UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE LETRAS MODERNAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS E LITERÁRIOS EM INGLÊS

MARLENE DEZIDÉRIO ANDREETTO

Por que textos de divulgação são mais difíceis para aprendizes de leitura com necessidades específicas do que textos científicos? - Um estudo direcionado pelo *corpus.*

São Paulo
2013

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE LETRAS MODERNAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS E LITERÁRIOS EM INGLÊS

Por que textos de divulgação são mais difíceis para aprendizes de leitura com necessidades específicas do que textos científicos? - Um estudo direcionado pelo *corpus.*

Marlene Dezidério Andreetto

Dissertação apresentada ao Programa de Pós-Graduação em Estudos Linguísticos e Literários em Inglês do Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, para a obtenção do título de Mestre em Letras.

De acordo:
Orientadora: Profa. Dra. Stella Esther Ortweiler Tagnin

Versão corrigida

São Paulo
2013

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catalogação na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

A558q

Andreetto, Marlene
Por que os textos de divulgação são mais difíceis para os aprendizes com necessidades específicas do que os textos científicos? Um estudo direcionado pelo corpus. / Marlene Andreetto ; orientadora Stella Tagnin. - São Paulo, 2013.
156 f.

Dissertação (Mestrado)- Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. Departamento de Letras Modernas. Área de concentração: Estudos Linguísticos e Literários em Inglês.

1. Linguística de Corpus. 2. Ensino de língua estrangeira. 3. Língua Inglesa. I. Tagnin, Stella, orient. II. Título.

FOLHA DE APROVAÇÃO

Marlene Dezidério Andreetto

Por que textos de divulgação são mais difíceis para aprendizes de leitura do que textos científicos? - Um estudo direcionado pelo *corpus.*

Dissertação submetida ao Programa de Pós-Graduação da Faculdade de Filosofia, Letras, e Ciências Humanas, Departamento de Letras Modernas, Área de Estudos Linguísticas e Literários em inglês, da Universidade de São Paulo, como requisito parcial à obtenção do grau de Mestre em Estudos Linguísticos e Literários em Inglês

Aprovada em:____/____/____

Banca Examinadora

Prof. Dr.:_____

Instituição:_____ Assinatura:_____

Prof. Dr.:_____

Instituição:_____ Assinatura:_____

Prof. Dr.:_____

Instituição:_____ Assinatura:_____

Dedico esta pesquisa à memória de meu pai que sempre considerou a educação um aspecto essencial na vida do cidadão, principalmente, para as mulheres.

AGRADECIMENTOS

À minha orientadora, Stella Tagnin, pela paciência e orientação.

Aos professores da Especialização e das disciplinas cursadas: Eduardo de Pietri e Marília Ferreira.

Especial agradecimento a Carmen Dayrell e Marilia Ferreira pelos valiosos comentários feitos no exame de qualificação.

Aos meus pais, Italo (*in memoriam*) e Olívia que sempre me apoiaram.

A meu filho Paulo, pelos comentários “ácidos”.

Agradeço, também Jonas de Muzio Junior pela ajuda com as tabelas e gráficos.

A minhas amigas, Rosangela e Marlene que sempre torceram por mim.

A Rozane Rebechi, pela ajuda impagável com a revisão, troca de ideias e consultoria.

Aos amigos do Comet: Danilo Murakami, Joacyr Oliveira, Nilson Barros e Sabrina Matuda.

A Sandra Elias pelos últimos retoques, não menos importantes.

Por último, a todos os meus amigos do forró que me ajudaram a relaxar. Em especial ao Zé de Loura, por seu ritmo gostoso, e senhor Cortez pela doação dos livros de sua editora, que foram de extrema importância.

SUMÁRIO

ÍNDICE DAS FIGURAS X
ÍNDICE DOS GRÁFICOS XII
ÍNDICE DAS TABELAS XIII
LISTA DE ABREVIATURAS XIV
RESUMO XV
ABSTRACT XVI
INTRODUÇÃO 1
CAPÍTULO 1. FUNDAMENTAÇÃO TEÓRICA 7
1.1 Gênero e análise 7
1.1.2 Sobre a diferença entre o discurso acadêmico e o de divulgação 9
1.1.3 Os gêneros e inglês para fins específicos 12
1.2 A teoria do esquema (*schema theory*) e do conhecimento prévio (*background knowledge)* 13
1.3 A noção de equivalência na aquisição do léxico 15
1.4 Linguística de *Corpus* 18
1.4.2 Conceito de gramática para a LC 19
1.4.3 Ferramentas computacionais de análise linguística 20
1.4.3.1 *AntConc 2.3.3w* 20
1.4.3.2 *Compleat Lexical Tutor:* 21
1.4.3.3 *WordSmith Tools* 24
1.4.4 Os utilitários das ferramentas: 24
1.4.4.1 As listas de palavras (*wordlists*) 24
1.4.4.2 As listas de palavras-chave *(Keyword List)* 25
1.4.4.3 Concordanciador *(Concordance)* 27
1.4.4.4 *Clusters* (combinações de palavras) 27
1.4.5 A ferramenta BootCat para compilação automática de corpus 28
1.4.6 A complexidade da linguagem falada 29
CAPÍTULO 2. METODOLOGIA 36
2.1 Ensaios com o *Lextutor* 36
2.2 O *corpus* de estudo e suas dificuldades 38
2.2.1 Critérios iniciais de compilação 40
2.2.1.1 Composição inicial do *corpus* 41

2.2.1.2 Balanceamento inicial do corpus....42
2.2.2 Critérios redefinidos....43
2.2.2.1 O *NEJM* e a escolha do *Health News*....43
2.2.2.2 A compilação do *WebMD*....44
2.2.3 A compilação manual....45
2.2.4 A compilação por meio do BootCat....45
2.2.5.1 O sub*corpus NEJM*....49
2.2.5.2 O sub*corpus Health News*....50
2.2.5.3 O sub*corpus WebMD*....50
2.3 A comparação das listas e as palavras-chave obtidas....51
2.3.1 A comparação das listas de palavras do NEJM e do Health News....54
2.3.1.1 Os verbos....55
2.3.1.2 Os substantivos....56
2.3.1.3 Os adjetivos....57
2.3.2 Comparação do *WebMD* e *NEJM*....57
2.3.2. Os verbos....59
2.3.2.2 Os substantivos....59
2.3.2.3 Os adjetivos....60
2.3.3 Comparação das listas do Health News e WebMD.....60
2.3.3.1 Os verbos....60
2.3.3.2 Os adjetivos....60
2.3.3.3 Os substantivos....60
2.4.1 *Stop list*....62
2.4.2 Linhas de concordância....63
2.4.3 *Clusters* (agrupamentos de palavras)....64
Capítulo 3. Resultados....67
3.1 Colocados de *was*....68
3.1.1 *Was + V -ed*....69
3.1.2 *Was + Adv-ly + V-ed*....70
3.1.4 *Was + V-ing*....73
3.2 Colocados de *were*....78
3.2.1 *Were + V-ed*....80
3.2.2 *Were + Adv –ly + V-ed*....80

3.2.3 *Were* + $V_{[irreg]}$ 81
3.2.4 *Were* + V_{-ing} 82
3.2.5 *Clusters* de *were* no *WebMD* 85
3.2.5.1 Comportamento do item lexical *likely* 89
3.2.5.2 *Be likely* em sequências de comparação 91
3.3 Os verbos de elocução 96
3.4.1 A investigação dos adjetivos cognatos 105
3.4.2 Os colocados dos substantivos cognatos 112
CONCLUSÃO 122
CONSIDERAÇÕES FINAIS 124
REFERÊNCIAS BIBLIOGRÁFICAS 127
ANEXOS 136

ÍNDICE DAS FIGURAS

Figura 1: Página do *WebMD* com um artigo sobre obesidade. 10
Figura 2: Página inicial do *The New England Journal of Medicine.* 11
Figura 3: Tela do *AntConc* com a palavra *cancer* como palavra de busca. 21
Figura 4: Tela do *Vocabprofile (Lextutor)* 23
Figura 5: As 30 primeiras palavras por ordem de frequência no *corpus* acadêmico *(NEJM* 25
Figura 6: Seleção das linhas de concordância tendo como palavra de busca *"used".* 34
Figura 7: Tela do *WSTools* demonstrando a comparação de duas listas (*NEJM x WebMD)* do *corpus* de estudo inicial ainda incompleto. 40
Figura 8:Tela da página do *site WebMD* 44
Figura 9: Palavras-chave obtidas a partir de 10 textos do *WebMD*.....47
Figura 10: Domínio escolhido (canto superior direito) e, abaixo, as *URLs* obtidas a partir das sementes.....48
Figura 11: As *URLs* obtidas para verificação e possível remoção. 49
Figura 12: As últimas palavras-chave do *NEJM* e o início das palavras negativas em comparação ao *HN.* 53
Figura 13: Linhas de concordância com *hospital-wide.* 64
Figura 14: Tela do *AntConc* mostrando os *clusters* obtidos a partir do *subcorpus NEJM* com o verbo *was.* 65
Figura 15: Continuação da Fig. 14 demonstrando as escolhas feitas. 66
Figura 16: Linhas de concordância a partir da busca "*was *ly *ed"* no *HN*.....71
Figura 17: Comparação das ocorrências de *was* + verbos irregulares no *NEJM* e *HN.* 71
Figura 18: Comparação do *NEJM* com *WebMD* 73
Figura 19: Comparação do *HN* com *WebMD* 73
Figura 20: Comparação do *HN* com *NEJM* 74
Figura 21: Linhas de concordância demonstrando as ocorrências de *was *ing* no *NEJM* 74
Figura 22: Linhas de concordância de *was + -ing* no *HN* 75
Figura 23: Linhas de concordância a partir da busca *was + *ing* no *WebMD* 78
Figura 24: Linhas de concordância a partir da busca *were *ly *ed* no *NEJM* 81
Figura 25: Linhas de concordância de *were *ing* no *NEJM* 83
Figura 26: Linhas de concordância demonstrando o padrão *were *ing* no *HN* 84
Figura 27: Linhas de concordância de *were * ing* no *WebMD*.....85
Figura 28: Linhas de concordância de *'as likely to'* no *HN.* 93
Figura 29: Linhas de concordância de *'as likely as'* no *HN.* 94
Figura 30: Linhas de concordância com *it * likely that* no *NEJM.* 95
Figura 31: Linhas de concordância com *it * likely that* no *HN.*.....95
Figura 32: Linhas de concordância com *It * likely that* no *WebMD.*.....95

Figura 33: Linhas de concordância com *it * unlikely that* no *NEJM.* 95
Figura 34: Linhas de concordância com *it * unlikely that* no *HN.*.....95
Figura 35: Linhas de concordância com *it * unlikely that* no *WebMD.*.....96
Figura 36: Seleção de linhas de concordância do verbo *says* no *WebMD*.....99
Figura 37: Seleção de linhas de concordância de *say* no *WebMD* 100
Figura 38: Seleção de linhas de concordância de *said* no *HN.* 102
Figura 39: Linhas de concordância demonstrando *said* seguido da próxima oração, sem o pronome *that* no *HN.* 102
Figura 40: Primeiras 35 linhas de concordância com o verbo *tells* no *WebMD*.....103
Figura 41: Seleção das linhas de concordância com a palavra de busca *adverse events.*.....107
Figura 42: Colocados de *adverse* no *NEJM.* 108
Figura 43: *Clusters* de *adverse events* no *NEJM.* 109
Figura 44: Seleção de linhas de concordância de *cardiac arrest* no *NEJM.* 110
Figura 45: *Clusters* de *virologic* no *NEJM*. 111
Figura 46: Seleção de colocados de *mortality* no *NEJM* 114
Figura 47: Linhas de concordância de *mortality* como núcleo em uma colocação substantiva no *NEJM.* 115
Figura 48: Linhas de concordância com *hospital-wide mortality rates* no *NEJM.*....116
Figura 49: Linhas de concordância *the efficacy of* no *NEJM.*.....117
Figura 50: Linhas de concordância de *efficacy* em posição anterior ao núcleo. 117
Figura 51: *Clusters* de *interval* no *NEJM* 119
Figura 52: Linhas de concordância de *confidence interval* no *NEJM* 119
Figura 53: Linhas de concordância de *cardiac* no *WebMD* 120
Figura 54: Seleção de linhas de concordância com *heart attack* no *WebMD.* 121

ÍNDICE DOS GRÁFICOS

Gráfico 1: Número de ocorrências dos padrões mais frequentes para *was* nos 3 sub*corpora.* 69
Gráfico 2: Distribuição de padrões de *were* nos três sub*corpora.* 79
Gráfico 3: Comparativo do número de ocorrências totais de *likely to/ that/than* nos três sub*corpora* 90
Gráfico 4: Distribuição dos usos comparativos de '*likely to'* nos três sub*corpora.* 92
Gráfico 5: Distribuição dos padrões comparativos de *likely than/as.* 94
Gráfico 6: Ocorrências de *say, says, said* e *tells* nos três sub*corpora.* 98
Gráfico 7: Ocorrências de *adverse, cardiac e virologic* nos três sub*corpora.* 106
Gráfico 8: Ocorrências de *mortality, efficacy, interval* e *arrest* nos três sub*corpora*. 113

ÍNDICE DAS TABELAS

Tabela 1: Porcentagem das palavras diferentes em um texto.....22
Tabela 2: Equivalência dos testes estatísticos (percentil, nível de porcentagem, valor de probabilidade e valor crítico) apresentada no manual da ferramenta.....26
Tabela 3: Porcentagens obtidas a partir da inserção dos diferentes gêneros de textos no utilitário *Vocabprofile (Lextutor)*.....37
Tabela 4: Resumo da composição inicial do *corpus.* 42
Tabela 5: Composição atual do *corpus* de estudo 51
Tabela 6: Lista de palavras-chave do sub*corpus NEJM* em comparação ao sub*corpus HN*.....56
Tabela 7: Listas de palavras-chave do *WebMD* em comparação ao *NEJM.*.....59
Tabela 8: Lista comparativa de palavras-chave do *HN* e *WebMD.* 62
Tabela 9: Número de ocorrências dos diferentes padrões identificados para *was* nos três sub*corpora.* 68
Tabela 10: Padrões de *were* em cada um dos três sub*corpora.* 79
Tabela 11: Número de ocorrências do padrão *were* + verbos irregulares.....82
Tabela 12: Comparação dos *clusters* de *was* nos três sub*corpora* 87
Tabela 13: Comparação dos *clusters* de *were* nos três sub*corpora*.....88
Tabela 14: Distribuição das ocorrências de *clusters* com *likely* nos três sub*corpora* 90
Tabela 15: Número de ocorrências dos usos de *likely to* em comparações.....91
Tabela 16: Número de ocorrências de *likely than/as* usado comparativamente nos três sub*corpora.* 93
Tabela 17: Número de ocorrências dos verbos *say, says, said* e *tells* nos três sub*corpora.* 97
Tabela 18: Número de ocorrências dos adjetivos-chave nos três sub*corpora.* 106
Tabela 19: Substantivos cognatos de origem latina por ordem de ocorrências a partir do *NEJM*.....112
Tabela 20: Exemplos de colocações substantivas em inglês e suas traduções em português. 115

LISTA DE ABREVIATURAS

LC	Linguística de *Corpus*
L1	Língua materna
L2	Língua estrangeira
NEJM	*New England Journal of Medicine*
LL	*Log likelihood*
HN	*Health News*
WST	*WordSmith Tools*
URL	*Uniform resource locator*
ESP	*English for specific purposes*

RESUMO

Com a crescente demanda por certificações como pré-requisito para o acesso aos cursos de pós-graduação para profissionais de diferentes áreas, a habilidade de leitura tornou-se um campo de grandes possibilidades para a pesquisa de ensino-aprendizagem para fins específicos. O objetivo desta pesquisa é fazer um levantamento das estruturas léxico-gramaticais que possam impedir a fluência da leitura dos diferentes gêneros relacionados à área médica. A partir daí, oferecer ao professor maiores subsídios na preparação e na organização do conteúdo programático em cursos de leitura instrumental, além de possibilitar a elaboração de atividades que possam auxiliá-lo nessa tarefa. Para tanto, foi compilado um *corpus* de estudo composto por três sub*corpora*: um acadêmico (*NEJM*) e dois de divulgação na área de saúde (*Health News e WebMD*), a partir de textos disponíveis na internet. A compilação do corpus foi feita de forma manual e automática, esta com a ajuda da ferramenta *BootCat*. A análise do *corpus* foi feita de forma semiautomática por meio do *AntConc 3.3 w* 2012. A metodologia baseada em *corpus* permitiu que o levantamento de alguns padrões léxico-gramaticais recorrentes fosse feito de uma forma adequada e eficiente, além de ter demonstrado ser possível a elaboração de um *corpus* individualizado, não muito laborioso, que possa atender às necessidades dos aprendizes de leitura com necessidades específicas. Diferenças significativas foram encontradas na comparação dos três sub*corpora*. Confirmando estudos anteriores, como, por exemplo, o de Biber (1998), a forte presença da voz passiva nos textos acadêmicos demonstrou ser uma característica importante para a organização de um conteúdo programático, bem como a forte presença dos verbos regulares. A ordem dos adjetivos nesse tipo de gênero foi outro tópico levantado. Nos textos de divulgação, a presença de características da linguagem falada, como a omissão do *that* nas orações substantivas e a sequência *to be likely,* apresentou um forte indício da dificuldade encontrada pelos aprendizes brasileiros de inglês na leitura desse tipo de gênero textual.

PALAVRAS-CHAVE: Linguística de *Corpus*. Leitura instrumental. Conhecimento prévio. Gêneros Textuais.

ABSTRACT

Given the growing number of certification in the academic area in English, reading has become of utmost importance for professionals in different areas. Brazilian professionals in the area of health are required to get a proficiency certificate to be accepted in the post graduate programme of an important university in São Paulo. Having that in mind, the aim of this study is to analyse the lexico-grammatical structures and items that can hinder reading comprehension of texts in the medical area. Results may aid teachers in getting more subsidies for the preparation and organisation of their reading syllabi in a course for specific purposes. To be able to accomplish that task we have compiled a corpus which is composed of three subcorpora: one with research articles from the *New England Journal of Medicine* online version and the two other ones with articles also published on the internet (*Health News* and *WebMD*). The compilation was both manual and automatic using the *Bootcat.* The corpus analysis was initially performed with the help of Wordsmith Tools 5.0 and continued with the AntConC 3.3.w 2012.

The use of Corpus Linguistics has allowed the researcher to observe more effectively some of the lexico-grammatical patterns which were typical and recurrent in the analised genres. Besides, this study has demonstrated that it is possible to build a specialized corpus without much effort so as to help teachers and students alike.

Significant differences were found as a result of the comparison of the three subcorpora. This study has confirmed the massive presence of the passive voice in academic texts (Biber, 1998). In the information genre, it has revealed its proximity to the spoken variety, which might be the cause for the difficulty encountered by most Brazilian students when reading this kind of articles.

KEYWORDS: Reading for specific purposes. Background knowledge. Genre. Corpus linguistics.

INTRODUÇÃO

"O acaso só favorece a mente preparada" (Louis Pasteur)

Na segunda metade do século XX, com o aumento do comércio entre os países, a necessidade de comunicação cresceu sobremaneira. E com isso, o ensino de idiomas, assim como as pesquisas sobre a aquisição de uma segunda língua, cresceu em igual proporção.

Como consequência do aumento dessa demanda da comunicação, surgiram novos métodos e diferentes abordagens, principalmente para o ensino de inglês como língua estrangeira.

Na década de 1970, o ensino da gramática caiu em desuso, pois, acreditava-se que o método de gramática e tradução não se mostrava eficaz para a comunicação verbal. Os exemplos utilizados eram extraídos dos livros clássicos do latim e do grego e não eram adequados à interação entre pessoas em atividades cotidianas. Seu objetivo principal era a leitura de textos literários (Richards e Rodgers,1995).

A partir desse momento, o método comunicativo, que enfatizava a comunicação oral, se fortaleceu. Existia uma tentativa de fazer uma ligação entre a vida real e a sala de aula. Nesse ambiente, o ensino de gramática perdeu seu lugar de destaque. O importante era se comunicar, ter fluência. A gramática deveria ser ensinada de uma maneira mais contextualizada, porém sem a menção de qualquer nomenclatura gramatical. Com isso, a acurácia deixou de ser priorizada. (Granger, 2010)

No final do século XX, outras formas de comunicação e troca entre os países deram lugar a novas metodologias e abordagens de ensino. Profissionais de diferentes especialidades, com diferentes necessidades, fizeram surgir o ensino de

línguas para fins específicos (*ESP)* [1]*.* Na área médica, o intercâmbio de experiências científicas torna essa necessidade premente.

Ao longo de mais de duas décadas de atuação em sala de aula como professora de inglês como língua estrangeira, pude experimentar as diversas práticas de atuação, bem como observar, ao longo desse período, as necessidades específicas demonstradas por aprendizes com perfis distintos.

Atualmente, trabalhamos com profissionais da área médica que necessitam de uma certificação de proficiência de leitura em língua inglesa para validar suas dissertações e teses. O teste para essa certificação é aplicado e elaborado por uma instituição brasileira de ensino de inglês. Esse teste é composto por quatro textos com diferentes níveis de dificuldade e com cerca de 500 palavras cada um. O tempo de duração do exame é de três horas e a avaliação é feita por meio de exercícios de compreensão (‘verdadeiro’ ou ‘falso’), exercícios de completar e de relacionar, dentre outros[2].

Esses profissionais da área médica dispõem de muito pouco tempo para se dedicar ao aprendizado de uma segunda língua, pois suas atividades continuam em paralelo. Portanto, o tempo dispendido para o aprendizado da leitura deve ser otimizado. Para tanto, procuramos escolher materiais que sejam relevantes e que os ajudem a ampliar essa habilidade no mínimo de tempo possível.

Além do tempo, outro aspecto de grande importância a ser levado em consideração é o fato de que esses candidatos, embora não sejam considerados iniciantes de fato no aprendizado da língua inglesa, não têm muitas horas de estudo. Já tiveram contato com a língua inglesa durante o período escolar e durante a graduação na escola de medicina por meio da literatura especializada. Porém, quando necessitam de um entendimento mais preciso, as dificuldades começam a aparecer.

[1] Em inglês: *English for Specific Purposes,* que em alguns casos é referido como “Inglês Instrumental”.
[2] No Anexo A, encontra-se um exemplar de um tipo de texto presente no exame, bem como os exercícios que o acompanham.

Trabalhamos com dois livros existentes no mercado, dedicados à leitura instrumental: *Leitura Instrumental em Língua Inglesa* (Souza et al, 2010) e *Preparatório TEAP* (Lanzoni, 2010). Este último recomenda a utilização de material de suporte gramatical em língua inglesa em paralelo, além do uso de dicionário.

Apesar de esses materiais serem bem elaborados e fundamentados em pesquisas na área de leitura instrumental, o próprio autor (Lanzoni, 2010) afirma que o livro não consegue contemplar todos os aspectos a serem abordados em um curso de preparação e atender às necessidades de cada grupo e/ou aprendiz.

Durante as atividades de leitura em sala de aula, percebemos que os alunos consideravam mais difícil a leitura de textos de divulgação de notícias da área de saúde do que ler textos da área médica de sua especialidade. Portanto decidimos investigar a razão dessa dificuldade. Seriam as estruturas das sentenças, os tempos verbais ou o vocabulário a causa do problema?

Partindo da visão de que a sala de aula é um excelente campo de observação, contanto que o professor se proponha a ser um pesquisador (Van Lier,1998), decidimos consultar os alunos por meio da leitura de textos de diferentes gêneros que nos fornecessem subsídios para a preparação de materiais e/ou atividades que pudessem otimizar as horas de estudo dos alunos.

Fizeram parte dessa consulta onze alunos classificados por meio de testes que compõem o material *New American Inside Out* (2009) para ensino de inglês geral. Esse material é constituído de estruturas graduadas, desde as consideradas mais fáceis até as mais elaboradas. De acordo com esses testes, os alunos são classificados como nível pré-intermediário e início de intermediário[3], em termos de conhecimentos e habilidades linguísticas. Alguns deles já tinham feito o teste de proficiência de leitura acima mencionado. Segundo depoimento desses alunos, os textos contidos nesses testes eram mais difíceis do que os artigos que leem em suas áreas de especialidade. Os alunos consultados foram: um psiquiatra, uma enfermeira graduada, um bioquímico, uma bióloga, um cirurgião ortopedista, um

[3] Cerca de 200 horas de estudo, nível de classificação (A2/B1) presente nessa série de livros didáticos que obedece aos critérios estabelecidos pelo Common European Framework (CFE), dividido em cinco categorias: A1, A2, B1, B2, C1, C2. Partindo do mais elementar para o mais avançado nível de inglês.

aluno de medicina, uma fisioterapeuta, um urologista, duas hematologistas e um engenheiro. Decidimos acrescentar o engenheiro, em primeiro lugar para que ele não se sentisse discriminado em relação aos outros, e em segundo porque acreditamos que obter o depoimento de um participante de uma área distinta poderia acrescentar informações importantes para nossa pesquisa.

Distribuímos, em diferentes ocasiões do curso, três tipos de textos para cada um, a saber: textos extraídos do periódico *New England Journal of Medicine,* doravante *NEJM*, e artigos extraídos do *site* da *BBC News Health*. O terceiro grupo de textos foi de exames anteriores preparados pela instituição de ensino que aplica os testes de proficiência em leitura.

Os textos do *NEJM* foram distribuídos de acordo com a área de especialidade de cada um. O psiquiatra recebeu um texto de psiquiatria; a enfermeira, um texto de pediatria, por ser sua área de atuação. Para o cirurgião ortopedista, um texto de cirurgia ortopédica e assim por diante.

O segundo tipo de texto (*BBC News Health*) distribuído para todos os participantes foi extraído do *site* de notícias da *BBC News* sobre diferentes áreas da saúde. Esses textos têm como objetivo informar o leitor sobre diferentes aspectos da área da saúde, como cuidar do diabetes, informações sobre transplantes e outros assuntos relacionados.

Os textos foram distribuídos em diferentes oportunidades e deveriam ser lidos como uma atividade extra, fora da sala de aula. Foi dito aos alunos para não se preocuparem com os resultados, pois esse não era o objetivo da atividade, que era obter informações em relação aos aspectos linguísticos apresentados em cada texto. Explicamos que a tarefa fazia parte de um estudo. Também lhes foi pedido que fizessem comentários de acordo com sua percepção quanto ao grau de dificuldade dos textos em relação à complexidade da estrutura, nível do vocabulário, organização do texto e como esses aspectos influenciavam o entendimento. Para que se pudesse ter uma noção do tipo de vocabulário que apresentava problemas, foi pedido que eles sublinhassem as palavras desconhecidas. As instruções que receberam foram: a) sublinhar palavras que oferecessem qualquer tipo de dificuldade e relatar o porquê; b) relatar a impressão geral que tiveram do texto,

tentando explicar o porquê e c) ler o último texto, dando uma opinião comparativa sobre os três diferentes tipos de textos lidos.

De acordo com os relatos (orais e escritos), os 11 alunos entenderam melhor o texto da sua área. A razão disso, segundo eles, era o fato de que os textos acadêmicos apresentam um vocabulário e uma estrutura gramatical repetitivas, que facilitavam a compreensão após várias leituras. Também relataram que o conhecimento sobre o assunto, tornava a atividade mais fácil. Em segundo lugar, vieram os textos da *BBC News,* que foram considerados mais difíceis do que os primeiros. As palavras sublinhadas foram adjetivos que expressam opinião, sentimentos e alguns substantivos que não são muito comuns, além de estruturas características da linguagem falada. O terceiro tipo de texto, de testes anteriores, foi considerado o mais difícil por conter um vocabulário mais complexo e “pegadinhas”.

Por isso, tendo como objetivo principal auxiliar alunos e professores a enfocar o que de fato é relevante ao aprendizado naquele momento, partindo das necessidades reais dos alunos, pretendemos demonstrar o levantamento feito de algumas das características léxico-gramaticais presentes nos textos acadêmicos e nos de divulgação que possam causar dificuldade de entendimento ao aluno.

A fim de cumprir esse objetivo, pretendemos responder às seguintes perguntas:

1) Que características linguísticas do artigo acadêmico facilitam a sua leitura por parte de alunos brasileiros aprendizes de inglês como língua estrangeira em comparação à leitura de artigos de divulgação científica?

2) De que maneira a Linguística de *Corpus* pode ajudar o professor na tarefa de identificação dos itens que podem causar estranhamento ao aluno?

Quanto à divisão, este trabalho estruturou-se da seguinte forma:

O “Capítulo 1” abordará algumas teorias que estão envolvidas na aquisição de leitura de uma língua estrangeira. Tentamos relacionar: a) a teoria do esquema; b) um modelo explicativo do processo mental envolvido na aquisição do léxico de uma segunda língua; c) a complexidade da gramática da oralidade e, por último: c)

como a Linguística de *Corpus* pode ajudar no levantamento dos padrões léxico-gramaticais e suas ferramentas de análise.

O “Capítulo 2” apresentará a explicação do processo de escolha e compilação do *corpus* de estudo, isto é, dos três sub*corpora,* bem como a metodologia que possibilitou o levantamento dos dados para a análise.

O “Capitulo 3” apresentará as análises quantitativas e qualitativas a partir da comparação dos três sub*corpora:* o primeiro composto por textos acadêmicos e dois outros por textos de divulgação, todos publicados na internet.

Em “Considerações Finais”, retomam-se, em linhas gerais, os objetivos da pesquisa, sua relevância e seu trajeto, salientando os resutados obtidos e suas possíveis aplicações.

O trabalho se completa com os anexos:

A) Amostra de um dos textos que consta do exame de proficiência a que os alunos se submetem e seus exercícios.
B) Linhas de concordância de *likely* nos sub*corpora HN* e no *WebMD.*
C) Linhas de concordância de *mortality* no *NEJM.*
D) Linhas de concordância com *cardiac arrest(s).*

CAPÍTULO 1. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, apresentamos os pressupostos teóricos que embasaram a presente pesquisa. Em primeiro lugar, a noção da teoria dos gêneros aplicada à leitura de uma língua estrangeira para fins específicos. Em seguida, discutiremos as implicações do conhecimento prévio *(background knowledge)* dentro da teoria do esquema *(schema theory).* Em terceiro lugar, uma explicação para o processo de aquisição e a construção de novos conceitos. Por último, a Linguística de *Corpus* e suas ferramentas de análise, que constituem a metodologia utilizada nesta pesquisa e a complexidade da linguagem falada.

1.1 Gêneros textuais e análise

Nas últimas décadas, a análise de gêneros textuais tornou-se uma das linhas de trabalho mais utilizadas para o estudo da escrita especializada em diferentes áreas do conhecimento. “A análise de gêneros é o estudo do comportamento linguístico” nos diferentes atos comunicativos (Bhatia, 2009). No caso deste trabalho, os artigos acadêmicos e artigos de divulgação científica são exemplos de atos comunicativos com propósitos diferentes. São diferentes gêneros textuais, ainda que possam tratar do mesmo assunto (artigos da área médica). Alguns autores, entre eles Bazerman (1988), Motta-Roth (2002) e Hyland (2004; 2009) consideram os gêneros ações retóricas de que os autores se valem para responder às situações que notadamente se repetem; os usuários consideram certas escolhas linguísticas como maneiras eficientes de representar o fazer das coisas em contextos familiares.

Os leitores de uma carta, por exemplo, reconhecem esse tipo de texto como tal, por sua estrutura de apresentação e escolhas linguísticas. O mesmo acontece com o reconhecimento de uma receita culinária. Conseguimos identificá-la por termos adquirido o conhecimento convencionado de que é essa a estrutura de

apresentação desse gênero, ou seja, uma fórmula que se repete. O conhecimento convencionado, conforme Bhatia (2009), se divide em três aspectos convencionais inter-relacionados: a) a recorrência de situações retóricas, b) os propósitos comunicativos compartilhados e c) as regularidades de organização esturutural. A análise de gêneros, portanto, é baseada na premissa de que as características de um grupo semelhante de textos dependem do contexto social de sua criação e uso. Suas características podem ser descritas de tal forma que relacionam um texto a seus similares (Bazerman, 1988; Motta-Roth, 2002).

A análise de gêneros não só salienta a natureza social da produção e da leitura, como também a linguagem presente nas realidades sociais, já que é por meio do uso recorrente de formas convencionais que indivíduos desenvolvem relações, estabelecem comunidades e "fazem as coisas acontecerem" (Hyland, 2009).

Os analistas de gênero (Swales, 1990; Bahtia, 2009 e outros) tentam descrever os textos dentro de um contexto social e textual contradizendo tendências de tratá-los isoladamente. "Em Linguística, a análise dos gêneros é essencialmente um exercício de classificação de 'atos tipificados de comunicação' baseados em sua forma e propósito (Hyland, 2009: 211)" [4].

Portando, cada comunidade tem sua forma de se comunicar. No contexto acadêmico, a análise de gêneros enfatiza "a importância da linguagem na produção de conhecimento por meio da interação entre pares, oferece uma perpectiva do texto como uma unidade socialmente e historicamente construída" (Bakhtin, 1986: 76-7, *apud* Motta-Roth, 2002).

Apesar de os artigos acadêmicos também terem o objetivo de informar, a organização dessa informação segue padrões diferentes daqueles na área da divulgação, que embora também tenham como objetivo transmitir conhecimento, o fazem respeitando a convencionalidade esperada por seu público alvo, que normalmente é leigo em uma dada área de especialidade. Por exemplo, um artigo

[4] No original: *In linguistics genre analysis is essentially an exercise in the classification of 'typified acts of communication' based on form and purpose.* (Todas as traduções deste trabalho são de autoria da pesquisadora).

acadêmico da área médica tem diferentes características de um artigo de divulgação da mesma área médica, a começar pelo número de palavras, que é mais reduzido, passando pela sintaxe e as escolhas lexicais. Segundo Bakhtin: “Cada um dos gêneros do discurso, em cada uma das áreas de comunicação verbal, tem sua concepção padrão de destinatário que o determina como gênero.” (1992: 321)

Muitas vezes esses textos são reescritos pelos próprios pesquisadores. Essa é uma prática que vem se tornando comum com a facilidade de publicações *online.*

Como resultado, o discurso científico sofrerá uma adaptação e uma transformação quando tiver que ser apresentado em outra esfera social para que haja uma compreensão do público em geral. Conforme Zamboni (2001: 83), vale dizer que:

> A atividade que transforma um discurso fonte em discurso-alvo, direcionado para um público específico no qual as adaptações devem estar a serviço da “compreensibilidade”. No caso do DC [discurso científico], a reformulação é uma atividade por meio da qual se transforma o discurso da ciência em um discurso do “cotidiano”.

1.1.2 Sobre a diferença entre o discurso científico e o de divulgação

Como observado intuitivamente pelos aprendizes consultados no início deste trabalho, o texto científico apresenta uma estrutura rígida que comporta partes claramente delimitadas:

> [segue] um padrão que poderíamos resumir no esquema: circunscrição do problema/ material e métodos/resultados/ discussão e conclusão. Esse esquema se subverte completamente ao passar para a divulgação, que privilegia, de modo quase unânime, os resultados, relegando a metodologia – item caro ao trabalho científico – a um plano inferior, quando não a suprime totalmente. (Zamboni, 2001: 17)

Ainda segundo a autora, também em outros níveis ocorrem escolhas. Nos níveis sintático, semântico e lexical, por exemplo, as nominalizações, as denotações e a nomenclatura especializada – ocorrências características do discurso científico – fazem surgir em seu lugar construções verbais, conotações e o léxico da linguagem cotidiana, como será demonstrado em nossas análises.

No contexto midiático, os artigos publicados *online* possuem características deteminadas pelo meio em que são publicados. Os de divulgação por exemplo, têm como objetivo transmitir informação de maneira que possa ser entendida pelo público em geral. A página do *site* bem como os artigos possuem elementos visuais chamativos, como fotos (cf. **Figura 1**), propagandas, etc. diferentemente dos artigos acadêmicos, que também apresentam gráficos, tabelas e em alguns casos fotos, porém de modo mais tradicional de apresentação e formatação de artigos, semelhante às suas publicações escritas (cf. **Figura 2**).

Demonstramos a seguir as páginas dos *sites WebMD* e *NEJM* com o intuito de ilustrar as diferentes contextualizações dos artigos apresentados. O artigo sobre obesidade no WebMD aparece no meio de uma série de outras informações e com vários *links,* enfim, mais poluído visualmente (cf. **Figura 1**):

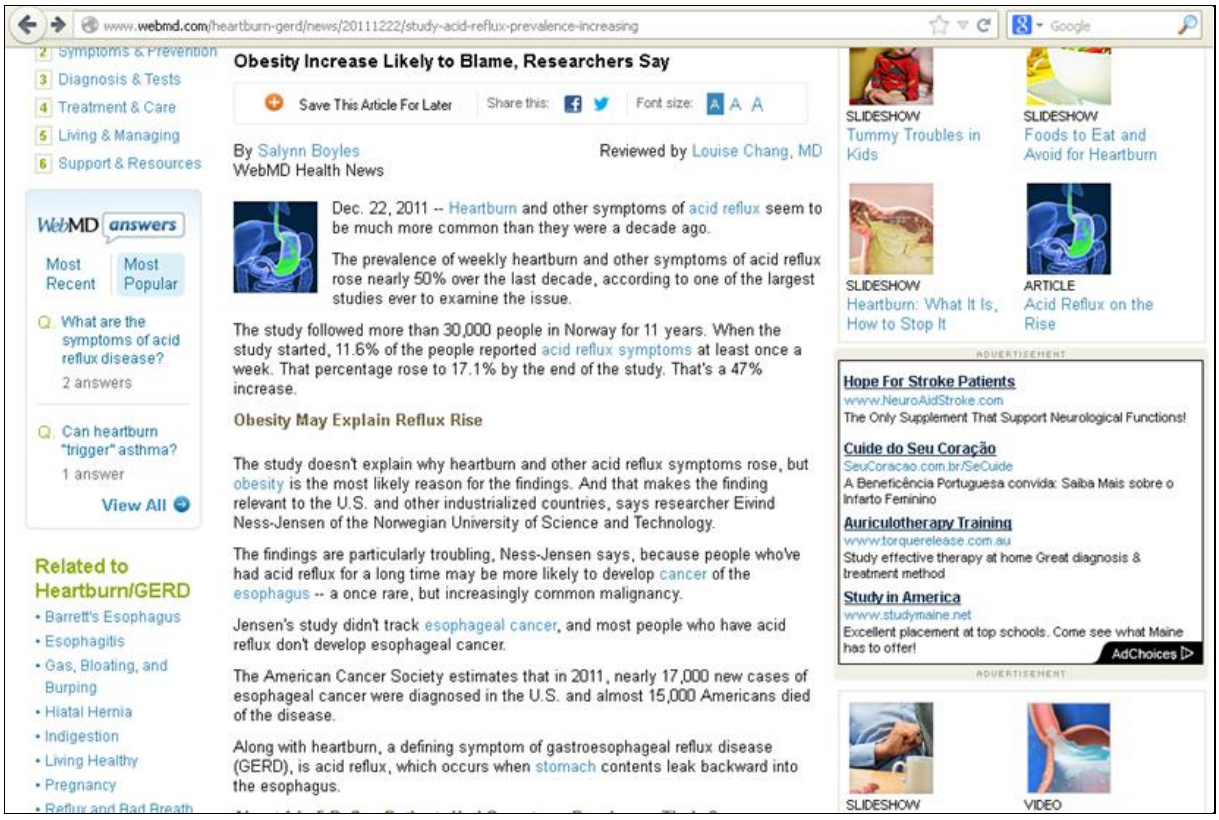


Figura 1: Página do *WebMD* com um artigo sobre obesidade.

Ao contrário, a página do *NEJM*, tem uma apresentação mais limpa visualmente e apresenta em suas abas a sequência de um artigo acadêmico: abstract, article, references, citing articles e letters (cf. **Figura 2**):

Figura 2: Página inicial do *The New England Journal of Medicine.*

Não só visualmente, como observado, mas em termos esturuturais, os artigos de divulgação publicados na internet para o público leigo em ciência sofrem uma transformação e deixam de ter as características iniciais de um texto científico adquirindo suas próprias regras e estilo. Isso acontece para que seu conteúdo seja compreendido pelo público-alvo:

> [Diversas áreas do conhecimento reconhecem que] a atividade de divulgação científica como uma prática de reformulação textual-discursiva, em cujo conjunto se insere a tradução, o resumo, a resenha, a paráfrase (em seu sentido amplo) bem como certas práticas pedagógicas de adaptar um determinado conteúdo para um determinado nível de audiência, de formular determinadas análises para um determinado grupo social, de reescrever determinadas mensagens publicitárias em função de certo público alvo, além de outras. (Zamboni, 2001: 50).

Segundo Pinheiro (2002), apesar de os gêneros textuais serem identificados por características próprias, eles não são estáticos, sofrem uma evolução da mesma forma que o mundo e, visto que, vão incorporando novos elementos. A autora chama de hibridismo o que resulta do confronto de dois sistemas de gêneros e que vai impingindo mudanças aos mesmos através dos tempos.

Bakhtin (1997:106) já argumentava que: "[...] o gênero é e não é o mesmo, sempre novo e velho ao mesmo tempo".

A partir desses conceitos de transformação e adaptação e de reescrita dos gêneros é que desenvolvemos esta pesquisa, tendo como objetivo identificar as características específicas de cada um dos gêneros - acadêmico e de divulgação – na área de saúde que causam a falta de fluidez na leitura de textos. Como mencinado, os elementos linguísticos que fazem parte dessa transformação são nosso objeto de estudo.

1.1.3 Os gêneros textuais e inglês para fins específicos

A abordagem de ensino de línguas para fins específi*cos (English for Specific Purposes, ESP*), segundo Hyland (2009), considera gênero, uma classe de eventos comunicativos empregados por uma comunidade de um discurso específico. Mais do que propriedade de uma cultura mais ampla, os gêneros pertencem às comunidades que os utilizam. A partir dessa premissa, os analistas investigam as práticas específicas daqueles grupos. Swales (1998: 20) observa que:

> [...] as comunidades discursivas desenvolvem suas próprias convenções e tradições para diferentes atividades verbais, tais como: conduzir reuniões, escrever relatórios e publicar suas atividades. Esses tipos de eventos comunicativos recorrentes são os gêneros que orquestram a vida verbal. [5]

Com base nesses preceitos, procuramos buscar neste trabalho as características que fazem os textos acadêmicos diferentes dos textos de divulgação e suas implicações para o ensino da leitura com o intuito de inserir esse aprendiz em suas comunidades.

[5] No original: *"discourse communities evolve their own conventions and traditions for such diverse verbal activities as running meetings, producing reports, and publishing their activities. These recurrent classes of communicative events are the genres that orchestrate verbal life."*

1.2 A teoria do esquema (*schema theory*) e do conhecimento prévio (*background knowledge)*

O trabalho de Keneth Goodman (1970, *apud* Brown, 1994), entre outros, deu início à pesquisa de questões relacionadas à pedagogia de leitura em língua estrangeira. O título do trabalho - "Leitura: um jogo de adivinhações psicolinguísticas" [6] - capta a essência do processo *bottom-up* (do particular para o geral) de decodificar o significado da página impressa. Os leitores devem reconhecer uma multiplicidade de signos linguísticos (letras, morfemas, sílabas, palavras, frases, elementos gramaticais, e marcadores de discurso), além de utilizar "mecanismos de processamento de dados linguísticos" para imprimir determinada ordem a esses signos. Esse processo direcionado por dados (*data-driven processing*) requer um sofisticado conhecimento da língua. O leitor deve fazer uma seleção a partir de toda essa informação para construir o significado do que está lendo. Para Goodman, ler envolve riscos – um jogo de adivinhações, pois o leitor precisa resolver um "quebra-cabeças", inferir significados e decidir o que reter, o que ignorar. A forma como os leitores constróem os significados consiste em um dilema para os professores de leitura instrumental em língua estrangeira. Existe sempre um questionamento na hora da decisão de o quê incluir, por onde começar e em que se deter para que o aprendiz possa se tornar um leitor proficiente. Deve-se começar por estratégias, pelo léxico, pelas estruturas? O processo é longo e difícil. Precisar a ordem dessa construção é objeto de vários estudos não só da linguística, mas também da psicologia e da sociologia.

Com o intuito de otimizar o aprendizado de estratégias de leitura, a teoria do esquema argumenta que um texto não carrega o significado em si mesmo. O leitor traz informação, conhecimento, emoção, experiência e cultura à palavra impressa (Clarke e Silberstein, 1977:136 -137, *apud* Brown, 1994: 284):

> As pesquisas mostram que a leitura é incidentalmente visual. O leitor contribui com mais informação do que a página impressa, isto é, o leitor entende o que lê, e é capaz de perceber o estímulo que vai além da representação gráfica e faz relações com grupos de conceitos armazenados em sua memória. A habilidade em leitura

[6] Título original: *Reading: A psycholinguistic guessing game.*

> depende de uma interação eficiente entre o conhecimento linguístico e o conhecimento de mundo.[7]

A teoria do esquema enfatiza que esse outro lado do processo de leitura, direcionado por conceito (*conceptually driven*), ou processo *top-down* (do geral para o particular), traz à tona o conhecimento prévio no momento de decidir o significado de uma palavra.

Nosso objetivo aqui não é nos aprofundarmos em estratégias de leitura (*skimming* e/ou *scanning*), mas sim entender as diferenças estruturais entre os diferentes gêneros, com o intuito de organizar a apresentação dessas características. Por essa razão, partimos do conceito de conhecimento prévio proposto por Clarke e Siberstein (1997) para essa identificação.

Nessa tentativa de entender como poderíamos auxiliar nosso aluno e a partir dos dados estatísticos obtidos em um ensaio com o *Lextutor*, descrito com maiores detalhes no capítulo 3, destinado à metodologia, pudemos notar que existe uma maior porcentagem de palavras de origem latina nos textos acadêmicos. Assume-se que, por serem de origem latina, essas palavras serão mais facilmente entendidas pelos aprendizes brasileiros por serem cognatas do português.

A partir do conhecimento linguístico prévio, isto é, dos cognatos, que inferimos sejam do conhecimento do leitor iniciante, passamos a pesquisar quais outros elementos peculiares nos gêneros a serem estudados poderiam ser a causa da dificuldade encontrada pelos aprendizes.

Acreditamos que se levarmos em conta a conceituação da teoria do esquema, juntamente com as noções de equivalência propostas por Pavlenko (2009) poderemos levantar os dados para saber “o quê” ensinar.

[7] No original: *Research has shown that reading is only incidentally virtual. More information is contributed by the reader than by the print on the page. That is, readers understand what they read because they are able to take the stimulus beyond its graphic representation and assign it membership to an appropriate group of concepts already depends on the efficient interaction between linguistic knowledge and knowledge of the world.*

1.3 A noção de equivalência na aquisição do léxico

Existem vários estudos sobre o conceito de equivalência na tradução. Os autores não chegaram a um consenso. Quine (1960), Holmes (1973) e Gentzler (1993), por exemplo, consideram que analisar uma tradução utilizando-se de listas de equivalência e não-equivalência é uma proposta muito vaga e limitante. Já Sager (1994:124) argumenta que "o fundamental a todas as teorias de tradução é o conceito de equivalência [...]" (*apud* Partington, 1998).

