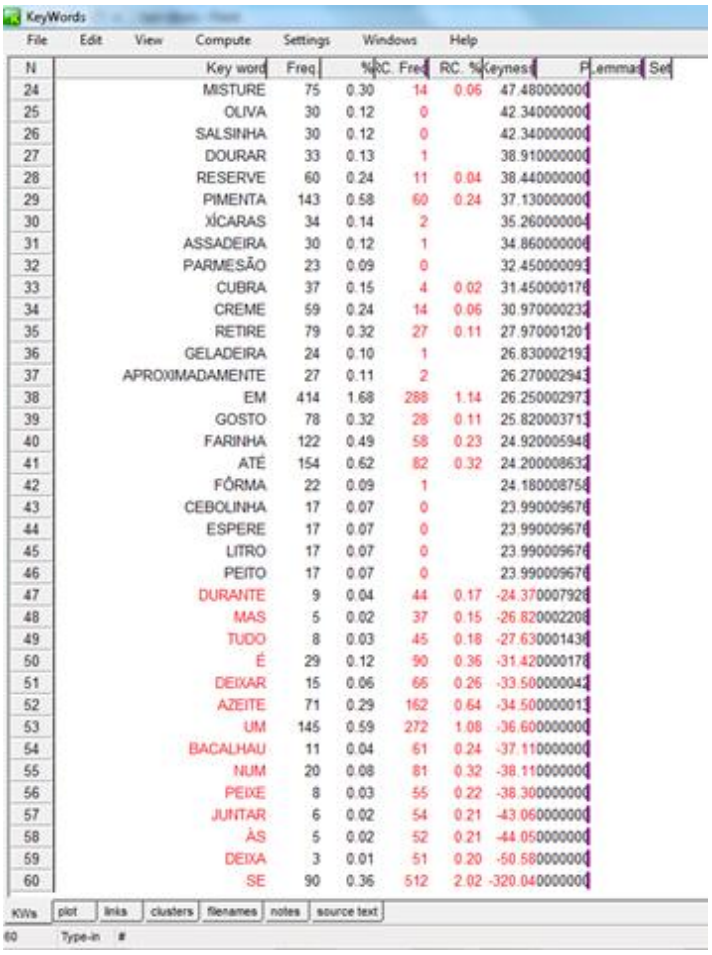


Figura 12: Palavras-chave positivas (até a linha 46) e negativas (linhas 47-60)

Na Seção 6, discutiremos as possibilidades de análise proporcionadas pelas palavras-chave.

5.3 Concordanciador

Uma das ferramentas mais importantes para a Linguística de *Corpus* são os concordanciadores, que geram linhas de concordância. Essas linhas apresentam todas as ocorrências de uma palavra ou expressão de busca em contexto. No geral, a palavra é apresentada no centro, isto é, com uma parte do contexto à direita e outro à esquerda. Muitos concordanciadores permitem reordenar os resultados pela ordem alfabética das palavras que ocorrem à direita ou à esquerda da palavra de busca. Esse processo facilita sobremaneira a identificação de padrões recorrentes, como podemos ver na Figura 13.

```
1 a) de manteiga 1/2 xíc. (chá) de pinoli sal a gosto Modo de fazer: Refogue a c
2  as finas 1 1/2 col. (sopa) de manteiga sal a gosto Preparo Disponha a farinha
3 da 1 pitada de páprica 2 gemas 3 claras sal a gosto Preparo Lave as batatas, f
```

```
 4   de purê de tomates 1 col. (chá) de mel sal a gosto Preparo Misture todos os i
 5 obalo, abadejo etc.) em pedaços grandes sal a gosto Preparo Tempere o peixe e
 6 nteiga, mexendo para misturar. Acerte o sal e a pimenta. Disponha o risoto nos
 7   Junte os cubinhos de tomate e acerte o sal e a pimenta. Disponha os filés no
 8 te a água para o cozimento, tempere com sal e a pimenta e deixe a panela semit
 9   que aparecem. Reserve. Junte a água, o sal e o açúcar numa fôrma de alumínio
10 coloque o salsão, a manteiga, o açúcar, sal e pimenta e acrescente a água até
11 batata ralada à manteiga e tempere com sal e pimenta. Faça panquecas, apertan
12 frescos, descascados e cozidos al dente sal e pimenta-do-reino moída 1 col. (s
13 tas médias descascadas e raladas grosso sal e pimenta-do-reino moída a gosto 1
14 da 4 col. (chá) de farelo de pão branco sal e pimenta-do-reino moída a gosto P
15 sa) picada Preparo Tempere os filés com sal e pimenta-do-reino moída a gosto.
16 a salsinha e as azeitonas e tempere com sal e pimenta-do-reino moída a gosto
17 mente) 1 col. (sopa) de salsinha picada sal e pimenta-do-reino moída a gosto P
18 ol. (chá) de páprica doce 1 ovo, batido sal e pimenta-do-reino moída a gosto P
19 sopa) de óleo 2 col. (sopa) de manteiga sal e pimenta-do-reino moída a gosto P
20 icada 3 col. (sopa) de salsinha, picada sal e pimenta-do-reino moída a gosto P
```

Figura 13: Linhas de concordância para *sal*, ordenadas pela 1ª e 2ª palavras à direita

A Figura 13 apresenta uma seleção das 845 linhas de concordância geradas pelo *WST* para a palavra *sal* num *corpus* de culinária, em que podemos identificar os seguintes padrões à direita da palavra de busca: *sal a gosto, o sal e a pimenta, sal e pimenta-do-reino moída a gosto*.

6) Tipos de pesquisa e o que revelam

São vários os tipos de pesquisa que podemos desenvolver com o uso de *corpora*, tanto comparáveis quanto paralelos. Alguns desses tipos são discutidos a seguir.

***6.1) Corpora* comparáveis – busca por equivalentes**

Os *corpora* comparáveis, com textos originalmente escritos em cada língua, são extremamente úteis para a tradução por atestarem o uso autêntico de palavras, termos ou expressões. Dessa forma, prestam-se para a busca de equivalentes (TAGNIN, 2007; PHILIP, 2009). Assim, se o tradutor estiver em dúvida se pode traduzir para o inglês, num texto sobre hotelaria, um apartamento num hotel por *apartment*, ele pode recorrer a um *corpus* especializado e verificar a ocorrência de *apartment*. A Figura 14 apresenta os primeiros 5 resultados do sub*corpus* de Turismo-Hotelaria no *Corpus* Técnico (CorTec) do Projeto CoMET.

```
1     u by adding an extra bed  The Apartment   We have one 2 bedroom suite
2     ast home is simple; A private apartment to call your home when visiti
```

```
3      fireplace.    GARDEN COTTAGE APARTMENT  Also, our Garden Cottage Apa
4     ENT  Also, our Garden Cottage Apartment sleeps up to six (queen bed i
5     house, the one-bedroom luxury apartment, the two bedroom cabin, or on
```

Figura 14: Algumas linhas de concordância para *apartment* no *subcorpus* Turismo-Hotelaria[16] do CorTec

A partir da leitura das linhas de concordância na Figura 14, é fácil notar que não se trata de apartamento em hotel, mas de apartamento em prédio. Nesse caso, cabe uma nova busca, dessa vez no corpus em português para identificar os contextos em que *apartamento* ocorre (cf. Figura 15).

```
1  partamentos Business center     No Novo Apartamento Premium você encontra:  31m2 3
2  quecível.   Núpcias White   * Diária em apartamento luxo superior * Café da manhã
3  uxo superior * Café da manhã servido no apartamento * 02 botões de rosas vermelhas
4  o) Bufê de café da manhã Café e chá  no apartamento Balcão de concierge Câmbio, na
5   t quente e frio  1 vaga de garagem por apartamento   Ampla sala com varanda  C
6   acionamento/desligamento de energia no apartamento Campainha, olho m gico e Rota
7  as at, 31/12/2011.        Tarifa Balcão Apartamento Standard Single Double  Triplo
8  rd Single Double  Triplo  Tarifa Acordo Apartamento Standard Single  Double  Tripl
9   de até 12 anos é free, quando no mesmo apartamento dos pais, sendo o máximo de 1
10 s pais, sendo o máximo de 1 criança por apartamento duplo;  Prazo de cancelamento:
```

Figura 15: Algumas linhas de concordância para *apartamento* no *subcorpus* Turismo-Hotelaria do CorTec

Nas linhas 7 e 8 da Figura 15, aparece a palavra *single*, como um tipo de apartamento. Buscando essa palavra no *corpus* em inglês encontramos nas linhas 1, 2 e 5 (vide Figura 16), a palavra *room*.

```
1   bath.   * Rooms have a Queen bed and 2 single beds in the bedroom and   a pull out
2  e Kitchens.  Couples Room Options:-   2 Single beds or 1 King bedÊ    Special Weekl
3   Square Inn was built in the 1860s as a single family home, and today its integrity
4  ek.   Corporate rates are available for single business travelers Sunday-Thursday.
5  t pets at the Inn.   Room rates are for single or double occupancy, $30.00 for each
6  VATE BATHROOMS:     Rates & taxes apply Single/Double occupancy (except Garden Cott
7  es in San Antonio are available for the single business traveler, starting at $117
8   of your day.   Corporate rates for the single weekday business traveler.   Staying
9   Key West guest room rates are based on single or double occupancy and are subject
10 o (2) for one (1) spa service: book any single spa service and get another of equal
```

Figura 16: Algumas linhas de concordância para *single* no *subcorpus* Turismo-Hotelaria do CorTec

[16] Esse *corpus* foi compilado por Navarro (2012) para sua dissertação de mestrado.

Embora *single* não se coloque com *room*, ou seja, não ocorre *single room*, observa-se que corresponde, no contexto, a 'apartamento' em português, pois encontramos, por exemplo, nas linhas 5, 6, e 9 da Figura 16 uma referência a *single or double occupancy*, ou seja, apartamentos para ocupação por uma ou duas pessoas. Dessa forma, o tradutor se assegura de que a melhor tradução de *apartamento* para o inglês é *room*.

Outra forma de averiguar equivalências é pela frequência da palavra no *corpus*. Uma busca no mesmo sub*corpus* de Turismo-Hotelaria em português, no CorTec, mostra que a palavra *apartamentos* aparece na 31ª posição, com 3.044 ocorrências.

Já no *corpus* em inglês, a primeira palavra de conteúdo é *room*, na 16ª posição, com 5.920 ocorrências.

Pos.	Palavra	Freq.
1	and	40699
2	the	38540
3	a	24787
4	of	21234
5	to	20020
6	with	14526
7	in	13984
8	for	12720
9	is	9924
10	your	9532
11	our	9108
12	or	7500
13	you	7354
14	are	6608
15	at	6580
16	room	5920
17	on	5300
18	from	4772
19	an	4194
20	will	3786

Figura 17: Lista parcial de palavras do *subcorpus* de Turismo-Hotelaria do CorTec

Apartment aparece apenas 64 vezes no singular e 52 no plural, num total de 116 ocorrências, o que já é indicativo de que não pode ser equivalente a *apartamento* em português.

Vejamos outro exemplo com o *subcorpus* de Insuficiência Renal. A palavra *renal* aparece na 15ª posição com 1.267 ocorrências, enquanto *rins* e *rim* aparecem 148 e 124 vezes, respectivamente.

No corpus em inglês, *renal* aparece 1.726 vezes, *kidney* e *kidneys*, 383 e 55 vezes, respectivamente.