Da mesma forma, na área de ensino de línguas estrangeiras, a consciência dos conceitos de equivalência, equivalência parcial ou nenhuma equivalência são de crucial importância, pois existem situações em que o professor precisa se valer desses conceitos para dar explicações sobre as estruturas ou sobre o léxico que o aprendiz não conhece, e com isso, auxiliá-lo na construção de sentidos.

Não só na área de tradução, mas também na área do ensino de línguas estrangeiras, o entendimento de como a aquisição de novos conceitos em uma segunda língua se processa na mente do aprendiz tem sido objeto de estudo de vários teóricos. Kroll e Stewart (1994) e De Groot (1992) propõem modelos de explicação para esse processo. Porém, adotaremos o Modelo Hierárquico Modificado[8], doravante *MHM*, proposto por Pavlenko (2009). A autora faz questão de afirmar que não desprestigia nenhum dos modelos anteriores. Ao contrário, todos contribuíram para que o *MHM* fosse desenvolvido. Ainda segundo a autora, o *MHM* traz uma camada mais profunda para o nosso entendimento do léxico mental bilíngue, postulando que as diferentes relações de equivalências conceituais têm diferentes implicações no aprendizado de vocabulário da L2. Segundo Pavlenko (2009), existem três tipos de processos mentais que explicam o aprendizado do léxico em uma segunda língua: equivalência total, equivalência parcial e falta de equivalência total.

O primeiro caso, equivalência total, é quando acontece uma simples transferência lexical de uma palavra da L1 para L2, em que os conceitos coincidem, exemplo mesa em português e *table* em inglês. A equivalência conceitual facilita o aprendizado do vocabulário da L2 pela transferência positiva; a principal tarefa

[8] No original: *The Modified Hierarchical Model*

nesse contexto é o estabelecimento das conexões entre palavras da L2 e conceitos já existentes. Como no caso de 'mortalidade' em português e *mortality* em inglês.

O segundo tipo é o da equivalência parcial, que ao mesmo tempo pode facilitar o aprendizado através de uma sobreposição (transferência positiva) também complicar, no caso de o aprendiz supor uma completa equivalência e haver uma transferência negativa. Podemos citar como exemplos os falsos cognatos e as palavras abstratas como, por exemplo, *privacy* e *frustration* em inglês. Essas últimas, apesar de serem palavras, aparentemente, equivalentes por terem formas parecidas nas duas línguas (português e inglês), tem conceituação diferente, pois culturalmente, o sentido dessas palavras é diferente. Nesses casos, a principal tarefa no aprendizado da L2 é a reestruturação conceitual (*conceptual restructuring*), ou seja, um processo em que "o aprendiz necessita fazer um ajuste na estrutura da categoria e seus limites de acordo com as restrições impostas pela categoria linguística da língua alvo (Pavlenko ,2009: 141)".[9]

Em terceiro lugar, a autora apresenta o conceito da falta de equivalência total ou não-equivalência, ao mesmo tempo em que complica o aprendizado à medida que os aprendizes precisam desenvolver novas categorias, pode também facilitar, por meio da ausência de representações que competem. Esse é o caso de estruturas não existentes na L2. Citamos como exemplo de uma estrutura problemática para aprendizes brasileiros alguns usos do *present perfect* em inglês, pela sua total falta de equivalência.

No caso da falta de equivalência total, a autora (Pavlenko, 2009: 141) sugere o desenvolvimento conceitual (*conceptual development*), ou seja , um processo de desenvolvimento de novas representações "multimodais que permitam que os falantes mapeiem novas palavras aos referentes do mundo real similares aos falantes nativos da língua-alvo".[10] Podemos citar, como exemplo, a falta de equivalência total, o caso da voz passiva em inglês em comparação ao português, em que temos dois casos de voz passiva: a analítica e a sintética. A voz passiva

[9] No original: […] *conceptual restructuring, that is adjustment of the category structure and boundaries in accordance with the constraints of the target linguistic category.*
[10] No original: […] *conceptual devolpment, that is development of new multimodal representations that allow speakers to map new words onto real-world referents similar to native speakers of the target language.*

analítica normalmente não causa problemas por ter a mesma estrutura nas duas línguas: verbo auxiliar (ser/estar) mais particípio passado do verbo principal. A falta da passiva sintética em inglês, que é construída com o pronome apassivador 'se' em português, causa estranhamento aos aprendizes brasileiros, pois ao sentirem necessidade do uso dessa construção em inglês, não a encontram em seu repertório lexical, causando um estranhamento.

Embora, muitas vezes, os aprendizes não tenham consciência do uso que fazem da passiva sintética em sua própria língua materna, não entendem quando devem produzir e/ ou traduzir essa equivalência em inglês, como por exemplo:

"Acredita-se que ele tenha ido embora." em português e "*It is believed he has gone*." em inglês.

Além da falta de passiva sintética nesse exemplo em português, o tempo verbal utilizado na oração substantiva é o subjuntivo que é inexistente em inglês, o que não é um tópico facilmente entendido pelos aprendizes. Ainda segundo Pavlenko (2009), essa tarefa de desenvolvimento de um conceito pode ser mais fácil no caso de novos objetos e mais desafiante nos casos de abstração.

Nesta pesquisa, podemos citar como exemplo de uma equivalência e transferência positiva os cognatos de origem latina, e como equivalência parcial com uma transferência negativa, o falso cognato *confidence.* Um exemplo de total falta de equivalência conceitual e estrutural é a expressão *to be likely,* que por essa razão causa dificuldade ao aprendiz brasileiro. Discutiremos a presença de outras construções que não apresentam equivalência entre o português e o inglês e que causam estranhamento ao aprendiz brasileiro.

Da mesma forma que os conceitos de equivalência, equivalência parcial e não-equivalência podem auxiliar o professor na explicação das diferenças entre as línguas, a convencionalidade é outro recurso nessa tarefa, pois são conceitos intimamente ligados. A convencionalidade abrange tudo que é utilizado por uma comunidade. "As convenções linguísticas são os "jeitos" aceitos pela comunidade que fala determinada língua" Tagnin (2005: 17- 20).

A convencionalidade é classificada por Tagnin (2005:17) em três níveis: sintático, semântico e pragmático. “O nível sintático compreende a combinabilidade dos elementos, sua ordem e sua gramaticalidade.” No semântico “observa-se a convencionalidade na relação não motivada entre a expressão e seu significado”. Como exemplo, a autora questiona o significado das expressões *kick the bucket* e *bater as botas*. O terceiro nível, o pragmático “abrange o uso da língua em situações de interação dos falantes”.

Assim, a Linguística de *Corpus* e suas ferramentas auxiliam o pesquisador na observação das convencionalidades da língua por meio de exemplos extraídos de amostras da linguagem em uso.

1.4. Linguística de *Corpus*

Denomina-se Linguística de *Corpus* (LC) a disciplina que se ocupa da investigação de unidades convencionais da língua. Embora ainda não haja um consenso para a classificação da Linguística de *Corpus*, alguns a definem como teoria, outros como metodologia e ainda outros como abordagem, decidimos por abordagem, em um primeiro momento, já que a Linguística de *Corpus* apresenta uma nova forma de abordar a língua a partir de uma grande quantidade de textos em formato eletrônico (*corpus*). *Corpus*, de acordo com Tagnin (2005: 21) é:

> [...] uma coletânea de textos, necessariamente em formato eletrônico, compilados e organizados segundo critérios ditados pelo objetivo de pesquisa a que se destina. O formato eletrônico permite que esses textos sejam investigados e analisados automaticamente, com o uso de ferramentas computacionais específicas.

Outros autores como McEnery e Wilson (1997) definem a Linguística de *Corpus* como o “estudo de uma língua baseado em exemplos autênticos extraídos do uso” [11]. Em um segundo momento, consideramos a LC uma metodologia, já que utilizaremos ferramentas de análise para a observação de dados. A observação dos dados por meio de *corpora* revela evidências empíricas e sistemáticas da linguagem que poderiam passar despercebidas de outra forma (Cobb, 1997). Outro aspecto importante é que os dados são apresentados com base em testes estatísticos, o que

[11] No original: *(...) the study of language based on examples of ‘real life’ language use.*

dá legitimidade aos resultados obtidos e confere aos estudos linguísticos um caráter científico.

Pesquisas sobre os padrões de uso da linguagem não são uma atividade recente. Segundo Sardinha (2004), datam da Antiguidade, quando os textos eram analisados manualmente. A grande vantagem que o advento do computador trouxe a esse tipo de análise foi a de colocar a matemática a favor da linguística. Tanto no que se refere a cálculos complexos quanto aos mais simples aplicados na observação de grande quantidade de textos em poucos milésimos de segundo. “A afinidade entre a Linguística de *Corpus* e a Estatística é indiscutível.” (Vasilévski, 2007).

Pela importância dada à frequência das palavras e à probabilidade e por trabalhar com dados, existe um conceito errôneo de que a LC é uma abordagem não-qualitativa, porém as noções de análise quantitativa e qualitativa são extremamente importantes para esse tipo de estudo (Halliday, 1991).

A análise linguística por meio de grandes *corpora* revela um nível de sistematização da linguagem que seria difícil somente pela intuição (O’ Keeffe et al., 2007).

1.4.1 Conceito de gramática para a LC

Para o ensino de línguas estrangeiras, principalmente para *ESP*, a questão da possibilidade e da probabilidade são aspectos importantes. Segundo a visão de Halliday (1991), existem muitas combinações gramaticalmente possíveis mas não utilizadas com frequência.

Algumas regras gramaticais são determinísticas, isto é, são regras que sempre se aplicam (Carter; McCarthy, 2006). Por exemplo, o adjetivo em inglês, na maioria dos casos aparece antes do substantivo, ou a terceira pessoa do indicativo presente dos verbos plenos sempre termina em –s (*she works*, não *she work).*

Outras regras são probabilísticas, o que significa que elas estabelecem o que é mais provável e menos provável de se aplicar em certas circunstâncias específicas. Por exemplo, na grande maioria dos casos, um pronome relativo *(who/*

which, that) deve ser usado para se referir ao sujeito de uma oração adjetiva. Contudo, em estilos informais falados, o pronome relativo pode ser omitido. Não é uma regra que o pronome relativo tenha de ser omitido; em inglês, é possível que seja omitido. Portanto, as regras que regem os usos são probabilísticas, isto é, indicam o que é mais provável na maioria dos casos, não que sejam obrigatórias. É dessa forma que as regras gramaticais ou padrões léxico-gramaticais são considerados pela LC, já que são baseadas nas observações do que é mais ou menos provável nos diferentes contextos reais.

1.4.2 Ferramentas computacionais de análise linguística

Para proceder à análise de um *corpus*, tendo como base os princípios da LC, pode-se fazer uso de uma ou mais ferramentas de exploração automática de informações. Para a presente pesquisa, utilizamos três ferramentas de análise, a saber: *AntConc* 2.3.3w, *Compleat Lexical Tutor* e *Wordsmith Tools*.

1.4.2.1 *AntConc 2.3.3w*[12]

O *AntConc* 2.3.3.w (2012) é uma ferramenta leve para análise de *corpus*, pois não ocupa muito espaço no computador, além disso, é gratuita, o que facilita o acesso de alunos quanto de professores. Foi desenvolvida por Anthony Laurence especialmente para ser usada em sala de aula. Possui, entre outros, um concordanciador, que é um utilitário normalmente encontrado nos programas de análise linguística e é usada para que possamos ver o contexto e o co-texto das palavras que estamos observando. O co-texto aparece na forma de linhas de concordâncias. As concordâncias são partes de frases, sentenças ou parágrafos nas quais a palavra desejada (chamada nódulo ou palavra de busca) aparece centralizada . A **Figura 3** apresenta uma tela do *AntConc* com a palavra de busca

[12] Disponível em: www.antlab.sci.waseda.ac.jp. Último acesso em 20.01.2013.

cancer . Se observarmos o co-texto ao lado esquerdo, notamos duas colocações: *prostate cancer* (linhas: 9, 10, 12, 13 e de 15 a 26) e *breast cancer* (linhas: 29 a 35).

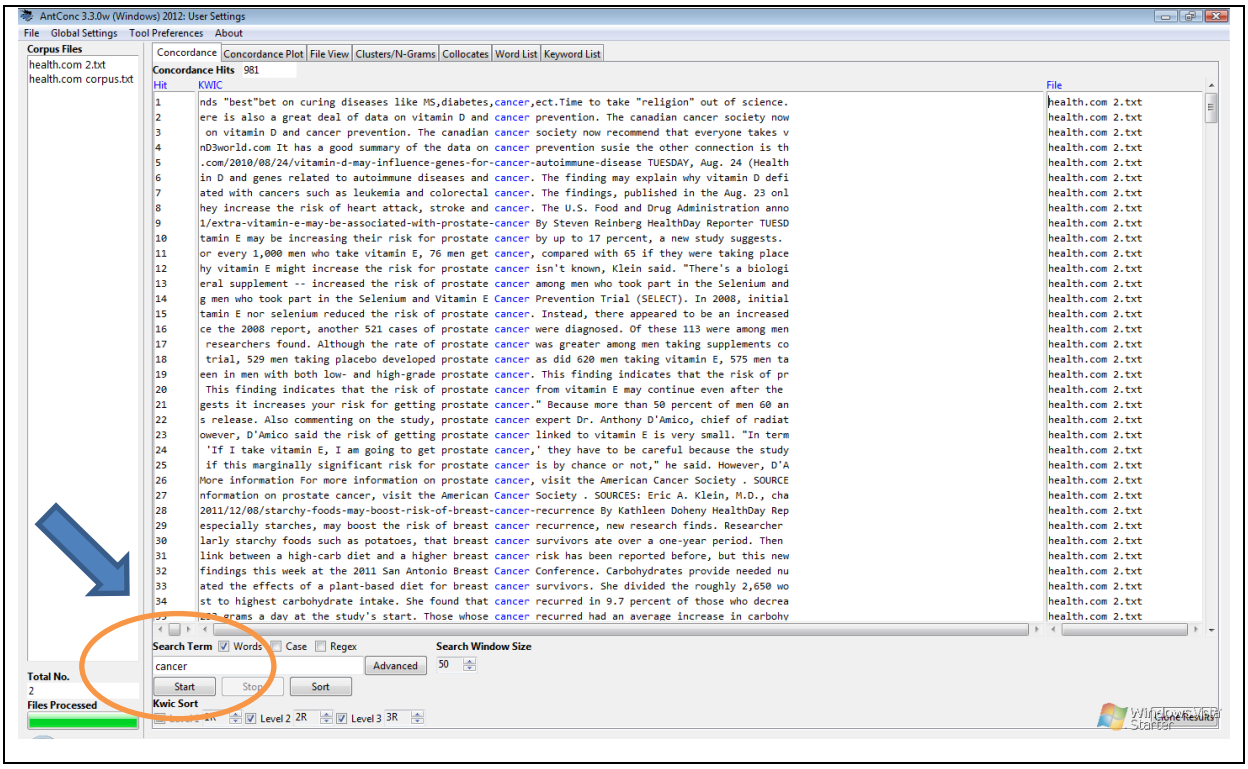


Figura 3: Tela do *AntConc* com a palavra *cancer* como palavra de busca.

1.4.2.2 *Compleat Lexical Tutor*[13]*:*

O *Compleat Lexical Tutor* (doravante *Lextutor*) é um conjunto de ferramentas *online* utilizado para o aprendizado direcionado por dados (*Data-driven learning* / DDL[14]), o que se denomina "dados por descoberta". Essa ferramenta foi desenvolvida pela Universidade de Quebec e está à disposição na internet gratuitamente. Dentre as várias possibilidades que oferece, há um utilitário denominado *Vocabprofile,* que permite verificar a proporção de palavras corridas (*tokens*) de um dado texto tendo como referência o *Brown Corpus*[15]. Possui também uma lista das mil palavras mais frequentes na língua de estudo, outra de 1001-2000,

[13] Disponível em: www.lextutor.ca. Último acesso em 23.01.13.
[14] "DDL: sigla para *data driven learning,* termo criado por Tim Johns (1991) para designar o ensino, especialmente de línguas estrangeiras, em que o aprendiz tem acesso a dados linguísticos (concordâncias) extraídos de linguagem autêntica, para examiná-los, levantar hipóteses, fazer inferências e generalizações".
[15] O *Brown Corpus* é composto por textos jornalísticos em inglês americano e tem um milhão de palavras.

uma terceira de 2001-3000 (West,1953), além de uma lista de palavras acadêmicas *(Academic Word List)* (cf. **Figura 4,** lado esquerdo)**.** De acordo com as diretrizes apresentadas na página, essa ferramenta (*Lextutor*) foi desenvolvida tanto para estudo individual de alunos, pesquisadores e professores, porém acreditamos que sua interface não seja tão amigável sem que haja uma preparação adequada para essa utilização, necessita-se de prática e estudo para que se possa usufruir dos recursos por ela oferecidos.

O conceito teórico básico que embasa a criação da ferramenta é a frequência das palavras. Segundo Thomas Cobb (1997), se adotarmos um critério aleatório na organização do conteúdo programático, podemos incorrer no risco de priorizar itens que sejam irrelevantes para o aluno naquele momento do aprendizado. Ainda conforme o autor, se o objetivo imediato do aprendiz for a leitura de textos acadêmicos em inglês com uma boa fluência e expandir seu vocabulário ainda mais; ele deve aprender as 2000 palavras mais frequentes das famílias de palavras em inglês, isto é, as palavras-chave, suas inflexões e derivações, pois essas palavras constituem, em termos gerais 80% das palavras diferentes (*word types*) em inglês. A **Tabela 1**[16] apresenta dados do *Brown corpus* (Nation, 1990, 2001).

Palavras Diferentes *(word types)*	**Porcentagem das palavras de um texto.**
86.741	100 %
43.831	99,0
6.000	89,9
5.000	88,6
4.000	86,7
3.000	84
2.000	**79,7**
1.000	72,0
10	23,7

Tabela 1: Porcentagem das palavras em um texto.

Essa tabela demonstra que em inglês, somente algumas palavras são diferentes (*word types*) em qualquer texto. Dez palavras , ou seja, 23,7% da página impressa são palavras repetidas, tipo *the* e *of.* Somente 1000 famílias de palavras

[16] A tabela foi copiada da página: *Vocab Research Resources.* Disponivel em www.lextutor.ca/research. Acesso em 27.01.2013

constituem 70% e 2000, 80%. Seriam as primeiras palavras a serem memorizadas. Se o conteúdo lexical não é programado, o aluno vai encontrar as palavras de que necessita aprender “naturalmente”, o que levaria mais tempo. Ao contrário, se ele tiver uma lista de palavras com vocabulário específico poderá maximizar seu tempo de estudo.

A partir dessa concepção apresentada, demonstramos na **Figura 4** um exemplo de tela do *Lextutor* que nos ajudou a identificar a porcentagem das palavras de origem latina em um ensaio feito no utilitário com um texto do *The New England Journal of Medicine.* Explicaremos com maiores detalhes as diferenças entre as porcentagens no capítulo 2, destinado à metodologia.

O *Vocabprofile* foi de suma importância para esta pesquisa já que foi por meio dessa ferramenta que começamos a fazer ensaios sobre a diferença entre os gêneros e percebemos a presença da alta porcentagem dos cognatos de origem latina nos textos acadêmicos (cf. **Figura 4**, canto direito do lado de baixo), fato que não é desconhecido, mas pôde ser comprovado estatisticamente com a ajuda da ferramenta.

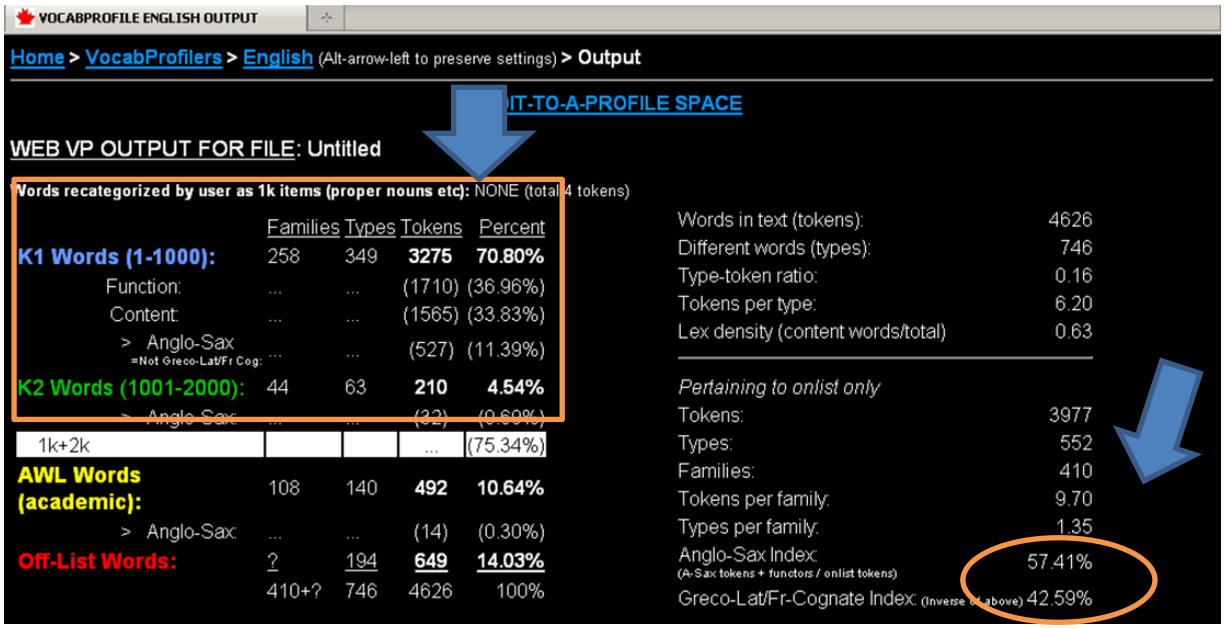


Figura 4: Tela do *Vocabprofile (Lextutor)*

O conceito do conhecimento prévio, discutido anteriormente, aliado à frequência de palavras são conceitos importantes no auxílio do professor em suas escolhas e na condução desse estudo.

1.4.2.3 *WordSmith Tools*

O *WordSmith Tools* versão 5.0 (desenvolvido por Mike Scott, 2008). No caso desta pesquisa, o utilitário *"Compare 2 lists"* foi utilizado inicialmente para se obter as palavras-chave de cada um dos três sub*corpora,* que será melhor detalhado no capítulo 3, destinado à metodologia.

As três ferramentas acima citadas geram listas de palavras *(worlists)*, listas de palavras-chave *(keywords)*, linhas de concordância *(Concord ou Concordance) e* combinações de palavras *(clusters).*

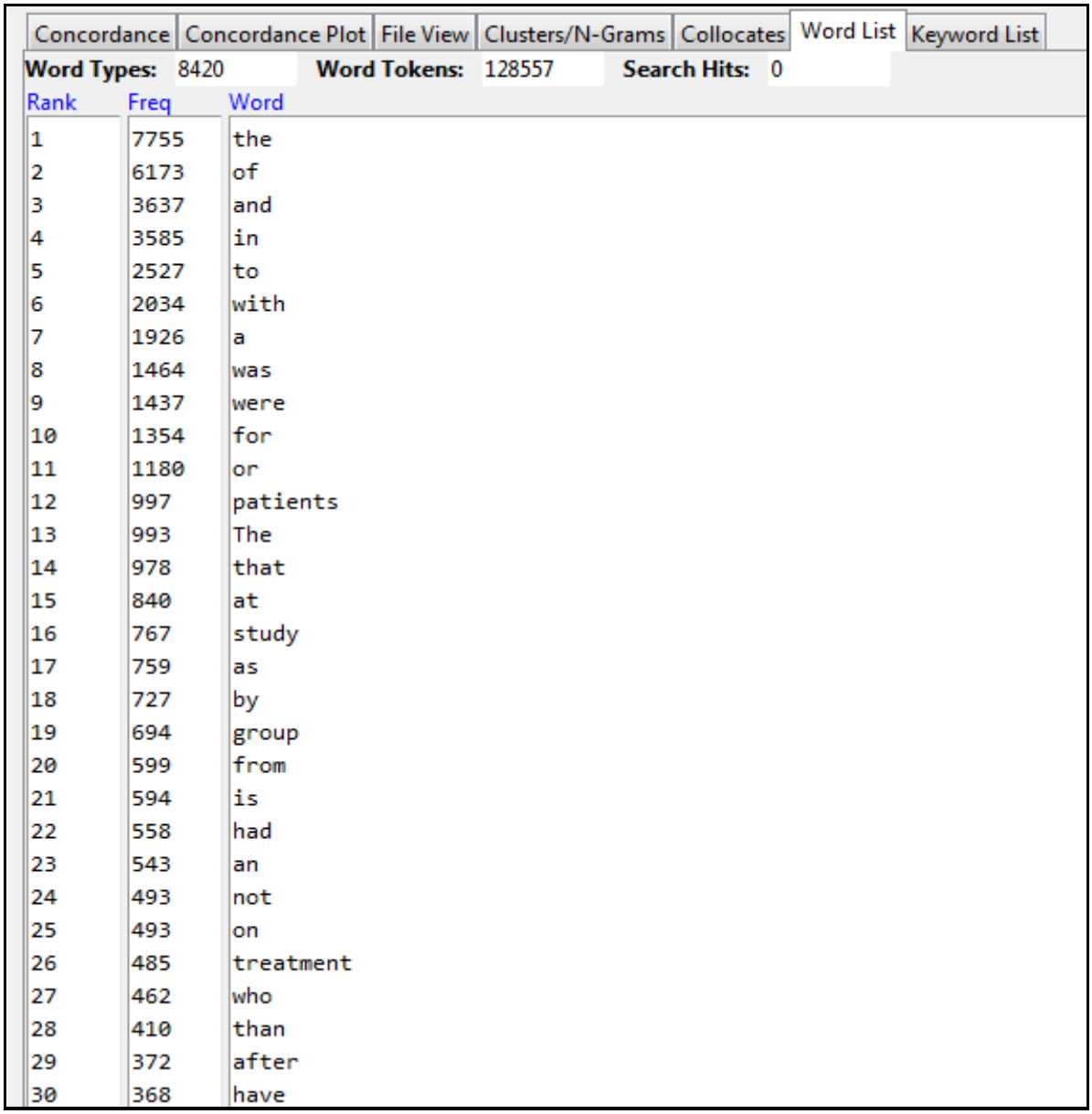
Apresentamos aqui uma síntese das principais funções dos utilitários, tanto do *AntConc, Lextutor* e *WST*: a *WordList, KeyWord*, *Concord(ance)* e *Clusters/N-Grams* que foram os recursos usados nesta pesquisa:

1.4.3 Os utilitários das ferramentas:

1.4.3.1 As listas de palavras (*wordlists*)

Como já mencionado, uma lista de todas as palavras de um *corpus* ou texto costuma ser parte integrante de quase todos os programas de análise linguística. Sua função é contar o número de palavras existentes em dado *corpus,* listá-las e mostrar a frequência com a qual essas palavras aparecem no texto. As ferramentas *AntConc* e *WST* mostram a lista de palavras tanto em ordem alfabética como por ordem de frequência. O uso desse utilitário se faz relevante para podermos perceber os itens lexicais e estruturais que caracterizam aquele *corpus*. A partir da lista de palavras pode-se conhecer o que se destaca naquele texto e sua função principal. Como podemos observar na **Figura 5** *(AntConc)*, no corpus acadêmico, a lista de palavras apresenta palavras de conteúdo, palavras funcionais e verbos entre as mais frequentes.

No caso dos textos acadêmicos, por exemplo, podemos notar o léxico pertinente à área de saúde: *patients, study, group, treatment* e outras (cf. **Figura 5**).



Rank	Freq	Word
1	7755	the
2	6173	of
3	3637	and
4	3585	in
5	2527	to
6	2034	with
7	1926	a
8	1464	was
9	1437	were
10	1354	for
11	1180	or
12	997	patients
13	993	The
14	978	that
15	840	at
16	767	study
17	759	as
18	727	by
19	694	group
20	599	from
21	594	is
22	558	had
23	543	an
24	493	not
25	493	on
26	485	treatment
27	462	who
28	410	than
29	372	after
30	368	have

Figura 5: As 30 primeiras palavras por ordem de frequência no *corpus* acadêmico *(NEJM).*

1.4.3.2 As listas de palavras-chave *(Keyword List)*

As listas de palavras-chave são obtidas por meio da comparação entre um *corpus* de estudo e um *corpus* de referência. Salienta-se o que é proporcionalmente mais recorrente no *corpus* de estudo, ou seja, as palavras-chave. “O resultado do contraste é uma lista de palavras-chave, ou palavras cujas frequências são

estatisticamente diferentes no *corpus* de estudo e no *corpus* de referência." (Berber Sardinha, 2004: 96).

Os programas listam esses itens por ordem de chavicidade (*keyness*), ou seja, o que é mais ou menos frequente estatisticamente no *corpus* de estudo. A chavicidade das palavras é determinada pela prova estatística presente na ferramenta. No caso desta pesquisa, foi utilizado o padrão da ferramenta *AntConc* - *log-likelihood*[17]- recomendado pelo autor e que apresenta em suas telas o Valor Crítico *(critical value)* das palavras. Esse valor indica a significância estatística resultante do teste *log-likelihood,* descrito em detalhes no manual da ferramenta. Um teste de significância estatística é uma forma de medir a importância da combinação das palavras através de um procedimento estatístico. A medida dessa importância é a probabilidade de a combinação ocorrer por acaso, tendo em vista os números da amostra extraída do *corpus* (Vasilévski, 2007).

Como utilizamos essa ferramenta na maior parte da pesquisa, descrevemos esses valores de acordo com o referido manual que utiliza o *log-likelihood* como medida estatística utilizando os seguintes valores de significância:

Percentil 95	Nível 5%	$p < 0.05$	Valor crítico = 3.84
Percentil 99	Nível 1%	$p < 0.01$	Valor crítico = 6.63
Percentil 99,9	Nível 0.1%	$p < 0.001$	Valor crítico = 10.83
Percentil 99,99	Nível 0.01%	$p < 0.0001$	Valor crítico= 15.13

Tabela 2: Equivalência dos testes estatísticos (percentil, nível de porcentagem, valor de probabilidade e valor crítico) apresentada no manual da ferramenta.

Como podemos observar, essa tabela faz uma equivalência entre os testes estatísticos: percentil, nível de porcentagem, valor de probabilidade, presentes em outras ferramentas de análise, como o *WST*. Essa última apresenta o valor de *p e o AntConc,* o Valor Crítico. Portanto, quanto maior o Valor Crítico; mais significativa será a diferença entre os dois números de ocorrências da sequência entre dois *corpora*.

[17] Disponível em: http://ucrel.lancs.ac.uk/llwizard.html. Acesso em 23.01.2013.

Dayrell (2010:148) apresenta uma explicação mais detalhada dessa comparação:

> Nas ciências sociais, é comum adotar 0,05 como o valor de corte, ou seja, o ponto acima do qual a probabilidade de se aceitar um resultado como significativo quando este tenha ocorrido ao acaso é no máximo de 5 em 100 chances (5%) (Hinton,1995: 38; Oakes,1998: 9). Em outras palavras, um valor de *p* maior ou igual a 0,05 indica que a diferença *não* é estatisticamente significativa. Já para um p < 0,05, a diferença entre as frequências observadas nos dois *corpora* é estatisticamente significativa, ou seja, o analista pode afirmar com certeza que as frequências observadas são diferentes.

1.4.3.3 Concordanciador *(Concordance)*

Esse utilitário também costuma fazer parte dos programas de análise linguística e é usado para que possamos ver o contexto e o co-texto das palavras observadas. Em uma linha de concordância os dados a serem analisados são apresentados de forma organizada para facilitar a visualização. Essas linhas apresentam a palavra de busca centralizada, inserida em seu contexto, facilitando a análise, a partir de uma leitura horizontal dentro do contexto ou uma leitura vertical, para observar as recorrências na busca de determinado padrão:

> [A frequência e a coocorrência de itens lexicais são...] como peças-chave da léxico-gramatica, respondem por características vitais do discurso, tais como a maneira como ‘as coisas’ são ditas e como são montadas em sequência todas as construções de uma língua. (Berber Sardinha, 2010:304).

Por isso, em nossa pesquisa a frequência das palavras é um dado importante, já que a partir dela observamos que tipo de palavras são características de cada gênero de texto e de que modo o padrão da linguagem é utilizado, e desse modo enfatizar o que é pertinente para compor o material de estudo do aprendiz.

1.4.3.4 *Clusters* (combinações de palavras)

Sob a ótica da LC, *clusters* são palavras que se repetem uma na companhia das outras em sequência. Elas representam uma relação mais forte do que os colocados porque funcionam como unidades de multi-palavras, grupos ou frases e colocações; apesar de serem palavras que ‘andam juntas’ podem não constituir uma

expressão fixa. Scott (2010) afirma que não basta que as palavras apareçam juntas para formarem uma verdadeira unidade. Essas combinações precisam produzir sentidos. Biber et al. (1999) denominam o mesmo agrupamento de palavras *lexical bundles*. Diferentemente da abordagem de Sinclair (1996a), Biber et al (1999) tendem a uma visão puramente quantitativa dos modelos de "feixes lexicais", sem se preocupar tanto com a relação entre a forma e o significado (O´Keeffe et al., 2007).

Os *clusters* ajudam na identificação de padrões fraseológicos nas linhas de concordância, especialmente se o pesquisador tiver centenas de linhas. A palavra de busca aparecerá na composição dos *clusters,* já que eles são baseados em linhas de concordância.

Nesta pesquisa, os *clusters* foram importantes para a verificação dos padrões linguísticos que serão descritos no capítulo 3, destinado aos resultados.

1.4.4 A ferramenta *BootCat* para compilação automática de corpus

O *BootCat*[18] (Baroni e Bernardini, 2004) é uma ferramenta que nos permite compilar um *corpus* a partir de textos extraídos da *web*. A vantagem dessa ferramenta é que a construção do *corpus* é muito mais rápida, o que facilita o trabalho de um professor no momento de compilar textos específicos para serem analisados em sala de aula. Para Varantola (2003), essa é a única fonte de construção viável de *corpora ad hoc* para um fim específico. Segundo a autora, *corpora* 'descartáveis' são fontes essenciais para profissionais que normalmente trabalham com linguagens especializadas e necessitam de *corpora* que sejam fáceis de construir e precisam ser atualizados com frequência.

Esse método de compilação tem sido utilizado tanto para a construção de *corpora* especializados para trabalhos de terminologia e tradução como para a construção de grandes *corpora* de língua geral. (Kilgarriff et al., 2011)

[18] *BootCat* é a forma abreviada de ***Boot****strapping* ***C****orpora* ***a****nd* ***T****erms.* Disponível em: http://sslmit.unibo.it/~baroni/bootcat.html. Acesso em 20.02.2012.

Uma restrição do *BootCat* é a limitação dos tipos de textos captados. A ferramenta não processa páginas com *frames*[19] e com direcionamentos, e ainda pode haver *links* repetidos entre os obtidos na busca, redundância que pode ser resolvida com a eliminação dos mesmos no momento de verificação das *URLs* (Localizador padrão de recursos) [20]. Daremos maiores detalhes do processo de compilação no capítulo 3, destinado à metodologia, quando explicarmos a captação de um de nossos sub*corpora (Health News)* e parte do segundo *(WebMD).*

1.4.5 A complexidade da linguagem falada

A análise multidimensional foi um dos primeiros trabalhos de análise conduzida por Biber (1988;1995) para tentar entender as diferenças entre textos escritos e a linguagem falada. Esse tipo de análise considera a distribuição de aspectos linguísticos em um *corpus* eletrônico para identificar os padrões de associações, ou seja, um conjunto de características linguísticas que tendem a coocorrer nos textos. A coocorrência[21] de itens lexicais, juntamente com a frequência das palavras são conceitos e fontes de dados importantes em LC. Cada um desses grupos é chamado por Biber de "dimensão".

> As dimensões tem um conteúdo tanto linguístico como funcional. O conteúdo linguístico de uma dimensão compreende um grupo de características linguísticas (tais como: passivas, nominalizações, e locuções prepositivas) que coocorrem com uma marcada alta frequência nos textos (Biber, 1989: 7) [22]

Para o autor, é preciso reconhecer a existência de um espaço multidimensional para poder captar as relações entre os gêneros. Esse tipo de estudo tem como aspectos centrais, as características lexicais e gramaticais com ênfase estatística.

---

[19] *Frames,* em informática, são divisões da tela do *browser* em diversas telas (ou "quadros"). Com isso, torna-se possível apresentar mais de uma página por vez: por exemplo, um índice principal em uma parte pequena da tela, e os textos relacionados ao índice em outra parte. http://br.answers.yahoo.com .Acesso em 18.01.2013

[20] Um *URL (Uniform Resource Locator)* é o endereço de um recurso (um arquivo, uma impressora, etc.) disponível em uma rede, seja a internet ou uma rede corporativa (intranet).

[21] Palavras que ocorrem lado a lado em um corpus ou texto.

[22] No original: *"Dimensions have both linguisitc and functional contente. The linguistic content of a dimension comprise a group of linguistic featurs (such as passives, nominalizatons, prepositional phrases) that co-occur with a markedly high frequency in texts".*

As cinco dimensões propostas por Biber (1988) apresentam uma descrição da variação dos gêneros em inglês. O modelo foi desenvolvido a partir da análise da distribuição da coocorrência de 67 características linguísticas em 481 textos escritos e falados do inglês contemporâneo.

Relatamos a seguir um resumo das características que coocorrem associadas a cada uma das cinco dimensões[23] propostas por Biber (1988) [24]:

– Dimensão 1 - **produção com interação** vs. **produção informacional**: os registros que melhor representam o modo de produção com interação são as conversas, tanto ao telefone quanto face a face; os registros que melhor representam a produção informacional são documentos oficiais, reportagens jornalísticas, e a prosa acadêmica.

Dentro de cada dimensão aplica-se a Análise Fatorial. Segundo Biber (2009:827 *apud* Sardinha, 2011):

> Cada dimensão representa diferentes conjuntos de características linguísticas que coocorrem. Esses padrões de coocorrência são identificados quantitativamente utilizando-se uma técnica estatística conhecida como análise fatorial; cada conjunto de características que coocorrem é chamado de fator. Em uma análise fatorial, um grande número de variáveis originais – neste caso, características linguísticas - são reduzidas a um conjunto de variáveis derivadas e subjacentes: os fatores.[25]

Ainda segundo Sardinha (2011), os fatores podem ser positivos e negativos, que em estatística, não significa que as variáveis são menos frequentes; simplesmente significam que quando as características positivas ocorrem com alta frequência em um texto, aquele mesmo texto terá uma baixa frequência de fatores negativos e vice-versa. Em estatística é uma questão de polaridade. Como descritos, os testes aplicados são bastante sofisticados. Nesta pesquisa levamos em

[23] A tradução da nomenclatura das dimensões é de Pacheco (1997:66) *apud* Berber Sardinha (2004b).
[24] Embora a apresentação das dimensões tenha sido em Biber, 1988, aqui a consulta foi feita em Biber, 1989.
[25] No original: *'Each dimension represents a different set of co-occurring linguistic features. These co-occurrence patterns are identified quantitatively, using a statistical technique known as factor analysis; each set of co-occurring features is referred to as a factor. In a factor analysis, a large number of original variables -- in this case the linguistic features -- are reduced to a small set of derived, underlying variables: the factors.'* (Biber 2009: 827).

consideração os resultados desses estudos e não entraremos em detalhes quanto a esses testes aplicados.

– Dimensão 2 - **preocupações narrativas** vs. **não-narrativas**: os registros que melhor demonstram uma preocupação com a narração são os registros de ficção, enquanto que os que melhor exprimem uma orientação não-narrativa são os registros de rádio e TV, passatempos, e documentos oficiais.

Características linguísticas: verbos no passado, pronomes na 3ª pessoa, verbos no perfeito, verbos públicos[26], negação sintética, orações no particípio presente.

(Fator negativo) verbos no presente, adjetivos atributivos.

– Dimensão 3 - **referências explícitas** vs. **referências dependentes do contexto**: os registros que apresentam referências explícitas em maior grau são documentos oficiais, cartas profissionais, resenhas jornalísticas e a prosa acadêmica. Já os registros de rádio e TV, conversas telefônicas e face a face, e ficção romântica exprimem referência dependendo da situação.

Características linguísticas:

(Fatores positivos) orações *WH*, coordenação frasal, nominalizações.

(Fatores negativos) locuções adverbiais de tempo e de lugar, advérbios.

– Dimensão 4 - **expressão explícita de persuasão** vs. **não-explícita**: os registros de caráter mais persuasivo são as cartas profissionais, os editoriais, e a ficção romântica. Por outro lado, os registros nos quais a persuasão é menos explícita são os de rádio e TV, resenhas jornalísticas, e ficção de aventura.

Características linguísticas:

Infinitivos, modais de predição, verbos de *persuasão*[27]*,* subordinação condicional, modais de necessidade, auxiliares separados, modais de possibilidade.

[26] Verbos públicos (admitir, concordar, argumentar, reclamar, declarar, negar, explicar, insistir, mencionar, prometer, protestar, responder, relatar, dizer, sugerir, (= *suasive*), escrever, jurar, xingar).

Fatores negativos:

Sem características complementares.

– Dimensão 5 **- informação abstrata** vs. **não-abstrata**: os registros que veiculam informação mais abstrata são os acadêmicos, os documentos oficiais, e os religiosos. Já as conversas telefônicas e face a face, e ficção romântica apresentam informação menos abstrata.

Características linguísticas:

Passivas sem agente, conjunções, orações de particípio passado e outras subordinadas adverbiais.

Fatores negativos:

Sem características complementares.

Em seu primeiro trabalho com a análise multidimensional Biber (1988), haviam cinco dimensões como mencionado, mais tarde a sexta dimensão foi incluída:

- Dimensão 6 – **elaboração informacional *online*:** os registros nos quais a elaboração da informação é mais imediata: são palestras, entrevistas, e palestras espontâneas, enquanto os registros em que a informação é elaborada de antemão são os de ficção (mistérios, científica e geral.) Pacheco Oliveira (1997:66 apud Sardinha 2004b).

Embora a análise multidimensional proposta por Biber (1988), segundo Berber Sardinha (1999), não esteja situada dentro dos limites da Linguística de Corpus desde o início, consideramos essa classificação como importante fundamentação teórica para os estudos léxico-gramaticais desenvolvidos por Biber a partir desse trabalho e embasamento teórico para esta pesquisa.

[27] Verbos *suasive* (concordar, pedir, ordenar, mandar, insistir, instruir, propor, recomendar, estipular, sugerir). Segundo Oliveira (1997), o mesmo que:`públicos´.

Como mencionado, as diferenças léxico-gramaticais entre a linguagem escrita e a linguagem falada têm sido uma preocupação expressa nos trabalhos de Biber (1988, 1992, 2006, 2009), um dos principais estudiosos nessa área. Em seu estudo, Biber (2009) utiliza uma abordagem direcionada por *corpus (corpus-driven)* para identificar os padrões das sequências de multi palavras na linguagem oral e na linguagem escrita. Uma abordagem direcionada por *corpus* é um tipo de pesquisa mais indutiva, na qual as construções linguísticas emergem da análise de um *corpus.* Essa abordagem difere das práticas linguísticas convencionais na medida em que mantém os pressupostos iniciais a um patamar mínimo no que se refere às construções linguísticas que são empregadas para a análise (Biber, 2009).

Em oposição à definição anterior, a abordagem baseada em *corpus (corpus based)* significa que o linguista pré-define seu objeto de estudo e consulta o *corpus* como meio de validação das características iniciais (Moon, 1998, *apud* Biber, 2009). Por exemplo, o pesquisador investiga a preposição utlizada depois do item lexical *used* em inglês, e descobre que existem várias outras possibilidades (cf. **Figura 6**) além de *used to* (que não aparece na seleção): *used for* (linhas 45 a 62), *used in* (linhas 66 a 69), dentre outros usos. Demonstramos a seguir, uma pesquisa *corpus-based*, em que o pesquisador busca uma resposta no *corpus*.

Figura 6: Seleção das linhas de concordância tendo como palavra de busca *"used"*.

Nesse trabalho, Biber (2009:300), discorre sobre a aplicação de uma metodogia "radicalmente" direcionada por *corpus* ao estudo de linguagem formulaica[28] em inglês. O autor afirma que a gramática do "discurso falado é fundamentalmente diferente da gramática da escrita".[29] Em especial, as orações (incluindo as subordinadas) são fundamentos do discurso falado, enquanto, especialmente, os sintagmas nominais e frases preposicionadas são características da escrita informacional. Para Biber (2009), a escrita informacional é tanto a escrita jornalística quanto a acadêmica, que é denominada escrita informacional de exposição.

[28] Linguagem formulaica, segundo Biber, são sequências de palavras que se repetem formando padrões, tanto na linguagem escrita quanto na oral. (2009). Exemplo: *to be likey to/ than.*
[29] No original: *[...] the grammar of speech is fundamentally different from the grammar of writing […]*

No caso do presente estudo, podemos afirmar que a abordagem utilizada foi direcionada pelo *corpus*, já que no início da pesquisa não tínhamos afirmações *a priori* quanto ao que poderíamos encontrar. As descobertas foram sendo feitas a partir das análises das palavras-chave e suas linhas de concordância. Demonstraremos com maiores detalhes as listas de palavras-chave no capítulo 2, destinado à metodologia, e as linhas de concordância no capítulo 3, destinado aos resultados.

Hyland (2004) resume que a pesquisa de Biber (1988) demonstra como a prosa escrita é caracterizada por grupos de características gramaticais: substantivos, palavras longas, adjetivos atributivos e frases preposicionais, pois, segundo Biber (1988), a função dessas palavras é apresentar informação densa e compacta, enquanto os pronomes de segunda pessoa, perguntas diretas, verbos no presente, verbos privados (*feel, think*), e omissão do pronome *that* são menos frequentes devido a seu caráter menos interativo (Hyland, 2004: 215).

CAPÍTULO 2. METODOLOGIA

Este capítulo contempla o procedimento de compilação do *corpus* de estudo, bem como suas dificuldades e os critérios para a seleção dos textos que o constituem. Em segundo lugar apresentamos o método utilizado para o levantamento dos dados por meios da comparação dos três sub*corpora*: *NEJM, Health News* e *WebMD.*

2.1 Ensaios com o *Lextutor*

Como citamos anteriormente, Van Lier (1989) propõe que a sala de aula seja um ambiente para pesquisa de campo, e para que isso aconteça é preciso que o professor esteja atento às necessidades do grupo ao qual está ensinando. Certamente, a experiência auxilia esse professor em suas tomadas de decisões. A observação é um fator importante nesse processo, assim como a disposição de operar mudanças no conteúdo programático à medida que novos desafios se apresentam (Hyland, 2004; Gavioli, 2005).

Com o intuito de preparar atividades que fossem adequadas aos alunos de leitura instrumental, começamos um experimento para nos guiar em direção à escolha de um conteúdo que fosse mais apropriado àquele grupo.