Essa semelhança no número de ocorrências de *renal* em português e em inglês pode, entretanto, induzir-nos a conclusões apressadas, fazendo-nos supor que sejam equivalentes. De fato, em alguns contextos as palavras são equivalentes, como, por exemplo, 'doença renal' e *renal disease*, como se vê na Figura 18.

```
1 ogia  Assunto específico: Insuficiência renal crônica  DISTRIBUIÇÃO  Tipo: Impresso
2 ente, nem sempre o tratamento da doença renal, quando existe, resulta em reversão com
3   podem determinar a natureza da doença renal. Eventualmente, métodos adicionais, com

1   General Considerations  Patients with renal disease can present in a variety of way
2 s on urinalysis.  Once it is clear that renal disease is present, the primary goals a
3 of the renal damage. The major types of renal disease can be grouped according to the
```

Figura 18: Algumas linhas de concordância atestando a equivalência de 'doença renal crônica' e *renal disease*

Vejamos agora uma concordância para *kidney* na Figura 19.

```
1 -dose methotrexate  Chronic: Polycystic kidney disease, medullary sponge kidney,
2 cystic kidney disease, medullary sponge kidney, medullary cystic kidney disease
3 dullary sponge kidney, medullary cystic kidney disease  Interstitial Disease  Acute:
4 erstitial disorders, such as polycystic kidney disease, in which multiple bilateral
5 rtension, and analgesic abuse.  Myeloma kidney, in which urinary immunoglobulin light
```

Figura 19: Primeiras cinco linhas da concordância para *kidney* no *subcorpus* de Insuficiência Renal no CorTec

Podemos observar, na Figura 19, duas ocorrências de *polycystic kidney disease* (linhas 1 e 4), uma de *medullary cystic kidney disease* (linha 3), uma de *medullary sponge kidney* (linha 2) e uma de *myeloma kidney* (linha 5). Embora *renal disease* seja uma colocação frequente, quando se trata de uma doença específica (*polycystic kidney disease* e *medullary cystic kidney disease*), temos *kidney* em vez de *renal*. Além disso, *kidney* é usado para se referir ao órgão, como em *medullary sponge kidney* e *myeloma kidney*, que é também como são usados *rim* e *rins* em português, como vemos na Figura 20.

1 ções de fazer avaliações estruturais do rim com detalhes finos, como dimensões de c
2 . Na estenose unilateral, a presença do rim contralateral íntegro exerce um efeito
3 . Na estenose bilateral, na estenose do rim único ou transplantado ou na coarctação
4 ocedimento11. Na situação particular do rim contraído unilateral, a decisão entre r
5 u eventualmente tentar revascularizar o rim pequeno deve levar em conta a dosagem d

1 rítica com disfunção renal importante e rins ecograficamente normais, podendo diagno
2), que normalmente são eliminados pelos rins. Embora controversas, existem evidência
3 is de parênquima pela ultra-sonografia. Rins menores do que 7cm, hiperecogênicos, se
4 ato gastrointestinal e excretados pelos rins. Na insuficiência renal crônica (I
5 irinas são normalmente excretadas pelos rins, níveis elevados dessas moléculas s

Figura 20: Algumas linhas de concordância para 'rim' e 'rins' no *subcorpus* de Insuficiência Renal no CorTec

Voltando à Figura 19, observamos, nas linhas 1 e 4 , a ocorrência de *polycystic kidney disease.* Para chegarmos ao equivalente desse termo em português, podemos fazer uma busca, no CorTec, usando a opção "Começando com policístic", para obter resultados nos dois gêneros (masculino e feminino) e números (singular e plural).

1 paciente selecionado tinha doença renal policística bilateral como doença primária
2 cas e tardia ou ausente na doença renal policística do adulto. À semelhança do que
3 istos renais simples ou da doença renal policística do adulto,4 desenvolvendo-se, s
4 os pacientes portadores de doença renal policística, os binefrectomizados e aqueles
5 penas os grupos de portadores de doença policística ou com diagnóstico indeterminad

Figura 21: Algumas linhas de concordância para *policístic** no *subcorpus* de Insuficiência Renal no CorTec

Os resultados apresentados na Figura 21 confirmam 'doença renal policística' como equivalente de *polycystic kidney disease*. Em outras palavras, nesse caso, *renal* em português corresponde a *kidney* e não a *renal* em inglês.

Outro termo que chama a atenção na Figura 18, é 'insuficiência renal crônica' (linha 1). Será que 'insuficiência' deve ser traduzida como *insufficiency*? Supondo que 'crônica' corresponda a *chronic*, podemos gerar uma concordância para essa palavra, que fornece, entre outros, os seguintes exemplos reproduzidos na Figura 22.

1 ly examinations reveals the presence of chronic renal disease. Timing may be partic
2 a patient with such findings who has chronic renal insufficiency, for example, th
3 , then the most likely causes of chronic renal insufficiency in a patient wit
4 at patients with most forms of acute or chronic renal failure may be oliguric (urine
5 AS: Measurement of renal function in chronic renal disease. Kidney Int 38:167, 19
6 ion rate determination in patients with chronic renal disease. JAMA 199:252, 1967

```
7                Vascular calcification in chronic renal failure.      i
8  desc="head">  Vascular calcification in chronic renal failure. The prevalence
9  tory included type 2 diabetes mellitus, chronic renal failure requiring hemodialysis
10 ith earlier studies showing that stable chronic renal failure patients adjust their
```

Figura 22: Seleção de linhas de concordância para *chronic* no *subcorpus* de Insuficiência Renal no CorTec

Temos, nesse excerto, duas ocorrências de *chronic renal insufficiency* (linhas 2 e 3) e cinco de *chronic renal failure* (linhas 4, 7, 8, 9 e 10). Cabe-nos decidir qual a forma mais usual. Um levantamento de todas as ocorrências apresenta os resultados reportados na Tabela 1.