Além disso, outro objetivo era verificar se havia algum indício que pudesse nos auxiliar a descobrir a diferença entre os textos acadêmicos e os textos de divulgação que causavam problemas para os alunos. O próximo passo foi submeter quatro tipos de textos ao utilitário *Vocabprofile* da ferramenta *Lextutor*: a) um artigo acadêmico do *NEJM*; b) dois artigos da *BBC News Health;* c) um artigo da *BBC News World*; d) três textos que fizeram parte de exames anteriores de leitura instrumental a que os alunos se submetem, preparados por uma instituição de ensino.

O que chamou a atenção nesses dados (cf. **Tabela 3**) foi o fato de a porcentagem das palavras anglo-saxônicas, que foi de **57,41% *(Anglo-Sax Index)***, e a de palavras cognatas latinas, **42,59%**, **(*Greco-Lat/Fr-Cognate Index)*** no *NEJM.*

	NEJM	*BBC News Health*	*BBC News World*	Exames anteriores
Anglo-Sax Index	57,41%	67,63%	76,54%	68,21%
Grec./Lat./ FrCognate Index	42,59%	32,37%	23,46%	31,79%

Tabela 3: Porcentagens obtidas a partir da inserção dos diferentes gêneros de textos no utilitário *Vocabprofile (Lextutor).*

A segunda experiência com a ferramenta ocorreu a partir da inserção de dois textos obtidos do *site* da *BBC News Health* no *Vocabprofile* do *Lextutor.* Utilizamos dois textos pelo fato de serem menores e, dessa forma, obtivemos o mesmo número de palavras da experiência anterior. A análise desses dados demonstrou que a porcentagem de palavras de origem anglo-saxônica aumentou para **67,63 % *(Anglo-Sax Index)*** e a das cognatas latinas diminuiu para **32,37% (*Greco-Lat/Fr-Cognate Index)***. Se compararmos essas porcentagens com as anteriores, notamos que talvez estivesse aí um indício de que o número maior de palavras cognatas do latim facilitasse o entendimento dos alunos que não têm um vocabulário amplo, e por isso, encontravam maior dificuldade na leitura de textos de divulgação por apresentam um vocabulário menos conhecido.

Para confirmar essa tendência, investigamos se os artigos de notícias (*BBC News World*) que não eram relacionados à área de saúde mudariam esse quadro. Observou-se que houve uma inversão na porcentagem das palavras cognatas se comparadas aos textos acadêmicos da área médica propriamente dita.

As porcentagens obtidas a partir da inserção dos textos da *BBC News World* indicam que houve uma diminuição ainda maior no número de palavras cognatas latinas para **23,46% (*Greco-Lat/Fr-Cognate Index)*** contra **76,54% *(Anglo-Sax Index).*** Isso confirmava a hipótese de que quanto mais os textos se aproximam do vocabulário do cotidiano, menos palavras de origem latina são utilizadas. Essa diferença pode impedir a compreensão de um aluno que não tenha um conhecimento lexical amplo, pois uma das primeiras técnicas de ensino de leitura instrumental é o reconhecimento de palavras cognatas (Lanzoni, 2010; Souza et al., 2010).

Se observarmos a última coluna obtida a partir dos exemplares de testes anteriores, notamos que as porcentagens dos mesmos se assemelham aos textos publicados na *BBC Health News*; essa constatação nos levou a investigar que construções e/ou léxico seriam diferentes em cada um desses grupos de textos.

2.2 O *corpus* de estudo e suas dificuldades

A construção e a busca pela qualidade de um *corpus* de estudo constituem uma fase importante da pesquisa. Muitos são os autores que discorrem sobre a compilação e a montagem de um *corpus*. Sinclair (1991), Atkins et al. (1992), Berber Sardinha (2004), entre outros, afirmam que quanto maior o *corpus*, melhor. Porém, quando fazem essa afirmação, referem-se à representatividade da língua geral. Conforme esses autores, para se obter essa representatividade é necessário que o *corpus* seja composto de milhões de palavras. Mas para Bowker (2002), Gavioli (2005), Chambers (2007) e Aston e Corbett (2009), se o *corpus* for bem construído e com objetivos específicos, o pesquisador pode obter informações bastante importantes e úteis de um *corpus* pequeno e especializado, adequado às necessidades da pesquisa, ideia que também é compartilhada por Tribble (1997):

> Se quisermos investigar o léxico de um domínio específico (ex.: saúde), um *micro corpus* especializado pode ser mais eficiente do que um maior, de língua geral. Por exemplo, nos componentes escritos do *BNC Sampler* (1.000.000 de palavras) não aparece sequer uma ocorrência de *"cancers".* "Um micro corpus de artigos de saúde da Encarta (24.805 palavras) apresenta 33 exemplos contextualizados que são muito úteis." (Tribble 1997: 112-3) [30]

Isso posto, nosso *corpus* de estudo foi estabelecido segundo os critérios propostos por Gavioli (2005). É especializado não só porque representa tipos de textos especializados e tópico único, mas também porque foi criado para o fim específico de levantar dados a serem ensinados a um determinado grupo de aprendizes. Existe um debate acerca do que seja um corpus especializado.

[30]No original: *"If one wishes to investigate the lexis of a particular current domain (e.g. health) a specialist micro-corpus can often be more useful than a much larger general corpus. For example, in the written component of the BNC Sampler (1,000,000 words) there are no instances of "cancers". An Encarta micro-corpus of health articles (24,805 words) gives 33 usefully contextualized examples*."

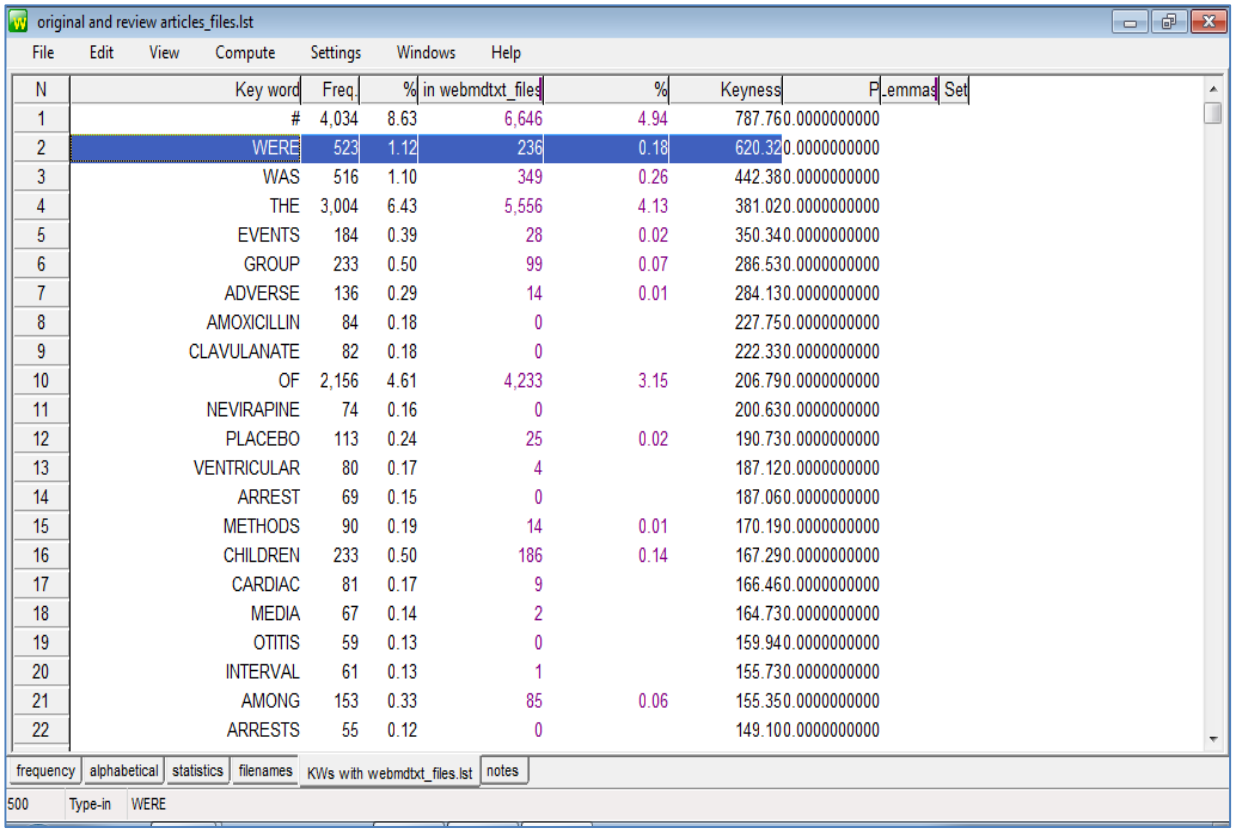
Segundo vários autores, entre eles Tribble (1997) e Gedhill (2000) (apud Gavioli, 2005) *corpora* especializados são aqueles que são compilados com o propósito de criar uma amostra de uma linguagem especializada. Podem ser similares (ex. ciência, medicina, negócios, filosofia), podem ser do mesmo tipo ou gênero: (artigos de pesquisa, cartas, capítulos de livros), ou até mesmo, ambos os tipos (ex. artigos de pesquisas médicas ou palestras científicas), além de textos de outras categorias especializadas como a linguagem jornalística e a linguagem acadêmica.

Podemos afirmar que o *corpus* compilado para este trabalho é especializado, uma vez que foi escolhido com o propósito definido de fazer um levantamento das principais características que diferenciam os artigos de pesquisa acadêmico-científicos dos artigos de divulgação da área de saúde publicados na internet. Se comparado aos *corpora* de língua geral *BNC, Coca, Cobuild* e outros com milhões de palavras, e tomando como parâmetro o tamanho sugerido por Berber Sardinha (2004), nosso *corpus* de estudo é considerado pequeno-médio, pois 'cada *subcorpus* tem aproximadamente 120.000 palavras.

Ainda segundo Gavioli (2005:63), "um *corpus* especializado 'captura' as características específicas da linguagem, tais como, especificidades de gêneros e registros de maneira mais efetiva do que um *corpus* de língua geral" [31].

No decorrer de nossa pesquisa pudemos constatar que essa premissa demonstrou ser verdadeira, já que em uma de nossas comparações, ainda com o *corpus* acadêmico incompleto (45.000 palavras), verificamos que a lista de palavras-chave não variou de maneira significativa em comparação à lista feita quando o corpus já estava completo, com 128. 557 palavras. Em ambos os casos, os verbos auxiliares *were* e *was* estão no topo das duas listas (cf. **Figura 7**), e os principais substantivos encontrados estão ligados à área de medicina e nomes de medicamentos.

[31] No original: *"[...] it seems now to be generally acknowledged that specialized corpora 'capture' local characteristics of language such as features of genre or register better than general corpora."*



N	Key word	Freq.	%	in webmdtxt_files	%	Keyness	P	Lemmas	Set
1	#	4,034	8.63	6,646	4.94	787.76	0.0000000000		
2	WERE	523	1.12	236	0.18	620.32	0.0000000000		
3	WAS	516	1.10	349	0.26	442.38	0.0000000000		
4	THE	3,004	6.43	5,556	4.13	381.02	0.0000000000		
5	EVENTS	184	0.39	28	0.02	350.34	0.0000000000		
6	GROUP	233	0.50	99	0.07	286.53	0.0000000000		
7	ADVERSE	136	0.29	14	0.01	284.13	0.0000000000		
8	AMOXICILLIN	84	0.18	0		227.75	0.0000000000		
9	CLAVULANATE	82	0.18	0		222.33	0.0000000000		
10	OF	2,156	4.61	4,233	3.15	206.79	0.0000000000		
11	NEVIRAPINE	74	0.16	0		200.63	0.0000000000		
12	PLACEBO	113	0.24	25	0.02	190.73	0.0000000000		
13	VENTRICULAR	80	0.17	4		187.12	0.0000000000		
14	ARREST	69	0.15	0		187.06	0.0000000000		
15	METHODS	90	0.19	14	0.01	170.19	0.0000000000		
16	CHILDREN	233	0.50	186	0.14	167.29	0.0000000000		
17	CARDIAC	81	0.17	9		166.46	0.0000000000		
18	MEDIA	67	0.14	2		164.73	0.0000000000		
19	OTITIS	59	0.13	0		159.94	0.0000000000		
20	INTERVAL	61	0.13	1		155.73	0.0000000000		
21	AMONG	153	0.33	85	0.06	155.35	0.0000000000		
22	ARRESTS	55	0.12	0		149.10	0.0000000000		

Figura 7: Tela do *WSTools* demonstrando a comparação de duas listas (*NEJM x WebMD)* do *corpus* de estudo inicial ainda incompleto.

2.2.1 Critérios iniciais de compilação

O primeiro passo em direção à compilação do *corpus* que atendesse ao propósito da análise foi verificar a origem dos textos que compunham o teste de proficiência a que os alunos se submeteriam. Constatamos que esses eram ora trechos de artigos acadêmicos retirados da internet, ora textos de divulgação e outras vezes, ainda, um resumo de um artigo editado e adaptado por quem elaborava a prova. Nenhum deles apresentava qualquer tipo de referência. Apesar de serem aplicados a profissionais da área médica, não eram textos exclusivamente daquela área. Por vezes, eram textos da área de ciências sociais, filosofia e outras, o que nos levou à hipótese de que estaria aí um indicativo para a dificuldade de compreensão enfrentada pelos alunos no momento em que deparavam com textos que não fossem de sua área de especialidade, pois, como já mencionamos, as características de um texto obedecem a várias convenções determinadas pela comunidade que lê e escreve aquele gênero textual. De tal sorte que os textos sobre linguística aplicada apresentam peculiaridades e usos que são diferentes dos textos

de biologia, que por sua vez, são diferentes dos de engenharia, e assim por diante (Hyland, 2008).

Para identificar a procedência dos textos dos exames e a fim de conhecer melhor a que gênero pertenciam, colocamos trechos dos mesmos no *Google* e entramos em contato com o *site* do periódico *The New England Journal of Medicine*. Percebemos que era um dos periódicos que tinham sido adaptados para fazer parte dos referidos exames*.* Na busca por textos adequados para a composição do *corpus,* constatamos que o periódico era dividido em diversas seções: editoriais, artigos acadêmicos, resenhas, correspondência (cartas) e relatos de casos. Como cada uma dessas seções apresentava características específicas de seu gênero, teríamos de decidir qual seria mais adequada ao nosso propósito.

Para viabilizar o estudo, escolhemos somente os artigos de pesquisa acadêmica, ali publicados, por considerá-los uma amostra interessante a ser estudada, já que esses textos apresentavam o mesmo formato daqueles lidos pelos alunos em suas respectivas áreas de especialidade. Não restringimos as disciplinas, pois nossos alunos advêm de diferentes especialidades médicas, e nem separamos por partes (*abstract, background, methods, results)*; consideramos os artigos como um bloco único, pois nosso interesse não era determinar as especificidades de cada seção e sim levar em conta a linguagem utilizada nos artigos como um todo. Desse modo, compilamos textos de diferentes disciplinas: cardiologia, cirurgia, dermatologia, entre outras especialidades.

2.2.1.1 Composição inicial do *corpus*

Inicialmente o *corpus* consistia somente de textos em inglês retirados de duas fontes distintas: do *NEJM,* composto por artigos acadêmicos, e dos artigos retirados do *site* da *BBC News Health,* com notícias da área de saúde. Naquele momento, o *BBC News Health* foi incluído por trabalharmos com aquele gênero de textos em sala de aula, a título de variedade. Incluímos, também, um terceiro *subcorpus* constituído pelos textos que compunham uma amostra do exame de leitura instrumental, acessados por meio do *site* da instituição que aplica esses exames. Naquele momento, contávamos com 15 textos de cada gênero, a saber: textos do

NEJM (189.939 palavras), textos da *BBC* (20.137 palavras) e textos do exame de leitura instrumental, com 9.853 palavras conforme **Tabela 4**:

Fonte	***NEJM***	***BBC News Health***	**Exames anteriores**
Tipo	Artigos acadêmicos	Artigos de notícias de saúde	Textos de exames anteriores
No. de palavras	189.939	20.137	9.853

Tabela 4: Resumo da composição inicial do *corpus.*

Notamos, então, que a representatividade do *corpus* estava comprometida já que os artigos acadêmicos científicos coletados eram de um periódico americano, enquanto o nosso *corpus* usado para comparação era de notícias sobre saúde, veiculados por um *site* britânico: a *BBC Health News.* No caso, estaríamos trabalhando com duas variantes do inglês, uma americana e outra britânica.

Para que os resultados de um c*orpus* sejam relevantes, não basta compilar qualquer tipo texto para construí-lo (Sinclair,1991). Os textos precisam representar uma área específica. Portanto, consideramos que os resultados da pesquisa seriam mais relevantes se tivéssemos uma amostra com a mesma variante do inglês. Por essa razão, escolhemos continuar nossa coleta mantendo o *NEJM* e buscamos outras fontes da variante americana que pudessem oferecer uma comparação, seguindo o critério de uniformidade de variante do inglês.

2.2.1.2 Balanceamento inicial do corpus

Naquele momento, o corpus ainda não estava balanceado, pois o número de palavras dos três sub*corpora* era desigual. Segundo critérios estabelecidos pela LC, o pesquisador pode determinar o critério de balanceamento que seja mais adequado para sua análise. No caso deste estudo, pudemos notar, já de início, que não poderíamos adotar o critério de número de textos, pois tínhamos 15 textos de cada

gênero e um número de palavras totalmente desigual. A partir dessa observação, adotamos o critério de número de palavras para balancear o *corpus* de estudo.

2.2.2 Critérios redefinidos

Neste item, descrevemos os novos critérios de compilação adotados para o *corpus* de estudo.

2.2.2.1 O *NEJM* e a escolha do *Health News*

Ao redefinirmos nosso critério de composição do *corpus* de estudo, buscamos uma amostragem representativa do inglês americano, como já mencionado. Dessa forma, na nova composição, continuamos com o *NEJM* por ser um periódico respeitado no meio acadêmico e ter um fator de impacto alto.[32] Para a composição do segundo sub*corpus,* buscamos um *site* que divulgasse notícias de saúde. Procuramos um que tivesse o mesmo formato do britânico do *BBC Health News,* porém em versão americana. Esperava-se que a *CNN Health News* tivesse o mesmo formato, observamos que a divulgação das notícias desse *site* possuía características diversas daquelas apresentadas pela *BBC News Health.*

Enquanto o *site* da *BBC News Health* apresenta características de artigos jornalísticos escritos com manchete e texto corrido, na mídia americana encontramos a transcrição das notícias que seriam lidas na televisão. Os textos preparados para a leitura eram adaptações dos artigos advindos de outro *site* sobre saúde: o *Health News.com*[33]. Utilizamos, então, para a composição do segundo *subcorpus* de estudo os textos publicados naquele *site,* porém a coleta dessa vez não foi manual como anteriormente. Utilizamos a ferramenta de compilação automática, *Bootcat*, item 1.5.5.

[32] Estatística sobre a relevância de publicações científicas por meio do Fator de Impacto. Disponível em : http://www. periódicos.capes.gov.br. Acesso em: 23 Nov. 2012.
[33]Disponível em: http://healthnews.com

2.2.2.2 A compilação do *WebMD*

O terceiro *subcorpus*, o *WebMD,* foi compilado a partir da busca de textos que tivessem como objetivo informar o público leigo sobre assuntos de saúde. É importante mencionar que esse *site* não é composto somente por textos de divulgação, mas também publica artigos acadêmicos, na seção *Doctors*, que só pode ser acessado após cadastramento.

Devido às características apresentadas, consideramos esses artigos de divulgação uma amostragem adequada, já que nosso propósito era pesquisar as diferenças entre os gêneros destinados ao público em geral e os artigos de pesquisa científica. Podemos notar pela própria apresentação da publicação (cf. **Figura 8**) que se destina a um público não especializado. Traz fotos e sua organização é como a encontrada em revistas. Além isso, utiliza um *slogan* com características publicitárias com o intuito de chamar a atenção: *Better information. Better health.* O *site* apresenta ainda: vídeos, *slides*, *quizzes* e aconselhamentos.

Figura 8:Tela da página do *site WebMD*

Ao clicar em *WebMD Home* temos acesso a duas seções: *WebMD Feature* e *WebMD Health News.* Como vínhamos trabalhando com notícias sobre saúde, escolhemos somente os artigos da seção *WebMD Health News,* que são revisados por autoridades médicas, como no caso do artigo acima, observamos o nome do profissional e seu título: *reviewed by Louise Chang, MD*.

A compilação foi iniciada manualmente e, mais tarde, com o intuito de obter um número de palavras semelhante ao dos outros dois *subcorpora*, completamos esse número por meio da busca utilizando a ferramenta *BootCat.* Aqui, salientamos uma das grandes vantagens da compilação de corpus por meio de uma ferramenta automática: o *BootCat* elimina automaticamente todas as figuras e outros itens que não interessam à pesquisa e armazena os textos em *txt*., que é o formato aceito pelas ferramentas de análise *WST* e *AntConc.* Esse processo de compilação facilita sobremaneira a composição de um *corpus ad hoc,* seja para uso em sala de aula ou para tradutores.

2.2.3 A compilação manual

Neste item relatamos somente o processo de compilação a partir de textos encontrados na *web.*

1º passo: Identificação dos textos adequados para a pesquisa.

2º passo: Verificação da possibilidade de copiá-los, pois, como já dissemos, alguns textos são protegidos e restritos a assinantes.

3º passo: Copiar, colar e salvar em dois arquivos: um em *Word* e outro em *txt.*

Nesse processo de compilação, a dificuldade foi encontrar textos que estivessem disponíveis para cópia.

2.2.4 A compilação por meio do *BootCat*

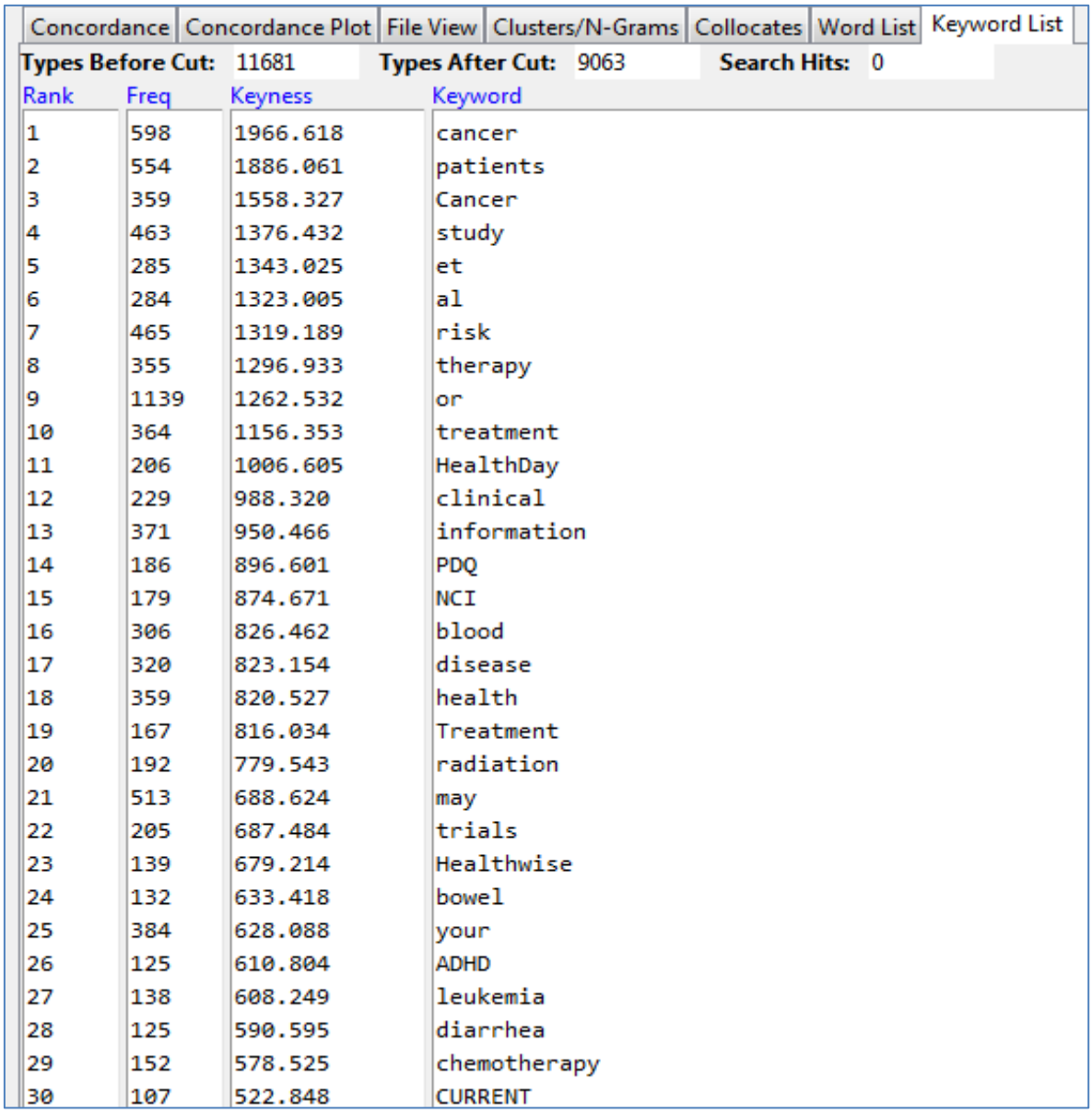
Conforme mencionado anteriormente, o *BootCat* é uma interface que serve como guia na criação de *corpora* simples a partir da *web.* Uma das grandes

vantagens dessa ferramenta é o fato de ser gratuita e de o cadastramento ser relativamente fácil. Segundo Baroni e Bernardini (2004), *corpora* representativos podem ser montados com poucas "sementes iniciais" (*seeds*) (entre 5 e 15), ou seja, palavras de um domínio especializado.

Por sugestão da autora (Bernardini, 2010), para se obter as sementes iniciais, o pesquisador deve compilar cerca de 10 textos de sua área de interesse, copiar, colar o arquivo e salvar em *txt.* Em seguida, deve-se gerar uma lista de palavras e, ao compará-la a um *corpus* de referência, obter uma lista de palavras-chave. Desse modo, o pesquisador obterá os itens lexicais da área a ser estudada e que comporão as sementes para o início da compilação do *corpus* específico. A lista também pode ser intuitiva, isto é, se o pesquisador já tiver conhecimento do léxico que deseja utilizar como sementes.

A **Figura 9** apresenta uma lista de palavras-chave obtidas a partir de 10 textos coletados do *site WebMD*. O *corpus* de referência utilizado foi o de um *corpus* jornalístico, disponibilizado no Projeto Comet[34], composto por textos de diferentes jornais americanos.

[34] Projeto Comet: Corpus Multilingue para o ensino e tradução.



Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Types Before Cut: 11681 **Types After Cut:** 9063 **Search Hits:** 0

Rank	Freq	Keyness	Keyword
1	598	1966.618	cancer
2	554	1886.061	patients
3	359	1558.327	Cancer
4	463	1376.432	study
5	285	1343.025	et
6	284	1323.005	al
7	465	1319.189	risk
8	355	1296.933	therapy
9	1139	1262.532	or
10	364	1156.353	treatment
11	206	1006.605	HealthDay
12	229	988.320	clinical
13	371	950.466	information
14	186	896.601	PDQ
15	179	874.671	NCI
16	306	826.462	blood
17	320	823.154	disease
18	359	820.527	health
19	167	816.034	Treatment
20	192	779.543	radiation
21	513	688.624	may
22	205	687.484	trials
23	139	679.214	Healthwise
24	132	633.418	bowel
25	384	628.088	your
26	125	610.804	ADHD
27	138	608.249	leukemia
28	125	590.595	diarrhea
29	152	578.525	chemotherapy
30	107	522.848	CURRENT

Figura 9: Palavras-chave obtidas a partir de 10 textos do *WebMD.*

Como é possível observar, a lista de palavras é composta por várias palavras que indicam um léxico que se refere à área de saúde. Exemplos: *cancer, patients, study, risk*, citando alguns.

Outra grande vantagem dessa ferramenta é que ela permite que o domínio da *web* seja determinado pelo pesquisador . É possível fazer buscas somente nos *sites* de seu interesse. Em nosso caso, por exemplo, como já sabíamos de antemão que iríamos construir um *subcorpus* com textos a partir do *Health.com,* o domínio

específico foi escolhido e como resultado, só obtivemos *URLs* advindas daquele *site* (cf. **Figura 10**, canto superior direito).

O passo seguinte foi verificar se as possibilidades apresentadas traziam textos que pudessem compor o *corpus*. Essa escolha é determinada pelo pesquisador (cf. **Figura 10**, no retângulo marcado em vermelho).

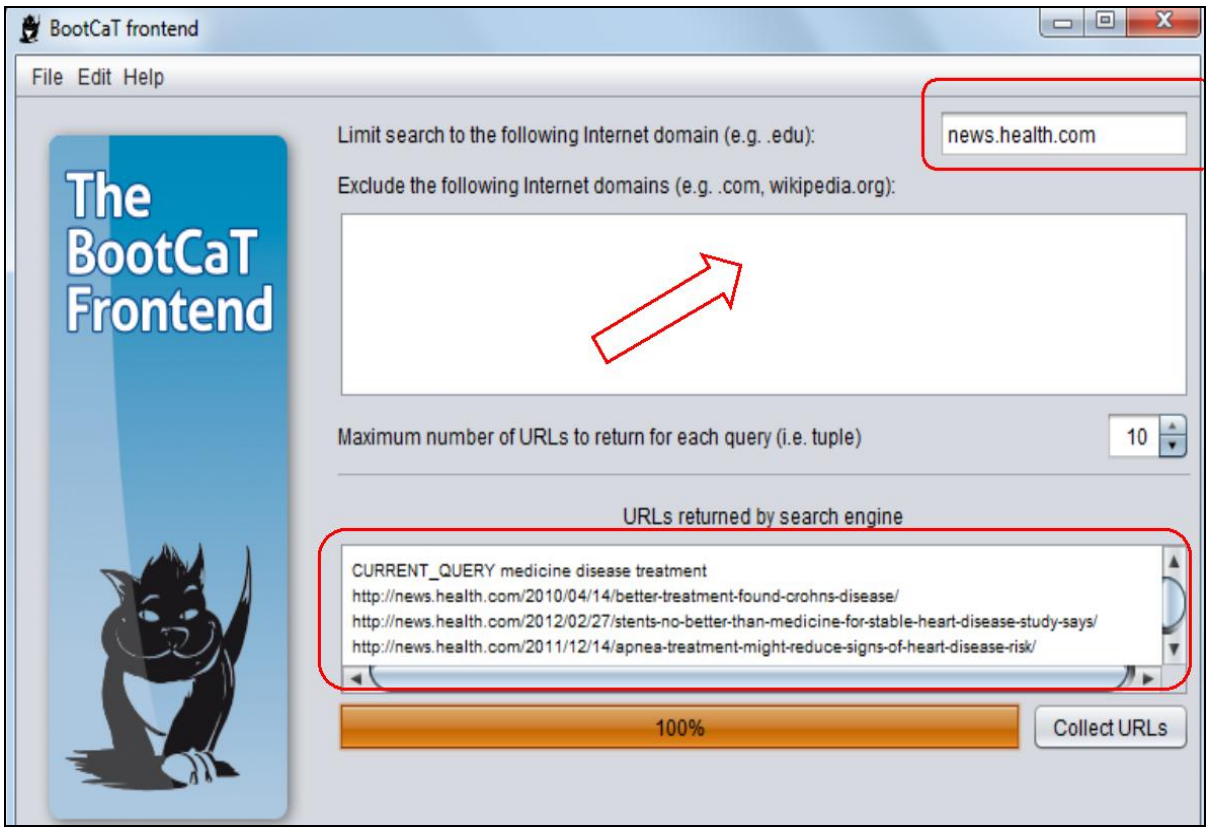


Figura 10: Domínio escolhido (canto superior direito) e, abaixo, as *URLs* obtidas a partir das sementes.

A terceira e grande vantagem da utilização do *BootCat* para a compilação de um *corpus* é que, diferentemente de um *corpus* compilado manualmente, com essa ferramenta não é necessário fazer buscas texto a texto na *web.* É possível obter uma grande quantidade de *URLs* de uma só vez. Como podemos observar na **Figura 11**, obtivemos 62 *URLs* em uma iteração. Iteração, segundo Baroni e Bernardini (2004), é o processo de busca por textos desde a entrada das sementes na ferramenta até a obtenção dos textos necessários para um estudo. Descartamos

a primeira opção por se tratar da página inicial do *site* e mantivemos as demais, ficando, pois, com 61 textos. Dessa forma, dependendo do número de palavras ou número de textos que o pesquisador/professor necessita, será preciso repetir o processo até que o mesmo obtenha o número de textos que julgar suficiente para a composição de seu corpus de estudo. Segundo Bernardini (2011) , é possível conseguir um número razoável de *URLs* com uma ou duas iterações, como no caso dessa compilação.

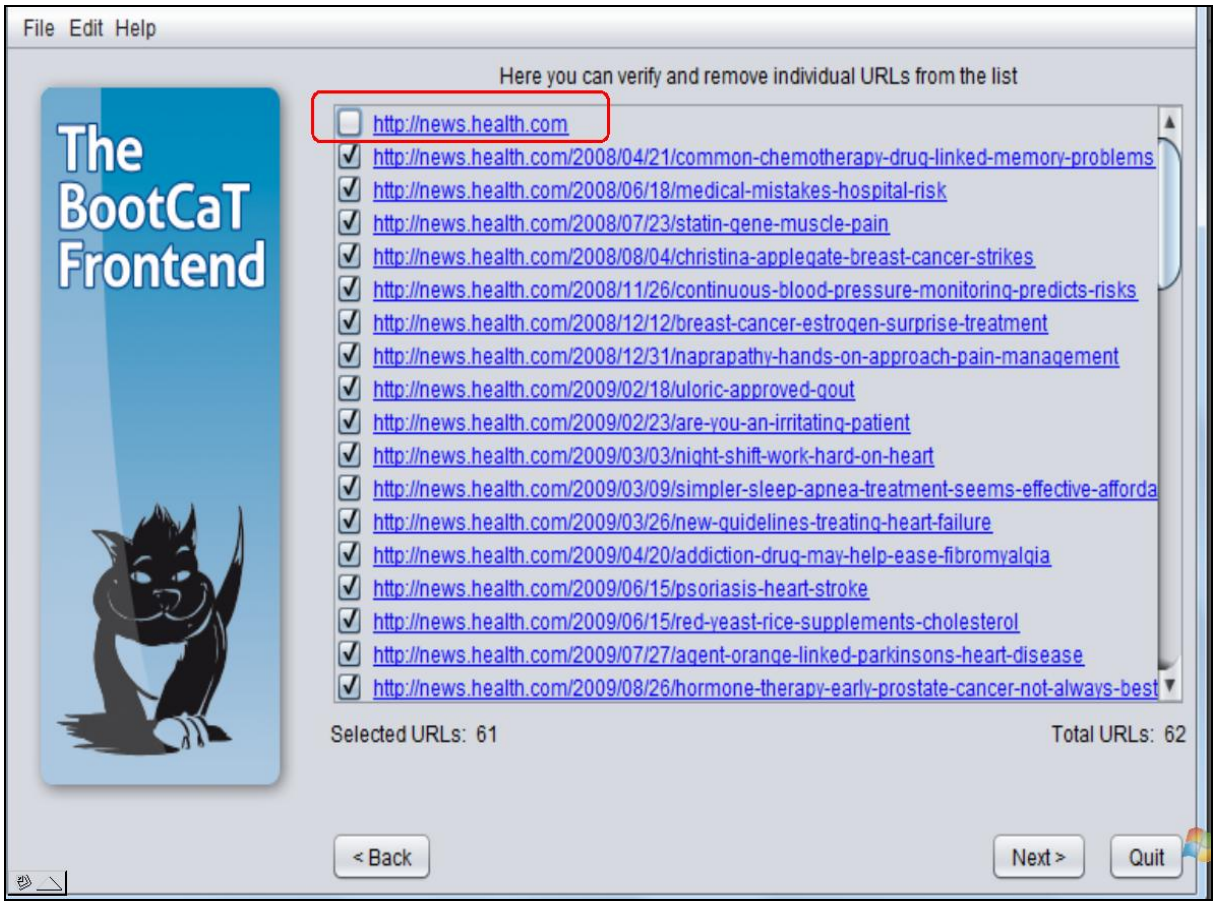


Figura 11: As *URLs* obtidas para verificação e possível remoção.

2.2.5 A questão do acesso aos textos

Discutimos neste item as possibilidades de acesso aos textos escolhidos.

2.2.5.1 O sub*corpus NEJM*

Ao iniciarmos a construção do *corpus* acadêmico, deparamos com a questão do acesso. O *site* comercializa suas publicações mais recentes, portanto nem todos os artigos estão disponíveis para o acesso público. Para que seja possível obter

certos artigos é preciso que se compre aquela publicação. Porém, é possível conseguir publicações anteriores por meio de cadastramento. Utilizamos, então, os textos que já estavam em domínio público. Essa coleta não obedeceu a um critério cronológico, mas todos os textos selecionados foram publicados a partir do ano 2000.

2.2.5.2 O sub*corpus Health News*[35]

O sub*corpus* com artigos da *Health.com* foi compilado com duas iterações do *Bootcat* e, como já possuíamos as "sementes" de pesquisas anteriores, facilitou sua compilação.

2.2.5.3 O sub*corpus WebMD*

O sub*corpus WebMD* foi coletado manualmente em sua maior parte e com duas iterações no *Bootcat* para completar o número de palavras que consideramos adequado a nossa análise. As "sementes" foram as mesmas utilizadas para compor o *Health News*. Um aspecto positivo para professores e pesquisadores é que não há restrição de acesso a esse *site.* Todos os artigos estão totalmente disponíveis (com exceção da seção *Doctors*), já que são textos de divulgação para o público leigo. Uma vez inscritos, os usuários podem receber, por meio de correio eletrônico, sugestões de novos temas relacionados ao assunto de seu interesse, o que facilita a tarefa do professor na escolha de textos atualizados a serem incluídos em suas aulas, formando um arquivo de sugestões de textos a serem trabalhados sem consumir muito tempo.

Depois de redefinidos os critérios, a composição do *corpus* ficou conforme apresentado na **Tabela 5:**

Sub*corpus*	**Número de palavras**

[35] Disponível em: http://www.healthnews.com . Último acesso 23.01.2013.

NEJM	124.445 (manual)
HEALTH NEWS	128.077 (semiautomático)
WebMD	125.815 (manual e semiautomático)

Tabela 5: Composição atual do *corpus* de estudo.

2.3 A comparação das listas e as palavras-chave obtidas

Neste item, apresentamos a metodologia de trabalho por meio da comparação dos três *subcorpora (NEJM, Health News e WebMD*) que compõem nosso *corpus* de estudo. A manipulação de dados é feita de forma semiautomática, com o software *WST* para a comparação inicial das listas de palavras, e da ferramenta de análise *AntConc* para a análise das linhas de concordância e *clusters.*

O levantamento dos dados a serem analisados foi dividido em três partes:

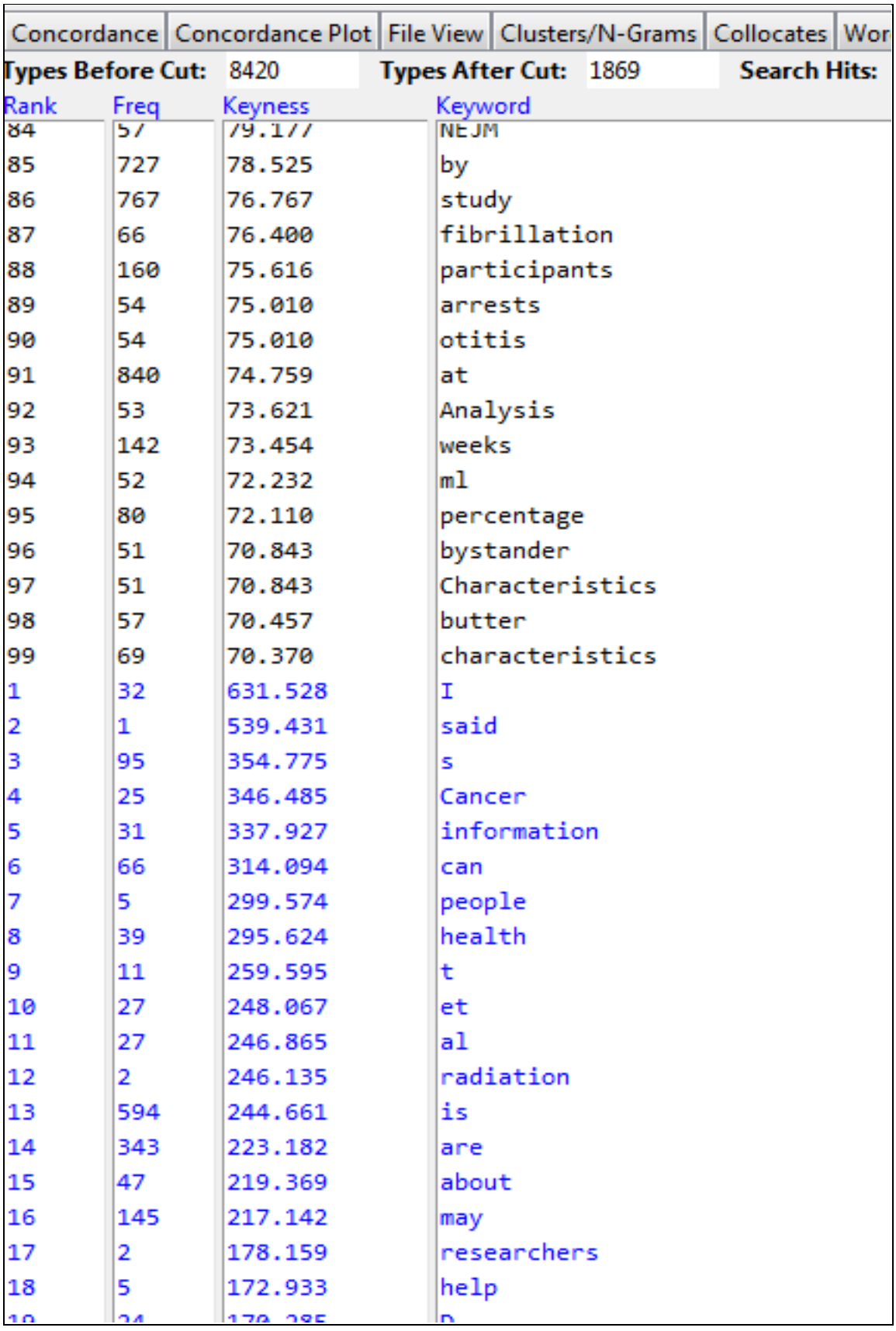
1) Comparação do *NEJM com o Health News .*
2) Comparação do *WebMD com o NEJM.*
3) Comparação do *Health News com o WebMD*

O primeiro passo foi fazer o levantamento de uma lista de palavras de cada um deles. Uma lista de palavras *(Wordlist)*, como já definido anteriormente no item 1.5.4.1, é uma lista que contém todas as palavras dos textos escolhidos, e suas frequências. Segundo Scott e Tribble (2006), listas de palavras, normalmente, são o ponto de partida para a compreensão das características do texto.

Como o objetivo do nosso estudo é verificar as possíveis diferenças entre os três diferentes gêneros, começamos pela utilização da função chamada "*Compare 2 lists*", disponível no *WST*. De acordo com o manual do próprio autor (Scott, 2010), esse utilitário permite comparar textos de diferentes fontes. Dessa comparação resultam as palavras-chave. Esse levantamento é feito de forma automática e obedece a critérios estatísticos determinados pela ferramenta. Essas palavras aparecem organizadas de acordo com sua frequência, isto é, o número de vezes em que elas ocorrem em um *corpus* e/ou texto, bem como sua chavicidade (*Keyness)*. Para ser considerada chave ela precisa ser significativamente mais frequente

(palavras-chave positivas) ou menos frequente (palavras-chave negativas) na comparação entre dois *corpora*.

As ferramentas também podem apresentar as palavras-chave negativas, mediante escolha. Palavras-chave negativas são aquelas que quando comparadas a *corpus* de referência são estatisticamente menos frequentes no corpus de estudo. Essas palavras aparecem com uma cor diferente nas ferramentas de análise *WST* e *AntConc*. As que foram encontradas na comparação do *NEJM* com *HN* estão demonstradas na **Figura 12.**



Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Wor

Types Before Cut: 8420 **Types After Cut:** 1869 **Search Hits:**

Rank	Freq	Keyness	Keyword
84	57	79.177	NEJM
85	727	78.525	by
86	767	76.767	study
87	66	76.400	fibrillation
88	160	75.616	participants
89	54	75.010	arrests
90	54	75.010	otitis
91	840	74.759	at
92	53	73.621	Analysis
93	142	73.454	weeks
94	52	72.232	ml
95	80	72.110	percentage
96	51	70.843	bystander
97	51	70.843	Characteristics
98	57	70.457	butter
99	69	70.370	characteristics
1	32	631.528	I
2	1	539.431	said
3	95	354.775	s
4	25	346.485	Cancer
5	31	337.927	information
6	66	314.094	can
7	5	299.574	people
8	39	295.624	health
9	11	259.595	t
10	27	248.067	et
11	27	246.865	al
12	2	246.135	radiation
13	594	244.661	is
14	343	223.182	are
15	47	219.369	about
16	145	217.142	may
17	2	178.159	researchers
18	5	172.933	help
19	24	170.285	D

Figura 12: As últimas palavras-chave do *NEJM* e o início das palavras negativas em comparação ao *HN.*

Em nosso *corpus* de estudo, chama a atenção o fato de as duas primeiras palavras-chave negativas serem o pronome *I* e o verbo *said*, que são características da linguagem oral, como discutiremos no capítulo 3, destinado aos resultados.

O teste estatístico utilizado para a obtenção das palavras-chave em nosso estudo foi o logaritmo de verossimilhança (*log-likelihood).* Tanto o WST como o *AntConc* possibilitam a escolha de dois testes estatísticos para o levantamento das palavras-chave: o *log-likelihood,* , que é a opção padrão e o teste utilizado nesta pesquisa, e o qui-quadrado (*chi-square*), que é definido por Vasilévski (2007: 208) como:

> O teste estatístico qui-quadrado baseia-se na comparação das frequências observadas, tipicamente apresentadas em uma tabela de contingência, e as frequências esperadas caso a associação de palavras em estudo fosse apenas uma consequência do acaso.

Mencionamos aqui as duas ferramentas (*WST* e *AntConc*)*,* pois iniciamos o trabalho com o *WST* e ao longo do processo fomos fazendo um trabalho paralelo e comparativo das duas ferramentas a título de experimentação. A razão dessa conduta foi verificar se os resultados apresentados variavam de maneira significativa. Ao constatarmos que isso não acontecia, optamos por usar o *AntConc* por entendermos ser essa ferramenta mais indicada para uso em sala de aula. Apesar de o *WST* ser uma ferramenta com mais opções e detalhamentos estatísticos, o *AntConc* apresenta vantagens, já que é uma ferramenta disponível gratuitamente com uma interface mais simples e, consequentemente, mais atraente em termos visuais.