Tabela 1: Ocorrências de *renal failure* e *renal insufficiency* no *corpus* de Insuficiência Renal no CorTec

Failure	Ocorrências	**Insufficiency**	Ocorrências
Chronic renal failure	189	Chronic renal insufficiency	25
Renal failure	387	Renal insufficiency	77
Total	**576**	**Total**	**102**

Em face disso podemos, com segurança, optar por *chronic renal failure* como a melhor tradução de *insuficiência renal crônica*.

Salientando a utilidade e confiabilidade dos *corpora* comparáveis para o fazer tradutório, Bernardini et al (2013) advogam a incorporação desse tipo de *corpus* e de programas computacionais para sua exploração ao conjunto de ferramentas que os tradutores já usam, no intuito de garantir um trabalho de melhor qualidade.

Os exemplos que apresentamos referem-se todos a linguagens técnicas. Uma vez estabelecidos os equivalentes, esses podem vir a constituir glossários dessas áreas, como foi feito, por exemplo, em Perrotti-Garcia & Rebechi (2007), Teixeira & Tagnin (2008) e Tagnin (2013).[17] A busca por equivalentes é também abordada nos capítulos de Moraes, Pagano et al. e Rebechi, incluídos neste volume.

6.2) *Corpora* paralelos

Corpora paralelos, como já vimos, são compostos por textos originais e suas respectivas traduções. Essas podem ser várias traduções na mesma língua ou em várias

[17] Para trabalhos que abordam o uso de *corpora* no fazer terminológico, ver Tagnin & Bevilacqua (2013).

línguas. Podem até consistir de várias versões intermediárias de um mesmo texto até o texto final, o que permite acompanhar o processo de revisão.

Contamos com dois corpora paralelos que contemplam exclusivamente o par português-inglês: o COMPARA e o CorTrad, do Projeto CoMET.

Apresentamos, a seguir, algumas possibilidades de explorar esse tipo de *corpus*.

6.2.1) Possibilidades de tradução

Talvez o uso mais frequente que se faz de *corpora* paralelos é verificar como determinada palavra ou expressão já foi traduzida.

Vejamos, por exemplo, algumas traduções para o verbo modal *can* no COMPARA.[18]

1. That's another thing that **can** happen to you that's worse than a pain in the knee.	Isso é mais uma coisa que pode acontecer com você que é pior do que uma dor no joelho.
2. The trouble is, as cognitive behaviour therapists point out, it **can** take forever to discover the suppressed traumatic experience, even supposing there was one.	O problema é que, como um terapeuta de comportamento cognitivo faz questão de esclarecer, leva-se muito tempo para descobrir uma experiência traumática suprimida, supondo-se que haja uma.
3. I mean, how trivial **can** you get?	Quero dizer, quer coisa mais banal do que essa?
4. It looks a bit tramp-like, but I **can** hardly bear to have it cut, each strand is so precious.	Dá um ar meio de andarilho, mas não posso suportar a idéia de vê-la cortada, pois cada mecha é preciosa demais.
5. My hands are quite small too, but with long, shapely fingers; like a pianist's, not that I **can** play any keyboard except an IBM one.	Minhas mãos são pequenas, mas com dedos longos, esbeltos como os de um pianista, se bem que o único teclado que eu saiba usar seja o de meu IBM.
6. There's no difference between one first-class stamp and another, and there's a very limited number of ways in which you **can** tear them off the sheets and shove them across the counter.	Não há diferença entre um ou outro selo de primeira classe, e há um número muito limitado de maneiras de destacá-los das folhas e passá-los pela janela do guichê.

Figura 23: Algumas linhas de concordância para *can* no COMPARA.

[18] Todos os exemplos do COMPARA são extraídos da versão 13.1.22.

Como se vê, nem sempre o modal *can* é traduzido pelo verbo *poder* em português. Com exceção das linhas 1 e 4, nas outras ocorrências o verbo é praticamente omitido:

2. *it **can** take forever to discover...* → leva-se muito tempo para descobrir…

3. *how trivial **can** you get?* → quer coisa mais banal do que essa?

5. *not that I **can** play any keyboard* → se bem que o único teclado que eu saiba usar...

6. *there's a very limited number of ways in which you **can** tear them off the sheets...* → há um número muito limitado de maneiras de destacá-los das folhas

Com esse tipo de pesquisa, o tradutor pode enriquecer seu arsenal de possibilidades de tradução e certamente encontrará opções que não lhe teriam ocorrido.

Uma área em que *corpora* paralelos são um recurso inestimável é a da tradução automática. O capítulo de Caseli neste volume discorre sobre o treinamento de sistemas de tradução automática a partir de *corpora* desse tipo.

6.2.2) Tradução de nomes próprios, palavras estrangeiras e termos culturalmente marcados

Outra possibilidade é verificar estratégias empregadas para a tradução de itens por vezes problemáticos, como nomes próprios, palavras estrangeiras e termos culturalmente marcados.

Um exemplo emblemático de palavra culturalmente marcada é a nossa 'cachaça', que recebe as mais variadas traduções. Entre elas, destacamos algumas na Figura 24.

PBAD2(770):	1. Januário tirou a rolha de sabugo com os dentes, deixou a **cachaça** cair quente goela abaixo.	Januario pulled the corn-cob stopper out with his teeth and poured the spirit hot down his throat.
PBJS1(106):	2. -- Se me permite, senhor Holmes, o melhor remédio para esta sensação matutina é uma boa **cachaça**.	«If you'll permit me, Mr. Holmes, the best medicine for this morning-after sensation is a good dose of *cachaça*.»
PBJS1(1322):	3. Watson emborcou outra garrafa de **cachaça** e desandou a rir:	Watson gulped another bottle of rum and burst into laughter.
PBMR1(1622):	4. Um ano depois da tragédia, eu substituía o uísque pelo finalô, imediatamente substituído pelo gim e	A year after the tragedy, I had replaced whiskey with brandy and water, which was immediately

	mesmo pela **cachaça** com limão.	supplanted by gin, and then by sugarcane rum with lime.

Figura 24: Algumas linhas de concordância para 'cachaça' no COMPARA

No primeiro exemplo, *cachaça* é traduzida por *spirit*; no terceiro, por *rum*; e no quarto, por *sugarcane rum.* Apenas no segundo é mantida a forma original. Além disso, observe-se que os exemplos 2 e 3 são da mesma obra (PBJS1[19]), mas a tradução não é padronizada. Temos aqui dois problemas: o emprego de *rum*, que é um produto diferente da nossa 'cachaça',[20] e a falta de padronização (ver REBECHI, 2012 para um estudo detalhado dessas traduções).

Na Figura 25, temos uma ocorrência de expressão cultural (i.e. festa de São João) e uma de nome de local, mais especificamente, de um cemitério no Rio de Janeiro, o São João Batista.

PBMA2(139):	Virgília tinha agora a beleza da velhice, um ar austero e maternal; estava menos magra do que quando a vi, pela última vez, numa festa de **São João**, na Tijuca...	Virgília now had the beauty of age, an austere, maternal look. She was less thin than when I saw here the last time at a Saint John's festival in Tijuca…
PBRF2(685):	Disponível era receber suas ordens, assinar cheques, desligar as luzes, pagar o imposto predial, verificar a fechadura antes de dormir, providenciar o seguro de vida e o jazigo perpétuo no **São João** Batista.	willing to take orders, sign checks, turn the lights off, pay the real estate taxes, check to see the doors are locked before bed, willing to make arrangements for life insurance and the family plot at São João Batista.

Figura 25: Linhas de concordância para 'São João' no COMPARA

Podemos observar que, no primeiro caso, o nome da festa foi simplesmente traduzido por *Saint John's festival*. Embora exista um *Saint John's festival* em outros países, em geral com um denominador comum – a construção de uma fogueira –, a forma de celebrar varia de

[19] PBJS1 significa Português Brasileiro Jô Soares e o número 1 refere-se à primeira obra desse autor no *corpus*. Trata-se do romance *O Xangô de Baker Street*, traduzido por Clifford Landers.

[20] Costa (2006) salienta as altas tarifas alfandegárias a que a cachaça estava sujeita por ser traduzida como *rum* nos documentos de exportação. Desde 2009, "[o] nome "Mapa da Cachaça está protegido por lei [Decreto 6871/2009] e a sua produção está restrita ao Brasil" (http://www.mapadacachaca.com.br/artigos/as-diferencas-entre-rum-e-cachaca/). Com isso, a cachaça também recebeu tarifação própria. Além disso, já está dicionarizada (http://www.merriam-webster.com/dictionary/cachaça), havendo, assim, uma tendência de se manter a denominação "cachaça" nos textos traduzidos.

lugar para lugar e, certamente, nenhuma dessas se assemelha à brasileira. Cabe perguntar, será essa a melhor tradução para *festa de São João*? A tradução de termos culturalmente marcados é sempre um problema e não há um procedimento fixo para tal. A opção dependerá da posição do tradutor. Se quiser manter a especificidade da referência, aproximando o leitor à cultura de partida (no caso, a brasileira), pode manter o termo na língua original e, eventualmente, acrescentar uma pequena explicação, em nota de rodapé ou mesmo no próprio texto, como um aposto. Se o tradutor preferir facilitar a leitura para seu público-alvo, pode traduzir o termo ou mesmo empregar um termo que designe algo similar na cultura de chegada.[21] Nesse último caso, justifica-se a tradução de "festa de São João" por *Saint John's festival*. Já no segundo exemplo, o nome do cemitério foi mantido na língua original, o que denota uma intenção, por parte do tradutor, de 'levar' o leitor para a cultura brasileira.

Uma visão mais ampla sobre a tradução de marcadores culturais é apresentada por Zavaglia et al. (2012), em que os autores analisam as versões alemã, francesa, brasileira e portuguesa de um manual de refrigerador. Discutem, entre vários outros aspectos, como a cortesia é expressa em cada língua-cultura. No alemão e no francês, os manuais se dirigem ao destinatário de maneira mais formal, *Sie* e *vous*, respectivamente; enquanto os manuais brasileiros usam *você*. Em contrapartida, os manuais portugueses evitam dirigir-se ao destinatário por meio de um pronome. Além disso, o alemão, francês e português brasileiro empregam formas de cortesia como *bitte, veuillez*/*s.v.p.*, e *por favor*, ao passo que no português europeu não se encontram expressões similares. Segundo os autores, essa estratégia reflete as expectativas dos destinatários portugueses, que considerariam as formas de cortesia um obstáculo, desviando a atenção do conteúdo instrucional do manual. Dessa forma, caso se fosse traduzir um manual alemão para o público português, essa formas deveriam ser eliminadas, ao passo que, na direção inversa, fórmulas de cortesia teria de ser incluídas.

6.2.3) Comparação de diferentes versões de uma tradução

Para alguns textos, o COMPARA oferece duas traduções, em geral uma portuguesa e uma brasileira. Na Figura 26, temos algumas linhas de concordância das duas traduções de *Therapy*, de David Lodge.[22]

---

[21] Venuti (1995) chama a primeira estratégia de *estrangeirização*, e a segunda de *domesticação.* Schleiermacher já apresentava essa dicotomia em 1813, em sua palestra *Ueber die verschiedenen Methoden des Uebersezens.* O texto original e respectiva tradução para o português, realizada por Celso R. Braida, podem ser encontrados em Heidermann (2010, p. 38-101).

[22] David Lodge. *Therapy*. Londres: Secker & Warburg, pp. 3-97. 1995.

EBDL1T1: *Terapia*. Tradução de Maria do Carmo Figueira. Lisboa: Gradiva, pp. 11-88. 1995.