2.3.1 A comparação das listas de palavras do *NEJM* e do *Health News*

Neste item apresentamos as 50 primeiras palavras-chave obtidas a partir da observação da lista comparativa entre os dois *subcorpora (NEJM* e *Health News)* por meio do *WST.* Para efeito de organização, as palavras-chave encontradas foram divididas em três categorias: verbos, adjetivos e substantivos.

2.3.1.1 Os verbos

Os verbos *was e were* foram os primeiros em frequência (1464 e 1437 ocorrências respectivamente) no *NEJM,* porém a chavicidade de *were* é maior que a de *was,* indicando uma marca do discurso científico. Uma chavicidade alta (cf.**Tabela 6**) já demonstra um indício de diferença entre os dois gêneros comparados (acadêmico e o de divulgação). Ao observarmos seus colocados veremos que existe uma forte presença de voz passiva. Segundo Biber (1988), uma das características da prosa acadêmica é o uso da voz passiva, o que indica um distanciamento entre o autor e o leitor. Outra característica citada pelo autor é o uso dos verbos no passado nesse tipo de escrita. Corroborando essa afirmação encontramos *had* (linha 21) e outros verbos terminados em *–ed* que poderiam ser tanto verbos no passado quanto adjetivos: *received* (linha 30)*, observed* (linha 38) e *performed* (linha 39).

ORDEM	FREQUÊNCIA	CHAVICIDADE	PALAVRA-CHAVE
1	1437	961.159	Were
2	1464	742.262	Was
3	7755	730.999	The
4	694	610.191	Group
5	3585	420.763	In
6	6173	373.802	Of
7	342	361.698	Table
8	249	345.880	Figure
9	310	299.122	Events
10	215	298.652	Appendix
11	215	298.652	Supplementary
12	188	261.147	Nevirapine
13	320	239.848	P
14	324	227.676	Per
15	284	222.221	Analysis
16	151	209.751	Fev
17	191	209.286	Adverse
18	149	206.972	Protocol
19	354	199.968	rate

20	168	184.269	Baseline
21	558	180.881	Had
22	219	177.535	Placebo
23	140	176.215	Analyses
24	2034	166.981	With
25	115	159.744	Interleukin
26	152	154.189	Cardiac
27	243	147.856	Data
28	118	146.334	Coli
29	111	144.143	Interval
30	176	139.829	Received
31	155	138.126	End
32	278	137.317	Among
33	148	136.943	Ci
34	997	129.545	Patients
35	93	129.184	Arrest
36	92	127.795	Virologic
37	159	120.234	Mortality
38	114	119.012	Observed
39	139	118.814	Performed
40	197	116.017	Groups
41	83	115.293	Amoxicillin
42	82	113.904	Ems
43	94	113.897	Strain
44	372	113.378	After
45	93	112.551	Confidence
46	81	112.515	Clavulanate
47	291	110.671	Between
48	204	106.326	Primary
49	135	104.539	Vaccine
50	119	101.670	Efficacy

Tabela 6: Lista de palavras-chave do sub*corpus NEJM* em comparação ao sub*corpus HN.*

2.3.1.2 Os substantivos

As primeiras palavras com chavicidade alta observadas depois dos verbos *were* e *was* foram aquelas específicas da área médica, aqui apresentadas na mesma sequência da lista: *group, table, figure, events, analysis, interval, patients,*

arrest, mortality, confidence e outros. É muito interessante notar na **Tabela 6** que *neviparine* (linha 12), *interleukin* (linha 25), *amoxicillin* (linha 41), *clavulanate* (linha 46), que correspondem a nomes de medicamentos e/ou princípios ativos, não tiveram nenhuma ocorrência no *corpus* de divulgação *Health.com.* (cf. **Tabela 8**), demonstrando serem essas palavras uma característica do *corpus* de estudo acadêmico, corroborando a afirmação de Bowker (2002), para quem ao gerarmos uma lista de palavras-chave é possível identificar as peculiaridades de um texto.

2.3.1.3 Os adjetivos

Os adjetivos encontrados dentre as 50 primeiras palavras-chave foram*: adverse* (linha 17), *cardiac* (linha 26), *virologic* (linha 36.), que serão analisados no capitulo 3. Destacamos também as palavras em *–ed* que tanto poderiam ser verbos como adjetivos: *received* (linha 31), *observed* (linha 38) e *performed* (linha 39), que não serão analisados individualmente, mas que serão mencionados quando discutirmos os colocados de *was* e *were.*

2.3.2 Comparação do *WebMD* e *NEJM*

O mesmo processo foi utilizado na comparação das listas de palavras do *WebMD* e do *NEJM*. Foram observados os verbos, os substantivos e os adjetivos (cf. **Tabela 7**). Listamos, também, os pronomes que apareceram nessa comparação e que não haviam aparecido anteriormente, na primeira lista comparativa: *NEJM* x *HN* (cf. **Tabela 6**).

ORDEM	FREQ.	CHAVICIDADE	PALAVRA-CHAVE
1	739	1.217.910	Says
2	585	648.804	Heart
3	402	614.359	People
4	317	522.432	Md
5	292	459.592	Researchers
6	402	443.924	Health
7	280	439.983	News
8	283	396.441	He

9	219	360.923	Your
10	434	358.075	They
11	337	353.276	New
12	212	349.387	You
13	205	326.354	University
14	487	307.724	It
15	384	296.295	Blood
16	332	265.802	Can
17	642	265.676	Risk
18	160	252.686	Medicine
19	372	251.009	Their
20	208	250.176	Pressure
21	282	249.172	About
22	501	238.068	Disease
23	147	231.431	She
24	139	229.079	Tells
25	418	213.983	May
26	127	209.302	American
27	147	204.060	Attack
28	118	194.470	Display
29	129	194.211	Say
30	938	185.964	Is
31	479	185.875	Women
32	168	182.955	Research
33	111	182.934	Top
34	110	181.286	Kids
35	150	175.513	Calcium
36	106	174.693	Get
37	634	174.577	Are
38	112	174.291	Journal
39	104	171.397	Professor
40	113	168.367	What
41	245	166.471	Men
42	137	164.271	Problems
43	99	163.157	Don
44	180	159.046	Diabetes
45	96	158.213	Researcher
46	385	156.899	But
47	116	155.294	Help

48	512	151.594	More
49	108	148.096	Sources
50	89	146.677	Attacks

Tabela 7: Listas de palavras-chave do *WebMD*[36] em comparação ao *NEJM.*

2.3.2.1 Os verbos

Notamos que os dados apresentados nessa segunda lista (cf.**Tabela 6**) são muito diferentes dos que aparecem na primeira (cf.**Tabela 5**), que compara o *subcorpus* acadêmico (*NEJM*) e o de divulgação (*HN*). No topo da lista por chavicidade, aparece o verbo *says,* indicando que os sujeitos estão na terceira pessoa do singular, inferência possível mesmo sem a observação das linhas de concordância, o que nos dá indícios de uma diferença do gênero acadêmico e do de divulgação. Outro verbo-chave que aparece na terceira pessoa do presente é *tells* (linha 24)*. Say* aparece na linha 29.

Os modais *can* (linha 16), com 332 ocorrências, e *may* (linha 25), com 418 ocorrências, também aparecem nessa lista, porém não serão objeto deste estudo, apesar de serem dignos de nota, por apontarem para características da linguagem falada. Segundo Conrad e Biber (2009), esses modais são comuns tanto na língua falada quanto na escrita, porém não tão frequentes, no discurso acadêmico científico, o que é corroborado pelo número de ocorrências nas listas de palavras-chave.

Os outros verbos que aparecem na lista são: *is, get, are, help,* que também não serão analisados neste trabalho.

2.3.2.2 Os substantivos

Como podemos observar, o primeiro substantivo em ordem decrescente de chavicidade é *heart,* que dá indicações do tema mais recorrente desse sub*corpus:* esses artigos tratam de doenças do coração. Esse substantivo foi seguido por *people* e *researchers.*

[36] Os resultados apresentados sempre se referem ao primeiro sub*corpus* citado.

2.3.2.3 Os adjetivos

A busca por adjetivos nessa segunda lista comparativa (*WebMD* e *NEJM*) demonstra que, dentre as 50 primeiras palavras-chave, encontramos somente dois deles*: new* (linha.11) e *American* (linha 26), o que normalmente não causa problemas aos aprendizes brasileiros.

2.3.3 Comparação das listas do *Health News* e *WebMD.*

No terceiro momento da investigação, ao contrastarmos as listas de palavras contidas no sub*corpus Health News* e no *WebMD* (cf.**Tabela 8**), mais uma vez, destacamos os verbos, os substantivos e os adjetivos, e pudemos verificar a semelhança de estilo com o aparecimento de pronomes pessoais e possessivos da mesma forma que na lista anterior.

2.3.3.1 Os verbos

Em relação a essa categoria gramatical, em primeiro lugar aparece *said* (linha 20) que, por sinal, é um dos poucos verbos que aparecem nessa lista, seguido de *use* (linha 35). O terceiro verbo encontrado foi *read* (linha 45).

2.3.3.2 Os adjetivos

Os adjetivos encontrados foram: *clinical* (linha16), *American* (linha18), *percent* (linha19), *acute* (linha 20), *lymphoblastic* (linha 33), *esophageal* (linha 36) e *nasal* (linha 38).

2.3.3.3 Os substantivos

Por aparecerem em grande número, vamos listar somente os 10 primeiros substantivos. A **Tabela 8** apresenta: *cancer* (linha 1), *radiation* (linha 3), *therapy* (linha 4), *treatment* (linha 5), *chemotherapy* (linha 6), *leukemia* (linha 7) , *trials* (linha 10), *summary* (linha 13), *lymphoma* (linha 14) e *university* (linha 15).

ORDEM	FREQ.	CHAVICIDADE	PALAVRA-CHAVE
1	981	644.596	Cancer
2	399	292.957	Said
3	247	290.908	radiation
4	373	235.658	therapy
5	531	211.722	treatment
6	177	205.881	chemotherapy
7	166	203.211	leukemia
8	641	179.883	I
9	502	175.322	Your
10	229	168.935	Trials
11	191	161.412	My
12	131	160.365	Adhd
13	130	159.141	Summary
14	124	151.796	Lymphoma
15	137	149.899	University
16	282	148.806	Clinical
17	592	142.145	This
18	111	135.882	American
19	180	134.519	Percent
20	132	131.748	Acute
21	521	129.941	You
22	114	129.636	Autism
23	104	127.313	Lymph
24	103	126.089	Hodgkin
25	125	113.674	Stage
26	520	111.915	Health
27	1145	103.372	Or
28	107	102.796	Tumors
29	145	98.271	Adult
30	79	96.709	Comments
31	85	94.719	obstruction
32	74	90.588	summaries
33	69	84.467	lymphoblastic
34	144	83.311	Bowel
35	308	80.154	Use
36	91	79.928	esophageal

37	65	79.570	Ph
38	71	77.939	Nasal
39	92	76.856	Marrow
40	79	75.659	Am
41	60	73.450	Comment
42	139	71.432	Diarrhea
43	56	68.553	Cavity
44	56	68.553	Society
45	80	67.714	Read
46	78	65.512	Sinus
47	53	64.881	Monday
48	53	64.881	Nodes
49	150	64.371	Cell
50	178	63.940	Surgery

Tabela 8: Lista comparativa de palavras-chave do *HN* e *WebMD.*

2.4 Utilitários das ferramentas de análise

Descrevemos aqui outros utilitários das ferramentas de análise*: a stop list,* as linhas de concordância e os *clusters.*

2.4.1 *Stop list*

A *Stop list* é uma função das ferramentas de análise que permite que sejam omitidos os itens que o pesquisador julgar irrelevantes. Tanto o *WST* quanto o *AntConc* oferecem essa possibilidade.

Ao analisar as palavras-chave de um *corpus* é possível identificar itens que não são de interesse para o estudo e que podem ser excluídos por meio de uma *stoplist* (cf. Bowker e Pearson, 2002). Por captar as *URLs* de forma automática, o *Bootcat* capta também todas as informações escritas. Dessa forma, informações desnecessárias para a pesquisa, como datas, nomes de clínicas e outros itens, demonstrados abaixo*,* acabam aparecendo nas listas de palavras-chave. Por essa razão, utilizamos esse recurso para eliminar as palavras que julgamos não serem relevantes ao resultado de nossa pesquisa, pois faziam parte das informações adicionais do texto. Assim, do sub*corpus HN,* as seguintes palavras foram colocadas em uma *stoplist: Information, et, al, Health, J, PDQ, NCI, Healthwise, all, Dr,*

Information, ALL, Last, Updated, Oncol, Web, Incorporated, Reporter, Clin, Copyright, Med. Já do sub*corpus WebMD: O, r, n, h, rdr, url, l, webmd, current, com, u, t, http.*

Pelo fato do sub*corpus NEJM* ter sido compilado manualmente, não utilizamos uma *stoplist* ao proceder à análise, pois no momento da escolha dos textos pelo processo manual, o pesquisador pode copiar e colar somente as partes do texto que lhe sejam convenientes. Nesse caso, somente o corpo dos artigos foi copiado, portanto não há informação irrelevante.

2.4.2 Linhas de concordância

Vários autores, entre eles Gavioli (2005), Hyland (2004), Anderson e Corbett (2009), afirmam que as linhas de concordância são um poderoso instrumento à disposição da pesquisa e do aprendizado, pois a partir delas é possível observar as características específicas da língua em estudo por meio dos colocados. O termo *collocation* foi utilizado pela primeira vez pelo linguista britânico J.R. Firth para se referir a casos de "coocorrência léxico-sintática, ou seja, palavras que normalmente 'andam juntas'". (Tagnin, 2005:37). Biber et al (1998) também afirmam que a observação dos 'colocados' é o início da investigação dos significados e das variações entre os diferentes tipos de registro: formal e informal, escrito e falado.

Uma das maneiras mais comuns de exploração de um *corpus* é por meio de linhas de concordância. Elas apresentam todas as ocorrências da palavra de busca ou de uma frase de um *corpus* em uma lista de linhas desconectadas do texto, o que permite que o pesquisador perceba padrões de uso da língua que poderiam não ser percebidas por meio de leitura sequencial. Quando lidas na horizontal, apresentam ocorrências de **uso** da língua, e quando lidas na vertical, apresentam evidências do **sistema** da língua (Hyland, 2004). Como demonstrado na **Figura 12**, ao observarmos o co-texto do substantivo *mortality* na horizontal, percebemos que esse item não só tem a função de substantivo nas linhas 1, 4, 5, 6, 9, 10, 11, 12, 13, 15, 16 e 24; como também compõe o grupo de palavras modificadoras de *rates* nas linhas 2, 3, 7, 14, 17, 18, 19, 20, 23, 27 e 28. Se a observarmos na vertical,

podemos perceber um padrão de adjetivo composto por hífen que o modifica. Como por exemplo: *hospital-wide mortality.*

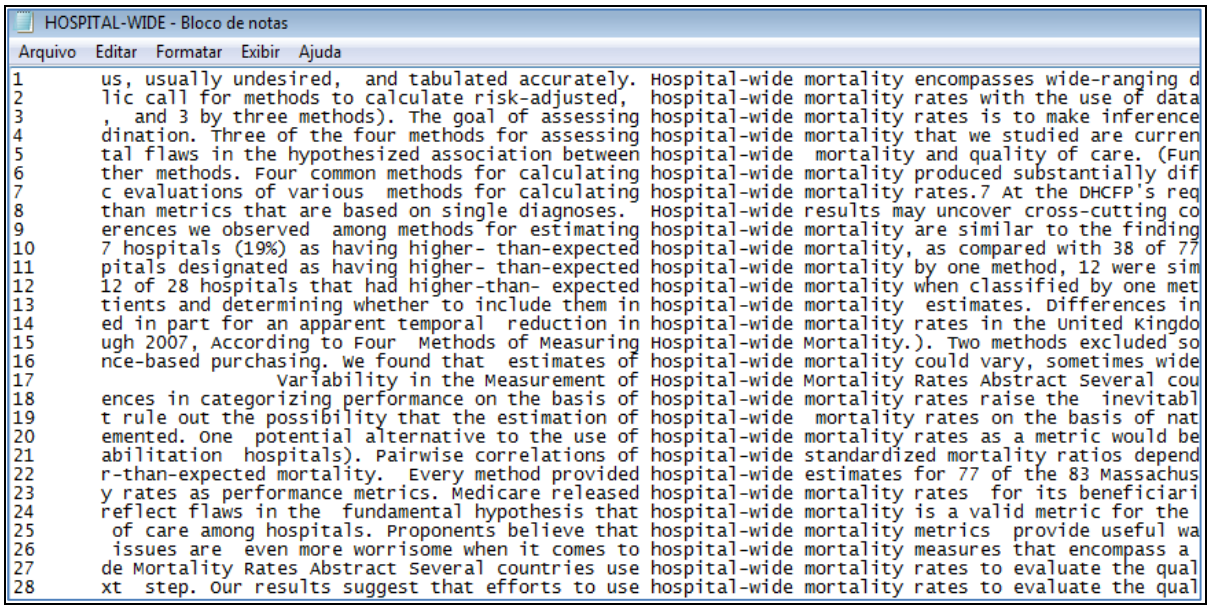


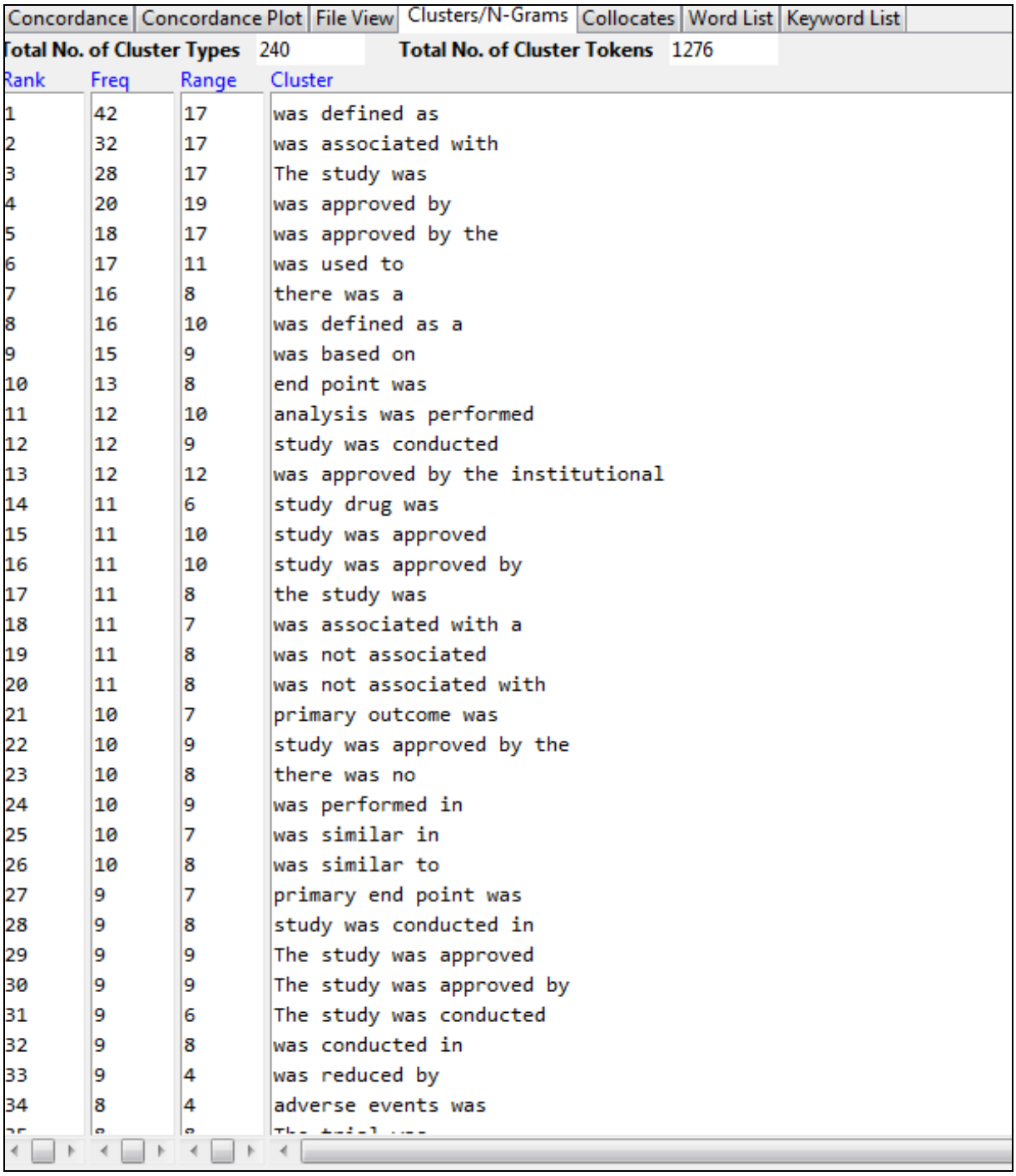
Figura 13: Linhas de concordância com *hospital-wide.*

A fim de buscar as diferenças entre os gêneros: acadêmico e de divulgação, partimos para a observação das linhas de concordância dos primeiros verbos, substantivos e adjetivos encontrados entre as 50 primeiras palavras-chave de cada um dos três *subcorpora.* Os itens pesquisados foram: *was/were, says, said, tells,* adjetivos e substantivos cognatos, por partirmos do princípio (*schema theory*) que o aprendizado se processa a partir da identificação de itens que sejam familiares aos aprendizes. O terceiro passo foi analisar o comportamento de *likely* a partir dos *clusters* obtidos por ser uma estrutura que não encontra equivalência em português e normalmente causa problemas de entendimento para os aprendizes de inglês por ser confundida com o verbo *like.*

2.4.3 *Clusters* (agrupamentos de palavras)

Na verdade, os *clusters* resumem os resultados gerados nas linhas de concordância. A título de exemplo, colocamos aqui uma tela obtida com a utilização do *AntConc* em que podemos observar os padrões de *was.* Apresentamos aqui (cf.

Figura 14) somente os primeiros 34 *clusters* por ser essa a apresentação possível sem a rolagem da tela:



Rank	Freq	Range	Cluster
1	42	17	was defined as
2	32	17	was associated with
3	28	17	The study was
4	20	19	was approved by
5	18	17	was approved by the
6	17	11	was used to
7	16	8	there was a
8	16	10	was defined as a
9	15	9	was based on
10	13	8	end point was
11	12	10	analysis was performed
12	12	9	study was conducted
13	12	12	was approved by the institutional
14	11	6	study drug was
15	11	10	study was approved
16	11	10	study was approved by
17	11	8	the study was
18	11	7	was associated with a
19	11	8	was not associated
20	11	8	was not associated with
21	10	7	primary outcome was
22	10	9	study was approved by the
23	10	8	there was no
24	10	9	was performed in
25	10	7	was similar in
26	10	8	was similar to
27	9	7	primary end point was
28	9	8	study was conducted in
29	9	9	The study was approved
30	9	9	The study was approved by
31	9	6	The study was conducted
32	9	8	was conducted in
33	9	4	was reduced by
34	8	4	adverse events was

Figura 14: Tela do *AntConc* mostrando os *clusters* obtidos a partir do *subcorpus NEJM* com o verbo *was.*

Por meio desse utilitário, as pesquisas dos *clusters* podem ser ordenadas por ordem de frequência, pelo começo ou pelo fim da palavra de busca. É possível também selecionar número de palavras máximo e mínimo de cada *cluster*. Como demonstrado na **Figura 15**, podemos observar que o tamanho escolhido dos clusters foi: mínimo de 3 palavras e máximo de 5 palavras e que ocorressem no mínimo três vezes. Não ordenamos o utilitário por posição da palavra de busca, que é outra possibilidade.

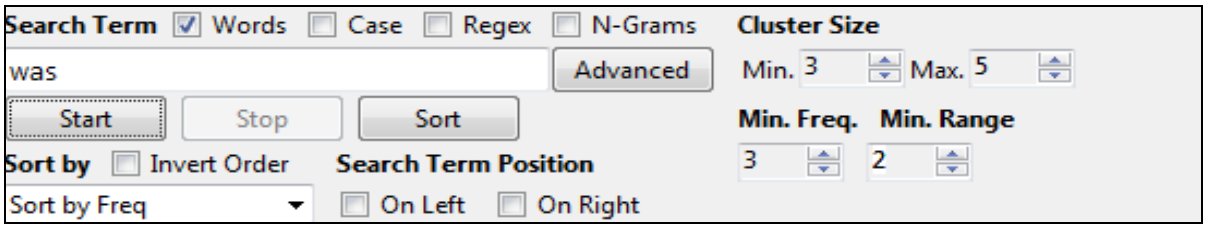


Figura 15: Continuação da Fig. 14 demonstrando as escolhas feitas.

Como mencionado no início deste capítulo, a pesquisa partiu de uma necessidade apresentada pelos alunos e de um experimento com a ferramenta *Lextutor.* Decidimos, a partir daí, pesquisar as diferenças entre os textos acadêmicos e os de divulgação. O primeiro critério que norteou os passos foi partir do princípio apresentado pela *schema theory* (do conhecido para o desconhecido) e observar as diferenças entre os *corpora*, estabelecendo como ponto de corte as 50 primeiras palavras-chave encontradas. Dentre elas, os verbos *was/were,* os substantivos *interval, patients, mortality, confidence* e *efficacy,* os adjetivos cognatos *adverse, cardiac* e *virologic*, bem como o comportamento de *likely*.

CAPÍTULO 3. RESULTADOS

Neste capítulo discutiremos as palavras selecionadas dentre as 50 palavras-chave resultantes da comparação dos três sub*corpora NEJM, HN* e *WebMD.* Apresentaremos a análise dos colocados dos verbos *was, were, say, says, said* e *tells,* bem como os colocados dos principais adjetivos e substantivos cognatos presentes nessas listas a partir do gênero acadêmico científico.

Como mencionado no capítulo 2, os primeiros itens com chavicidade alta na lista de palavras-chave do sub*corpus* acadêmico foram os verbos *was* e *were*. A fim de obtermos maiores dados do co-texto desses verbos, observamos seus colocados. A pesquisa foi-se delineando à medida que novos dados eram observados, o que corrobora o princípio da "serependidade"[37] apresentado por Higgins (1991, apud Partington 1998).

A relação desse princípio com a pesquisa baseada em *corpus* é colocada de maneira muito clara por Partington (1998:1):

> O pesquisador tem uma intuição sobre a língua e verifica essa intuição no *corpus*. Esse processo de verificação geralmente sugere outros caminhos pelos quais a pesquisa vai-se enveredar, quase sempre inesperados no início do processo (esse é o chamado principio da serendipidade, cunhado por Higgins 1991: 2).[38]

Para o autor, "a intuição e os dados caminham lado a lado[39]". O que significa que, normalmente, o pesquisador depara com informações das quais não suspeitava no início do trabalho e vai descobrindo novas possibilidades à medida que a pesquisa avança. Esse foi o percurso deste estudo: observamos os colocados de uma palavra e encontramos outras combinações que apresentaram padrões de utilizações típicas do gênero em estudo. Portanto, os achados aqui descritos podem

[37] Serependidade , também serendipismo, é um anglicismo que se refere às descobertas afortunadas feitas, aparentemente , por acaso. Palavra criada pelo escritor britânico Henry Walpole (1754) a partir do conto persa infantil "Os três príncipes de Serendip".

[38] No original*: "A researcher has an intuition about the language, checks this against the data the corpus provides, and this checking process frequently suggests other avenues of research to be taken, often unsuspected at the start of the process (the so-called "serendipity" principle , Higgins 1991:2).*

[39] No original: *Intuition and data collection work hand in hand*:"

ser de grande valia no momento em que o professor fizer suas escolhas no sentido de priorizar o conteúdo a ser trabalhado com os alunos.

3.1 Colocados de *was*

Os padrões léxico-gramaticais encontrados para *was* foram: (i) *was* + verbos regulares terminados *em –ed,* (ii) *was* + *Adv* $_{-ly}$ + *V* $_{–ed}$, (iii) *was* + V $_{[irreg]}$, e (iv) *was* +*V*$_{-ing}$. A **Tabela 9** apresenta o número de ocorrências dos diferentes padrões identificados para *was* nos três sub*corpora*:

	NEJM	***Health News***	***WebMD***
***(i)was + V*$_{-ed}$**	647	103	96
***(ii)was + Adv*$_{-ly}$ *+ V*$_{-ed}$**	23	10	6
(iii)was +V $_{[irreg.]}$	28	22	24
***(iv)was + V*$_{-ing}$**	7	20	12

Tabela 9: Número de ocorrências dos diferentes padrões identificados para *was* nos três sub*corpora.*

O **Gráfico 1** apresenta o número de ocorrências dos padrões mais frequentes para *was* nos 3 sub*corpora* presentes na **Tabela 9** para melhor visualização dos dados.

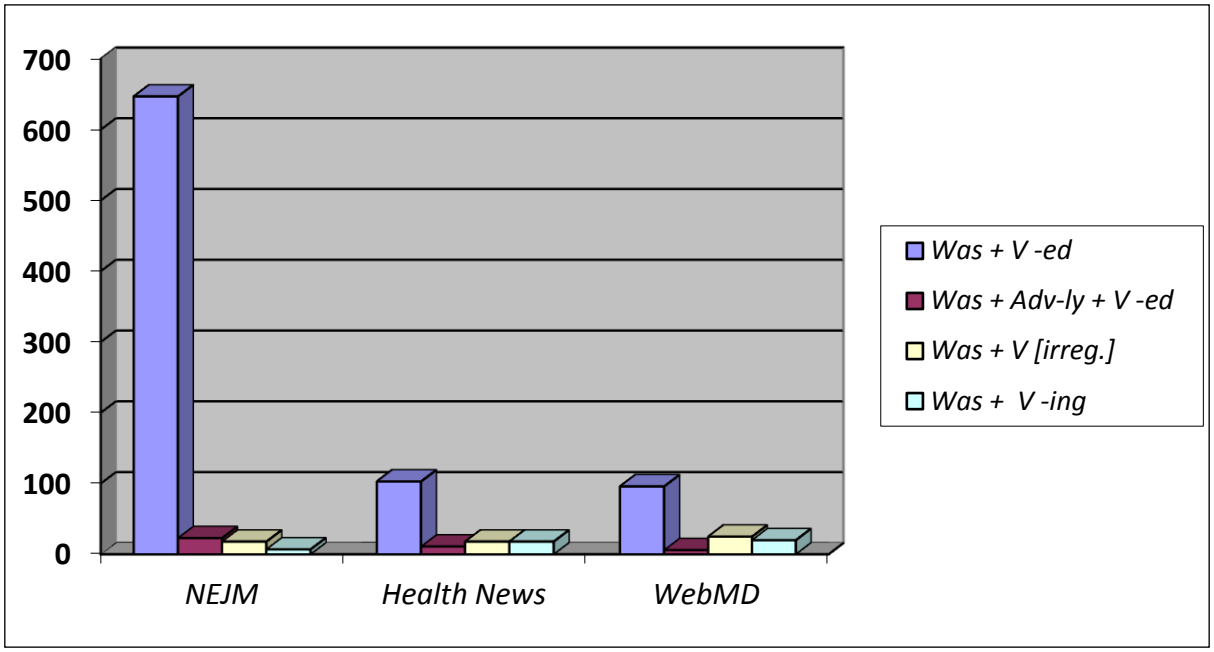


Gráfico 1: Número de ocorrências dos padrões mais frequentes para *was* nos 3 sub*corpora.*

3.1.1 *Was + V -ed*

Como demonstrado na **Tabela 9**, o padrão formado por *was* +V*-ed*, que corresponde à forma da voz passiva, foi o mais frequente no *NEJM*, com 647 ocorrências, seguido de 103 ocorrências no *HN* e 96 no *WebMD*. Esses resultados demonstram que, como afirmam vários estudos (Biber e Conrad, 2009; Carter e McCarthy, 2006), a voz passiva é um padrão recorrente no discurso acadêmico, já que existe a necessidade de mudar o foco do praticante para a ação, os processos e eventos descritos. Segundo esses autores, o uso da voz passiva caracteriza-se por demonstrar um distanciamento entre o escritor e o leitor, além de ser muito mais comum na escrita acadêmica do que no discurso oral. Biber e Conrad (2009) sugerem que a omissão do agente acontece por várias razões:

1) Quando o leitor conhece o agente, frequentemente *"researchers*" (pesquisadores);

2) A voz passiva permite que o objeto da pesquisa seja usado como sujeito da sentença, dando mais importância ao objeto do que ao pesquisador;

3) A passiva confere ao discurso um sentido de maior objetividade por não mencionar ações humanas.

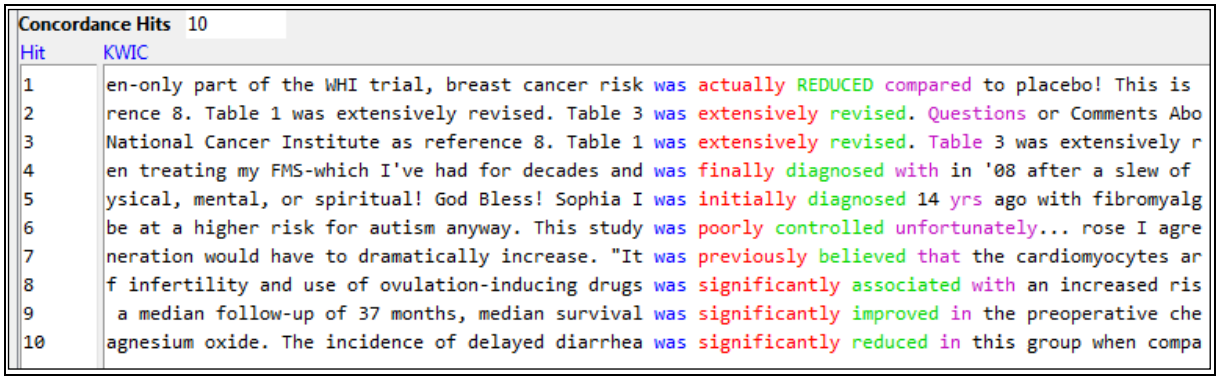
Essas afirmações são confirmadas pelas palavras-chave encontradas na comparação entre o *WebMD* e o *NEJM* (cf. **Tabela 6)** em que *researchers* (linha 5) aparece com 292 ocorrências no *WebMD*, demonstrando que em um gênero de divulgação que busca ser mais interativo e próximo ao leitor é importante que o conteúdo apresentado seja legitimado por uma autoridade. Embora os pronomes não sejam o foco do nosso trabalho, é interessante ressaltar que as palavras-chave demonstraram essa característica interativa do gênero de divulgação com a forte presença de pronomes pessoais em oposição à presença da voz passiva no gênero acadêmico.

3.1.2 *Was + Adv$_{-ly}$ + V$_{-ed}$*

O segundo padrão examinado foi *was + Adv$_{-ly}$ +V$_{-ed.}$* Constatou-se que essa combinação também é muito mais frequente no sub*corpus* acadêmico, com 23 ocorrências contra 10 no *HN* e 6 no *WebMD* (cf. **Tabela 9**).

A **Figura 16** apresenta as linhas de concordância demonstrando a ordem *was+ Adv$_{-ly}$ + V$_{-ed}$*. Os advérbios terminados em *–ly* sempre aparecem em posição intermediária, isto é, entre o verbo e o particípio. Apesar de não ser uma questão problemática para fins de leitura, a ordem das palavras em inglês causa bastante estranheza aos alunos brasileiros, pois esse aspecto obedece a uma regra em inglês que nem sempre coincide com a do português (Bowker, 2002). Como resultado, quando o aprendiz não encontra um equivalente em seu léxico mental, a leitura perde em fluidez (Pavlenko,1997 *apud* 2009). A conscientização desse fato auxiliaria os aprendizes em eventos futuros semelhantes.

Outra questão importante a ser mencionada é que essa estrutura só ocorre com verbos regulares (terminados em *-ed*).



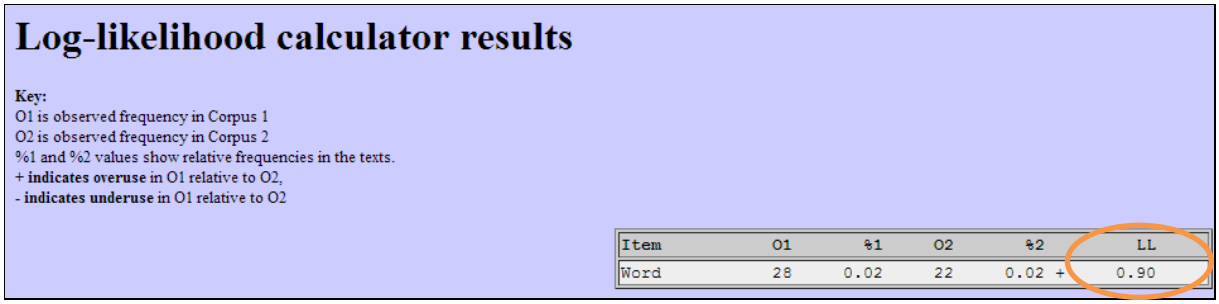
Concordance Hits 10

Hit	KWIC
1	en-only part of the WHI trial, breast cancer risk was actually REDUCED compared to placebo! This is
2	rence 8. Table 1 was extensively revised. Table 3 was extensively revised. Questions or Comments Abo
3	National Cancer Institute as reference 8. Table 1 was extensively revised. Table 3 was extensively r
4	en treating my FMS-which I've had for decades and was finally diagnosed with in '08 after a slew of
5	ysical, mental, or spiritual! God Bless! Sophia I was initially diagnosed 14 yrs ago with fibromyalg
6	be at a higher risk for autism anyway. This study was poorly controlled unfortunately... rose I agre
7	neration would have to dramatically increase. "It was previously believed that the cardiomyocytes ar
8	f infertility and use of ovulation-inducing drugs was significantly associated with an increased ris
9	a median follow-up of 37 months, median survival was significantly improved in the preoperative che
10	agnesium oxide. The incidence of delayed diarrhea was significantly reduced in this group when compa

Figura 16: Linhas de concordância a partir da busca "*was *ly *ed"* no *HN.*

3.1.3 *Was* + $V_{[irreg]}$

Quanto aos verbos irregulares, observa-se que a proporção das ocorrências encontradas é relativamente parecida nos três sub*corpora*. Em maior número no *NEJM* , 28 ocorrências, no *HN* com 22 ocorrências, e um pouco acima no *WebMD*, com 24 ocorrências (cf. **Tabela 10**). A presença da voz passiva com *was* + $V_{[irreg]}$ nos três sub*corpora* está balanceada, sem demonstrar uma diferença significativa. O *log-likelihood* (LL) [40] obtido na comparação é 0.90 (cf. **Figura 17),** não atinge o mínimo esperado que é 3.84 para que a diferença seja estatisticamente significativa.[41]



Log-likelihood calculator results

Key:
O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ **indicates overuse** in O1 relative to O2,
- **indicates underuse** in O1 relative to O2

Item	O1	%1	O2	%2	LL
Word	28	0.02	22	0.02 +	0.90

Figura 17: Comparação das ocorrências de *was* + verbos irregulares no *NEJM* e *HN.*

[40] Conforme detalhado no item 1.5.4.2.
[41] Fizemos o cálculo de LL entre o número de ocorrências mais baixo e o mais alto para demonstrar que a diferença entre elas não é significativa em termos estatísticos.

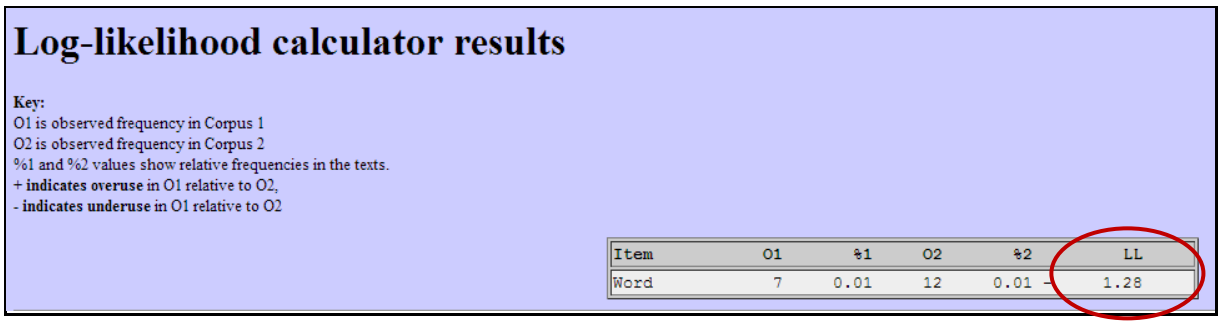
	NEJM	***HN***	***WebMD***
SEEN	**6**	**4**	**7**
FOUND	**5**	**3**	**7**
BORN	0	**4**	2
DONE	0	0	2
KNOWN	**7**	2	2
MADE	0	0	1
MET	1	0	1
SET	2	0	1
THOUGHT	0	0	1
TOLD	0	**4**	0
PUT	0	**3**	0
UNDERTAKEN	0	1	0
GIVEN	2	1	0
TAKEN	2	0	0
UNKNOWN	2	0	0
FELT	1	0	0

Tabela 10: Número de ocorrências do padrão *was* + verbos irregulares nos três sub*corpora.*

A menor presença dos verbos irregulares é um fato importante de ser mencionado, já que, no âmbito do ensino de inglês como segunda língua, consome-se muito tempo na tentativa de fazer os aprendizes memorizarem listas e listas de particípios dos verbos irregulares. No entanto, notamos aqui que esses verbos são um pouco mais frequentes nos textos acadêmicos, com uma pequena diferença entre os outros dois sub*corpora*. Isso não significa ignorar os verbos irregulares, mas é preciso levar em consideração esses levantamentos, já que os dados foram gerados por meio das ferramentas que apresentam dados estatísticos relevantes que podem auxiliar o professor ao fazer as escolhas de o "quê" ensinar em um dado momento do aprendizado.

3.1.4 *Was + V-ing*

O quarto padrão identificado foi *was +V-ing* nos três *subcorpora* (cf.**Tabela 9**). Essa estrutura é mais utilizada nos dois sub*corpora* de divulgação do que no sub*corpus* acadêmico: 20 ocorrências no *NEJM,* 12 no *WebMD* e somente 7 no *NEJM.* Nota-se que a diferença entre o *HN* e o *WebMD* (divulgação) não é significativa (cf. **Figura 18)**: o LL obtido na comparação é 1.28, conforme descrito anteriormente, não atinge o mínimo esperado que é 3.84 para que a diferença seja estatisticamente significativa. Porém, essa diferença passa a ser significativa se compararmos o *HN* e o *NEJM* (cf. **Figura 20**) em que o valor do *log-likelihood* (LL) é de 8.93, o que indica uma diferença altamente significativa.



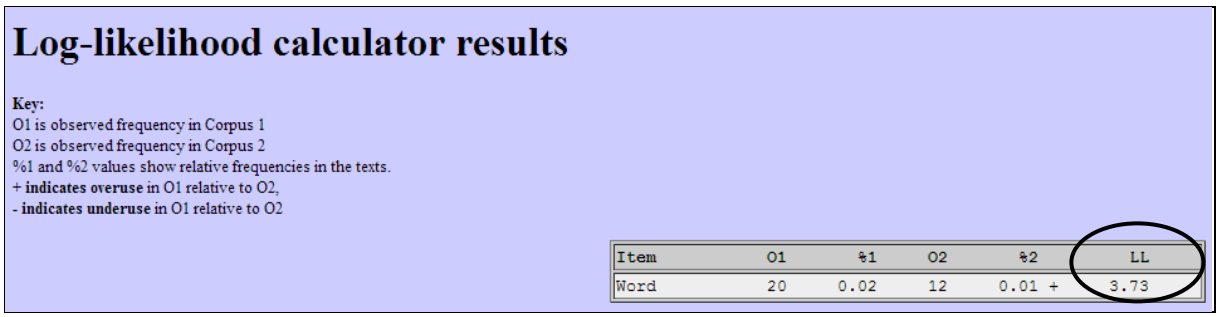
Log-likelihood calculator results

Key:
O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ **indicates overuse** in O1 relative to O2,
- **indicates underuse** in O1 relative to O2

Item	O1	%1	O2	%2	LL
Word	7	0.01	12	0.01 -	1.28

Figura 18: Comparação do *NEJM* com *WebMD*

Quanto a **Figura 19**, notamos a diferença entre as ocorrências no *HN* em comparação ao WebMD (dois subcorpora de divulgação) está muito próxima do valor de significância estatística, mas não chega a atingi-lo.

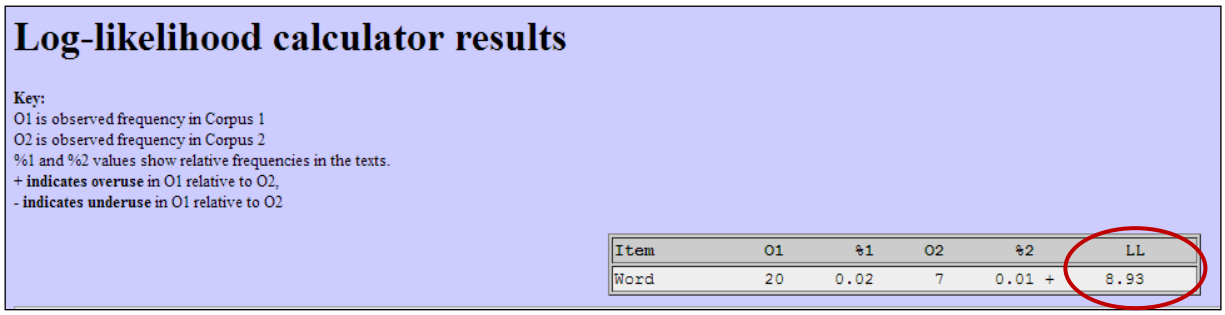


Log-likelihood calculator results

Key:
O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ **indicates overuse** in O1 relative to O2,
- **indicates underuse** in O1 relative to O2

Item	O1	%1	O2	%2	LL
Word	20	0.02	12	0.01 +	3.73

Figura 19: Comparação do *HN* com *WebMD.*



Log-likelihood calculator results

Key:
O1 is observed frequency in Corpus 1
O2 is observed frequency in Corpus 2
%1 and %2 values show relative frequencies in the texts.
+ **indicates overuse** in O1 relative to O2,
- **indicates underuse** in O1 relative to O2

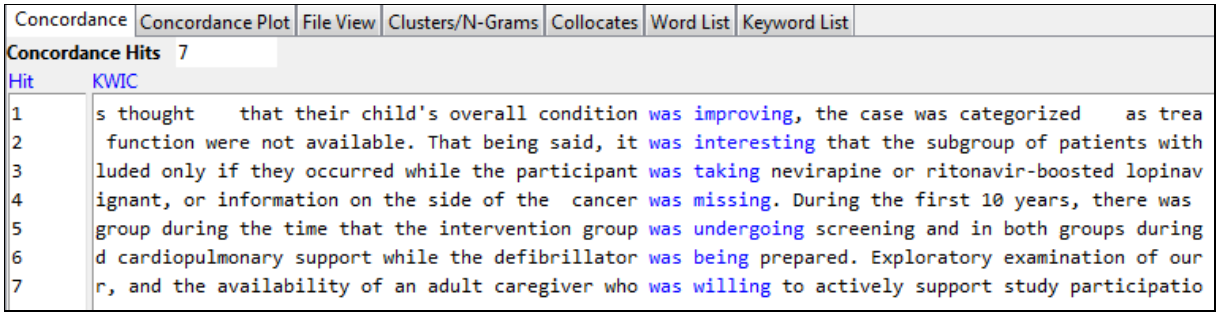
Item	O1	%1	O2	%2	LL
Word	20	0.02	7	0.01 +	8.93

Figura 20: Comparação do *HN* com *NEJM*

O *past progressive* ou *past continuous*, denominação da estrutura, frequente no discurso indireto da linguagem falada, é uma característica das narrativas. Esse tempo verbal é utilizado para enfatizar uma ação contínua ou um evento repetido, ou simplesmente para "dramatizar" o discurso indireto (Carter e McCarthy, 2006). Embora, tenhamos encontrado uma frequência alta dessa estrutura nos artigos de divulgação e suas características de uso demonstram um nível de informalidade maior do que no discurso acadêmico. Acreditamos que esse tipo de estrutura por ser semelhante ao português, não causaria problemas para um aprendiz brasileiro.