EBDL1T2: *Terapia*. Tradução de Lídia Cavalcante-Luther. São Paulo: Scipione, pp. 11-115, 1997.

EBDL1T1(900):	«And how's the *Angst* ?»	-- E o *Angst* , como vai?
EBDL1T2(900):	«And how's the *Angst* ?»	-- E como anda a *Angst* ?
EBDL1T1(903):	Don't pretend you don't know what *Angst* is.	Não finjas que não sabes o que é o *Angst* .
EBDL1T2(903):	Don't pretend you don't know what *Angst* is.	não queira dar uma de que não sabe o que é *Angst* .
EBDL1T1(1358):	Reading through that last entry reminded me of Amy's odd question, «How's your *Angst*?» and I looked the word up.	Ao reler a entrada de ontem, lembrei-me da estranha pergunta da Amy -- «como está o teu *Angst*?» -- e fui à procura da palavra.
EBDL1T2(1358):	Reading through that last entry reminded me of Amy's odd question, «How's your *Angst*?» and I looked the word up.	Ao ler aquela última passagem, lembrei-me daquela pergunta esquisita de Amy: «Como vai a sua *Angst* ?» e fui checar a palavra no dicionário.

Figura 26: Excertos de duas traduções de *Therapy*, de David Lodge, no COMPARA.

Nas duas primeiras e nas duas últimas linhas notamos que a tradutora portuguesa, responsável pela obra EBDL1T1, decidiu que *Angst*, uma palavra em alemão que significa 'medo', seria uma palavra masculina, enquanto a tradutora brasileira (vide EBDL1T2) optou por considerá-la feminina. Outra diferença que se nota é que a tradução portuguesa usa a 2a. pessoa do singular ("Não finjas") enquanto a brasileira emprega a 3a. pessoa ("não queira dar").

Uma pesquisa que comparou traduções foi a de Gonçalves (2006) que construiu um *corpus* paralelo com os contos *Dubliners*, de James Joyce, e suas respectivas traduções por Hamilton Trevisan e José Roberto O'Shea para estudar as diferenças de abordagem dos dois tradutores.

Neste volume, contamos com um capítulo que também compara traduções: Magalhães e Blauth analisam alguns aspectos relativos ao estilo de seis tradutores da obra *Heart of Darkness*, de Joseph Conrad. Essa pequena amostra já indica a gama de possibilidades que esse tipo de comparação pode gerar.

6.2.4) Revisões de texto traduzido

O CorTrad apresenta uma ou mais versões da tradução de um texto. Isso permite acompanhar o processo de revisão pelo qual passa a tradução. Na Figura 29, temos alguns exemplos em que houve alterações de uma versão para outra, em geral de vocabulário.

Original	**Primeira versão**	**Segunda versão**	**Versão publicada**
I strike it **home** and regain equilibrium.	Golpeio o ponto de origem e recupero o equilíbrio.	Golpeio o ponto de origem e recupero o equilíbrio.	Golpeio o alvo e recupero o equilíbrio.
My father died a lingering, painful death of cancer -- in a Catholic old men's **home**.	Meu pai teve com uma morte lenta, dolorosa por causa do câncer, em um asilo católico.	Meu pai teve uma morte lenta e dolorosa por causa do câncer, em um asilo católico.	Meu pai teve uma morte lenta e dolorosa por causa do câncer, em um asilo católico.
Bill waited for Joe near the gap, and walked **home** with him.	Bill esperou por Joe, perto da vala, para voltarem juntos para casa.	Bill esperou por Joe, perto da vala, para voltarem juntos para casa.	Bill esperou por Joe perto do desfiladeiro para voltarem juntos.
Bill saw it but said nothing, and, seeing their father coming **home** from work, they hurried on.	Bill observou mas não disse nada, e, vendo o pai voltar do trabalho, ele e o irmão correram.	Bill observou mas não disse nada, e, vendo o pai voltar do trabalho, ele e o irmão correram.	Bill viu tudo, mas não disse nada; ao avistarem o pai voltando do trabalho, correram.

Figura 27: Algumas linhas de concordância para *home* no *subcorpus* literário do CorTrad

Notamos, na Figura 27, *home* empregada com vários sentidos. Na primeira linha, tem o sentido de "alvo", que é recuperado na última revisão. Na segunda, refere-se a "asilo", sentido mantido em todas as versões. Já nas duas últimas, refere-se a "casa", que está presente nas duas primeiras versões da terceira linha, mas fica implícita na última, ou seja, "voltarem juntos para casa" dá lugar a apenas "voltarem juntos". Na quarta linha, a palavra *home* não é traduzida em nenhuma das versões.

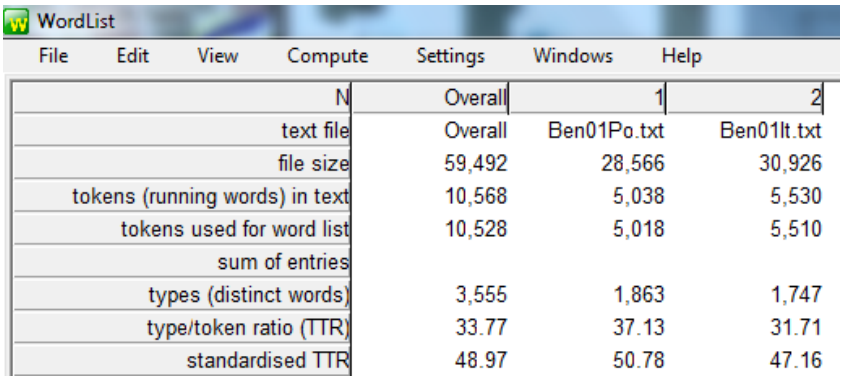
Apesar de a busca ter sido feita para *home*, a observação das várias versões revela outras mudanças. Na segunda linha, por exemplo, podemos ver que *a* "*lingering, painful death*" foi primeiramente traduzida por "uma morte lenta, dolorosa" e já na primeira revisão alterado para "uma morte lenta e dolorosa", com o acréscimo do conectivo "e", produzindo, assim, uma forma mais natural de se expressar em português.