Os exemplos encontrados no *NEJM* (cf. **Figura 21**) (linhas 1, 4, 5 e 7), correspondem à descrição de uma ação em curso no passado. Já nas linhas 3 e 6, a estrutura descreve eventos que aconteciam ao mesmo tempo. E na linha 2, há um exemplo que não corresponde a essa estrutura, nesse caso: *was* + *adj* (terminado em *–ing): interesting.*

A abordagem direcionada por dados nos aponta para uma diferença entre os gêneros de divulgação que pode ser uma questão de autoria ou de normas da publicação.



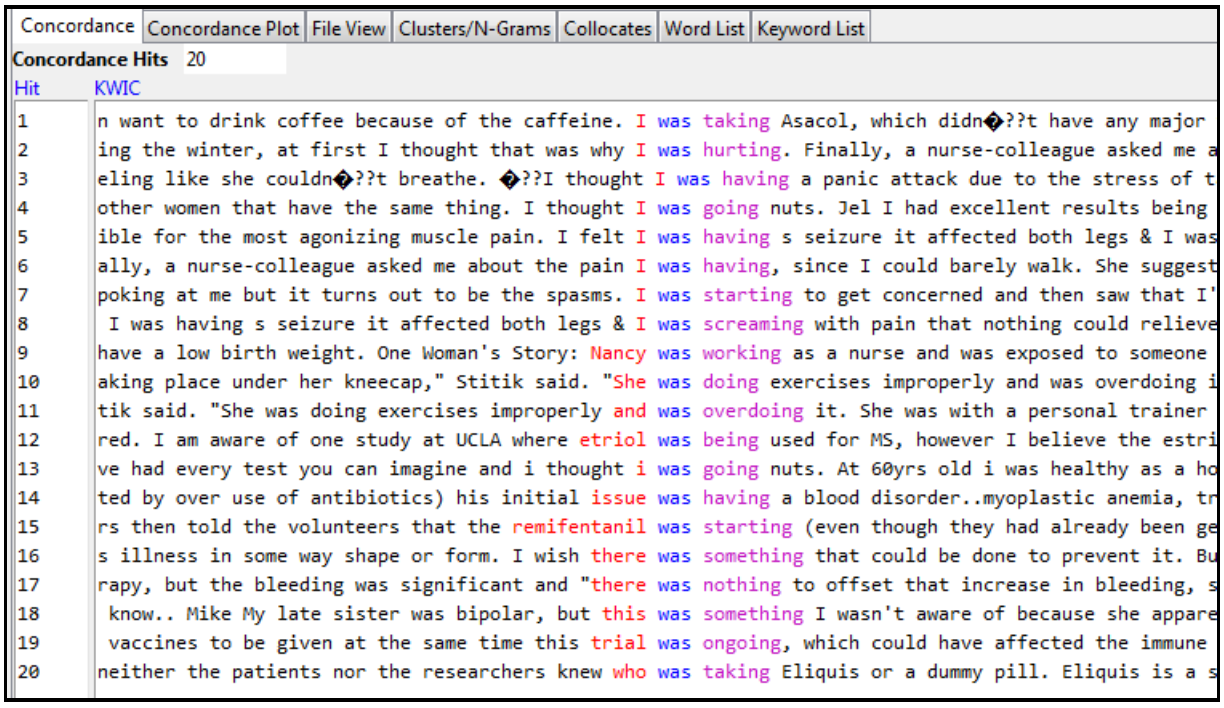
Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Concordance Hits 7

Hit	KWIC
1	s thought that their child's overall condition **was improving**, the case was categorized as trea
2	function were not available. That being said, it **was interesting** that the subgroup of patients with
3	luded only if they occurred while the participant **was taking** nevirapine or ritonavir-boosted lopinav
4	ignant, or information on the side of the cancer **was missing**. During the first 10 years, there was
5	group during the time that the intervention group **was undergoing** screening and in both groups during
6	d cardiopulmonary support while the defibrillator **was being** prepared. Exploratory examination of our
7	r, and the availability of an adult caregiver who **was willing** to actively support study participatio

Figura 21: Linhas de concordância demonstrando as ocorrências de *was *ing* no *NEJM.*

A análise das linhas de concordância geradas a partir da busca da sequência *was* +-*ing* no *HN* (cf. **Figura 22**) apresenta 20 ocorrências desse padrão e essa tela foi a que forneceu maiores subsídios que demonstram as diferenças entre os gêneros estudados. Observamos que a maioria das linhas (17 das 20) apresenta o *past progressive* com a função de narrar eventos em curso no passado, o que era o esperado. O que chamou a atenção, porém, foram principalmente as sentenças apresentadas nas linhas 4 e 13: *I thought I was going nuts.* Nesse exemplo, podemos observar as três características que indicam oralidade, segundo Conrad e Biber (2009:148): a) a oração principal traz o verbo *think;* b) o sujeito da oração principal é o mesmo da oração substantiva; e c) o pronome pessoal é o mesmo nas duas orações. Os autores classificam esse tipo de estrutura como "gramática complexa da conversação", o que certamente causaria problemas de entendimento a um aprendiz que não tivesse conhecimento dessas peculiaridades da língua.


Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Concordance Hits 20

Hit	KWIC
1	n want to drink coffee because of the caffeine. I was taking Asacol, which didn't have any major
2	ing the winter, at first I thought that was why I was hurting. Finally, a nurse-colleague asked me a
3	eling like she couldn't breathe. I thought I was having a panic attack due to the stress of t
4	other women that have the same thing. I thought I was going nuts. Jel I had excellent results being
5	ible for the most agonizing muscle pain. I felt I was having s seizure it affected both legs & I was
6	ally, a nurse-colleague asked me about the pain I was having, since I could barely walk. She suggest
7	poking at me but it turns out to be the spasms. I was starting to get concerned and then saw that I'
8	I was having s seizure it affected both legs & I was screaming with pain that nothing could relieve
9	have a low birth weight. One Woman's Story: Nancy was working as a nurse and was exposed to someone
10	aking place under her kneecap," Stitik said. "She was doing exercises improperly and was overdoing i
11	tik said. "She was doing exercises improperly and was overdoing it. She was with a personal trainer
12	red. I am aware of one study at UCLA where etriol was being used for MS, however I believe the estri
13	ve had every test you can imagine and i thought i was going nuts. At 60yrs old i was healthy as a ho
14	ted by over use of antibiotics) his initial issue was having a blood disorder..myoplastic anemia, tr
15	rs then told the volunteers that the remifentanil was starting (even though they had already been ge
16	s illness in some way shape or form. I wish there was something that could be done to prevent it. Bu
17	rapy, but the bleeding was significant and "there was nothing to offset that increase in bleeding, s
18	know.. Mike My late sister was bipolar, but this was something I wasn't aware of because she appare
19	vaccines to be given at the same time this trial was ongoing, which could have affected the immune
20	neither the patients nor the researchers knew who was taking Eliquis or a dummy pill. Eliquis is a s

Figura 22: Linhas de concordância de *was + -ing* no *HN.*

A gramática da linguagem falada parece mais simples se comparada à escrita acadêmica, pois as palavras e sentenças na conversação são, normalmente, mais

curtas e menos elaboradas, porém algumas estruturas gramaticais mais complexas são mais comuns na conversação do que na escrita. (Conrad e Biber, 2009:148)

Outra característica da oralidade encontrada nas sentenças acima foi a omissão do *that.* Na conversação, as orações substantivas com *that (that-noun clauses*) são comuns depois de expressões que incluem um dos quatro verbos: *think, guess, know* e *say,* normalmente tendo como sujeito o pronome de primeira pessoa *I ,* além de omitir o pronome *that* (Conrad e Biber, 2009).

Quanto ao uso da expressão idiomática *going nuts,* acrescentaríamos uma quinta característica, pois levantamos a hipótese de que a presença dessa expressão seja mais um indício da dificuldade encontrada pelos aprendizes na leitura desses textos de divulgação, já que a análise da mesma estrutura (*was* + -*ing*) no *NEJM* não apontou para nenhuma expressão coloquial.

Os exemplos das linhas 16 e 18 apresentam estruturas diferentes, ou seja, o a terminação -ing não se refere a um verbo, mas apresentam a omissão do *that*. Segundo Biber (2009), o padrão *there was* + *something* ou *nothing* são características da linguagem falada e normalmente são empregados sem o *that.*

Vejamos novamente o exemplo da linha 16*:*

I wish (that) there was something that could be done to prevent it.

No exemplo, somente o primeiro pode ser omitido.

Podemos perceber também a omissão da partícula *that* , após *wish*, no exemplo acima.

E na linha 18:

[...] but this was something (that) I wasn´t aware of because […] também temos a omissão do *that*, também no exemplo acima. Além de apresentar a expressão coloquial *be aware of,* outra característica da linguagem falada.

A análise desse padrão demonstrou algumas características dignas de nota. Segundo Conrad e Biber (2009: 127), os pronomes indefinidos *(somebody,*

something, someone, etc.) são modificados por orações adjetivas. Os autores afirmam também que essas estruturas são muito comuns na conversação com ou sem pronome relativo, mas não são frequentes na escrita acadêmica, por envolverem um pronome indefinido, o que iria contra a objetividade pretendida pelo discurso científico. Dessa forma, os exemplos acima citados corroboram essas características de oralidade encontradas no *subcorpus* de divulgação *HN* e não encontradas no *subcorpus* composto por textos científicos do *NEJM.*

Além das características de oralidade já apresentadas sobre o *HN*, o exemplo da linha 10 (cf. **Figura 22**):

[...] taking place under her kneecap," Stitik said. " She was doing exercises improperly and was overdoing it [...] apresenta um relato com discurso direto entre aspas, forma usada nas narrativas literárias para designar um diálogo.

No terceiro *subcorpus (WebMD),* composto por textos de divulgação da *web,* encontramos também características de oralidade, porém um pouco diferentes daquelas demonstradas no *HN*, o que indica que cada gênero, mesmo que sejam próximos, trazem características de autoria e do veículo em que são publicados (Swales, 1998). Na **Figura 23** não observamos nenhuma omissão do *that* como no *subcorpus* analisado anteriormente. Temos 12 ocorrências do padrão *was* + $V_{\text{-}ing}$, porém somente 9 linhas correspondem ao *past progressive.* Os exemplos nas linhas 1 e 5 correspondem ao padrão *was* + *-ing* (adjetivo) e a linha 9 traz a sequência *was* + *something big* seguido.

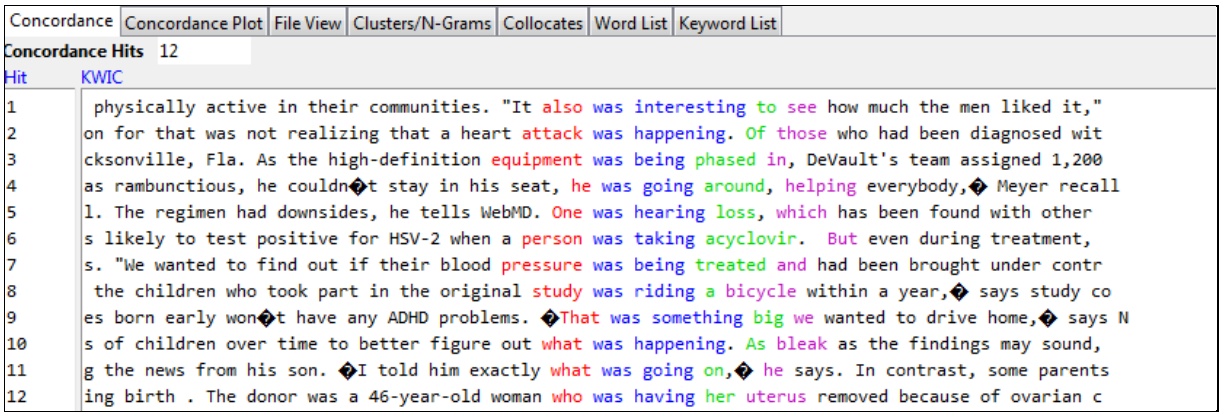


Figura 23: Linhas de concordância a partir da busca *was* + **ing* no *WebMD*

Citamos aqui mais alguns exemplos de expressões coloquiais e verbos frasais presentes no discurso de divulgação que são características da linguagem oral e que poderiam acarretar problemas de entendimento caso o aluno não tenha esse conhecimento:

Na linha 10: *"...to better figure out what was happening."*

Na linha 11: *"I told him exactly what was going on".*

3.2 Colocados de *were*

Os números das ocorrências dos padrões identificados para *were* nos três sub*corpora* estão descritos na **Tabela 10.** O padrão (i) refere-se aos verbos terminados em *-ed,* o padrão (ii) à sequência *were* + *Adv* $_{-ly}$ *+V* $_{-ed}$,, o padrão (iii) *were* + V $_{[irreg]}$ e (iv) *were* + *V*$_{-ing.}$

	NEJM	***HN***	***WebMD***
(i) Were +V_{-ed}	718	80	122
(ii) Were +Adv - $_{ly}$ + V_{-ed}	63	10	8
(iii)Were + $V_{-[irreg]42}$	41	20	26
(iv) Were + V_{-ing}	23	10	20

Tabela 10: Padrões de *were* em cada um dos três sub*corpora.*

Esses mesmos dados são apresentados no **Gráfico 2** para melhor visualização:

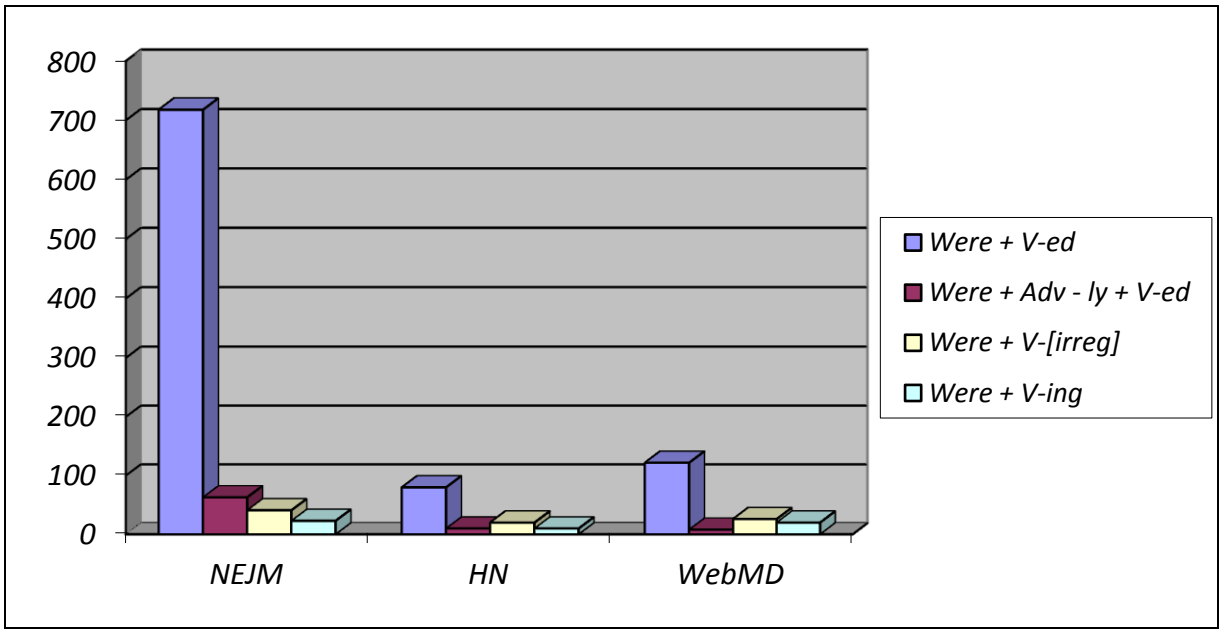


Gráfico 2: Distribuição de padrões de *were* nos três sub*corpora.*

[42] Incluímos nessa contagem todos os verbos irregulares que aparecem como colocados de *were,* inclusive os que ocorreram só uma vez.

A distribuição dos padrões no *NEJM* (cf. **Gráfico 2**) apresenta, claramente, a diferença do uso de *were +V -ed* no sub*corpus* acadêmico em comparação aos dois subcorpora de divulgação.

3.2.1 *Were + V-ed*

O padrão *were* + V*-ed* é muito mais frequente no *NEJM,* com 718 ocorrências, menos frequente no *HN,* com 80 ocorrências e no *WebMD,* com 122 ocorrências. Notamos que, apesar de o *HN* apresentar um número menor de ocorrências, os dois *subcorpora* de divulgação apresentam melhor distribuição em relação ao número de ocorrências, não demonstrando uma diferença significativa.

3.2.2 *Were + Adv –ly + V-ed*

Dentre os padrões formados por *were*, a sequência *were + Adv -ly + V-ed* também é muito mais frequente no *NEJM* (63) do que nos *subcorpora HN* (10) *e WebMD* (8), indicativo de um padrão predominante de colocação adverbial na voz passiva em textos acadêmicos.

Como já mencionado anteriormente, as regras de colocação adverbial no inglês obedecem a um sistema mais rígido do que as do português. Embora esse fato não seja tão problemático para a leitura, chamar a atenção para essa peculiaridade auxilia o aprendiz a tomar consciência da falta de equivalência entre as línguas. Os estudos com *corpora* ajudam nessa tarefa, pois facilitam a visualização. Além disso, trabalhar com textos especializados (Pavlenko, 2003) permite a familiarização com estruturas pertinentes à área de especialidade de cada aprendiz.

A título de curiosidade mencionamos a ocorrência do advérbio *randomly + assigned,* que é justamente a colocação adverbial com 6 ocorrências em 10, na sequência *was + Adv-ly + V-ed,* nas linhas de concordância do *NEJM* (cf. **Figura 24)**.

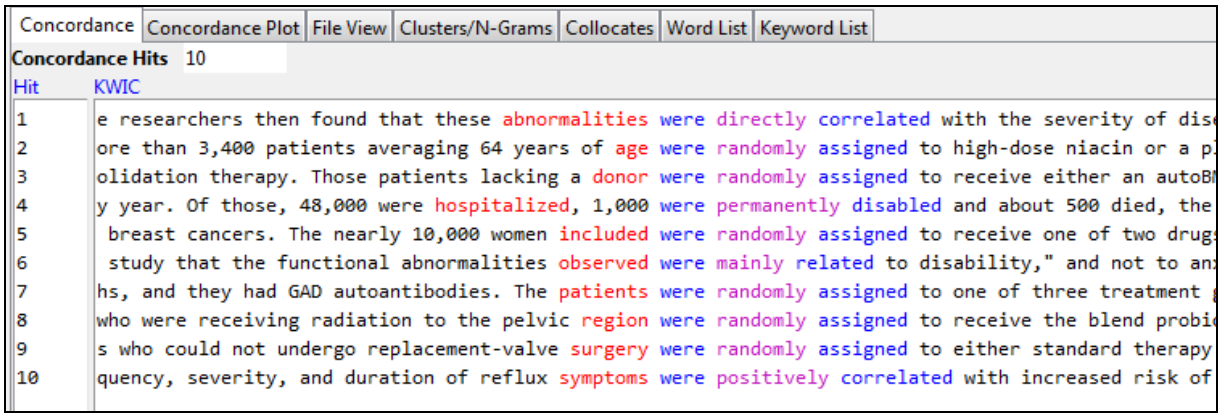


Figura 24: Linhas de concordância a partir da busca *were *ly *ed* no *NEJM.*

Esse advérbio é o mais utilizado nos textos acadêmicos, o que explica sua tradução literal (randomicamente) para o português e consequente uso desse empréstimo pelos profissionais da área médica atualmente, apesar de seu sinônimo: 'aleatoriamente' ser ainda o mais utilizado, com 3.820.000 ocorrências no *Google,* contra 88.700 para 'randomicamente'. O Novo Dicionário da língua portuguesa de Aurélio Teixeira, em sua primeira edição (1975), registra 'randômico' com o mesmo sentido de 'aleatório', e, em suas edições posteriores, acrescenta 'randomização' como sinônimo de 'acidentalização'. O dicionário Houaiss, além dos dois termos citados, já registra o verbo 'randomizar', sancionando, assim, a incorporação desses anglicismos à língua portuguesa.

3.2.3 *Were + V [irreg]*

As ocorrências do padrão de *were + V[irreg]* estão listadas na **Tabela 11**. Como podemos observar, esse padrão aparece em número muito menor que o dos regulares. Como no caso do padrão *was +V -ing*, o foco da aprendizagem, pelo menos no início, pode enfatizar somente os verbos irregulares mais recorrentes, e, assim, otimizar a leitura do aluno.

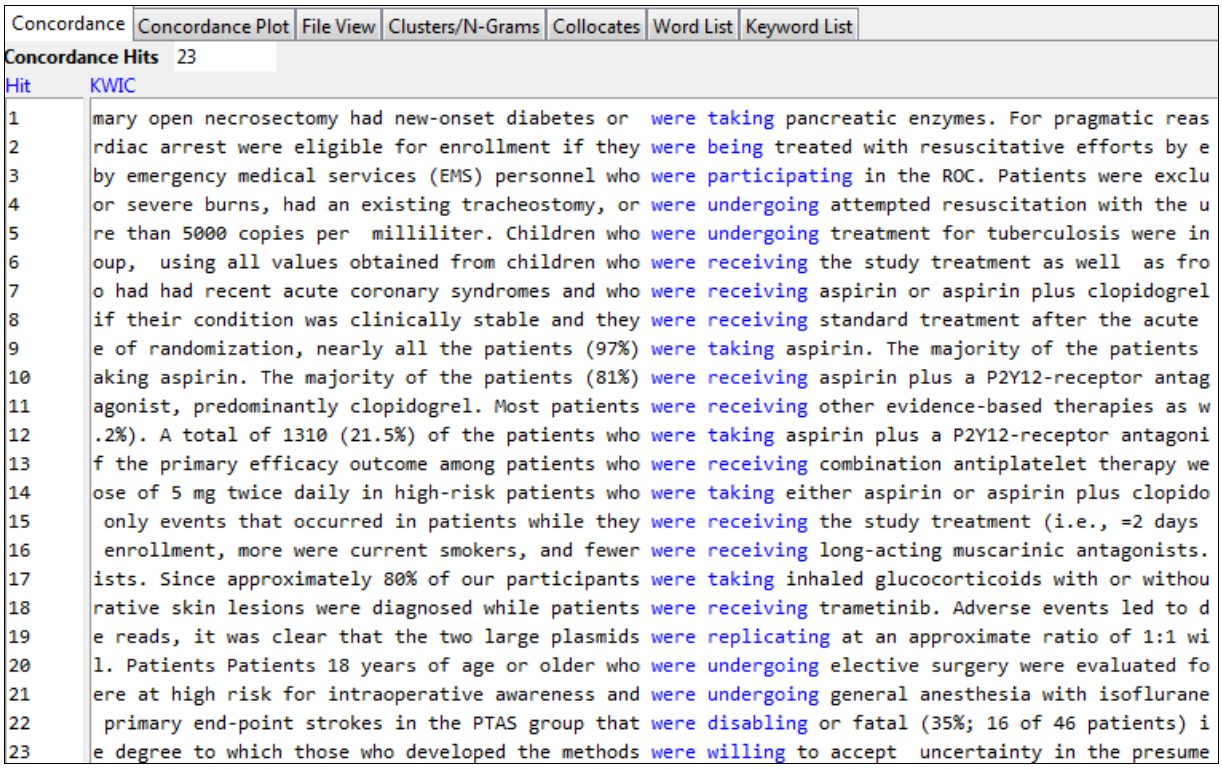
	NEJM	***HN***	***WebMD***
GIVEN	8	1	1
SEEN	7	2	2
MADE	6	2	1
LOST	6	1	0
FOUND	4	2	5
DONE	0	1	3
BORN	0	6	9
KNOWN	1	0	0
MET	2	0	0
SET	2	0	0
TOLD	1	2	2
BROKEN	2	0	0
CHOSEN	1	0	0

Tabela 11: Número de ocorrências do padrão *were* + verbos irregulares.

3.2.4 *Were* + V_{-ing}

A **Figura 25** apresenta as linhas de concordância com o padrão *were* + V_{-ing} analisado no *NEJM.* Aqui, essa sequência também corresponde ao *past progressive,* relatando, no caso desse subcorpus de medicina, resultados de tratamentos aplicados em pacientes.

O pronome relativo *who* é o colocado mais frequente desse padrão e funciona como sujeito das orações adjetivas, referindo-se a pessoas. Segundo Conrad e Biber (2009), essa estrutura é muito comum na conversação, afirmação corroborada pela análise. Já o pronome *that,* que introduz as orações adjetivas é mais comum na conversação e raro na escrita. Isso foi confirmado pela análise, pois encontramos somente um exemplo de *that* na linha 22, referindo-se a um grupo de estudo *(PTA group).*



Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Concordance Hits 23

Hit	KWIC
1	mary open necrosectomy had new-onset diabetes or **were taking** pancreatic enzymes. For pragmatic reas
2	rdiac arrest were eligible for enrollment if they **were being** treated with resuscitative efforts by e
3	by emergency medical services (EMS) personnel who **were participating** in the ROC. Patients were exclu
4	or severe burns, had an existing tracheostomy, or **were undergoing** attempted resuscitation with the u
5	re than 5000 copies per milliliter. Children who **were undergoing** treatment for tuberculosis were in
6	oup, using all values obtained from children who **were receiving** the study treatment as well as fro
7	o had had recent acute coronary syndromes and who **were receiving** aspirin or aspirin plus clopidogrel
8	if their condition was clinically stable and they **were receiving** standard treatment after the acute
9	e of randomization, nearly all the patients (97%) **were taking** aspirin. The majority of the patients
10	aking aspirin. The majority of the patients (81%) **were receiving** aspirin plus a P2Y12-receptor antag
11	agonist, predominantly clopidogrel. Most patients **were receiving** other evidence-based therapies as w
12	.2%). A total of 1310 (21.5%) of the patients who **were taking** aspirin plus a P2Y12-receptor antagoni
13	f the primary efficacy outcome among patients who **were receiving** combination antiplatelet therapy we
14	ose of 5 mg twice daily in high-risk patients who **were taking** either aspirin or aspirin plus clopido
15	only events that occurred in patients while they **were receiving** the study treatment (i.e., =2 days
16	enrollment, more were current smokers, and fewer **were receiving** long-acting muscarinic antagonists.
17	ists. Since approximately 80% of our participants **were taking** inhaled glucocorticoids with or withou
18	rative skin lesions were diagnosed while patients **were receiving** trametinib. Adverse events led to d
19	e reads, it was clear that the two large plasmids **were replicating** at an approximate ratio of 1:1 wi
20	l. Patients Patients 18 years of age or older who **were undergoing** elective surgery were evaluated fo
21	ere at high risk for intraoperative awareness and **were undergoing** general anesthesia with isoflurane
22	primary end-point strokes in the PTAS group that **were disabling** or fatal (35%; 16 of 46 patients) i
23	e degree to which those who developed the methods **were willing** to accept uncertainty in the presume

Figura 25: Linhas de concordância de *were *ing* no *NEJM.*

Nesse sub*corpus,* os exemplos não apresentam grande variação de padrões, ou seja, são mais lineares, conforme apontado no depoimento dos alunos em sala de aula.

Diferentemente do que sugere a gramática *English grammar in use* (Murphy, 1994), *who* e *that* não são usados indiscriminadamente, mas sim utilizados mais ou menos dependendo do contexto gramatical e registro (formal ou informal) (Conrad e Biber, 2009: 121).

A **Figura 26** apresenta as 10 ocorrências de *were* +V_{-ing} no *HN.* Em termos gerais, o padrão se repete aqui com o *past progressive* na função de descrever uma situação que estava ocorrendo no passado. A omissão do *that* também se repete nesses exemplos retirados do *HN* nas linhas 5, 6, e 7, mostrando indícios de oralidade. Na linha 9, observamos uma pergunta que relata um evento. Porém, o discurso direto está ocupando o lugar do discurso indireto, o que seria o mais esperado em um relato de caso, indicando maior proximidade com o leitor (Biber,

1998). Diferentemente dos colocados de *was,* não encontramos aqui nenhuma sequência com uma expressão coloquial.

Dentre os colocados de *were* encontramos o pronome *who* introduzindo orações adjetivas nas linhas 8 e 10.

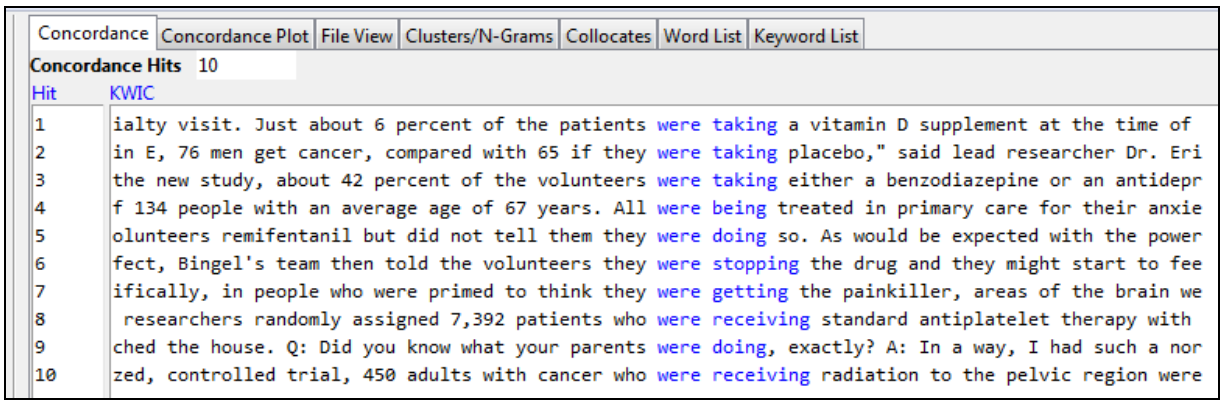


Figura 26: Linhas de concordância demonstrando o padrão *were *ing* no *HN.*

A **Figura 27** apresenta as linhas de concordância de *were* + V_{-ing} no *WebMD* narrando situações no passado. A observação dessas linhas não traz muita novidade se comparado ao que já foi discutido anteriormente. Encontramos 5 ocorrências de *who* introduzindo orações adjetivas e nenhum exemplo com *that,* demonstrando que até mesmo em artigos destinados ao público leigo o *who* é o preferido. Todas essas informações são importantes para o conhecimento do professor e dos alunos.

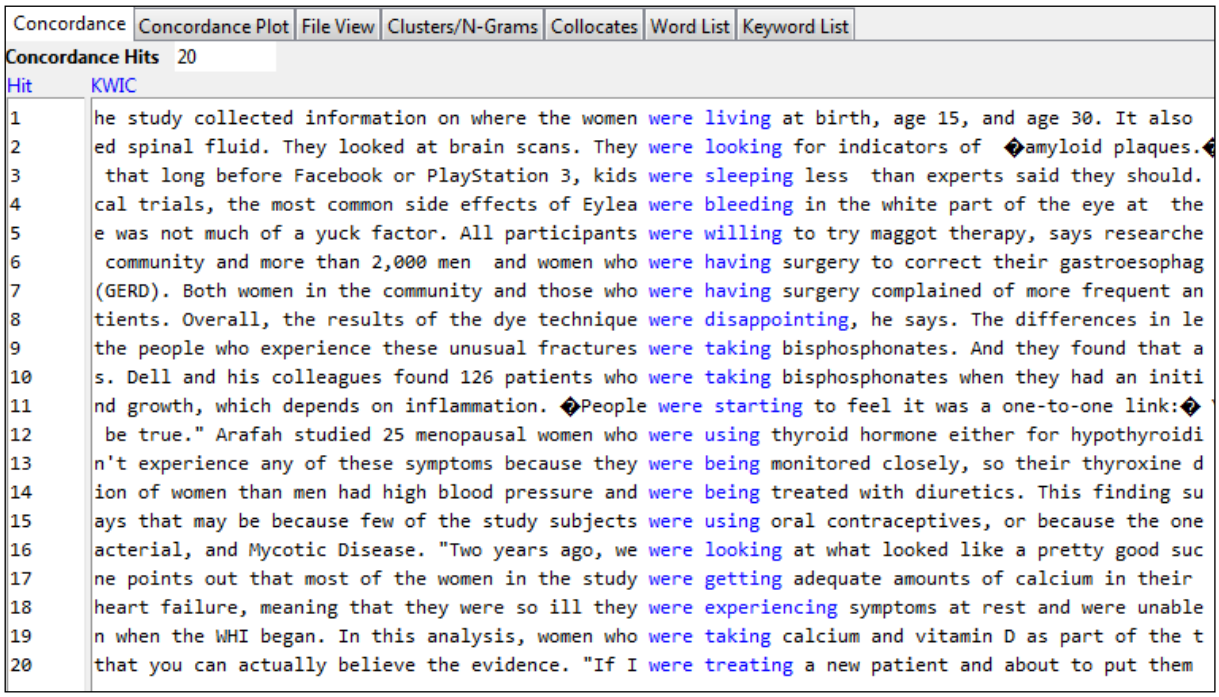


Figura 27: Linhas de concordância de *were * ing* no *WebMD.*

3.2.5 *Clusters* de *were* no *WebMD*

Apresentamos, neste item, as tabelas dos clusters de *was* e *were* a título de comparação. Por questão de metodologia, observamos os *clusters* dos dois verbos-chave, porém os de *was* não apresentaram nenhum indício de estruturas que pudessem causar problemas para a leitura. Desse modo, eles não foram analisados, porém apresentamos a **Tabela 12** para observação.

O *cluster* mais frequente formado com *were* nos três *subcorpora* é *who*, introduzindo orações adjetivas (cf. **Tabela 13**). A presença da combinação *to be likely* entre os 20 primeiros colocados no *HN* e no *WebMD* chamou a atenção, indicando uma diferença linguística entre os sub*corpora*, demonstrando que a expressão tende a ser mais usada nos textos de divulgação. Concluímos que essa diferença poderia ser problemática para os alunos, já que essa expressão não encontra equivalente em português, e é um tipo de modalidade característica do inglês.

O termo ‘modalidade’ refere-se à atitude do falante ou escritor em relação a um ponto de vista, ou a um estado do mundo. Em especial, os modais são usados

para dizer se algo é real e verdadeiro, ou se é mais objeto de especulação do que de conhecimento definitivo (Carter e McCarthy, 2006: 638).[43]

Por ser uma estrutura que indica modalidade e não encontra equivalente em português, e por isso passível de prejudicar a compreensão de um texto por parte de um aprendiz brasileiro (Pavlenko, 1997, 2003, *apud* Pavlenko, 2009), apresentamos, nesta seção, a análise do seu comportamento.

[43] No original: *The term 'modality' refers to a speaker´s or a writer´s attitude towards, or point of view about, a state of the world. In particular, modals are used to say whether something is real or true, or whether it's the subject of speculation rather than definite knowledge*.

CLUSTERS DE WAS						
ORDEM	NEJM		HEALTH NEWS		WEBMD	
	FREQUÊNCIA	CLUSTER	FREQUÊNCIA	CLUSTER	FREQUÊNCIA	CLUSTER
1	75	was not	40	I was	29	study was
2	66	study was	16	it was	28	was not
3	53	was performed	16	there was	22	was associated
4	50	was a	16	was not	21	was associated with
5	46	that was	11	was a	18	who was
6	44	was defined	10	risk was	16	The study was
7	42	was defined as	10	study was	15	there was
8	34	there was	10	was the	15	was not involved
9	32	was associated	9	that was	14	was associated with a
10	32	was associated with	9	was diagnosed	13	was a
11	28	and was	8	and was	13	was not involved in
12	28	The study was	8	was diagnosed with	12	was not involved in the
13	28	was similar	8	was no	11	it was
14	28	was the	7	there was no	11	was no
15	27	was used	7	was associated	11	was the
16	26	it was	7	was associated with	10	who was not
17	25	analysis was	7	was on	9	that was
18	24	which was	7	which was	9	who was not involved
19	23	was conducted	7	who was	7	association was
20	22	was also	6	the risk was	7	There was

Tabela 12: Comparação dos *clusters* de *was* nos três sub*corpora.*

CLUSTERS DE WERE						
ORDEM	NEJM		HEALTH NEWS		WEBMD	
	FREQUÊNCIA	CLUSTER	FREQUÊNCIA	CLUSTER	FREQUÊNCIA	CLUSTER
1	75	who were	27	who were	45	who were
2	60	that were	16	they were	29	were more
3	58	were not	14	were not	24	they were
4	43	events were	12	were more	17	were more likely
5	43	patients were	10	findings were	15	were also
6	36	There were	9	patients were	14	were born
7	33	were randomly	8	children were	14	were followed
8	32	analyses were	8	were more likely	14	women were
9	32	were observed	7	there were	13	were more likely to
10	32	Were randomly assigned	7	were more likely to	11	that were
11	30	were performed	6	The findings were	11	were about
12	30	were similar	6	were born	11	were followed for
13	29	and were	6	were randomly	10	patients were
14	26	were enrolled	6	were randomly assigned	9	and were
15	24	were no	6	were randomly assigned to	9	were able
16	24	were randomly assigned to	6	were treated	9	were able to
17	24	were used	6	were treated with	9	were the
18	23	were included	5	findings were published	9	women who were
19	22	There were no	5	findings were published in	8	levels were
20	22	were reported	5	findings were published in the	8	Participants were

Tabela 13: Comparação dos *clusters* de *were* nos três sub*corpora.*

3.2.5.1 Comportamento do item lexical *likely*

Atualmente, os dicionários direcionados a aprendizes de inglês, por exemplo o Macmillan, (2002), já utilizam *corpora*[44] para sua construção. Porém, mesmo com um trabalho criterioso, não conseguem abranger todas as instâncias de um item lexical por questão de espaço que limita a versão impressa. O dicionário acima citado, por exemplo, apresenta várias entradas para o item *likely.*

O verbete *likely* aparece nesse dicionário (Macmillan, 2002: 828) como:

likely adj **** * ****[45]

1. *probably going to happen, or probably true. The most likely cause of the fire was cigarette. a)* ***likely to do smth.*** *The study shows some people are more likely to suffer from back problems. b)* ***+ that*** *It seems likely that the interest rates will increase. c)* ***more than likely*** *(= almost certain) It´s more than likely I´ll see Ben this week.*

2. *[only before noun] suitable, or almost certain to be successful: She´s a likely candidate for the job.*

a likely story *(spoken) used for telling someone that you do not believe what they have just said.*

Apesar de existir a tentativa de um trabalho minucioso e completo, nenhum dicionário consegue atender todas as necessidades específicas de alguns aprendizes. Em contrapartida, o estudo com um *corpus* especializado na área de saúde, por exemplo, permite ao pesquisador/professor identificar quais os usos mais recorrentes na área de especialidade pertinente a seus alunos.

Assim, a partir da identificação de um dos *clusters (were more likely),* (cf. **Tabela 13)** mais recorrentes no *WebMD* e no *Health News,* ambos de divulgação, apresentamos, a seguir, os padrões de *likely* que apareceram em cada um dos três sub*corpora.*

[44] As definições desse dicionário derivaram do *World English Corpus (Bloomsbury Corpus of World English)* de 200 milhões de palavras.

[45] Os verbetes em vermelho seguidos de estrelas neste dicionário indicam a frequência de uso dessas palavras. Assim, um verbete com três estrelas indica que ele aparece entre as 2.500 palavras mais comuns e básicas do inglês, duas estrelas são as um pouco menos comuns e uma estrela palavras com menos frequência, mas com certa frequência digna de nota.

A **Tabela 14** apresenta as ocorrências encontradas para os padrões gramaticais formados por *likely.* Dentre esses padrões as sequências mais frequentes foram: *likely to* + infinitivo; *likely that* e *likely than,* sendo que *likely to* é muito mais utilizado nos três sub*copora* do que os demais padrões, com 30 ocorrências no *NEJM*, 64 no *HN* e 98 no *WebMD.*

	NEJM (51)	***Health News (142)***	***WebMD (225)***
likely to + infinitivo	*30*	*64*	*98*
Likely that	*6*	*2*	*3*
likely than	*2*	*8*	*15*
more likely to	*6*	*45*	*53*
more likely than	*1*	*6*	*13*
less likely to	*5*	*7*	*19*
less likely than	*1*	*2*	*2*
as likely as	*0*	*3*	*7*
as likely to	*0*	*5*	*15*

Tabela 14: Distribuição das ocorrências de *clusters* com *likely* nos três sub*corpora.*

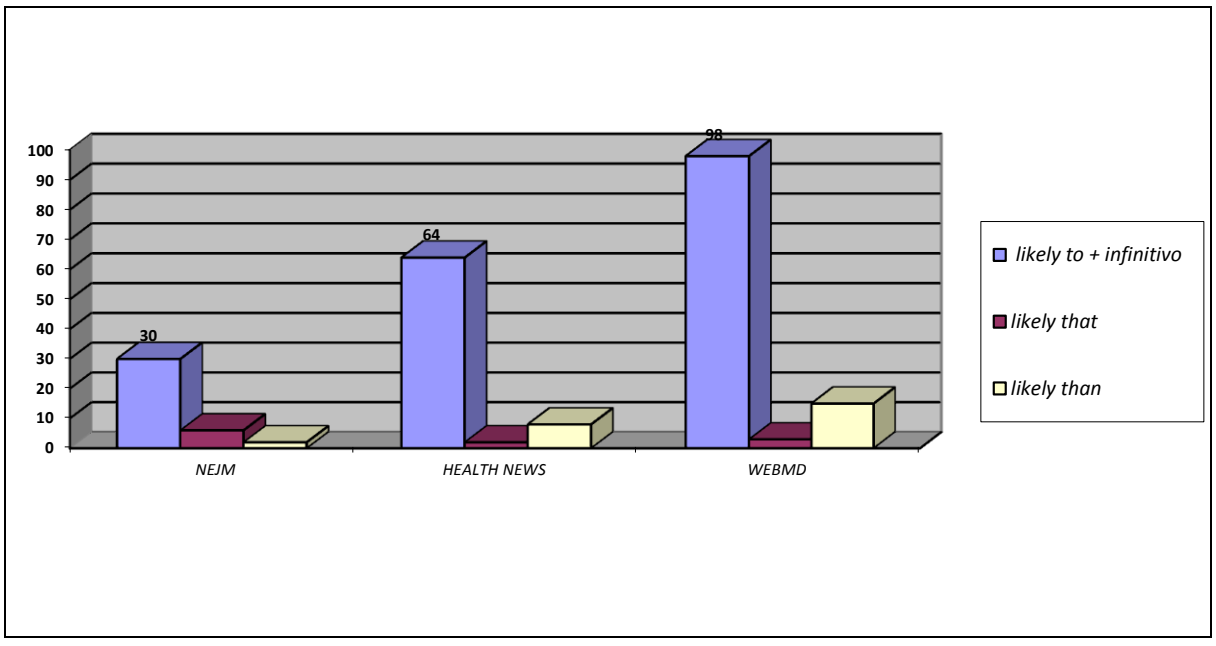


Gráfico 3: Comparativo do número de ocorrências totais de *likely to/ that/than* nos três sub*corpora.*

Agrupamos esses padrões em três categorias: *likely to*, *likely that* e *likely than*. No geral, *likely to* + infinitivo é o padrão mais comum nos três *subcorpora*, porém é muito mais utilizado no *WebMD* (98 ocorrências), com uma diferença significativa (LL 15.16) para baixo em relação ao *HN* (64 ocorrências). Na comparação do *WebMD* e *NEJM*, a diferença é altamente significativa (LL 49.92). Isso indica que essa estrutura é mais utilizada nos dois *subcorpora* de divulgação e muito menos no *subcorpus* acadêmico, o que, para fins didáticos, é um dado importante. Essa descoberta fornece subsídios para que essa estrutura seja trabalhada de maneira diferenciada com alunos que necessitam ler textos de divulgação, além de apresentar outro indício das dificuldades enfrentadas pelos aprendizes de leitura, já que essa estrutura não encontra equivalência em português.

3.2.5.2 *Be likely* em sequências de comparação

Ao observarmos os colocados dessa sequência notamos que as dificuldades se ampliam à medida que mais componentes entram nessa estrutura. O uso comparativo de *be likely to* + infinitivo, por exemplo, é muito mais recorrente nos dois sub*corpora* de divulgação (cf. **Tabela 15**):

	NEJM	***HN***	***WebMD***
more likely to	*6*	*45*	*53*
less likely to	*5*	*7*	*19*
as likely to	*0*	*5*	*15*

Tabela 15: Número de ocorrências dos usos de *likely to* em comparações.

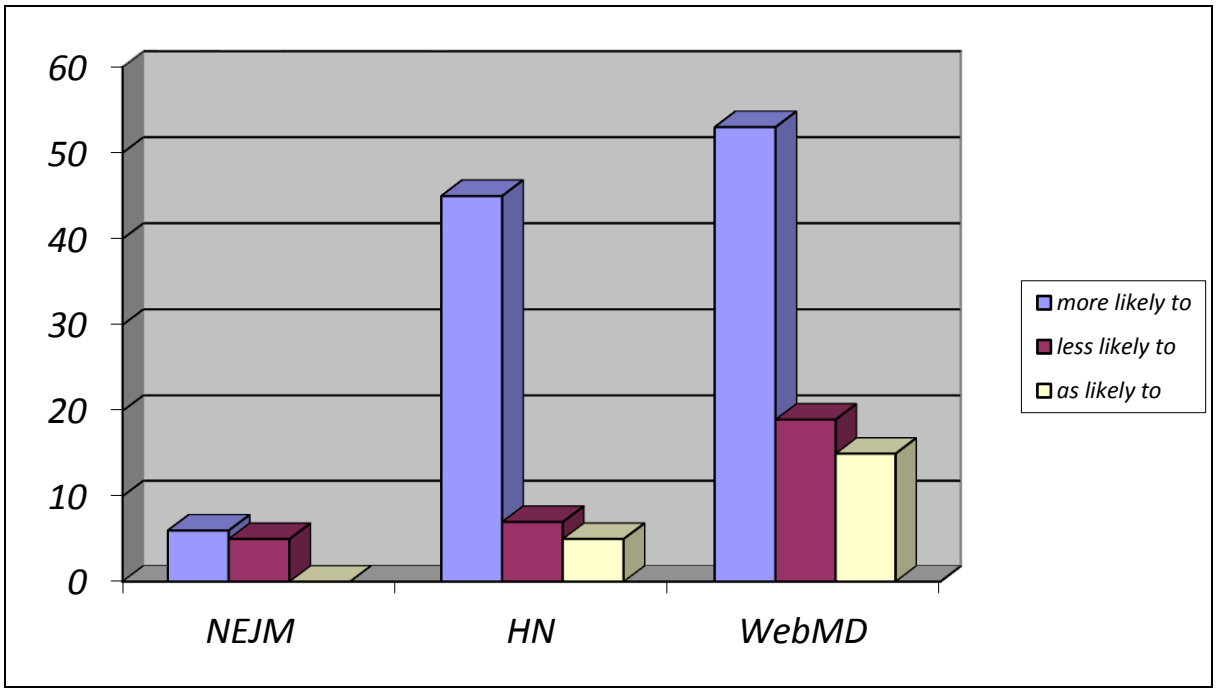


Gráfico 4: Distribuição dos usos comparativos de '*likely to'* nos três sub*corpora.*

A sequência *more likely to* tem 45 occorências no *HN*, com uma ligeira diferença para cima no *WebMD* (53 ocorrências), e com uma enorme diferença para baixo no *NEJM,* com somente 6 ocorrências, o que em termos estatísticos representa diferença bastante significativa (LL 51.00). Em segundo lugar, aparece *less likely to* com 19 no *WebMD*, 7 no *HN* e 5 no *NEJM,* e em terceiro, *as likely to* apresenta 15 ocorrências no *WebMD*, que é três vezes mais do que no *HN,* com 5 ocorrências, e nenhuma no *NEJM*, indicando que a comparação com esse tipo de estrutura não é comum em textos acadêmicos, pois nesse tipo de texto não houve nenhuma ocorrência. A possível razão para essa diferença seria a tentativa de maior objetividade no discurso científico (Zamboni, 2001), já que *likely* indica possibilidade.

Os modificadores de *as likely to* chamaram a atenção pela mesma razão mencionada anteriormente: a falta de equivalência da estrutura entre o português e o inglês, o que acarretaria grande dificuldade para aprendizes brasileiros. O entendimento da sequência (cf. **Figura 28)** *just as likely to* + infinitivo (linhas 1 e 2), bem como a sequência que aparece nas linhas: 4 (17 *times as likely to*) e 5 (*almost*

twice as likely to + infinitivo) também causariam problemas pela mesma falta de equivalência.

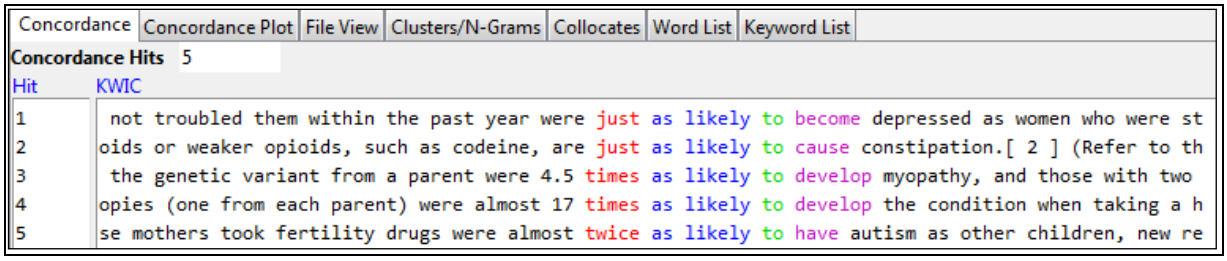


Figura 28: Linhas de concordância de *'as likely to'* no *HN.*

Uma possível explicação para a maior presença dessas sequências no discurso de divulgação é que existe menos lugar para possibilidade no discurso da ciência (Conrad e Biber, 2009; Carter e McCarthy, 2006 e Zamboni, 2001), portanto não há muito espaço para esse tipo de estrutura nos artigos acadêmicos.

Outro padrão recorrente, com 13 ocorrências no sub*corpus* de divulgação *WebMD,* é *more likely than* (cf. **Tabela 16**). Possívelmente, isso se deve ao fato de ser o *WebMD* aquele que mais se aproxima do público leigo, devido a sua interação direta com o leitor, e ter a possibilidade de levantar hipóteses menos asssertivas e demonstrar opinião sobre o assunto relatado (Conrad e Biber, 2009: 106).

	NEJM	***HN***	***WebMD***
more likely than	*1*	*6*	***13***
less likely than	*1*	*2*	*2*
as likely as	*0*	*3*	*7*

Tabela 16: Número de ocorrências de *likely than/as* usado comparativamente nos três sub*corpora.*

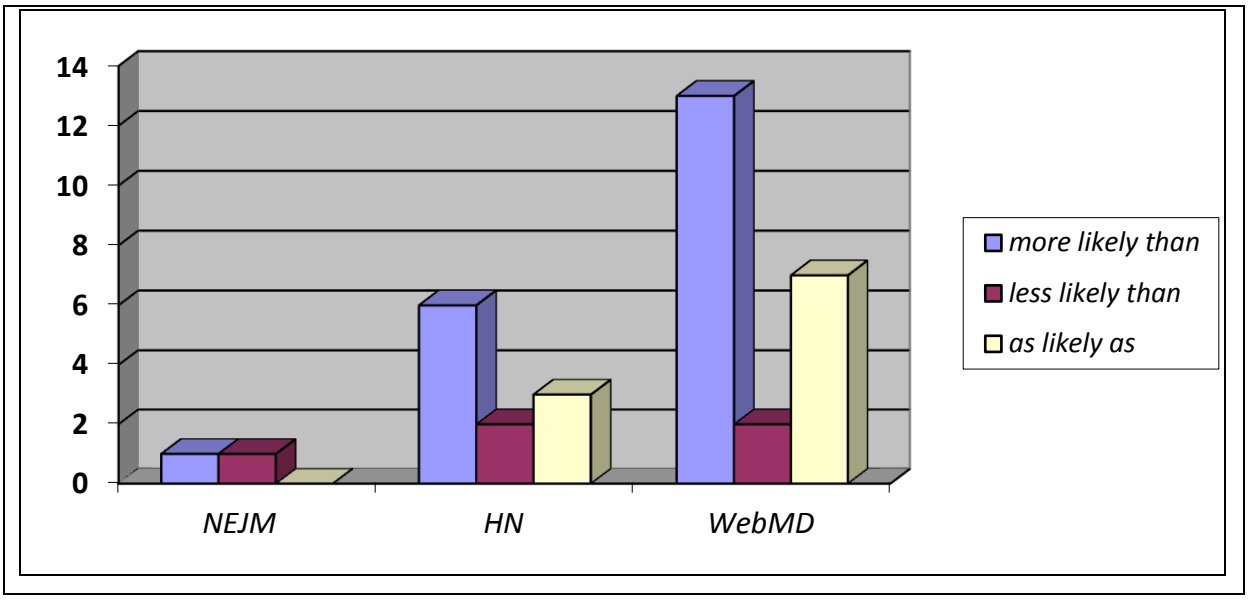


Gráfico 5: Distribuição dos padrões comparativos de *likely than/as.*

Não foi encontrada nenhuma ocorrência de *as likely as* no *NEJM.* No *WebMD* foram 7 ocorrências e no *HN*, 3 ocorrências (cf. **Tabela 16**). É interessante notar que essa estrutura está sendo empregada para descrever resultados, o que não acontece no discurso acadêmico, pois os resultados devem ser apresentados de maneira mais precisa com dados estatísticos.

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Concordance Hits 3

Hit	KWIC
1	eks but not full term] children." Boys were twice as likely as girls to suffer behavioral problems,
2	s before their second birthday were roughly twice as likely as their peers to develop ADHD by the ti
3	nancy (known as gestational diabetes) were twice as likely as their peers to meet the criteria for

Figura 29: Linhas de concordância de *'as likely as'* no *HN.*

O advérbio *roughly* (cf. **Figura 29,** linha 2) como colocado da sequência *twice as likely as* acrescenta a noção de imprecisão ao discurso, além de ser uma sequência que causaria dificuldade para o entendimento.

A estrutura *It+ is+ likely+ that– clause* é usada para avaliar uma ideia. O adjetivo expressa a atitude do autor ou sua avaliação, e esse conceito é expresso na sentença com *that* (Conrad e Biber , 2009: 106). Essas estruturas são mais comuns na escrita acadêmica; ainda segundo os autores, são raras na conversação. São as formas significativas de se demonstrar atitudes na escrita.

Conforme demonstrado nas **Figuras 30, 31 e 32,** as descobertas corroboram a ideia citada anteriormente: maior ocorrência dessa estrutura no *NEJM*, embora não com uma diferença significativa.

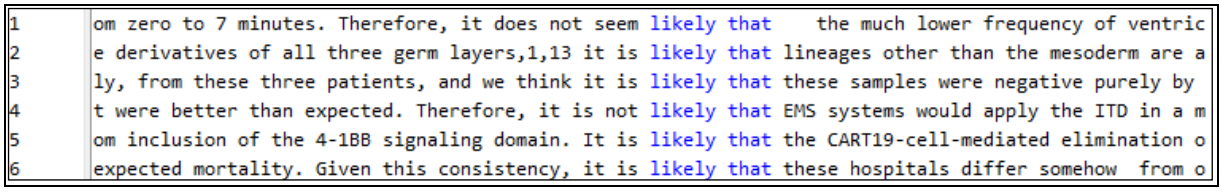


Figura 30: Linhas de concordância com *it * likely that* no *NEJM.*

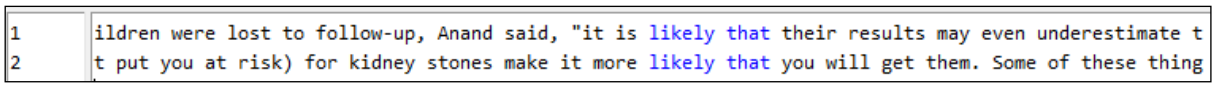


Figura 31: Linhas de concordância com *it * likely that* no *HN.*

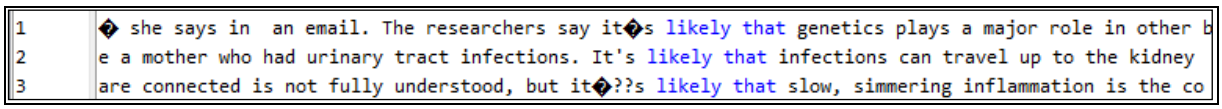


Figura 32: Linhas de concordância com *It * likely that* no *WebMD.*

A dúvida é expressa com *it seems/is unlikely (*cf. **Figura 33***),* que é demonstrada pelos exemplos retirados do sub*corpus NEJM:*

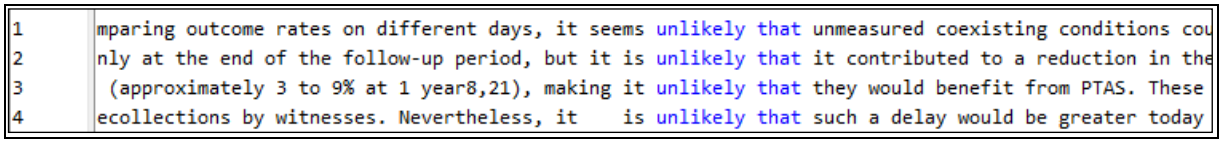


Figura 33: Linhas de concordância com *it * unlikely that* no *NEJM.*

As **Figuras 34** e **35** apresentam as ocorrências de *unlikely that* nos dois sub*corpora* de divulgação. Apesar de Conrad e Biber (2009) afirmarem que essa estrutura é mais comum na escrita acadêmica, não encontramos uma diferença significativa do uso dessa sequência entre os três sub*corpora*, ao contrário, o número e ocorrências é a metade no *WebMD*:

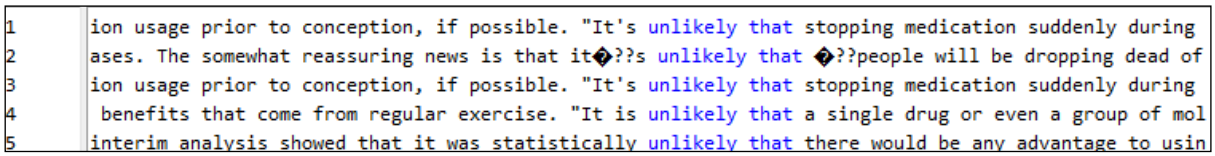


Figura 34: Linhas de concordância com *it * unlikely that* no *HN.*

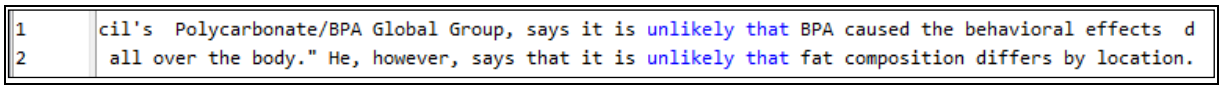


Figura 35: Linhas de concordância com *it * unlikely that* no *WebMD.*

Nos *subcorpora* de divulgação o uso de *it´s unlikely that* está relacionado com relatos e no *HN* (linhas 1, 3 e 4) aparece em uma sentença dentro de aspas indicando discurso direto, corroborando a hipótese de mais proximidade com o leitor por meio de uma narrativa com característica de um diálogo.

Como mencionado no início deste capítulo, pudemos comprovar o princípio do serependismo, pois, a partir das primeiras palavras-chave e análise de seus colocados, foi possível comprovar estudos anteriores, como a presença da voz passiva em textos acadêmicos e a utilização em maior escala dos verbos no particípio passado terminados em –ed. Ao mesmo tempo fizemos novas descobertas, traçando paralelos entre recursos utilizados na linguagem falada e no discurso mais informal de divulgação, o que, segundo Biber (2009), consiste na complexidade da oralidade.

3.3 Os verbos de elocução

Conforme apontado no capítulo 3, na comparação das listas entre o *WebMD* e o *NEJM*, os verbos encontrados por ordem de chavicidade foram: *says* e *tells,* seguidos de *say.* Já na comparação entre *Health News* e o *WebMD* (os sub*corpora* de divulgação), o verbo com a chavicidade mais alta foi *said para o HN*. Isso aponta para uma diferença no estilo de relatar entre os dois sub*corpora*. Parece que o *HN* adota um estilo narrativo mais formal, por relatar seus fatos usando o passado, que é o esperado do discurso indireto. Já o *WebMD* adota um discurso mais informal com o uso de sujeitos na 3ª pessoa do singular e, principalmente, verbos no presente (cf. **Tabela 17**), o que o aproxima da narração falada. (Biber, 1989).

Esses verbos de elocução não apareceram como palavras-chave na lista do *NEJM* e demonstraram pouquíssimas ocorrências no geral (cf. **Tabela 17**). A única ocorrência de *said* foi na estrutura formal ‘*That being said’,* que é uma estrutura fixa para a organização do discurso, e mais formal.

O número de ocorrências de cada um dos *subcorpora* está explicitado na **Tabela 17:**

	NEJM	***HN***	***WebMD***
SAY	2	42	129
SAYS	0	94	**832**
SAID	1	**399**	48
TELLS	0	5	137

Tabela 17: Número de ocorrências dos verbos *say, says, said* e *tells* nos três sub*corpora.*

O **Gráfico 6** apresenta as ocorrências dos verbos de elocução para melhor visualização

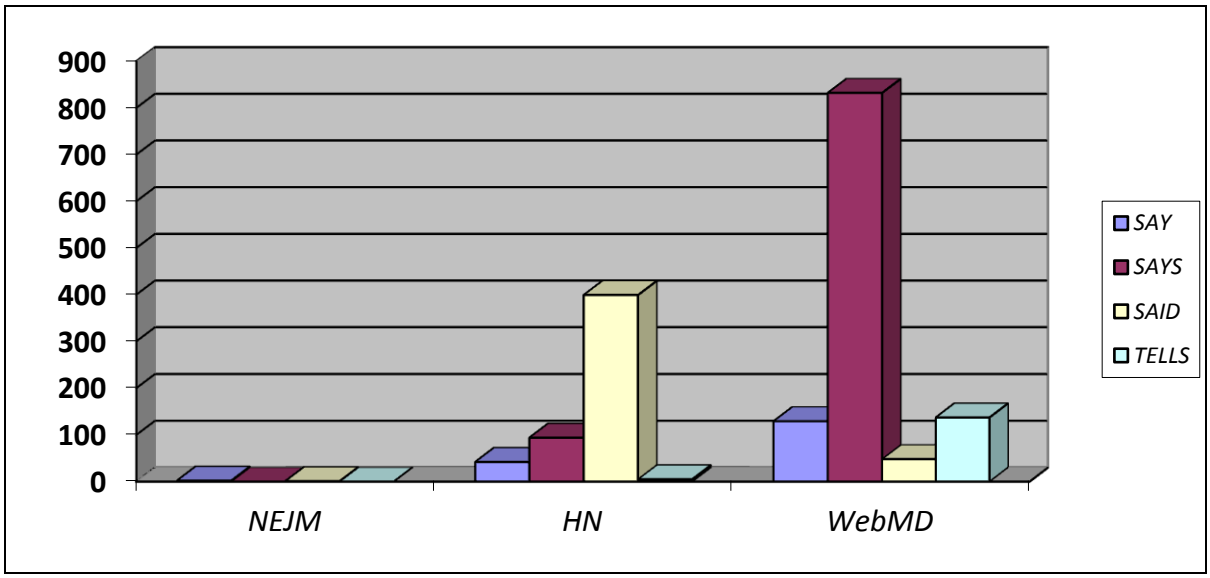
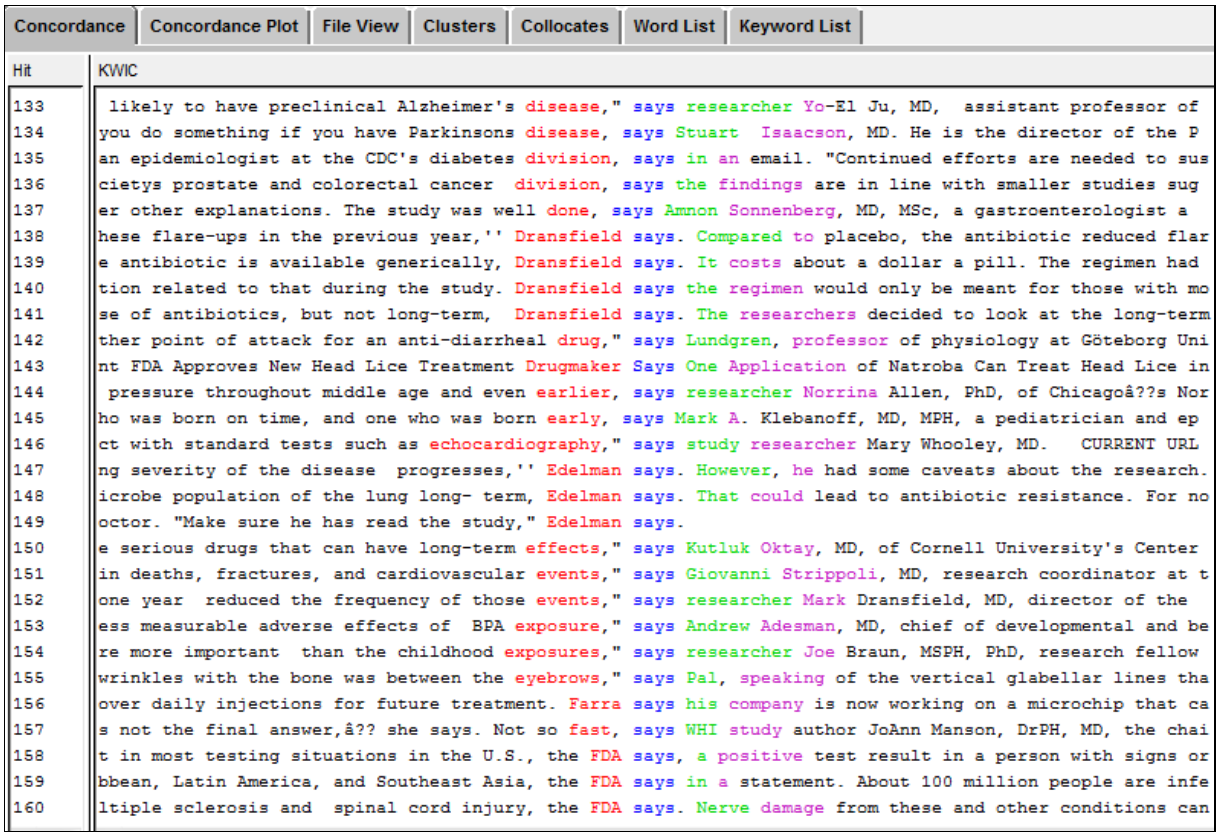


Gráfico 6: Ocorrências de *say, says, said* e *tells* nos três sub*corpora.*

Inicialmente, acreditamos que esses verbos de elocução não representariam um obstáculo à compreensão da leitura dos textos pelo nosso público alvo. Porém, ao exame mais detalhado do co-texto observamos que existiam maiores implicações em relação ao uso do discurso direto e indireto nesses textos.

Os colocados de *says,* por exemplo, nos textos de divulgação *WebMD* apresentam, na maioria dos exemplos (cf. **Figura 36**), o nome de uma autoridade no assunto (ex.: Linha 133: *researcher* Yo-El Ju, MD, *assistant professor* e linha 134: Stuart Isaacson (*director*)) , sua titulação (ex.: Linha 144: *PhD*, linha 145: *MD*, *MPH*) ou o nome de uma instituição de renome, *FDA* (linhas 158,159,160). Isso remete ao fato de que o estudo das características do texto deve ser combinado com o estudo das práticas e culturas de certa comunidade (Hyland, 2004), pois a menção dos nomes de autoridades médicas e instituições renomadas dão credibilidade à informação relatada no artigo. O que é corroborada pela afirmação "[...] (a) autoridade dentro das comunicações científicas pode emprestar autoridade a outras áreas de comunicação que conferem credibilidade à ciência e aos cientistas." (Bazerman, 2007: 33)



Hit	KWIC
133	likely to have preclinical Alzheimer's disease," says researcher Yo-El Ju, MD, assistant professor of
134	you do something if you have Parkinsons disease, says Stuart Isaacson, MD. He is the director of the P
135	an epidemiologist at the CDC's diabetes division, says in an email. "Continued efforts are needed to sus
136	cietys prostate and colorectal cancer division, says the findings are in line with smaller studies sug
137	er other explanations. The study was well done, says Amnon Sonnenberg, MD, MSc, a gastroenterologist a
138	hese flare-ups in the previous year,'' Dransfield says. Compared to placebo, the antibiotic reduced flar
139	e antibiotic is available generically, Dransfield says. It costs about a dollar a pill. The regimen had
140	tion related to that during the study. Dransfield says the regimen would only be meant for those with mo
141	se of antibiotics, but not long-term, Dransfield says. The researchers decided to look at the long-term
142	ther point of attack for an anti-diarrheal drug," says Lundgren, professor of physiology at Göteborg Uni
143	nt FDA Approves New Head Lice Treatment Drugmaker Says One Application of Natroba Can Treat Head Lice in
144	pressure throughout middle age and even earlier, says researcher Norrina Allen, PhD, of Chicagoâ??s Nor
145	ho was born on time, and one who was born early, says Mark A. Klebanoff, MD, MPH, a pediatrician and ep
146	ct with standard tests such as echocardiography," says study researcher Mary Whooley, MD. CURRENT URL
147	ng severity of the disease progresses,'' Edelman says. However, he had some caveats about the research.
148	icrobe population of the lung long- term, Edelman says. That could lead to antibiotic resistance. For no
149	octor. "Make sure he has read the study," Edelman says.
150	e serious drugs that can have long-term effects," says Kutluk Oktay, MD, of Cornell University's Center
151	in deaths, fractures, and cardiovascular events," says Giovanni Strippoli, MD, research coordinator at t
152	one year reduced the frequency of those events," says researcher Mark Dransfield, MD, director of the
153	ess measurable adverse effects of BPA exposure," says Andrew Adesman, MD, chief of developmental and be
154	re more important than the childhood exposures," says researcher Joe Braun, MSPH, PhD, research fellow
155	wrinkles with the bone was between the eyebrows," says Pal, speaking of the vertical glabellar lines tha
156	over daily injections for future treatment. Farra says his company is now working on a microchip that ca
157	s not the final answer,â?? she says. Not so fast, says WHI study author JoAnn Manson, DrPH, MD, the chai
158	t in most testing situations in the U.S., the FDA says, a positive test result in a person with signs or
159	bbean, Latin America, and Southeast Asia, the FDA says in a statement. About 100 million people are infe
160	ltiple sclerosis and spinal cord injury, the FDA says. Nerve damage from these and other conditions can

Figura 36: Seleção de linhas de concordância do verbo *says* no *WebMD*

As mesmas características são apresentadas na observação dos sujeitos de *'say'* (cf. **Figura 37**): *co-authors* (linha 3), *the authors* (linha 4), nomes de autoridades (linhas 8 e 9), *experts* (linhas 16-23), *institutions* (linha 30) *e FDA officials* (linha 34) para citar alguns exemplos:

Hit	KWIC
1	their 30s and 40s, we are shocked and amazed and say, 'That is too young.' But I am fearful we are goi
2	you developed gestational diabetes, follow up and say, â??What can I do to prevent heart disease and di
3	ensity diminished. Yosipovitch and his co-authors say they arent sure why. Previous research has shown
4	the nervous system to release fluids, the authors say. Vital Information: Severe diarrhea and vomiting
5	She says that studying the effects of caffeine, say, on the cardiovascular system, would be challeng
6	conditions. But we need more data before we can say one causes the other," says March Seabrook, MD, a
7	may be time to take a fresh look at patient care, say the two studies. "New health care models must be
8	r James J. Sejvar, MD, of the CDC, and colleagues say movement disorders may be underrecognized effects
9	dy, the researchers write. Cappola and colleagues say the heart benefits of treating thyroid problems h
10	ecommendation' The findings need to be confirmed, say the researchers. "While awaiting confirmation of
11	al detectives looking for a heart disease culprit say very high levels of a blood protein called lipopr
12	o your office and they had lots of wrinkles you'd say 'Aha! That's a great person for me to target to g
13	would take up about 3.5 hours of a doctor's day, say the researchers, who included J. Lloyd Michener,
14	roblems are a stumbling block, Cutter and DeParle say. For instance, Cutter says, "12% of children have
15	release large volumes of water and electrolytes , say Ove Lundgren, MD, PhD, and others from Göteborg U
16	es where the esophagus meets the stomach. Experts say that can be a problem people are born with, but
17	rn at Night Are Men Suffering in Silence? Experts say the findings could signal a couple of things. The
18	chanics of Acid Reflux May Differ Other experts say the findings could also mean that different mecha
19	alcium May Increase Heart Attack Risk But Experts Say Evidence Is Not Convincing By Salynn Boyles WebMD
20	ate Drugs Studied; No Reason to Stop Use, Experts Say (continued) Real Risk or Fluke? WebMD asked two i
21	itment from both you and your family, the experts say. Some say they would reserve it for patients who
22	head off some overexertion injuries? Many experts say yes. The basic snow shovel hasn't changed much si
23	le Inexpensive Way to Screen Bones Other experts say that while it was a new idea to connect wrinkles
24	nslating research into practice." What Your Feet Say About Your Health Diabetes-Related Amputations Do
25	tiffness, as well as improving physical function, say the researchers. "NSAIDs are among the most frequ
26	d to high cholesterol. But the new AAP guidelines say drug treatment "should be considered" only for ki
27	matory arthritic condition in check, doctors here say. Still, the study does not prove cause and effect
28	control in these situations," she says. "Often, I say that the child doesnt need a time out -- sometim
29	e risk factors, may help avoid kidney infections, say the researchers. View Article Sources SOURCES: Sc
30	ational Cancer Institute, and other institutions say they have discovered two genetic variations that
31	ommended Care Most people who wear contact lenses say they know about the wear and care recommendations
32	-identify as one or the other, and the answer may say something about their personalities, a study show
33	Our Destiny? Is it all about the X? I would not say that all is in the 'X,' but as a geneticist I wou
34	cy brand. Who Is at Increased Risk? FDA officials say in a news release that the tampons were manufactu

Figura 37: Seleção de linhas de concordância de *say* no *WebMD.*

A observação das linhas de concordância de *says* e *say* remete à afirmação feita por Zamboni (2002: 52) sobre o discurso de divulgação :

> A presença numerosa de nomes de prestígio empresta ao discurso de divulgação um “efeito real” do discurso da ciência, ao mesmo tempo em que lhe assegura a autoridade e a seriedade da voz que ‘diz a verdade’.

Citamos como exemplos desse procedimento, as linhas 17, 18 e 26 (cf. **Figura 37**) no sub*corpus WebMD*:

Linha 17: ***Experts*** *say the findings...*

Linha 18: ***Other experts*** *say the findings…*

Linha 26: ***The new APP guidelines*** *say drug treatment “should be considered”*

Retomando os colocados de *say* no *WebMD,* dessa vez sob a ótica da léxico-gramática, encontramos quatro padrões:

a) *say* + oração substantiva com *that* (4 ocorrências: linhas 16, 23, 28, 32)

b) *say* + oração substantiva com omissão do *that* (12 ocorrências: 3,6, 8, 9, 11,17, 18, 19, 20, 26, 30, 31)

c) *say* + objeto (*something*)

d) sujeito + *say*

A omissão do *that* mais uma vez aparece demonstrando a proximidade dos textos de divulgação à linguagem falada. Conforme já apontado, a falta de equivalência dessa estrutura com a língua materna do aluno pode ser um obstáculo à fluidez na leitura de um texto destinado ao público leigo.

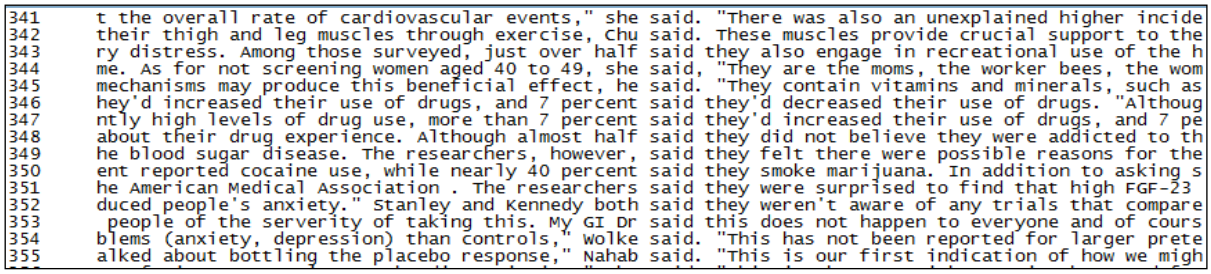
. Outro fato a ser mencionado é que a omissão do *that*, embora não seja uma regra determinada pela gramática da língua, é mais empregada no discurso de divulgação, reforçando a questão da probabilidade no uso da língua.

A diferença entre a gramática como estrutura e a gramática como escolha é uma questão interessante quando se faz um estudo dos gêneros. Segundo Carter e McCarthy (2006: 6), a gramática como estrutura (determinística) está mais preocupada em ditar as regras de construção de sentenças ou orações de forma adequada. Por outro lado, a gramática como escolha (probabilística) envolve vários tipos de elipse ou omissões, fenômeno que só pode ser entendido na observação do co-texto ou do contexto, como é o caso do *that*, já mencionado anteriormente.

Apesar de estarmos analisando gêneros muito próximos, isto é, os dois sub*corpora* de divulgação, notamos que eles carregam características linguísticas particulares, pois enquanto o *WebMD* utiliza como principais verbos de elocução *say e says,* no presente, um dos verbos-chave no *HN* é *said*, com 399 ocorrências (cf. **Tabela 7**, capítulo 2). Isso indica que a característica do estilo apresentado nesse *site* é utilizar o verbo no passado para relatar os depoimentos, diferentemente do que é apresentado no *WebMD*, com um estilo muito mais interativo e próximo do leitor pelo uso do discurso direto (Biber,1988).

A forma de organizar os relatos no *HN* apresenta características de relato das falas mais de acordo com as normas da gramática do discurso indireto. E quando

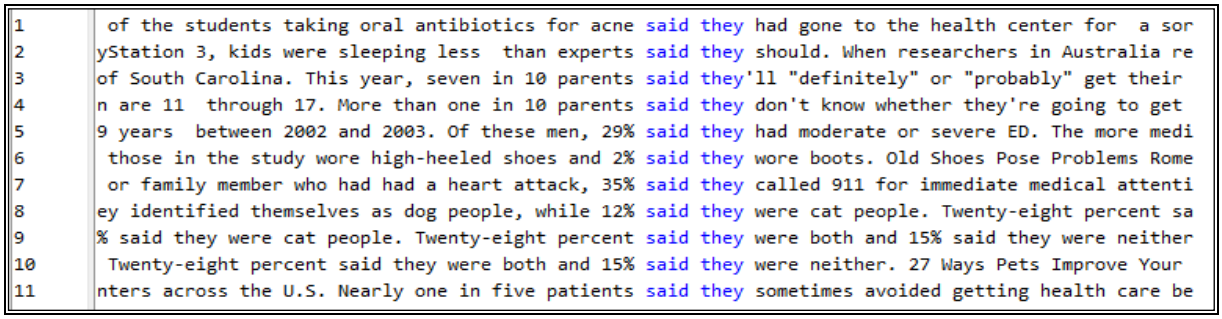
apresenta o discurso direto, ele aparece entre aspas, seguido do verbo de elocução no passado -- *said* -- (cf. **Figura 38**, linhas 341, 344, 345, 354 e 355), o que é o esperado desse tipo de discurso na escrita (Carter e McCarthy, 2006).



341	t the overall rate of cardiovascular events,"	she	said. "There was also an unexplained higher incide
342	their thigh and leg muscles through exercise,	Chu	said. These muscles provide crucial support to the
343	ry distress. Among those surveyed, just over half	said	they also engage in recreational use of the h
344	me. As for not screening women aged 40 to 49,	she	said, "They are the moms, the worker bees, the wom
345	mechanisms may produce this beneficial effect,	he	said. "They contain vitamins and minerals, such as
346	hey'd increased their use of drugs, and 7 percent	said	they'd decreased their use of drugs. "Althoug
347	ntly high levels of drug use, more than 7 percent	said	they'd increased their use of drugs, and 7 pe
348	about their drug experience. Although almost half	said	they did not believe they were addicted to th
349	he blood sugar disease. The researchers, however,	said	they felt there were possible reasons for the
350	ent reported cocaine use, while nearly 40 percent	said	they smoke marijuana. In addition to asking s
351	he American Medical Association . The researchers	said	they were surprised to find that high FGF-23
352	duced people's anxiety." Stanley and Kennedy both	said	they weren't aware of any trials that compare
353	people of the serverity of taking this. My GI Dr	said	this does not happen to everyone and of cours
354	blems (anxiety, depression) than controls," Wolke	said.	"This has not been reported for larger prete
355	alked about bottling the placebo response," Nahab	said.	"This is our first indication of how we migh

Figura 38: Seleção de linhas de concordância de *said* no *HN.*

Da mesma forma que nos verbos descritos anteriormente, a omissão do pronome *that* (cf. **Figura 39)**, nas linhas 346 a 353, também aparece no *HN:*



1	of the students taking oral antibiotics for acne	said	they	had gone to the health center for a sor
2	yStation 3, kids were sleeping less than experts	said	they	should. When researchers in Australia re
3	of South Carolina. This year, seven in 10 parents	said	they	'll "definitely" or "probably" get their
4	n are 11 through 17. More than one in 10 parents	said	they	don't know whether they're going to get
5	9 years between 2002 and 2003. Of these men, 29%	said	they	had moderate or severe ED. The more medi
6	those in the study wore high-heeled shoes and 2%	said	they	wore boots. Old Shoes Pose Problems Rome
7	or family member who had had a heart attack, 35%	said	they	called 911 for immediate medical attenti
8	ey identified themselves as dog people, while 12%	said	they	were cat people. Twenty-eight percent sa
9	% said they were cat people. Twenty-eight percent	said	they	were both and 15% said they were neither
10	Twenty-eight percent said they were both and 15%	said	they	were neither. 27 Ways Pets Improve Your
11	nters across the U.S. Nearly one in five patients	said	they	sometimes avoided getting health care be

Figura 39: Linhas de concordância demonstrando *said* seguido da próxima oração, sem o pronome *that* no *HN.*

As mesmas características são inferidas para o uso de *tells,* que é uma das 50 palavras-chave (cf.**Tabela 6**, linha 24, capítulo 2) na comparação do *WebMD* com o *NEJM.*

A partir da tela que apresenta as linhas de concordância para *tells* no *AntConc*, (cf. **Figura 40**), fica claro o padrão desse verbo: *tells* + objeto, o que nos parece uma estratégia visual caso o professor queira utilizar uma concordância na sala de aula. Aqui ficam muito claros os sujeitos e os objetos de *tells:*

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Concordance Hits 137

Hit	KWIC
1	my of Pediatrics. The groundbreaking new advisory tells parents and pediatricians to screen high-ris
2	tayed above 140 over 90 had the highest,\u2022?? Allen tells WebMD. \u2022??The longer people can delay the o
3	l, your pediatrician should talk to you," Altmann tells WebMD. "In some instances, a low-fat diet an
4	as been inactive for at least six months, Aranow tells WebMD. And certain rheumatoid arthritis medi
5	n exciting first step, albeit preliminary," Atala tells WebMD. "This is the first published report
6	r daily with groups with the lowest intake, Aune tells WebMD. The amount of fiber eaten by those wh
7	sling, PhD, of the University of Texas at Austin, tells WebMD. He believes this is because earlier s
8	same social experience they get at school." Bass tells WebMD that chronically ill children need bot
9	r Absorbency," Kimberly-Clark spokesman Bob Brand tells WebMD. "We are looking for about 300 boxes,
10	re are canned foods that are good for you," Braun tells WebMD. "Those are healthy foods and you sho
11	ore sensitive to estrogen than others, Braunstein tells WebMD. As the body mass index , a measure of
12	an should do a thorough physical exam, Braunstein tells WebMD, making sure there are no tumors, thyr
13	nz, MD, of the University of Southern California tells WebMD. \u2022Young women seem to be benefiting mo
14	could reflect brain changes happening," Carrillo tells WebMD. The study finding is not surprising,
15	in terms of not realizing it was that high," Chan tells WebMD. But, he says, to put it in perspectiv
16	g. As bleak as the findings may sound, Van Cleave tells WebMD she sees a bright spot: ''The most imp
17	usceptible to the effects of caffeine,\u2022 Cornelis tells WebMD. \u2022More research on the precise functi
18	evels that produce pleasurable effects,\u2022 Cornelis tells WebMD. So what does this mean? \u2022Clearly the
19	was an increased risk of heart disease," Daniels tells WebMD. "This held true even after adjusting
20	ation figures from the U.S. Census Bureau. Davis tells WebMD that 1,558 people polled were white, 2
21	her who has studied the issue for almost a decade tells WebMD that the jury is still out on the trea
22	rsity of Oxford professor of rheumatic diseases , tells WebMD. \u2022??What this study says is that we sh
23	ads to even more shortness of breath, Dransfield tells WebMD. "A typical person with moderate to se
24	or vitamin D, or exposure to sunlight," Ezekowitz tells WebMD Commenting on the study, Andrew Smith,
25	ce the risk of developing osteoporosis, Ezekowitz tells WebMD that: Everyone should exercise their e
26	managing their disease on a daily basis,\u2022 Farra tells WebMD. Drug-Delivery Chip Preferred Over Dai
27	Center for Reproductive Medicine and Fertility , tells WebMD that the team used standard PGD techni
28	l director of the National Parkinson Foundation, tells WebMD. He reviewed the results. Although Oku
29	sor of gerontology at the University of Georgia, tells WebMD. The information is especially importa
30	third cause of death within two decades," Gershon tells WebMD. Energy-Boosting Foods for COPD Calcul
31	ymptoms and others don't? Gut sensitivity, Gibson tells WebMD. "It's an abnormal response to a stimu
32	Jennifer Logue, MD, of the University of Glasgow, tells WebMD via email. \u2022??We also already knew tha
33	s fractures are fairly easy to treat," Goodwillie tells WebMD. Resting six to eight weeks is typical
34	mething every day and it does have a cost," Grant tells WebMD. "We know there are other approaches t
35	ind confirmation from another laboratory," Guilak tells WebMD. While Hedrick's team is interested in

Figura 40: Primeiras 35 linhas de concordância com o verbo *tells* no *WebMD*

A esse respeito, Pinheiro (2002: 268) afirma que “A fugacidade dos gêneros na atualidade imprime um caráter volúvel aos gêneros, pois permite a variabilidade, a troca, a mudança, a mixagem, a hibridação”.

Existe uma transferência de características de gêneros que Bazerman (1988) chama de discurso híbrido, em que o recurso estilístico de um gênero empresta características a outro como forma de legitimação. O gênero de divulgação precisa dessa legitimação para obter credibilidade. O discurso da ciência é tomado como legítimo, assim, para um artigo de divulgação ter credibilidade, é preciso que nomes de autoridades médicas sejam mencionados.

Essa premissa foi confirmada na observação dos sujeitos dos verbos de elocução encontrados nos dois subcorpora de divulgação.

A apresentação desses verbos, usados para depoimentos, fazem um contraste com a voz passiva utilizada nos artigos acadêmicos, em que o agente não deve aparecer. Nos textos de divulgação, em contraste, encontramos as características da linguagem falada como recurso de legitimação e ao mesmo tempo de interação com o público leitor.

Como apontado anteriormente, a palavra *researchers* é omitida no discurso acadêmico como forma de omitir o agente, uma vez que esse é de conhecimento do leitor. No discurso de divulgação, ao contrário, essa agentividade[46] aparece explicitada e confere um *status* de autoridade à informação veiculada.

Da mesma forma que os verbos *say, says e tells,* os pronomes não são o foco deste estudo. Porém, são dignos de menção por ocuparem um lugar de destaque nas palavras-chave (cf. **Tabela 7 e 8, cap. 2**), demonstrando ser uma marca significativa nos textos de divulgação indicando interação com o leitor, o que se contrapõe ao discurso acadêmico, que tenta ser mais distante e impessoal (Biber, 1988; Hyland 2004).

Exemplos dos pronomes encontrados na lista de comparação entre os *corpora* de divulgação *HN* e *WebMD (*cf. **Tabela 8**):

Pronomes pessoais: ‘*he´*(l.8), ‘*they´* (l.10), ‘*you´* (l.12), ‘*it´* (l.14), ‘*she´* (l.23).
Pronomes possessivos: ‘*your´* (l.9), *‘their´* (l.19).

O uso do discurso direto e de verbos no presente, segundo Biber (1988), indica um estilo mais interativo de texto, que procura se aproximar do leitor com o uso de vocabulário mais coloquial. Pode estar aí um indício da maior dificuldade encontrada pelos alunos na leitura de textos que não sejam acadêmicos: eles

[46] Agentividade em linguística é o conceito relativo à pratica da ação verbal, não necessariamente a condição do sujeito da oração.

apresentam características da variante falada, o que por vezes fica distante do conhecimento desses alunos que não tem grande exposição à segunda língua.

A importância da conscientização das características específicas de cada gênero não se restringe ao aprendizado da leitura, apesar de ser esse nosso foco. Pode-se afirmar que essa importância se estende a outras áreas do conhecimento, como à escrita acadêmica e à tradução.

3.4 As colocações nominais a partir dos substantivos e adjetivos cognatos

Um sintagma nominal é um grupo de palavras com um substantivo como núcleo e outros elementos que o modificam, os modificadores (Lanzoni, 2010). A compreensão da estrutura dos sintagmas é importante tanto para os leitores que não conhecem o significado de todas as palavras da sentença, quanto para aqueles que não costumam ter problemas com o vocabulário, mas apresentam dificuldades na compreensão articulada do texto. Em português, o adjetivo pode preceder o substantivo ou ser posposto a ele. Em inglês, os adjetivos precedem os substantivos, a não ser em casos específicos como os de adjetivos predicativos que seguem os verbos (Biber, 1998). Isso quer dizer que o núcleo vem sempre depois do modificador, razão pela qual a observação do comportamento desses sintagmas é importante para o ensino instrumental de leitura.

3.4.1 A investigação dos adjetivos cognatos

Neste item discutiremos os três primeiros adjetivos que apareceram nas listas de comparação entre os três *subcorpora* (cf. **Tabela 6, cap. 2):** *adverse, cardiac* e *virologic,* que são cognatos (de origem latina) em português. Normalmente, esses adjetivos não causam obstáculos para o entendimento da leitura. Mesmo que o aprendiz seja iniciante, consegue entender os textos de sua área de especialidade, pois a terminologia é recorrente e o leitor-aprendiz está familiarizado com esse léxico. Por essa razão, os cardiologistas não encontram problemas ao lerem os

artigos sobre cardiologia, os ortopedistas conseguem entender os artigos sobre ortopedia e assim por diante.