6.3) *Corpora* personalizados

Quando podemos manipular os *corpora* inteiros, como no caso de construirmos nossos próprios corpora, temos acesso, com o *AntConc* ou o *WST*, por exemplo, a uma tabela com vários dados estatísticos, a partir dos quais podemos tirar algumas conclusões.

6.3.1) Diferença de extensão entre texto original e texto traduzido

Uma das afirmações correntes sobre traduções é que essas costumam ser mais longas do que seus originais. Por exemplo, em um estudo de *corpus* acerca de originais e traduções de obras literárias canônicas e populares, Viana et alii (2008) mostram que as traduções são maiores do que os respectivos textos originais. Contudo, apesar de o número de palavras crescer nos textos literários traduzidos, o número de palavras distintas é maior em português não obstante o fato de a obra ser original ou traduzida. A Figura 28 mostra os resultados para o primeiro capítulo do romance *Benjamin*, de Chico Buarque, no original em português (Ben01Po.txt) e em sua versão traduzida para o inglês (Ben01It.txt).


WordList

File Edit View Compute Settings Windows Help

N	Overall	1	2
text file	Overall	Ben01Po.txt	Ben01It.txt
file size	59,492	28,566	30,926
tokens (running words) in text	10,568	5,038	5,530
tokens used for word list	10,528	5,018	5,510
sum of entries			
types (distinct words)	3,555	1,863	1,747
type/token ratio (TTR)	33.77	37.13	31.71
standardised TTR	48.97	50.78	47.16

Figura 28: Estatísticas da ferramenta *WordList* do *WST*

Os números parecem confirmar que as traduções costumam ser mais longas do que os originais: há 5.038 palavras (*tokens (running words) in text*) em português em comparação a 5.528 na tradução em inglês.

6.3.2) Densidade lexical

Outra informação que pode ser extraída da Figura 28 é a densidade lexical, ou seja, a riqueza de vocabulário das duas versões. Quanto maior o índice, maior a diversidade lexical. A tabela apresenta 1.863 palavras distintas (*types (distinct words)*) para o português e 1.745 para o inglês, o que corresponde a um índice item/ocorrência (*type/token ratio*) de 37,13% para o português e 31,69% para o inglês, corroborando os achados de Viana et alii (2008) de

que textos em português têm uma diversidade vocabular maior do que aqueles em inglês. Porém, tratando-se de textos de tamanhos diferentes, deve-se utilizar o índice padronizado (*standardized TTR*), que dá 50,78% para o português e 47,16% para o inglês, confirmando a maior diversidade lexical do português.

7) Conclusão

O objetivo deste capítulo foi apresentar um panorama das possibilidades oferecidas pelo uso de *corpora* nos Estudos da Tradução, sempre com farta exemplificação. Começamos com uma definição de *corpus* no âmbito da Linguística de *Corpus* e passamos a descrever alguns dos tipos de *corpora* que mais se prestam à pesquisa e ao fazer tradutório. Além dos *corpora* disponíveis *on-line*, descrevemos os passos para a compilação de um *corpus* que se adeque aos objetivos de pesquisa do investigador. Posteriormente, discutimos alguns tipos de anotação que enriquecem os *corpora*. Na seção seguinte, abordamos as principais ferramentas para a pesquisa com *corpora*: lista de palavras, lista de palavras-chave e concordanciador. Por fim, discutimos alguns tipos de pesquisa possíveis com *corpora*, com ilustração de trabalhos já realizados.

Obviamente, o capítulo não pretendeu cobrir o assunto de forma exaustiva, pois muito dependerá do tópico que o pesquisador pretende estudar e de sua criatividade na exploração de *corpora*. Uma área que não foi aqui abordada foi a da compilação e exploração de *corpora* de aprendizes de tradução,[23] o que requereria um capítulo por si só. Outra área foi a da Tradução Audio-Visual, em que o uso de *corpora* ainda é bastante incipiente. Enfim, este capítulo pretendeu estimular o leitor a enveredar por esse caminho ainda pouco explorado e descobrir o quanto essa metodologia pode revelar – coisas que jamais poderiam ser descobertas de outra forma.

8) Referências bibliográficas

ALUÍSIO, S. et alii. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In: ARCHER, D.; RAYSON, P.; WILSON, A.; MCENERY, T. (Eds.). *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster: UCREL, 2003, p. 14-21. Disponível em: <http://ucrel.lancs.ac.uk/publications/CL2003/papers/aluisio.pdf>. Acesso em: 16 abril 2014.

[23] Vide Oliveira (2012) sobre o emprego de um editor de planilhas eletrônicas para o alinhamento de textos em um *corpus* de aprendizes de tradução.

ANTHONY, L. Advancing AntConc: Design and Performance Improvements for Multi-Language. *Japan Association for English Corpus Studies (JAECS) Annual Conference*, Sept. 29, 2012, Osaka University, Osaka, Japan, 2012.

_____ AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. In *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning* pp. 7-13, 2004.

BERNARDINI, S. & FERRARESI, A. Old needs, new solutions. Comparable corpora for language professionals. In SHAROFF, S.; RAPP, R.; ZWEIGENBAUM, P.; FUNG, P. (eds.) *Building and Using Comparable Corpora*, Springer, 2013.

BIBER, D. Representativeness in Corpus Design. *Literary and Linguistic Computing*, Vol. 8, No. 4, 1993, 243-257, 1993

BOWKER, L. & PEARSON, J. 2002. Term extraction. In *Working with Specialized Language – A practical guide to using corpora*. London & New York: Routledge, 165-176.