Como podemos observar na **Tabela 18**, esses adjetivos são muito mais frequentes no *subcorpus* acadêmico e muito pouco utilizados nos dois *subcorpora* de divulgação, o que confirma que a grande porcentagem de palavras de origem latina pode facilitar a leitura desses textos:

	NEJM	***HN***	***WebMD***
ADVERSE	244	14	10
CARDIAC	164	9	11
VIROLOGIC	101	0	0

Tabela 18: Número de ocorrências dos adjetivos-chave nos três sub*corpora.*

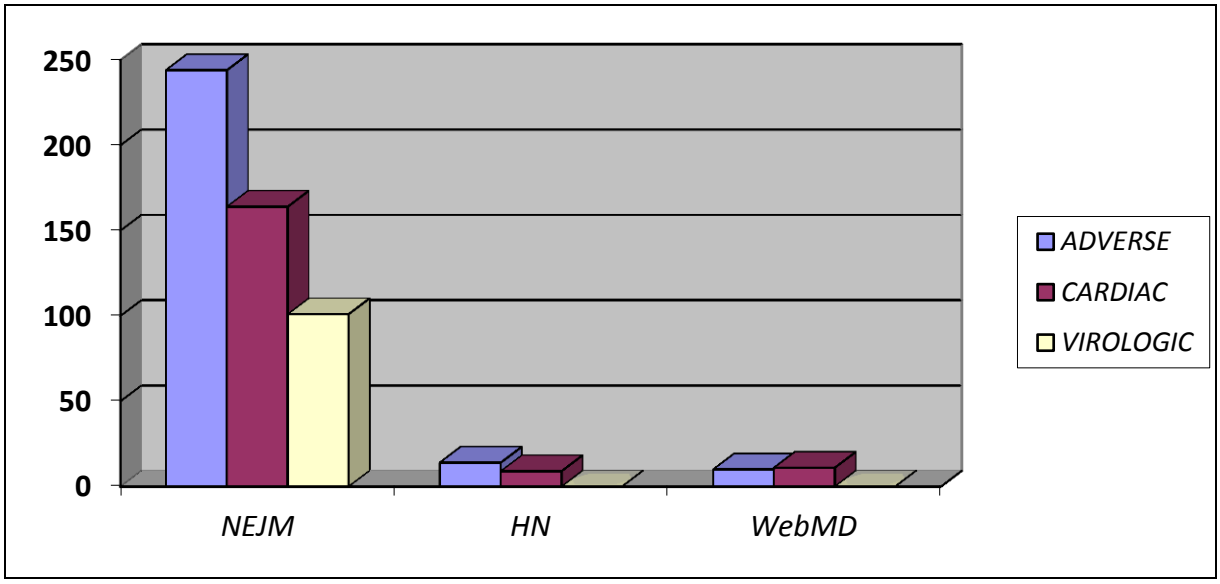
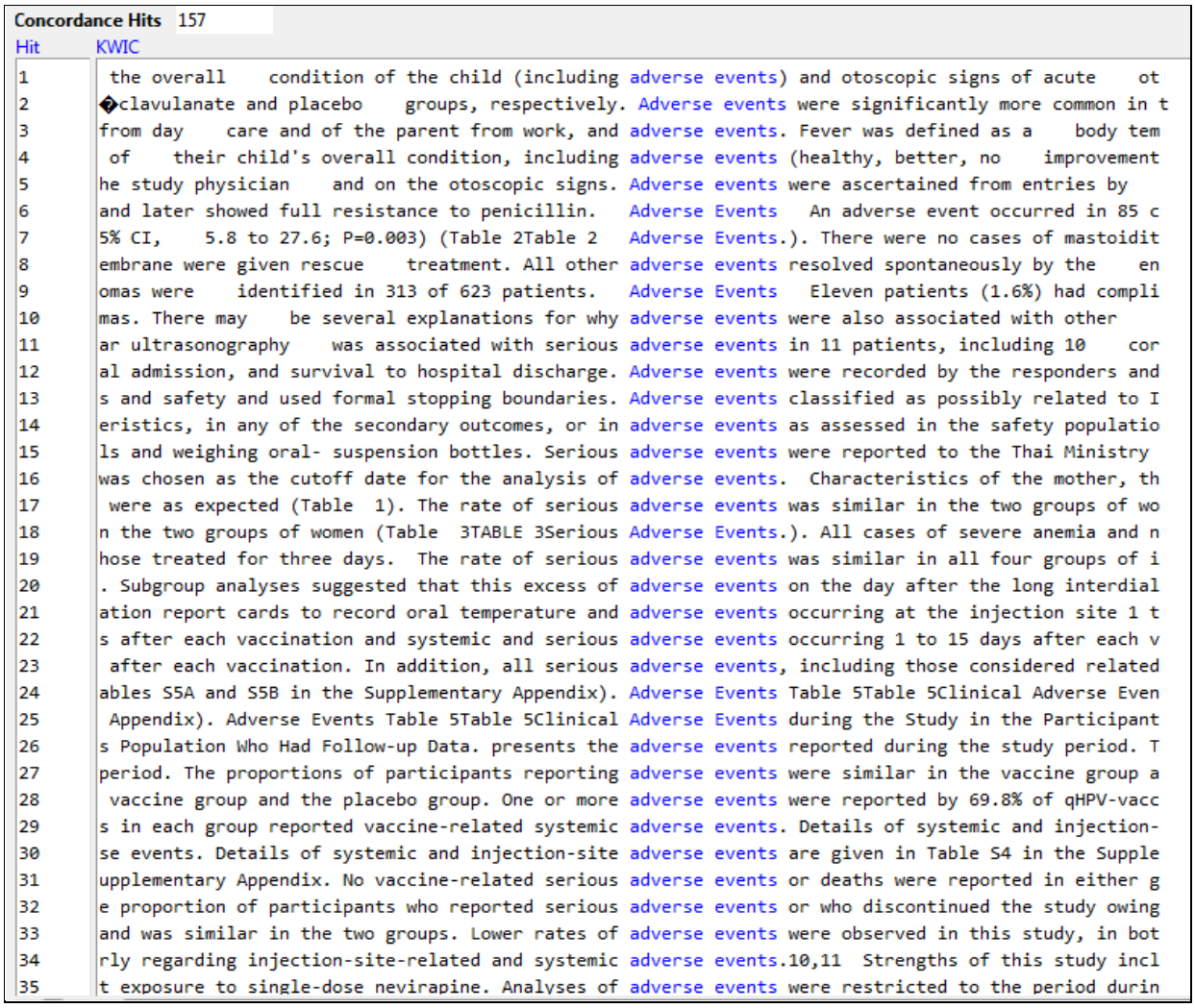


Gráfico 7: Ocorrências de *adverse, cardiac e virologic* nos três sub*corpora.*

Partindo desses dados, decidimos investigar mais a fundo os colocados desses adjetivos. O primeiro adjetivo investigado foi *adverse*. Já na primeira consulta às linhas de concordância no *NEJM,* pudemos verificar que o colocado mais frequente de *adverse* é *events*. Em uma busca por *adverse events*, encontramos

157 ocorrências dessa combinação (cf. **Figura 41**), demonstrando, portanto, uma colocação adjetiva. Vale observar que *events* também é uma palavra-chave:



Concordance Hits 157

Hit	KWIC
1	the overall condition of the child (including adverse events) and otoscopic signs of acute ot
2	clavulanate and placebo groups, respectively. Adverse events were significantly more common in t
3	from day care and of the parent from work, and adverse events. Fever was defined as a body tem
4	of their child's overall condition, including adverse events (healthy, better, no improvement
5	he study physician and on the otoscopic signs. Adverse events were ascertained from entries by
6	and later showed full resistance to penicillin. Adverse Events An adverse event occurred in 85 c
7	5% CI, 5.8 to 27.6; P=0.003) (Table 2Table 2 Adverse Events.). There were no cases of mastoidit
8	embrane were given rescue treatment. All other adverse events resolved spontaneously by the en
9	omas were identified in 313 of 623 patients. Adverse Events Eleven patients (1.6%) had compli
10	mas. There may be several explanations for why adverse events were also associated with other
11	ar ultrasonography was associated with serious adverse events in 11 patients, including 10 cor
12	al admission, and survival to hospital discharge. Adverse events were recorded by the responders and
13	s and safety and used formal stopping boundaries. Adverse events classified as possibly related to I
14	eristics, in any of the secondary outcomes, or in adverse events as assessed in the safety populatio
15	ls and weighing oral- suspension bottles. Serious adverse events were reported to the Thai Ministry
16	was chosen as the cutoff date for the analysis of adverse events. Characteristics of the mother, th
17	were as expected (Table 1). The rate of serious adverse events was similar in the two groups of wo
18	n the two groups of women (Table 3TABLE 3Serious Adverse Events.). All cases of severe anemia and n
19	hose treated for three days. The rate of serious adverse events was similar in all four groups of i
20	. Subgroup analyses suggested that this excess of adverse events on the day after the long interdial
21	ation report cards to record oral temperature and adverse events occurring at the injection site 1 t
22	s after each vaccination and systemic and serious adverse events occurring 1 to 15 days after each v
23	after each vaccination. In addition, all serious adverse events, including those considered related
24	ables S5A and S5B in the Supplementary Appendix). Adverse Events Table 5Table 5Clinical Adverse Even
25	Appendix). Adverse Events Table 5Table 5Clinical Adverse Events during the Study in the Participant
26	s Population Who Had Follow-up Data. presents the adverse events reported during the study period. T
27	period. The proportions of participants reporting adverse events were similar in the vaccine group a
28	vaccine group and the placebo group. One or more adverse events were reported by 69.8% of qHPV-vacc
29	s in each group reported vaccine-related systemic adverse events. Details of systemic and injection-
30	se events. Details of systemic and injection-site adverse events are given in Table S4 in the Supple
31	upplementary Appendix. No vaccine-related serious adverse events or deaths were reported in either g
32	e proportion of participants who reported serious adverse events or who discontinued the study owing
33	and was similar in the two groups. Lower rates of adverse events were observed in this study, in bot
34	rly regarding injection-site-related and systemic adverse events.10,11 Strengths of this study incl
35	t exposure to single-dose nevirapine. Analyses of adverse events were restricted to the period durin

Figura 41: Seleção das linhas de concordância com a palavra de busca *adverse events.*

O utilitário *collocates* (cf. **Figura 42**) e os *clusters* (cf. **Figura 43)** confirmaram a importância dessa colocação, além do aparecimento de outras sequências mais longas formando diferentes sintagmas nominais como: *major serious adverse events* e *rates of adverse events.* Para estudantes brasileiros a identificação dessa ordem sem preposições causa problemas de compreensão, como no caso de:

Major serious adverse events	Eventos adversos graves

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Total No. of Collocate Types: 20 **Total No. of Collocate Tokens:** 627

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
1	191	0	0	-1	adverse
2	137	1	136	8.21653	events
3	53	0	0	-1	Adverse
4	33	1	32	8.76659	Events
5	30	30	0	8.84208	serious
6	28	28	0	7.64739	major
7	28	0	28	7.71012	event
8	27	0	27	8.08342	cardiovascular
9	23	23	0	1.32643	of
10	13	13	0	5.53282	reported
11	11	11	0	3.76924	an
12	9	9	0	0.73601	and
13	7	0	7	6.17961	effects
14	6	6	0	4.94616	related
15	6	6	0	1.57657	for
16	5	5	0	0.72644	with
17	5	5	0	3.29870	Table
18	5	5	0	8.71655	Reported
19	5	5	0	8.90919	Major
20	5	5	0	5.73927	common

Figura 42: Colocados de *adverse* no *NEJM.*

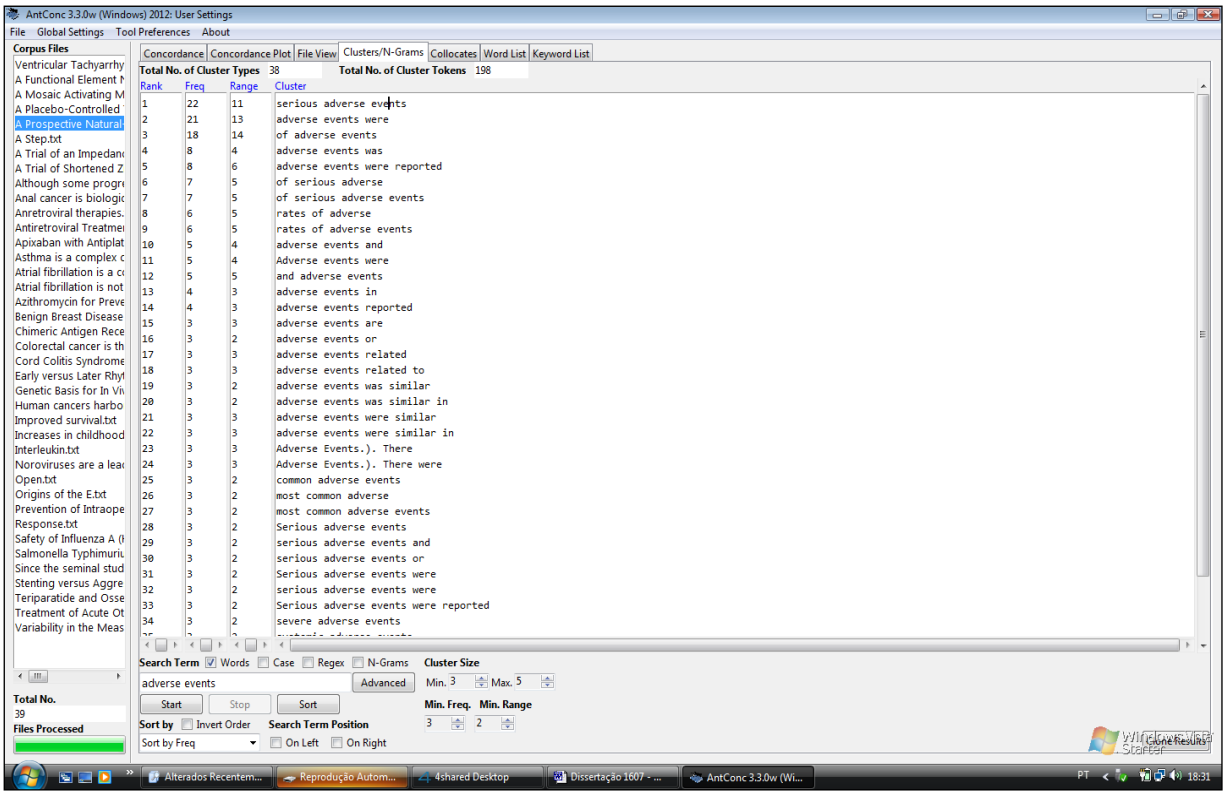


Figura 43: *Clusters* de *adverse events* no *NEJM.*

Como mencionado anteriormente, o segundo adjetivo cognato investigado foi *cardiac* com 164 ocorrências no *NEJM.* Seu principal colocado foi *arrest,* com 69 ocorrências no singular e 25 no plural. A partir da observação de seu co-texto, obtivemos a colocação *cardiac arrest.* Vale lembrar que *arrest* é um substantivo-chave no *NEJM* e será discutido com os substantivos cognatos.

Prosseguindo na análise do co-texto, encontramos como colocados à esquerda de *cardiac arrest* exemplos de adjetivos hifenizados: *out-of-hospital cardiac arrest(s)* (linhas 29-35; 23 ocorrências), *bystander-witnessed cardiac arrest(s)* (linhas 1, 2 e 3; 6 ocorrências) e *EMS-witnessed cardiac arrests* (2 ocorrências). Alguns desses exemplos[47] estão demonstrados na **Figura 44** e formam uma colocação nominal precedendo o núcleo. Assim, nas construções citadas temos *cardiac arrest* como uma colocação adjetiva e uma unidade de sentido estendida:

[47] O Anexo D apresenta a lista completa de ocorrências das colocações estendidas com *cardiac arrest(s).*

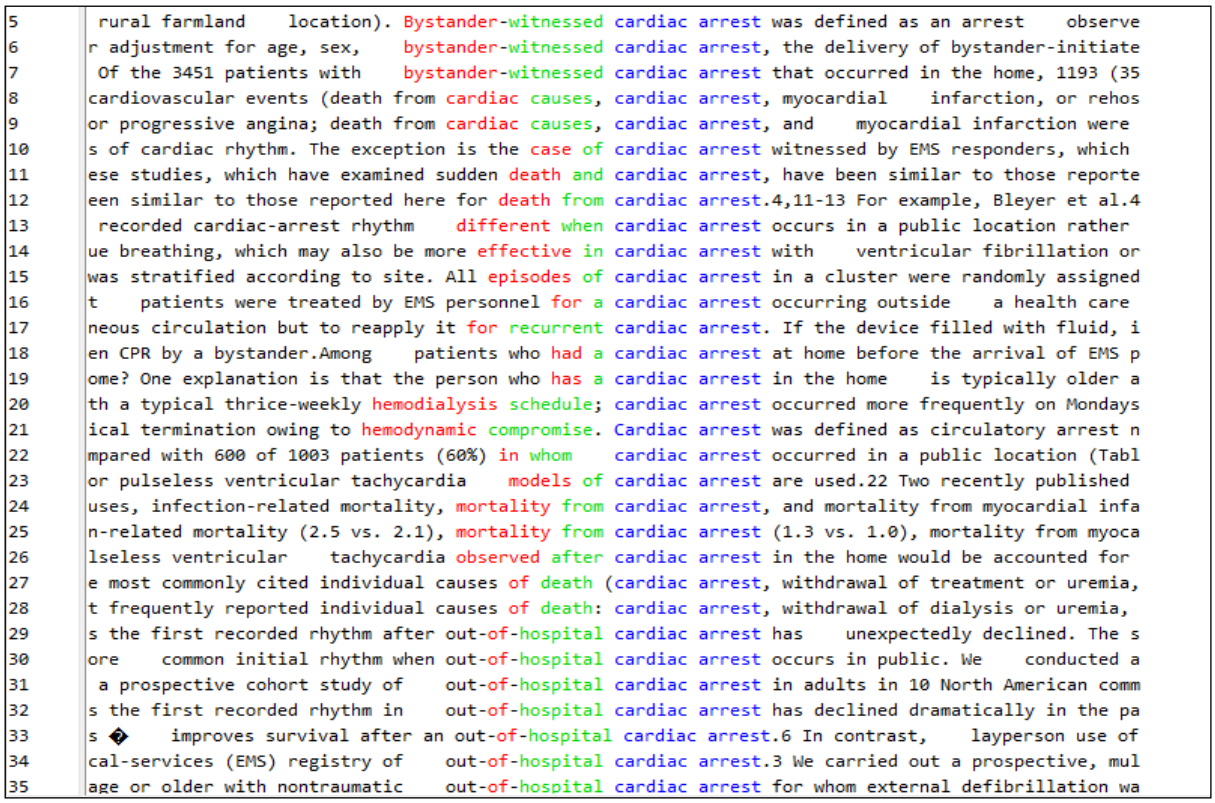


Figura 44: Seleção de linhas de concordância de *cardiac arrest* no *NEJM.*

O adjetivo *virologic*, terceiro analisado, apresentou 101 ocorrências no NEJM e nenhuma nos outros *subcorpora*. No *NEJM* as ocorrências originaram de apenas dois dos 39 textos que compõem o *subcorpus*, demonstrando ser um léxico específico dessa área.

No caso de *virologic*, foram encontradas as mesmas características dos adjetivos anteriores: faz parte de um sintagma nominal denso e com informação compactada, característica do gênero acadêmico (Biber,1998). Apresentamos os exemplos extraídos dos clusters encontrados a partir da busca de *virologic.* (cf. **Figura 45**).

Exemplos:

Linha 8: *a sustained virologic response.*

Linha 12: *an extended rapid virologic response*

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keywo

Total No. of Cluster Types 29 **Total No. of Cluster Tokens** 441

Rank	Freq	Range	Cluster
1	51	2	virologic response
2	36	3	virologic failure
3	32	1	sustained virologic
4	32	1	sustained virologic response
5	18	1	rapid virologic
6	18	1	rapid virologic response
7	17	1	a sustained virologic
8	17	1	a sustained virologic response
9	17	1	extended rapid virologic
10	17	1	extended rapid virologic response
11	16	1	an extended rapid virologic
12	16	1	an extended rapid virologic response
13	14	2	of virologic
14	13	1	of sustained virologic
15	13	1	of sustained virologic response
16	12	1	had an extended rapid virologic
17	12	2	of virologic failure
18	12	2	virologic failure or
19	9	3	had virologic
20	9	3	had virologic failure
21	8	1	had a sustained virologic
22	8	1	had a sustained virologic response
23	7	1	rate of sustained virologic
24	7	1	rate of sustained virologic response
25	7	2	virologic failure or death
26	6	2	of virologic failure or
27	6	1	rates of sustained virologic
28	6	1	rates of sustained virologic response
29	5	3	Virologic failure

Figura 45: *Clusters* de *virologic* no *NEJM*.

Como afirmado anteriormente, as colocações nominais são um aspecto importante na instrução de leitura de textos acadêmicos, pois contém informação densa e condensada caracterizada por grupos nominais, palavras longas, adjetivos atributivos e coligações adjetivas[48] (Biber, 1988; Carter e McCarthy, 2006: 331). Dessa forma, é necessário que se chame atenção para esses aspectos a fim de que a compreensão seja possível. Assim, um aprendiz com necessidades específicas vai se beneficiar de uma atividade com esse enfoque mais especializado.

[48] Segundo Tagnin (2005), coligações são combinações consagradas de elementos linguísticos em que o colocado é uma palavra gramatical, por exemplo em inglês, *used **to,*** em português, utilizado **por**. As coligações de regência podem ser com verbos, substantivos, adjetivos e advérbios.

No caso dos adjetivos, é sabido que precedem os substantivos em inglês. É importante que esse ponto seja enfatizado ao aluno para que ele possa perceber essa ordem diferenciada entre sua língua materna e a língua de estudo, no caso o inglês (Bowker, 2002). Mais uma vez, nesse caso o uso de um *corpus* pode demonstrar exatamente o tipo de colocação presente na área de estudo de cada um.

3.4.2 Os colocados dos substantivos cognatos

Neste item analisamos os substantivos cognatos que apareceram entre as 50 palavras na lista comparativa do *NEJM* com os outros sub*corpora*: *mortality* (177 ocorrências), *efficacy* (153 ocorrências), *interval* (114 ocorrências) e *arrest* por sua alta frequência, com 107 ocorrências.

Os dados comprovam a forte presença dos cognatos de origem latina nos textos acadêmicos, o que corrobora a hipótese inicial de que aqui esteja um indício de que o léxico não é um problema para a leitura instrumental de áreas de especialidade:

	NEJM	***HN***	***WebMD***
MORTALITY	177	24	18
EFFICACY	153	13	5
INTERVAL	114	1	1
ARREST	107	0	0

Tabela 19: Substantivos cognatos de origem latina por ordem de ocorrências a partir do *NEJM*.

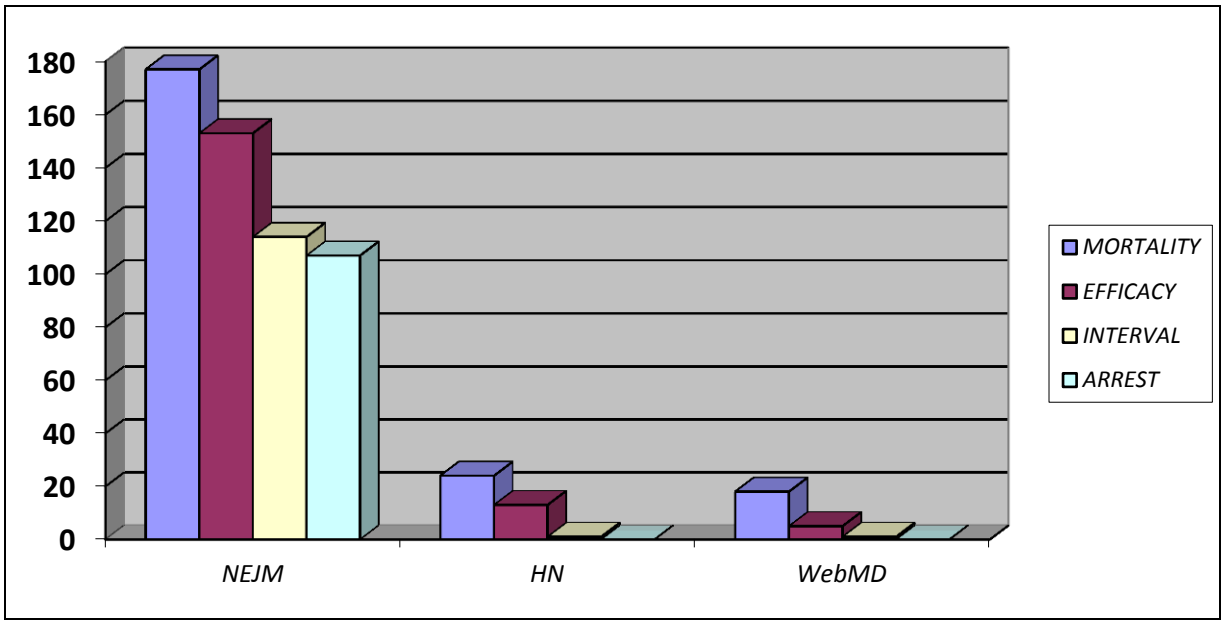


Gráfico 8: Ocorrências de *mortality, efficacy, interval* e *arrest* nos três sub*corpora*.

O primeiro dos substantivos que chamou atenção na lista comparativa dos subcorpora *NEJM* e *HN* foi *mortality*, por ter o maior número de ocorrências (177) em relação aos outros dois *subcorpora* (24 no *HN* e 18 no *WebMD*), o que por si só não causaria problemas de entendimento por ser de origem latina e um cognato do português. A questão apareceu quando observamos seus colocados e percebemos que o substantivo *mortality* apresentava duas funções; a) a de núcleo do sintagma, que era esperado; e b) como componente dos modificadores de outros substantivos.

A **Figura 46** demonstra os exemplos de *mortality* como substantivo nas linhas 77 e 79, e como modificador nas linhas 74, 75, 76, 79 e 80. A título de exemplificação apresentamos uma seleção dessas linhas. A lista das 177 ocorrências encontra-se no Anexo B.

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Concordance Hits 177

Hit	KWIC
74	ven more worrisome when it comes to hospital-wide **mortality** measures that encompass a diverse portf
75	ates Abstract Several countries use hospital-wide **mortality** rates to evaluate the quality of hospita
76	results suggest that efforts to use hospital-wide **mortality** rates to evaluate the quality of care m
77	hasing. We found that estimates of hospital-wide **mortality** could vary, sometimes widely, among meth
78	e of the four methods for assessing hospital-wide **mortality** that we studied are currently marketed
79	desired, and tabulated accurately. Hospital-wide **mortality** encompasses wide-ranging diagnoses, per
80	ethods to calculate risk-adjusted, hospital-wide **mortality** rates with the use of data from standard
81	methods. We also compared hospitals' standardized **mortality** ratios and classification of hospitals
82	(Figure 2FIGURE 2 Percentages of Hospitals with Mortality Rates Higher or Lower Than Expected for
83	lity ratios and classification of hospitals with **mortality** rates that were higher or lower than exp
84	ide mortality estimates. Differences in hospital **mortality** findings among methods also probably rel
85	ion criteria, two methods calculated in-hospital **mortality** rates (4.0% and 5.9%) that were twice th
86	t the benefit of endoscopic screening in reducing **mortality**31,32 from and the incidence33,34 of prox
87	ethods. A broader question is whether in-hospital **mortality** rates are valid quality indicators.4-6,2
88	trauma or advanced cancer), linking in-hospital **mortality** with hospital quality is clearly problem
89	was associated with an increase in postoperative **mortality**21,22 or in the amount of anesthetic agen
90	tients who were older and had higher in-hospital **mortality**, longer lengths of stay, and more respir
91	r and Dr. Foster methods had overall in- hospital **mortality** rates of 5.9% and 4.0%, respectively, pe
92	ng the study period, with an overall in-hospital **mortality** rate of 2.1% (Table 1TABLE 1 Hospital D
93	nt of anesthesia administered or in postoperative **mortality**.11,21,22 Both protocols in the B-Unaware
94	noses for which the linkages between in-hospital **mortality** and quality are problematic. In addition
95	the absence of an association between in-hospital **mortality** and quality or, given the study designs
96	to determine if there were differences in overall **mortality**; cause-specific **mortality**, with causes g
97	f discharges and hospitals, including in-hospital **mortality** rates; the age, sex, and race or ethnic
98	fectiveness of flexible sigmoidoscopy in reducing **mortality** related to distal colorectal cancer refl
99	ferent inclusion and exclusion criteria in their **mortality** algorithms. We therefore calculated numb
100	and sigmoid had little effect on the incidence or **mortality** results. Cancer Treatment, Screening Com
101	oup Table 2Table 2Colorectal-Cancer Incidence and **Mortality**. shows colorectal-cancer incidence and m
102	tal, and Proximal Colorectal-Cancer Incidence and **Mortality**.). The cumulative incidences of proximal
103	rtant decrease in colorectal-cancer incidence and **mortality**. The incidence of colorectal cancer was
104	it with regard to colorectal-cancer incidence and **mortality** was limited to the distal colon, but onl
105	reening. The observed reductions in incidence and **mortality** are similar to the results of the United
106	Mortality. shows colorectal-cancer incidence and **mortality** in the intervention group as compared wi
107	ond examination on colorectal-cancer incidence or **mortality**. In the United States, endoscopic screen
108	ely; P<0.001 for both comparisons). Incidence and **Mortality** According to Location and Stage of Cance

Figura 46: Seleção de colocados de *mortality* no *NEJM.*

Como é sabido, os substantivos aparecem na função de modificadores nas colocações substantivas. O importante aqui é demonstrar, em primeiro lugar, que a pesquisadora não tinha qualquer ideia pré-concebida ao iniciar a análise. A observação das linhas de concordância foi apresentando os padrões utilizados na língua. Em segundo lugar, que esse fenômeno ocorre tanto em português quanto em inglês, com a diferença de que, em português, para que o substantivo seja utilizado como modificador, é necessário o uso de uma preposição (de, do, das, etc.). Em inglês isso não se faz necessário, basta que se faça a inversão (Lanzoni, 2010) (cf. **Tabela 20**):

Hospital-wide mortality rates	Taxas de mortalidade hospitalar
In-hospital mortality rates	Taxas de mortalidade intra-hospitalar ou em pacientes internados
Colorectal-cancer incidence or mortality	Mortalidade ou incidência de câncer colorretal

Tabela 20: Exemplos de colocações substantivas em inglês e suas traduções em português.

Como demonstrado, ‘de mortalidade’ (*mortality*) e ‘de câncer’ (*cancer*) têm função adjetiva (modificadores).

A diferença encontrada nesses sintagmas em inglês foram os modificadores (adjetivos) compostos por hífen, que não encontram equivalência estrutural em português, como nos exemplos encontrados na **Figura 47**, em que temos:

Higher-than-expected mortality, lower-than-expected mortality e *as-expected mortality.*

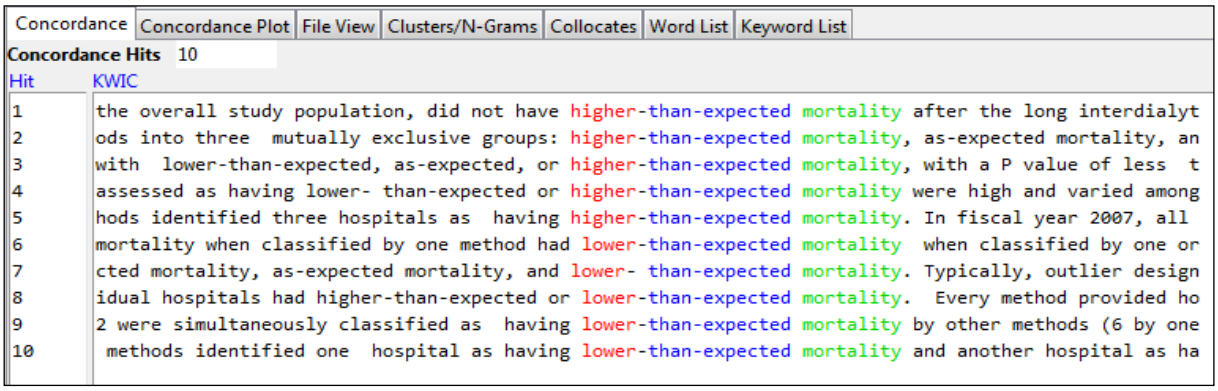


Figura 47: Linhas de concordância de *mortality* como núcleo em uma colocação substantiva no *NEJM.*

A **Figura 48** apresenta linhas de concordância geradas no *Antconc* com a palavra de busca *hospital-wide* no *NEJM,* em que podemos observar *mortality* como componente de um modificador do substantivo *rates*:

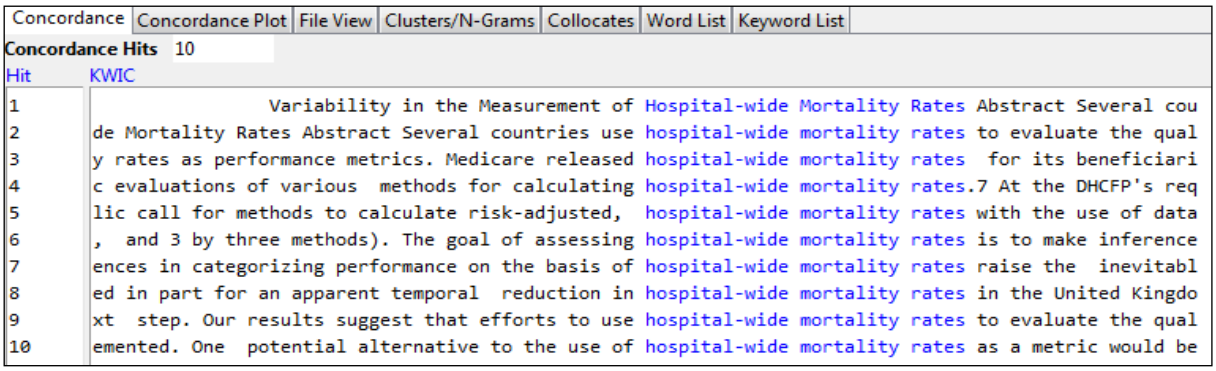


Figura 48: Linhas de concordância com *hospital-wide mortality rates* no *NEJM.*

Dessa forma, obtivemos os sintagmas nominais compostos por *mortality,* ora como núcleo do sintagma nominal, ora como modificador. Existem outros exemplos de substantivos e modificadores, porém por questão de espaço não apresentaremos todos.

O segundo substantivo em número de ocorrências na lista comparativa foi *efficacy.* Assim como *mortality, efficacy* aparece na função de substantivo e de modificador nos grupos nominais encontrados no *NEJM.* A **Figura 49** demonstra o substantivo como núcleo de uma sequência separado de seus modificadores por uma preposição, ou seja, um sintagma preposicional, chamado de pós-modificador em inglês (Carter e McCarthy, 2009).

Destacamos o exemplo do substantivo *efficacy* que aparece na **Figura 49** como núcleo de um sintagma preposicional na linha 15:

"Our findings are largely related to a reduction in the efficacy of neviparine-based antiretroviral therapy…"

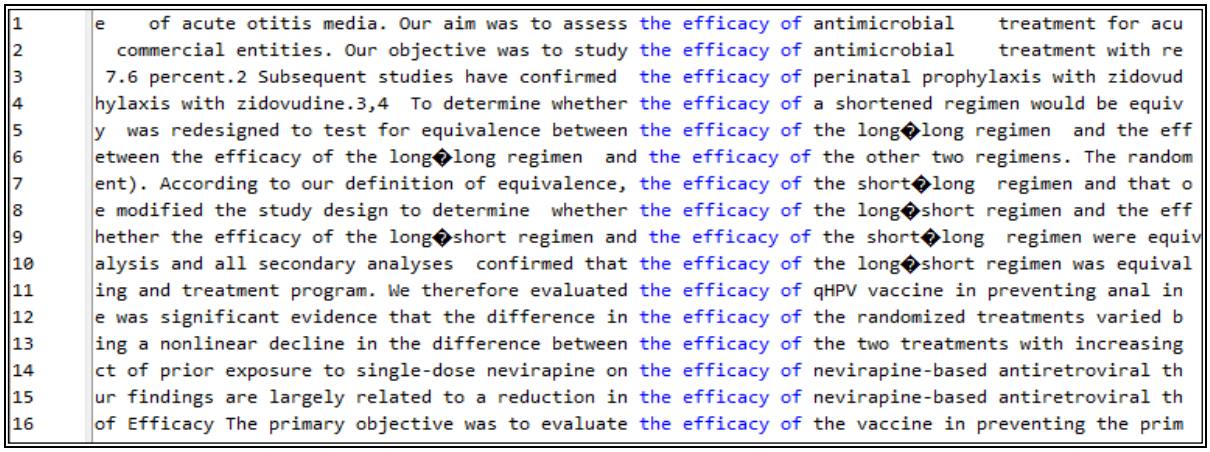


Figura 49: Linhas de concordância *the efficacy of* no *NEJM.*

Para que o aprendiz possa compreender a estrutura desses sintagmas nominais nos textos, é preciso aprender a identificar o núcleo dessa sequência. Se o professor ensina que normalmente os substantivos são o núcleo dos sintagmas, o aprendiz poderá ter problemas quando deparar com um substantivo em posição anterior (cf. **Figura 50**). Por isso, como no caso de *efficacy* e *mortality,* que aparecem nas duas posições comprovados pelas linhas de concordância, é importante que haja a conscientização dessa possibilidade.

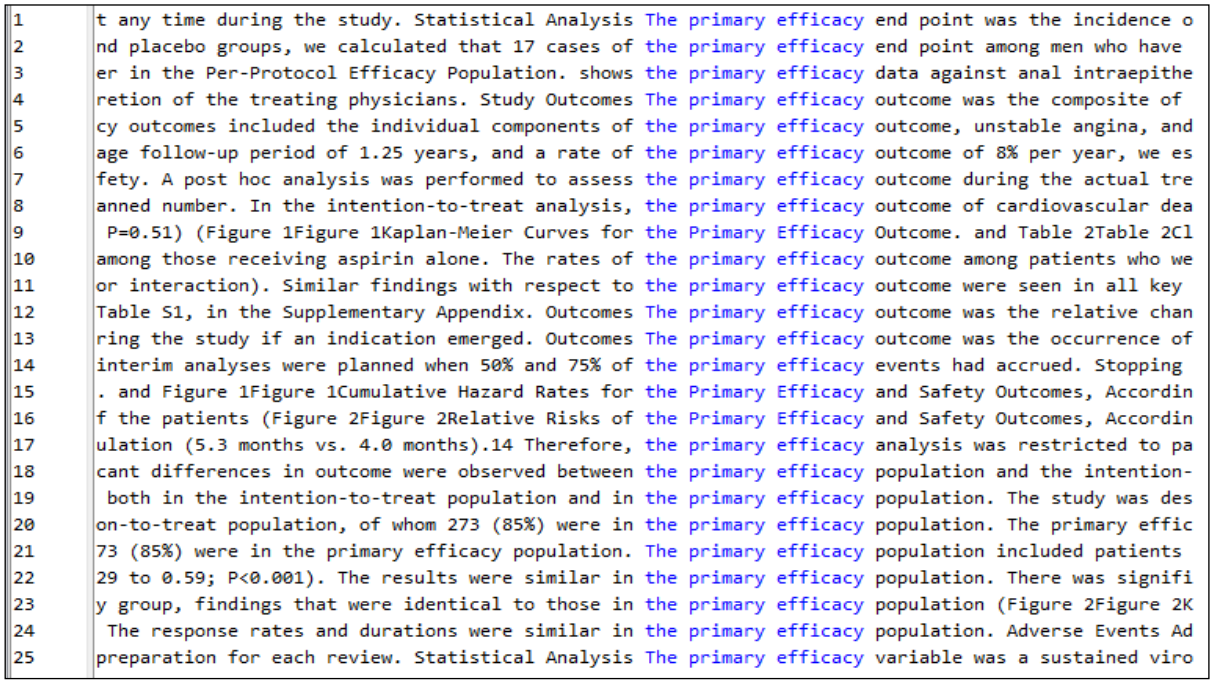


Figura 50: Linhas de concordância de *efficacy* em posição anterior ao núcleo.

O terceiro substantivo cognato na lista de palavras-chave é *interval*, com 114 ocorrências no *subcorpus NEJM.* Os *clusters* (cf. **Figura 51**) desse item demonstraram que seu colocado mais frequente é outro substantivo -- *confidence--,* formando um termo, *confidence interval*[49] (cf. **Figura 52**). Não existe um consenso em relação ao que se pode considerar e como definir um termo. Neste estudo, consideramos um termo a partir da recorrência de uso, por pertencer a um domínio específico e até mesmo pela presença de uma sigla ao lado da sequência (Krieger e Finatto, 2004).

O fato interessante desse achado é que não havia a intenção de se pesquisar o substantivo *confidence,* por se tratar de um falso cognato. O objetivo inicial era partir dos cognatos legítimos. Porém a LC permite esse serependismo, conforme explicado no início deste capítulo, que leva o pesquisador a achados inesperados. Outro fato de interesse é que a metodologia a partir das palavras-chave permite não só ao professor identificar termos da área de interesse, mas também permite a um aluno e/ou tradutor investigar um termo de uma área específica, como no caso de *confidence interval.*

[49] Em estatística, um intervalo de confiança (IC) é um intervalo estimado de um parâmetro estatístico. Em vez de estimar o parâmetro por um único valor, é dado um intervalo de estimativas prováveis. Quão prováveis são estas estimativas é determinado pelo coeficiente de confiança. Quanto maior a probabilidade de o intervalo conter o parâmetro, maior será o intervalo. Disponível em Wikipédia: acesso em 19.12.2012.

Rank	Freq	Range	Cluster
1	72	25	confidence interval
2	31	2	percent confidence interval
3	22	21	interval [CI
4	21	21	confidence interval [CI
5	10	6	confidence interval for
6	10	6	interval for
7	9	5	confidence interval for the
8	9	5	interval for the
9	7	5	interval between
10	5	5	percentage points (95% confidence interval
11	5	5	percentage points (95% confidence interval [CI
12	5	5	points (95% confidence interval
13	5	5	points (95% confidence interval [CI
14	3	3	confidence interval for the difference
15	3	3	interval for the difference
16	3	2	the interval

Figura 51: *Clusters* de *interval* no *NEJM.*

Concordance Hits 72

Hit	KWIC
18	2.4 to 7.0 percent), and 8.6 percent (95 percent confidence interval, 5.6 to 11.6 percent) in the
19	ere the same or similar: 7.8 percent (95 percent confidence interval, 4.6 to 10.9 percent), 4.8 per
20	al, 4.6 to 10.9 percent), 4.8 percent (95 percent confidence interval, 2.5 to 7.2 percent), and 8.6
21	2.5 to 7.2 percent), and 8.6 percent (95 percent confidence interval, 5.5 to 11.7 percent). Accord
22	ansmission or death were 7.6 percent (95 percent confidence interval, 4.9 to 10.2 percent), 5.6 per
23	l, 4.9 to 10.2 percent), 5.6 percent (95 percent confidence interval, 3.1 to 8.0 percent), and 8.9
24	3.1 to 8.0 percent), and 8.9 percent (95 percent confidence interval, 5.8 to 11.9 percent) in the l
25	the respective rates were 9.1 percent (95 percent confidence interval, 5.6 to 12.6 percent), 5.8 pe
26	l, 5.6 to 12.6 percent), 5.8 percent (95 percent confidence interval, 3.2 to 8.4 percent), and 9.2
27	3.2 to 8.4 percent), and 9.2 percent (95 percent confidence interval, 6.0 to 12.4 percent). In a st
28	ansmission in utero were 1.6 percent (95 percent confidence interval, 0.7 to 2.6 percent) for the l
29	ong maternal regimen and 5.1 percent (95 percent confidence interval, 3.2 to 7.0 percent) for the s
30	was marginal; the upper limit of the 95 percent confidence interval for the difference between the
31	for this end point of 85%, a lower bound for the confidence interval for vaccine efficacy greater t
32	lial neoplasia due to any HPV type was 25.7% (95% confidence interval [CI], -1.1 to 45.6) (Table 1Ta
33	primary end point, the hazard ratio, P value, and confidence interval have been adjusted for the mul
34	ratio for the nevirapine group, 3.6; adjusted 95% confidence interval [CI], 1.7 to 7.5; adjusted P=0
35	with a difference of 18.6 percentage points (95% confidence interval [CI], 3.7 to 33.6; nominal P=0
36	time of death. A point estimate and two-sided 95% confidence interval for the hazard ratio were calc

Figura 52: Linhas de concordância de *confidence interval* no *NEJM.*

Dentre os substantivos na lista das primeiras 50 palavras-chave do subc*orpus NEJM, arrest* aparece com 107 occorrências (cf. **Tabela 6).** Nosso intuito era analisar somente os substantivos cognatos, porém decidimos analisar esse substantivo por ter aparecido nessa posição de destaque.

O principal colocado de *arrest* é *cardiac* formando, assim, um termo dessa área de especialidade. Fizemos uma busca no outros dois sub*corpora* e não encontramos nenhuma ocorrência de *arrest.* Encontramos *cardiac* no *WebMD*, porém com outros colocados: *nurse, events, risks, problems, patients, rehabilition, surgeons, surgery, defibrillators* e *status.* (cf. **Figura 53**).