CARVALHO, L. *A tradução de binômios nos contratos de* common law *à luz da linguística de corpus*. Dissertação de mestrado. FFLCH-USP, 2007.

COSTA, A.T.P. *Brasil mostrando a sua cara: estratégias de tradução no material de divulgação cultural – um estudo baseado em corpus.* Dissertação (Mestrado em Linguística Aplicada) – Departamento de Línguas Estrangeiras e Tradução da Universidade de Brasília, Brasília, 2006.

GONÇALVES, L.B. *Dubliners sob a lupa da lingüística de corpus: uma contribuição para a análise e a avaliação da tradução literária*. Tese de doutorado. USP – FFLCH, 2006.

HALLIDAY, M.A.K. "Categories of the theory of grammar". *Word*, Vol. 17, No. 3, pp. 241-292, 1961.

HEIDERMANN, Werner (ed.). *Clássicos da Teria da Tradução.* Vol. 1: Alemão/Português. 2a. ed. revisada e ampliada. Florianópolis: UFSC, 2010.

LEECH, G. "The State of the Art in Corpus Linguistics". In: *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Ed. by AIJMER, K. & ALTENBERG, B., London & New York: Longman. 8–29, 1991.

MCENERY, T., XIAO, R. & TONO, Y. *Corpus-based language studies*. London & New York: Routledge, 2006.

NAVARRO, S. L. M. *Glossário bilíngue de colocações de hotelaria: um modelo à luz da Linguística de Corpus.* Dissertação de mestrado. USP - FFLCH, 2012.

OLIVEIRA, J. (2012). A linguística de corpus na formação de tradutores: compilação e análise de um corpus de aprendizes de tradução. Trabalhos em Andamento – *Anais do Xi Encontro de Linguística de Corpus (ELC 2012).* São Carlos - SP. Disponível em: http://nilc.icmc.sc.usp.br/elc-ebralc2012/anais/andamento/102956.pdf

OLOHAN, M. *Introducing corpora in translation studies*. Oxford: Routledge, 2004.

PERROTTI-GARCIA, A. J. & REBECHI, R.R. *Vocabulário para Química – inglês/português – português/inglês.* Série Mil & Um Termos. São Paulo: SBS Editora, 2007.

PHILIP, G. 2009. Arriving at equivalence: Making a case for comparable general reference *corpora* in translation studies. In: BEEBY, A., INÉS, P. R. & SÁNCHEZ-GIJÓN, P. (eds.*), Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate.* Amsterdam: John Benjamins, 59-73. Disponível em http://amsacta.cib.unibo.it/2124/1/ComparableCorpora.pdf

REBECHI, R.R. 'Cachaça' na tradução de obras literárias brasileiras para a língua inglesa. TradTerm, São Paulo, v. 20, dezembro/2012, p. 95-110.

SANTOS, D., TAGNIN, S.E.O. & TEIXEIRA, E.D. CorTrad and Portuguese-English translation studies: investigating colours. *VARIENG*, vol 12, 2012. Disponível em: http://www.helsinki.fi/varieng/series/volumes/12/santos_tagnin_teixeira/

SCOTT, M. *WordSmith Tools*. Oxford: Oxford University Press. 1996.

TAGNIN, S.E.O. A Identificação de equivalentes tradutórios em *corpora* comparáveis. In *Anais do I Congresso Internacional da ABRAPU*I: Belo Horizonte, 3 a 6 de junho de 2007. Disponível em: http://www.fflch.usp.br/dlm/comet/Novo/Stella_Abrapui%202007_artigo.pdf

TAGNIN, S.E.O. (org.). *Vocabulário para Fotografia – inglês/português.* Série Mil & Um Termos. São Paulo: SBS Editora, 2013.

TAGNIN, S. & BEVILACQUA, C. (org.). *Corpora na Terminologia*. São Paulo: HUB Editorial, 2013.

TAGNIN, S. E. O. & TEIXEIRA, e. D. British vs. American English, Brazilian vs. European Portuguese: how close or how far apart? - a corpus-driven study. In *Lodz Studies in Language 9*, Frankfurt am Main. 2004, p. 193-208.

TEIXEIRA, E.D. *A lingüística de corpus a serviço do tradutor: proposta de um dicionário de culinária voltado para a produção textual*. 2008. Tese (Doutorado em Estudos Linguísticos e Literários em Inglês), Faculdade de Filosofia, Letras e Ci~encias Humanas, Universidade de São Paulo, São Paulo

TEIXEIRA, E.D. & TAGNIN, S.E.O. *Vocabulário para Culinária – inglês/português*. Série Mil & Um Termos. São Paulo: SBS Editora, 2008.

VENUTI, L. *The translator's invisibility: a history of translation*. London: Routledge, 1995.

VIANA, V. *Corpora* para consulta *on*-line e *off*-line: um guia para o tradutor/pesquisador. In VIANA, V. e TAGNIN, S.E.O. (org.) Corpora *na tradução*. São Paulo: HUB Editorial.

VIANA, V.; ZYNGIER, S.; FAUSTO, F.; SILVEIRA, N. Original e tradução de obras canônicas e populares em perspectiva: A análise de corpus como um recurso à crítica literária. In: ZYNGIER, S.; VIANA, V.; SILVEIRA, N. (Eds.). *Ver & visualizar: Letras sob o prisma empírico*. Rio de Janeiro: Publit, 2008, p. 61-74.

ZANETTIN, F. 2012. *Translation-driven corpora: Corpus resources for descriptive and applied translation studies*. Manchester: St. Jerome, 2012.

ZAVAGLIA, A.; AZENHA, JR., J.; REICHMANN, T. Cultural markers in LSP Translation. In: BAUMANN, K.-D. (Ed.). *Fach Translat Kultur. Interdisziplinäre Aspekte der vernetzten Vielfalt*. Berlim: Frank & Timme, 2012. p. 785-808.