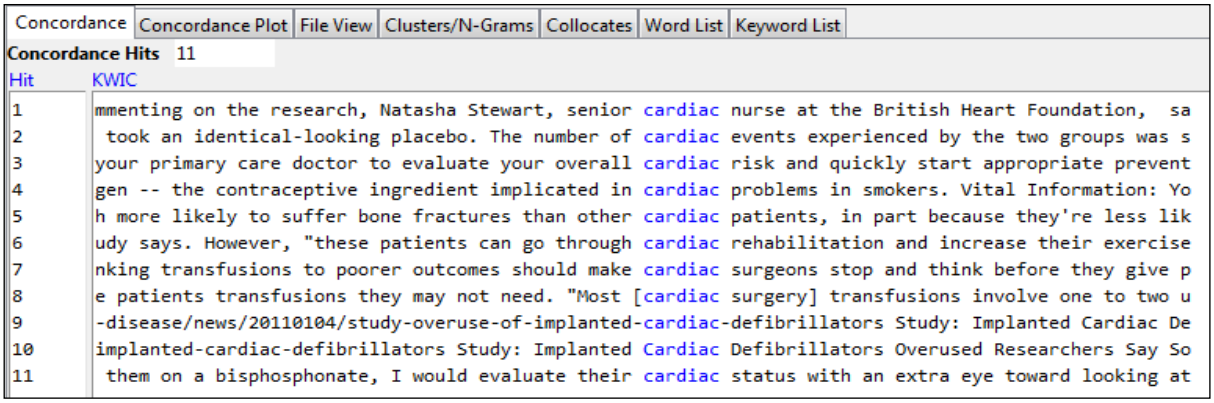
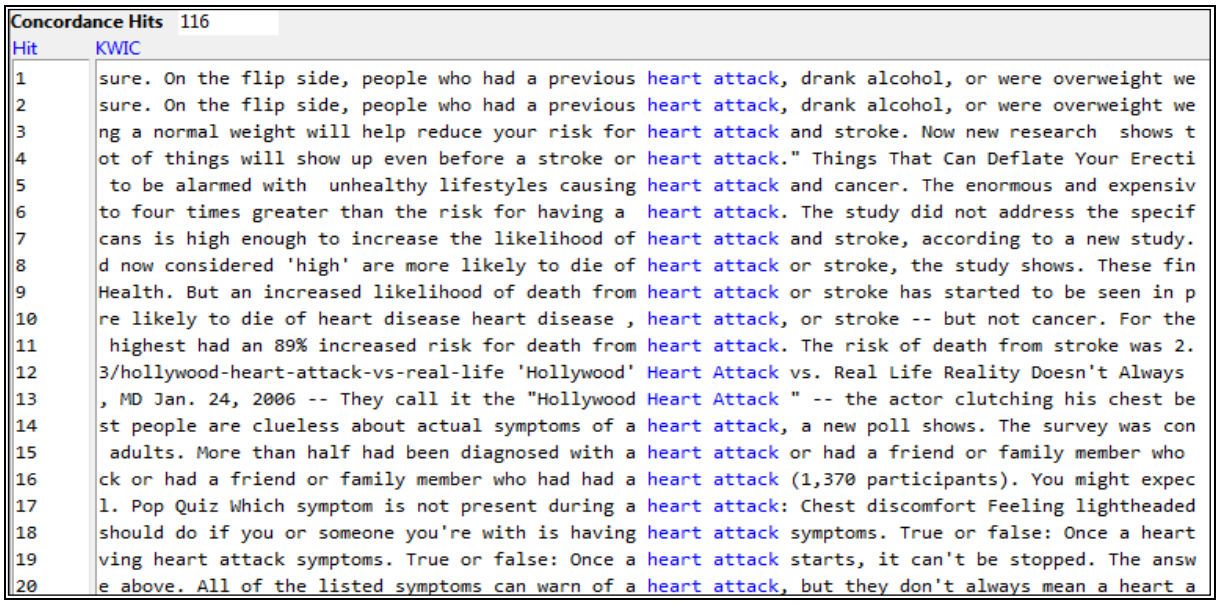


Figura 53: Linhas de concordância de *cardiac* no *WebMD.*

Por ser *cardiac* uma palavra de registro mais formal em inglês, procuramos nos sub*corpora* de divulgação o equivalente informal mais próximo: *heart,* já que os textos de divulgação se destinam a abordar os assuntos de saúde de maneira informativa ao público leigo em medicina. Outro fato é que *heart* aparece como uma palavra-chave (cf.**Tabela 6,** linha 10) no *WebMD*. Nessa busca, encontramos *heart attack* com 116 ocorrências (cf. **Figura 54**). Uma possível conclusão é que nos textos destinados ao público leigo, *heart attack* é utilizado como equivalente de *cardiac arrest.* Isso vem corroborar a argumentação de Zamboni (2001) de que a terminologia e a escolha do léxico nos diferentes gêneros são adaptadas ao público a que se destinam, havendo, portanto, uma reescrita do texto original. Citando Bakhtin que afirma: “Cada um dos gêneros do discurso, em cada uma das áreas de comunicação verbal, tem sua concepção padrão de destinatário que o determina como gênero.” (Bakhtin, 1992: 321)



Concordance Hits 116

Hit	KWIC
1	sure. On the flip side, people who had a previous heart attack, drank alcohol, or were overweight we
2	sure. On the flip side, people who had a previous heart attack, drank alcohol, or were overweight we
3	ng a normal weight will help reduce your risk for heart attack and stroke. Now new research shows t
4	ot of things will show up even before a stroke or heart attack." Things That Can Deflate Your Erecti
5	to be alarmed with unhealthy lifestyles causing heart attack and cancer. The enormous and expensiv
6	to four times greater than the risk for having a heart attack. The study did not address the specif
7	cans is high enough to increase the likelihood of heart attack and stroke, according to a new study.
8	d now considered 'high' are more likely to die of heart attack or stroke, the study shows. These fin
9	Health. But an increased likelihood of death from heart attack or stroke has started to be seen in p
10	re likely to die of heart disease heart disease , heart attack, or stroke -- but not cancer. For the
11	highest had an 89% increased risk for death from heart attack. The risk of death from stroke was 2.
12	3/hollywood-heart-attack-vs-real-life 'Hollywood' Heart Attack vs. Real Life Reality Doesn't Always
13	, MD Jan. 24, 2006 -- They call it the "Hollywood Heart Attack " -- the actor clutching his chest be
14	st people are clueless about actual symptoms of a heart attack, a new poll shows. The survey was con
15	adults. More than half had been diagnosed with a heart attack or had a friend or family member who
16	ck or had a friend or family member who had had a heart attack (1,370 participants). You might expec
17	l. Pop Quiz Which symptom is not present during a heart attack: Chest discomfort Feeling lightheaded
18	should do if you or someone you're with is having heart attack symptoms. True or false: Once a heart
19	ving heart attack symptoms. True or false: Once a heart attack starts, it can't be stopped. The answ
20	e above. All of the listed symptoms can warn of a heart attack, but they don't always mean a heart a

Figura 54: Seleção de linhas de concordância com *heart attack* no *WebMD.*

Como é possível observar, o discurso científico sofreu uma adaptação e uma transformação quando apresentado a outra esfera social, com o intuito de possibilitar a compreensão do público em geral, que é o destinatário dessas enunciações (Zamboni, 2001).

A análise à luz da LC demonstrou padrões lexicais e estruturais específicos dos textos estudados, como, por exemplo, a das colocações nominais em inglês, bem como o aparecimento de uma terminologia que possivelmente seria desconhecida e passaria despercebida para o aluno e também para o professor, como é o caso das colocações encontradas: *confidence interval, cardiac arrest* e *heart attack.*

CONCLUSÃO

De acordo com os resultados obtidos, concluímos que é possível fazer um levantamento dos itens a serem ensinados por meio de um *corpus* especializado. Os dados nos demonstraram que:

1) *Were* e *was* são as primeiras palavras-chave no corpus acadêmico, confirmando outros estudos que afirmam ser a voz passiva muito utilizada nesses textos. Além disso, confirmamos que os verbos regulares são os mais utilizados nesse tipo de literatura;

2) As colocações substantivas são mencionadas e ressaltadas como um ponto importante a ser ensinado nos livros para a leitura instrumental (*Preparatório TEAP e Leitura instrumental em língua inglesa*), mas o uso de um *corpus* especializado pode ajudar o professor a preparar suas atividades com um enfoque mais direcionado à necessidade dos aprendizes, sendo possível elaborar um conteúdo mais individualizado;

3) A omissão do *that* nos artigos de divulgação se aproxima da linguagem falada, e por essa razão causa estranhamento ao aprendiz brasileiro, que deve ser alertado sobre o fato para que sua leitura possa ser fluida. De acordo com Conrad e Biber (2009) essa estrutura consiste em uma complexidade da linguagem falada e, portanto, um assunto digno de nota;

4) Tanto a omissão do *that* quanto o item lexical *likely* e seus colocados *(more/less likely to/than)* não encontram equivalência em português, portanto são itens de importância na elaboração de um conteúdo programático;

5) Os verbos de elocução encontrados são elucidativos no sentido de demonstrar a diferença entre a estrutura da escrita e da linguagem falada. Embora os artigos de divulgação sejam textos escritos, eles trazem características da linguagem oral na tentativa de indicar maior informalidade e proximidade com o leitor leigo em assuntos da área médica.

No percurso da análise, evidenciamos a questão da falta de equivalência das estruturas, que, segundo alguns autores, é um conceito muito vago na tentativa de explicar as diferenças entre as línguas. Porém, acreditamos ser bastante importante sua utilização em sala de aula. Embora possa parecer óbvio para os linguistas, não é tão claro para um aprendiz de língua estrangeira.

A principal conclusão deste trabalho foi encontrarmos no discurso de divulgação características da linguagem falada na escrita e, por consequência, um distanciamento das estruturas mais lineares do discurso acadêmico a que os alunos estão acostumados. Acreditamos que esses achados são úteis no sentido auxiliar o professor a despertar a consciência dos aprendizes para as especificidades da segunda língua.

CONSIDERAÇÕES FINAIS

A partir das experiências em sala de aula e com o intuito de auxiliar aprendizes de leitura a se tornarem proficientes, pretendia-se descobrir as possíveis diferenças léxico-gramaticais entre os textos acadêmicos e os de divulgação, pois os primeiros não causavam dificuldade a esses aprendizes.

A fim de atingir esse objetivo, construímos um *corpus* de estudo que fosse uma amostragem dos diferentes gêneros -- acadêmico e de divulgação -- publicados na internet. A construção do corpus foi manual e semiautomática por meio do *BootCat.* Para a análise foram utilizadas as ferramentas computacionais existentes para esse fim: *WST e AntConc*. Fizemos uma comparação entre as listas de palavras dos três sub*corpora* e chegamos às palavras-chave de cada um deles.

A partir dessas comparações constatamos aspectos linguísticos já descobertos em outros trabalhos, tais como Biber (1988), Hyland (2004) e outros. Por exemplo, observamos que os verbos *were/ was* são as primeiras palavras-chave dos textos acadêmicos, indicando a forte presença da voz passiva nesses textos. Constatamos, também, a alta frequência dos particípios de verbos regulares como colocados de *were e was.* As colocações substantivas e adjetivas foram outra descoberta importante para a conscientização dos aprendizes das diferenças na constuição da ordem das palavras em inglês. A análise dos dados demonstrou que os substantivos de origem latina, que, a princípio, não causariam problemas para os alunos, possuem características especiais que mais uma vez demonstram falta de equivalência na questão da ordem em que seus modificadores são apresentados.

Nos textos de divulgação, a composição da sequência *were/was likely to/than*, é digna de nota pela sua falta de equivalência estrutural e conceitual com o português.

Outros aspectos que também mereceram destaque foram os verbos de elocução, já que por meio deles identificamos características da linguagem falada nos textos de divulgação utilizados como meio de interação com o leitor leigo.

Ainda nos textos de divulgação, descobrimos características de linguagem oral, como a omissão do *that,* que podem causar problemas para alunos iniciantes que não estiverem familiarizados com essas estruturas.

Acreditamos que não só as descobertas dessa pesquisa podem contribuir para um enfoque mais individualizado na preparação do conteúdo programático apresentado em sala de aula, como também a metodologia empregada pode auxiliar professores, aprendizes e tradutores a elaborarem seus próprios *corpora,* uma vez que apresentamos duas possibilidades de composição de um *corpus*: manual e automática, além de ferramentas de análise disponíveis *online* para uso imediato em sala de aula. Outra possibilidade seria o uso das linhas de concordância em sala de aula que apresentaria ao aprendiz a possibilidade de fazer suas próprias pesquisas e tornar-se também um pesquisador.

A compilação de um *corpus* com um fim específico, mesmo que pequeno, demonstra ser uma ferramenta valiosa, pois o professor e o aluno têm a possibilidade de explorar construções e exemplos que sejam condizentes com sua área de interesse (Gavioli, 2005).

Embora o foco deste trabalho tenha sido a identificação das diferenças entre o gênero acadêmico e de divulgação, acreditamos que os pontos levantados podem trazer benefícios também aos tradutores e aprendizes mais avançados que querem escrever na L2 e que, para isso, possam construir seu próprio *corpus* de estudo.

Existem outras possibilidades que não puderam ser analisadas neste trabalho. Detectamos em nossas primeiras investigações que existem diferenças entre a frequência dos modais *can*, *may* , *should* entre os sub*corpora.* Outra possibilidade são as estruturas compostas por *if* em inglês, que também podem causar dificuldade para o aluno brasileiro, visto que existe a falta de equivalência entre o português e o inglês.

Um possível desdobramento deste trabalho seria elaborar uma minigramática que utilizasse exemplos extraídos desse *corpus* para que os aprendizes com dificuldade no entendimento das estruturas do inglês pudessem nelas se concentrar,

sem a preocupação, pelo menos em estágios iniciais, de adquirir novos conceitos lexicais.

.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, W.; CORBETT, J. (2010) Teaching English as a friendly language: lessons from the SCOTS corpus. *English Language Teaching Journal,* 64(4), 414-423.

ATKINS, S., J. CLEAR, N. OSTLER (1992) Corpus Design Criteria. In: *Literary and Linguistic Computing*, 7 (1): 1-16.

BAKHTIN, M. (1981) *Marxismo e filosofia da linguagem* . Tradução de Michel Lahud e Yara Frateschi Vieira - São Paulo: Editora Hucitec.

BAKHTIN, M. (1986) *Speech Genres & Other Late Essays*. Tradução para o inglês de Vem McGee, Carol Emerson & Michel Hoquist: Austin, Texas: University of Texas.

BAKHTIN, M. (2005) *Problemas da Poética de Dostoiévski* .Tradução de Paulo Bezerra. São Paulo: Forense Universitária.

BARONI, M.; BERNARDINI, S. (2004) Bootstrapping corpora and terms from web. *Proceedings of the LREC 2004*, Lisboa 26-28 maio 2004. (págs.1313-1316).

BAZERMAN, C. (2005) *Gêneros textuais , tipificação e interação.* DIONISIO, A.; HOFFNAGEL, J. (orgs.). Tradução e adaptação de Judith C. Hoffnagel. São Paulo: Cortez.

BAZERMAN,C. (2007) *Escrita, gêneros e interação social*. DIONISIO, A.; HOFFNAGEL, J. (orgs.). Tradução e adaptação de Judith C. Hoffnagel. São Paulo: Cortez.

BAZERMAN, C. (2007) *Gêneros textuais, intertextualidade, e atividade: Teórico Consideração.* DIONISIO, A.; HOFFNEAGEL, J. (orgs.). Tradução e adaptação de Judith C. Hoffnagel. São Paulo: Cortez.

BERBER SARDINHA, T. (1999) Beginning Portuguese Corpus Linguisitcs: exploring a corpus to teach Portuguese as a foreign language. DELTA, São Paulo, v. 15, n. 1. Disponível em: <http://www.scielo.br>. Acesso em 12.01.2013.

BERBER SARDINHA, T**.** (2004a) *Linguística de Corpus*. Barueri: Manole.

BERBER SARDINHA, T. (2004b) *Informatividade, interatividade e narratividade na reunião de negócios* – Análise Multidimensional e palavras-chave *i DIRECT Papers* 52 ISSN 1413-442x.

BERBER SARDINHA, T. (2011) Como usar a Linguística de Corpus no ensino de língua estrangeira (por uma linguística de corpus educacional brasileira). In: VIANA, V.; TAGNIN, S. (orgs.) *Corpora no ensino de línguas estrangeiras*. - São Paulo: Hub Editorial.

BERBER SARDINHA, T. (2012) *Lexicogrammar.* The Encyclopedia of Applied Linguistics. Disponível em: <http://www.onlinelibrary.wiley.com>. Acesso em 11.01.2013.

BIBER, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

BIBER, D. (1989) *A Typology of English texts* Linguistics 27: 3-43.

BIBER, D. (1992) On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15 (2), 133-163.

BIBER, D. (1995) *Dimensions of Variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

BIBER*,* D. CONRAD, S.; REPPEN, R. (1998) *Corpus Linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

BIBER, D. et al. (1999) *Longman grammar of spoken and written English*. London: Longman.

BIBER, D. (2006) *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam/Philadelphia: John Benjamins.

BIBER, D. (2009). A corpus-driven approach to formulaic language in English)* Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14:3, 279-31.

BHATIA, V. K. (2009) A análise dos gêneros hoje. Tradução de Gomes Bezerra. In: BEZERRA, B.; RODRIGUES, B.; CAVALCANTE, M. (orgs.) *Gêneros e sequência textuais* (orgs.) Recife: EDUPE.

BOCK, A. M.; GONÇALVES, M. G. M.; FURTADO, O. (orgs.) (2001) *Psicologia Sócio-histórica* - São Paulo: Cortez.

BOWKER, L.; PEARSON, J. (2002) *Working with specialized language – A practical guide to using corpora.* London/New York: Routledge.

BROWN, D. (1994) *Teaching by principles: an interactive approach to language pedagogy* – New Jersey: Prentice Hall Regents.

CARTER, R.; MCCARTHY, M. (2006) *Cambridge Grammar of English: a comprehensive guide*. Cambridge: Cambridge University Press

CHAMBERS, A. (2007) *Corpora and Language Learning Graduate seminar.* University College Cork.

CHARTIER, R.; (2002) *Os desafios da escrita.*Tradução de Fulvia M. L. Moretto. Editora Unesp.

COBB, T. (1997) From Concord to Lexicon: Development and Test of a Corpus - Based Lexical Tutor. *PhD Thesis,* Department of Educational Technology, Concordia University, Montreal, Quebec, Canada.

CONRAD, S.; BIBER, D. (2009) *Real Grammar: A Corpus-Based Approach to English.* NY: Pearson Education.

COTS, J. M.; TUSÓN. A. (1994) Language in Education: An Interview with Leo Van Lier *Sintagma 6*, pp. 51-65 Universidad de Barcelona.

COXHEAD, A. (2000) A new academic word list. *TESOL Quarterly*, 34, 213-238.

DAYRELL, C. (2010) Corpora no ensino do inglês acadêmico. In: VIANA, V.; TAGNIN, S. (orgs.) *Corpora no ensino de línguas estrangeiras*. São Paulo: Hub Editorial.

DAVIS, C.; OLIVEIRA, Z. (2010) *Psicologia na Educação.* São Paulo: Cortez.

GAMA, A. N. M. et al. (2005) *Introdução à leitura em inglês III*. Rio de Janeiro: Gama Filho.

GAVIOLI, L. (2005) *Exploring Corpora for ESP Learning.* Amsterdam/Philadelphia: John Benjamins.

GRANGER, S. (2010) *Learner corpora: a multifarious resource for pedagogical lexicography* Comunicação pessoal ELC Porto Alegre.

HALLIDAY, M. A. K. (1991) Corpus studies and probabilistic gramar: In: AIJMER, K.; ALTENBERG, B. *English Corpus Linguistics*. London: Longman.

HYLAND, K. (2004) *Genre and second language writing*. Michigan: University of Michigan Press.

HYLAND, K. (2008) *English for specific purposes* 27 2-21: Disponível em <http://www.sciencedirect.com>. Acesso em: 20.10.2011.

HYLAND, K. (2009). *Genre analysis.* In: K. Malmkjaer (Ed. Rouledge Encyclopedia of Linguistics 3ª- edição. London: Routledge. p. 210-213.

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

JOHNS, T. (1991) *From print out to hand out: Grammar and teaching in the context of data-driven learning*, CALL Austria 10:14-34.

KILGARIFF, A., PVS. A.; POMIKÁLEK, J. (2011) *BootCatting Comparable Corpora* - Short papers of the 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, p. 123-126 Paris, 8-10 November 2011.

KRIEGER; M.G.; FINATTO, M.J. (2004) *Introdução à terminologia: teoria e prática.* São Paulo: Contexto.

LANZONI, H. (2010) *Preparatório TEAP – Test of English for Academic Purpose* – Ribeirão Preto, SP: Tese Prime – Sistemas de Avaliação Linguística.

LAURENCE, A. (2004) AntConc: A Learner and Classroom Friendly, Multi-Platform *Corpus* Analysis Toolkit. IWLeL 2004: *An Interactive Workshop on Language e-Learning*. Waseda: Japan.

LEE, B. (1985) Intellectual origins of Vygotsky´s semiotic analysis In: WERTSCH, J.V. *Culture, communication, and cognition: Vygotskian perspectives.*

MCENERY, T; WILSON, A. (1997) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

MOTTA-ROTH D. (2002) A construção social do gênero resenha acadêmica. In: Meurer, L.; Motta-Roth D. (orgs.). *Gêneros Textuais e práticas discursivas: subsídios para o ensino da linguagem.* Bauru: EDUSC.

MUNHOZ, R. (2000) *Inglês Instrumental: estratégias de leitura, módulo I*. São Paulo: Textonovo.

MURPHY, R., (1994) *English grammar in use*. Cambridge: Cambridge University Press.

NATION. P; LAUFER. B. (1999) A vocabulary size test of controlled productive ability. In*: Language Testing* 16:33-51. Tese de doutorado. Universidade de Haifa: Israel.

O'KEEFFE, A.; MCCARTHY. M.; CARTER, R. (2007) *From Corpus to Classroom.* Cambrigde: Cambridge University Press.

PARTINGTON, A. (1998) *Patterns and Meanings.* Amsterdam/Philadelphia: John Benjamins.

PAVLENKO, A. (2009) *The Bilingual Mental Lexicon – Interdisciplinary Approaches* Cromwell: Multilingual Matters – Great Britain.

PINHEIRO, N. F. (2002) A Noção de gênero para análise de textos midiáticos. In: MEURER, J. L.; MOTTA-ROTH D. (orgs.) *Gêneros Textuais e práticas discursivas: subsídios para o ensino da linguagem*. Bauru: EDUSC.

REPPEN, R.; FITZMAURICE, S. M.; BIBER, D. (2002) *Using corpora to explore linguistic variation*. Amsterdam/Philadelphia: John Benjamins.

RICHARDS, J. C.; ROGERS, T. S. (1995) *Approaches and Methods in Language Teaching.* Cambridge: Cambridge University Press.

SINCLAIR, J. (1991) *Corpus, Concordance, Collocation*. Oxford: OUP.

SINCLAIR, J. (1996a) *The search for units of meaning, Textus 9* (1): p. 75-106.

SCOTT, M.; TRIBBLE, C. (2006) *Textual Patterns. Key words and corpus analysis in language education.* Amsterdam/Philadelphia: John Benjamins: Studies in *Corpus* Linguistics, Vol.22.

SCOTT, M. (2008) *Wordsmith Tools version 5* Liverpool: Lexical analysis Software.

SCOTT, M. (2010) Manual: *WordSmith.pdf*. Disponível em : <http://www.lexically.org/wordsmith/step_by_stepguide_english.pdf>. Acesso em: 20.11. 2011.

SOUZA, A. et al. (2010) *Leitura em língua inglesa: uma abordagem instrumental.* São Paulo: Disal Editora.

SUE, K.; VAUGHAM, J. (2009) *New American Inside Out* – Intermediate Macmillan: Oxford UK.

SWALES, J. (1990) *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

SWALES, J. (1998) *Other floors, other voices: a textography of a small university building*. Mahwah, NJ: Lawrence Erlbaum.

TAGNIN, S. O. (2005) *O jeito que a gente diz:* expressões convencionais e idiomáticas*.* São Paulo: Disal Editora.

TORRES, M. L. (2007) *Inglês Instrumental para profissionais da saúde* – São Paulo: All Print Editora.

TRIBBLE, C.; JONES, G. (1990) *Concordances in the Classroom* Harlow. UK: Longman.

TRIBBLE, C. (1997) *Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching.* In Melia. J; B Lewandowska-Tomaszcyk (ed.) PALC '97 Proceedings Lodz: Lodz University Press.

VARANTOLA, K. (2003) Translators and disposable corpora In: F. Zanettin, S. Bernardini e D. Stewart, editores, *Corpora in Translator Education*, p. 55 –70, St Jerome.

VASILÉSKI, V. (2007) Linguística de *corpus*, linguística computacional e estatística: trio metodológico. In: GERBER, R. M.; VASILÉVSKI, V. (orgs.) *Um percurso para pesquisas com base em corpus*. Florianópolis: Editora da UFSC. p. 63-75).

WEIMER, M. G. (2002) *Learner-centered teaching: five key changes to practice.* San Francisco: Jossey-Bass.

WEST, M. (1953) *A General Service List of English Words*. London: Longman.

VAN LIER L. (1989) *The Language Classroom and the language learner.* London: Longman.

VIGOTSKY, L.S. (1996) *A formação social da mente*. São Paulo: Martins Fontes.

VIGOTSKY,L.S. (2001) *A construção do pensamento e da linguagem*. Tradução Paulo Bezerra. São Paulo: Martins Fontes.

ZAMBONI, L.M.S. (2001) *Cientistas, jornalistas e a divulgação científica; subjetividade e heterogeneidade no discurso de divulgação científica*. Campinas: Autores Associados.

Sites consultados:

<http://www.fflch.usp.br/dlm/comet/artigos/AntConc.pdf>. (Tutorial de como utilizar o *AntConc* 2.3.1 preparado pelo Professor Adauri Brezolin). Acesso em 08.09.2010.

<http://www.lextutor.ca>. Acesso em 23.01.2013.

<http://www.antlab.sci.waseda.ac.jp>. Acesso em 23.01.2013.

<http://www.lael.pucsp.br/direct>. Acesso em 20.11.2011.

<http://www.ucrel.lancs.ac.uk/llwizard.html>. Acesso em 23.01.2013.

<http://www.culturainglesasp.com.br>. Acesso em 22.01.2013.

<http://www.lexically.org/wordsmith/step_by_stepguide_english.pdf>. Acesso em 20.11. 2011.

<http://www.sciencedirect.com>. Acesso em : 20.10.2011.

<http://www.scielo.br>. Acesso em 12.01.2013.

<http:// www.onlinelibrary.wiley.com>. Acesso em 11.01.2013.

<http://www.google.com>. Wikipédia: acesso em 19.12.2012.

<http://www.sslmit.unibo.it/~baroni/bootcat.html>. Acesso em 20.02.2012.

<http://www.br.answers.yahoo.com>. Acesso em 18.01.2013.

<http://www.periodicos.capes.gov.br>. Acesso em: 23 Nov. 2012.

<http://www.healthnews.com>. Acesso em 23.01.2013.

<http://www.bbc.co.uk/news/health>. Acesso em 23.01.2013.

<http://www.nejm.org>. Acesso em 23.01.2013.

<http://www.webmd.com>. Acesso em 23.01.2013.

<http://www.bv.fapesp.br>. Acesso em 28.11.2011.

Dicionário:

MACMILLAN ENGLISH DICTIONARY FOR ADVANCED LEARNERS – Macmillan Publishers Limited. Oxford 2002.

ANEXOS

ANEXO A: AMOSTRA DE TESTES DE PROFICIÊNCIA EM LEITURA[50]

HIV INFECTION IN ADOLESCENTS

A The human immunodeficiency virus (HIV), which causes AIDS, ranks seventh among the leading causes of death for U.S. children 5 to 14 years of age and sixth for young people 15 to 24 years of age. Because the average period of time from HIV infection to the development of AIDS is 10 years, most young adults with AIDS were **likely** infected with HIV as adolescents. Almost 18 percent of all reported cases of AIDS in the United States have occurred in people between the ages of 20 and 29.

B In the United States through June 1999, 3,564 cases of AIDS in people aged 13 through 19 had been reported to the Centers for Diseases Control and Prevention. Many other adolescents are currently infected with HIV but have not yet developed AIDS. Data from the 31 states that conduct HIV case surveillance indicate that among adolescents aged 13 through 19: 46 percent were male; 54 percent were female; 28 percent were non-Hispanic white; 66 percent were non-Hispanic black/5 percent were Hispanic/ less than 1 percent were Asian/Pacific Islander or American Indian/Alaskan Native.

C Most adolescents recently infected with HIV are exposed to the virus through sexual intercourse or injection drug use. Through June 1999, HIV surveillance data suggest that nearly half of all HIV-infected adolescent males are infected through sex with men. A small percentage of males appear to be exposed by injection drug use and/or heterosexual contact. The same data suggest that almost half of all adolescent females who are infected with HIV were exposed through heterosexual contact and a very small percentage through injection drug use.

D CDC studies conducted every two years in high schools(grades nine through 12 consistently indicate that approximately 60 percent of students have had sexual intercourse by grade 12; half report use of a latex condom during last sexual intercourse, and about one-fifth have had more than four lifetime sex partners.

E Approximately two-thirds of the 12 million cases of sexually transmitted diseases (STDs) that are reported in the United States each year are in individual under the age of 25 and one-quarter are among teenagers. This is particularly significant because if either is infected with another STD, the risk of HIV transmission increases substantially. If one of the partners is infected with an STD that causes the discharges of pus and mucus such as gonorrhea or **chlamydia**, the risk of HIV transmission is three to five times greater. If one of the partners is infected with an STD that causes ulcers, such as syphilis or genital herpes, the risk of HIV transmission is nine times greater.

F Adolescents tend to think they are invincible, and therefore, to deny any risk. This belief may cause them to engage in risky behavior, to delay HIV-testing, and they test positive, to delay or refuse treatment. Doctors report that many young people, when they learn they are HIV-positive, take several months to accept their diagnosis and return for treatment. Health care professionals may be able to help these adolescents by explaining the information slowly and carefully, eliciting questions from them, and emphasizing the success of newly available treatments

[50] As amostras dos testes foram obtidas do site da Cultura Inglesa: www.culturainglesasp.com.br.

Mark your answers on Answer Sheet

I. The text is divided into sections A, B, C, D, E and F. Using <u>EACH LETTER ONLY ONCE</u>, identify which section of the text

1. Points out the overconfidence of adolescents in themselves as one of the main reasons for the spread of HIV.
2. shows that homosexual behavior is the most important cause of HIV infection among adolescent males in the USA.
3. presents evidence to the effect that there is great disparity among the different ethnic groups in respect of HIV infection.
4. refers to the high incidence of promiscuity among adolescents
5. argues that most young adults who die of AIDS must have been infected by the HIV when they were adolescents.
6. points to incresead risk of HIV transmission among people already suffering from other sexual diseases.

II. Are these adolescents True (T) or False (F) according to the text?

7. Adolescents are in general more likely to listen to medical advice than young adults.
8. More and more adolescents are confident today that, thanks to the advance of the medical science, they do not run the risk of dying of AIDS.
9. Adolescents infected by HIV develop AIDS much faster than persons from other age groups.
10. HIV infection among adolescents is steadily decreasing in the US.

III. Complete each of the following sentences by choosing a, b, c, or d

11. The word **<u>"likely"</u>** (underlined, line 5) in the text means
    a) certainly
    b) enjoyable
    c) probably
    d) none of the above

12. **<u>"Chlamydia"</u>** (underlined , line 35 is)
    a) one of the causes of HIV transmission.
    b) the name of a sexually transmitted disease.
    c) responsible for the high incidence of AIDS among teenagers.
    d) an early manifestation of HIV infection.

Anexo B Linhas de concordância de *likely* no *HN* e *WebMD:*

1) Lista das 94 ocorrências de *likely* encontradas no *HN :*

ne from each parent) were almost 17 times as **likely** to develop the condition when taking a
ne test could be used in the clinic, but not **likely** If further research confirms the findi
men 55 years and older were 1.37 times more **likely** to be diagnosed as having bipolar diso
not full term] children." Boys were twice as **likely** as girls to suffer behavioral problems
n were lost to follow-up, Anand said, "it is **likely** that their results may even underestim
take the epilepsy drug valproate may be more **likely** to have a child with autism than those
have suggested that people with MS were more **likely** to have been born in the spring (with
efit from taking D in regions where they are **likely** to be deficient. There is no proof
fast cereal, pasta sauce, and sandwich bread **likely** contain more salt than one would think
rs took fertility drugs were almost twice as **likely** to have autism as other children, new
statistics is that older women are both more **likely** to have fertility problems and to take
research has shown older moms are also more **likely** to have autistic children. In the stud
nough, it's like death and taxes -- you will **likely** get osteoarthritis," said Dr. Todd Sti
e said. "It is chronic overload [and] a very **likely** cause of osteoarthritis." People who i
eart -- regular exercise, eating well -- are **likely** to be the healthiest for their joints
ho had lower vitamin D levels were also more **likely** to have higher average blood sugar lev
mounts of vitamin D in their blood were more **likely** to have the syndrome than those with s
in the first trimester of pregnancy are more **likely** to have a congenital heart defect than
diabetes ." High school girls were far less **likely** than boys to meet the recommendation f
e board, students in higher grades were less **likely** than underclassmen to get enough exerc
f the children. Cholesterol levels were more **likely** to decline when changes were made to t
edication. Because HIV-infected children are **likely** to live well into adulthood, the resea
the tainted supplement were three times more **likely** to purchase it in local stores than ov
ing, because we know that ADHD children will **likely** have a learning disability. And we kno
women who drink caffeinated coffee are less **likely** to become depressed and the more

aged two to three cups per day were 15% less **likely** to develop depression, even after the
le who drank only decaf were no more or less **likely** to be depressed than women who drank
naire. Public awareness of excess sodium has **likely** risen since 2008, thanks to increased
o consumed very little sodium were also more **likely** than those with average intake to die
han 8,000 milligrams were roughly 50% more **likely** than the group with average excretion
??less than 2,000 milligrams were 37% more **likely** than the average group to die of heart
to die of heart-related causes and 29% more **likely** to be hospitalized for heart failure.
t or justify their use and the benefits will **likely** be less substantial than conventional
d, it might also be that caregivers are more **likely** to have the condition themselves and "
their second birthday were roughly twice as **likely** as their peers to develop ADHD by the
Young children are far more **likely** to experience attention and hyperactiv
nown as gestational diabetes) were twice as **likely** as their peers to meet the criteria fo
rn to such moms are as much as 14 times more **likely** to have ADHD by the age of 6, the stud
ercially grown fruit and vegetables are more **likely** to have attention deficit hyperactivit
Middle-aged women are roughly 40% more **likely** to become depressed if they experience
ubled them within the past year were just as **likely** to become depressed as women who
ns who were born in another country are less **likely** to have a stroke than white Americans,
ariant that causes cervical cancer, are more **likely** to develop cancers of the head and nec
ts with HPV-16 positive tumors were far more **likely** to test positive for oral HPV-16 infec
s with HPV-16 positive tumors were also more **likely** to test positive for all types of the
take the epilepsy drug valproate may be more **likely** to have a child with autism than those
rent with bipolar disorder are 14 times more **likely** than their peers to have bipolar-like
s themselves and are two to three times more **likely** to be diagnosed with an anxiety or moo
nts are bipolar, children are 3.6 times more **likely** to have bipolar disorder than children
be negative. Perhaps these people were less **likely** to stick to the drug. There is a bit o
buse and social shaming as children are more **likely** to have psychosocial health problems t
and a sense of masculinity failure were more **likely** to use illegal drugs and to engage in
tients with high cholesterol levels are less **likely** to die than those with lower cholester

early-stage chronic kidney disease are more **likely** to die if they have elevated levels of highest levels of FGF-23 were 4.3 times more **likely** to die than those with the lowest leve phetamine-use disorders were 76 percent more **likely** to develop Parkinson's disease than th syndrome, and that obese patients were more **likely** to have it than those who were of norm years -- affecting 214 children so far -- is **likely** driven by travelers returning from abr that about 80 percent of people with type 1 **likely** have GAD65 autoantibodies. GAD-alum g drugs, had less brain damage and were more **likely** to recover their motor skills than mor with the highest level of activity were more **likely** to have blood flow restored within two und that older women were significantly more **likely** to die from breast cancer than younger ion, kids with higher PFC exposure were more **likely** than their peers to have antibody leve ts with a history of heart disease will most **likely** not reduce their risk for developing c isease are 65 or older. Older women are more **likely** to die of heart attacks within a few w close relatives with heart disease are more **likely** to develop it themselves. Race: Heart ally if a lot of it is at the waist?are more **likely** to develop heart disease and stroke ev rt. Stroke People who have diabetes are more **likely** to have a stroke than people who do no s who also have high blood pressure are more **likely** to develop complications that affect t Americans and Native Americans are also more **likely** than those in other ethnic groups to h of breath, energy, and asthma symptoms will **likely** get better within the first 2 to 4 wee hen you quit smoking. Your kids will be less **likely** to start smoking. If your teen already your teen already smokes, he or she is more **likely** to quit smoking if you quit. 5 Underwe ight babies and problem pregnancies are less **likely** in women who quit smoking. Natural, lo ts with a history of heart disease will most **likely** not reduce their risk for developing c ho use estrogen replacement therapy are more **likely** to have gallstones that cause gestin (taken 10 to 14 days a month) is more **likely** to cause: Premenstrual-like symptoms, ors range between 50 and 75 Gy, enteritis is **likely** to occur.[5] In this summary, unless weaker opioids, such as codeine, are just as **likely** to cause constipation.[2] (Refer to irect effect on cancer treatment outcome are **likely** to expand.[10] Table 1. National Can ical goblet cells. Dysplasia is particularly **likely** to develop in the intestinal type muco

patients younger than 50 years (who are more **likely** to tolerate the side effects and benef
ican-Americans and Native Americans are more **likely** to get chronic kidney disease. Being m
you at risk) for kidney stones make it more **likely** that you will get them. Some of these
. Men between the ages of 30 and 50 are most **likely** to get kidney stones. Postmenopausal w
stem is and how rapidly it is declining. How **likely** it is that your infection will progres
stem is and how rapidly it is declining. How **likely** it is that your infection will progres
roblems with alcohol or other drugs are more **likely** to have serious substance abuse proble
will allow substance use, he or she is more **likely** to try drugs or alcohol. Keep your tee
your body rejects the donor organ. You will **likely** not face major physical limitations af
ge your behavior, making other problems more **likely**. Sharing needles to inject drugs makes
aring needles to inject drugs makes you more **likely** to get or spread diseases, such as HIV
s HIV or hepatitis B or C. You also are more **likely** to have unsafe sex when on drugs, and

1) Lista das 140 ocorrências de *likely* encontradas no sub*corpus WebMD*

o were older, female, or Hispanic were more **likely** to have poorly controlled blood pressu
drank alcohol, or were overweight were more **likely** to have had their blood pressure broug
ics for acne were more than three times more **likely** to complain of sore throats than peop
ents using topical antibiotics were no more **likely** to report having a sore throat than th
ably from clothing; some were skin fragments **likely** caused by scratching. Importantly, nor
chological health. Psychotherapy is far more **likely** to be successful than medication in br
ple, Davis says, black adults are much more **likely** to cite violence-related issues as big
dy: Crohn's Disease, Ulcerative Colitis Less **Likely** in Women Living in Sunnier Regions Li
archers Find The poorer your sleep, the more **likely** you may be to develop Alzheimer's dise
n five awakenings in an hour, they are more **likely** to have preclinical Alzheimer's diseas
ly, more than five times an hour, were more **likely** than the others to have these abnormal
omarkers. Ju can't say exactly how much more **likely** the diagnosis was in those who woke up
time in bed actually sleeping were also more **likely** to have preclinical Alzheimer's disea
r variations of two specific genes are more **likely** to consume caffeine, and to drink more

says in an email. The researchers say it's **likely** that genetics plays a major role in ot
variants is needed, however, and there are **likely** more caffeine genes to be identified
o were older, female, or Hispanic were more **likely** to have poorly controlled blood pressu
drank alcohol, or were overweight were more **likely** to have had their blood pressure broug
g against colorectal cancer. Women are less **likely** than men to get the disease, and they
and people in the military, who may be more **likely** to take a blow to the head. But PRK ha
s (PPIs) for two years or longer may be more **likely** to sustain a hip fracture. And the lo
four weeks. The result: Swabs were 95% less **likely** to test positive for HSV-2 when a pers
nge at the start of the study were 50% more **likely** than men with normal blood pressure to
tolic readings of 80 or higher were 79% more **likely** than those with lower diastolic blood
that both men and women were almost equally **likely** to develop the disease as a result of
ese children were more than three times more **likely** than non-obese children to be vitamin
n resistance. Obese children were also more **likely** than non-obese children to skip breakf
than planned were one-and-a-half times more **likely** to report problems with infertility th
g solid foods. They were also slightly more **likely** than men to be takingmedication to tre
than the female surgical patients, were more **likely** to have weak valves where the esophag
ogress to esophageal cancer. Women were more **likely** to have a problem called a hiatal hern
n going in for reflux surgery were also more **likely** to be obese than their male counterpar
that different mechanical processes are more **likely** to be driving acid reflux in women co
vious studies have found that women are less **likely** than men to be satisfied with the resu
consider TIAs serious, so they may be less **likely** to stay on recommended treatments," sh
n among women. Men are also three times more **likely** than women to develop head and
ren a chance to speak their mind -- are more **likely** to raise children who are disrespectfu
s a legitimate authority figure and are more **likely** to break the rules, says researcher R
he new study, authoritarian parents are most **likely** to raise children who are disrespectfu
new study shows that shingles is at least as **likely** to strike a person who's already had
these patients throughout their lives, it's **likely** the recurrence rate will be much highe
th tobacco and marijuana were three times as **likely** as nonsmokers to have clinically conf

Southerners May Be Less **Likely** to Have Crohn's Study Shows Women Who Study Shows Women Who Live in South Are Less **Likely** to Have Crohn's Disease and ve in the southern part of the U.S. are less **likely** to have ulcerative colitis and Crohn' hern latitudes at age 30 were about 50% less **likely** to have Crohn's disease than those wh lived in the South were about one-third less **likely** to have ulcerative colitis than those South at birth and age 15 were somewhat less **likely** to have the gut disorders than those Study: 1 in 4 People **Likely** to Develop COPD Researchers Say Risk o n in her mid-30s is more than three times as **likely** to develop the progressive and potent the average man is more than three times as **likely** to develop COPD as prostate cancer. G ason, while seven in 10 parents say they're **likely** to get their children immunized, accor n The more medications a man takes, the more **likely** he will experience moderate or severe The more medications the men took, the more **likely** they were to experience moderate or s ker and/or are on more medications are more **likely** to have ED. He is an assistant profess orkbook with counseling were about twice as **likely** to report improvements in their overal ive to 10 years longer than men and are less **likely** to develop certain diseases. Specifica hromosome. Women may live longer and be less **likely** to develop some diseases, but they ar between the sexes is anything but simple and **likely** involves both genes and hormones, s ght boost eye pressure, making glaucoma more **likely**. So say Geraldo Magela Vieira, MD, and nee joint replacement surgeries were half as **likely** as non-users to need repeat surgeries non-Hispanic black women were 3.4 times more **likely** to die from preterm-related causes men Who Ate Low-Fat Dairy Products Were Less **Likely** to Develop Type 2 Diabetes By t over a decade, middle-aged women were less **likely** to be diagnosed with type 2 diabetes i estosterone levels may make men over 60 more **likely** to suffer an osteoporosis bone fractur n with osteoporosis bone fractures were more **likely** than other men to have had low testost fractures in healthy older men, "it is most **likely** to be justified only in those with the teenage athletes are under-reported and more **likely** to affect girls than boys, according t (thigh bone): 4% Varsity athletes were most **likely** to get a stress fracture, accounting f ounting for 64% of the cases. Boys were most **likely** to get fractures from track, football,

dney infections were more than four times as **likely** to have recently had a urinary tract i
Kidney infection was nearly four times more **likely** to occur in women with a history of di
kidney infections were almost four times as **likely** to have had trouble holding in their u
o had kidney infections were nearly twice as **likely** to have had a new sex partner in the p
with kidney infections were almost twice as **likely** to have recently used a spermicide. Mo
th kidney infections were about 1.5 times as **likely** to have a mother who had urinary tract
other who had urinary tract infections. It's **likely** that infections can travel up to the k
n their blood now considered 'high' are more **likely** to die of heart attack or stroke, the
hose with higher blood lead levels were more **likely** to be older, black or Mexican American
blood levels greater than 2 mcg/dL were more **likely** to die of heart disease heart disease
h pain and physical restrictions may be more **likely** to have problems relating to their pee
. "Kids with chronic illnesses are 30% more **likely** to become depressed," she says. "And e
Children with chronic illnesses are 30% more **likely** to develop depression, but parents can
e effects of the treatment reported. That is **likely** to be important news to parents and do
concluded that allergy, GERD, or asthma "are **likely** to account" for chronic cough in 90% o
panic white children were more than twice as **likely** as non-Hispanic black children or Hisp
lthough those with both conditions were more **likely** to need to be admitted to the intensiv
and black children. Older children were more **likely** than younger children to have a chroni
ave the inflammatory skin condition are more **likely** to have high blood cholesterol levels
onnected is not fully understood, but it's **likely** that slow, simmering inflammation is t
hildren, obese children were almost 40% more **likely** to have psoriasis than were normal-wei
xtremely obese children were almost 80% more **likely** to have psoriasis than were normal-
e normal-weight children. Psoriasis was more **likely** to be severe in children who were obes
new study shows. Gulf War veterans were more **likely** than veterans who served outside the
se who served in the Gulf were 40 times more **likely** to have a diagnosis of chronic fatigue
eterans. The deployed veterans were twice as **likely** as nondeployed veterans to have a diag
a diagnosis of fibromyalgia, almost twice as **likely** to experience frequent indigestion, an
xperience frequent indigestion, and 40% more **likely** to have skin rashes. Obligation Isn't

Employees who worked overtime were 61% more **likely** to sustain a work-related injury than
ys. So the heavier the teen or man, the more **likely** he may get the condition. But "gynecom
ited trait that has a big impact on both how **likely** a woman is to develop breast cancer an
with dense breasts are up to six times more **likely** to develop breast cancer than other wo
hinner with age, and women are four times as **likely** as men to develop osteoporosis. In add
roke Symptoms May Differ in Women Women Less **Likely** to Report Classic Symptoms By
Francisco) -- Female stroke victims are less **likely** to report classic stroke symptoms than
possible explanation for why they aren't as **likely** to get a crucial stroke drug. The new
ound that female stroke victims are 33% less **likely** to report a classic stroke symptom whe
shows women who suffer strokes are much less **likely** to be given the clot-busting drug tPA
ported that women aged 45 to 54 are twice as **likely** to report having had a stroke than men
roke Gender Gap Middle-Aged Women 3 Times as **Likely** to Have Had a Stroke Than Men By
ith women aged 45 to 54 now three times more **likely** than men in that age group to report h
women aged 45 to 54 were more than twice as **likely** as their male counterparts to have had
ol. "We found that in general, men were more **likely** to have conventional cardiovascular ri
und that women who smoked were 10 times more **likely** to have a heart attack than women
dren with one of the disorders were twice as **likely** to have had a strep infection within t
of the neurological disorders were twice as **likely** to have had a strep infection in the t
le with mildly overactive thyroids were more **likely** to develop atrial fibrillation , a typ
ve to cut back on health care -- you're less **likely** to recover as fully as more fortunate
rriers to health care services were 30% more **likely** to need a second hospital visit for he
barriers to prescription drugs were 70% more **likely** to need a second hospital visit for he
barriers to prescription drugs were 55% more **likely** to suffer angina -- chest pain due to
Chang, MD July 18, 2008 -- Global warming is **likely** going to hurt U.S. health, especially
e systems, and people living in poverty will **likely** suffer more than others from global wa
oken bones. The arms and hands were the most **likely** body regions to sustain a bone fractur
roblems. Adults over 55 were 4.25 times more **likely** than younger people to have heart-rela
under 19. Children were nearly 15 times more **likely** than adults to be hurt because they we

anate juice on hardening of the arteries are **likely** largely due to its high antioxidant co
lure Boosts Bone Fracture Risk Patients Less **Likely** to Exercise, Raising Their Risk of Fra
8 -- People with heart failure are much more **likely** to suffer bone fractures than other ca
rdiac patients, in part because they're less **likely** to exercise, a new study says. However
g physicians that these patients may be more **likely** to have fractures" and that more heart
ith the highest Lp-PLA2 levels were 89% more **likely** to have heart attacks, angina , or byp
inic reported. These patients were also more **likely** to suffer surgical complications, incl
hem anyway, tended to be older and were more **likely** to have other serious health problems
outside of recommended guidelines were more **likely** to die in the hospital, had greater ri
lso already knew that obese people were more **likely** to have high cholesterol , high blood
e protein in their blood were eight times as **likely** as patients with the lowest levels to
d pressure while asleep. They were also more **likely** to have overall good control of their
pressure. The bedtime group was one-third as **likely** to have heart and blood vessel problem
es show that migraine sufferers are far more **likely** to have the heart abnormality than peo
267-5210492) (Figure 1B). This deletion most **likely** disrupts expression of the d-globin ge
tely upstream from the d-globin gene that is **likely** to mediate silencing of the ?-globin g
obin-gene regulation. For example, there are **likely** to be functional roles of other elemen
ivatives of all three germ layers,1,13 it is **likely** that lineages other than the mesoderm
rom these three patients, and we think it is **likely** that these samples were negative purel
ssociated with recurrent events were more **likely** than those not associated with recurre
med only at baseline, it is possible (indeed **likely**) that some atheromas evolved over t
uld require thousands of patients and is not **likely** to be performed. Our results indicate
m analysis showed that the findings were not **likely** to change with continuation of the stu
e better than expected. Therefore, it is not **likely** that EMS systems would apply the ITD i
tion, attribution of cause-specific death is **likely** to be of limited accuracy in the popul
gh-grade anal intraepithelial neoplasia most **likely** reduces the incidence of anal cancer.
acy population to identify the HPV type most **likely** to have caused a given lesion. The met
hich of multiple HPV types detected was most **likely** to have caused the lesion, three lesio

ated sexual activity, vaccination would most **likely** result in levels of efficacy similar t

e hypothesized that this population was most **likely** to benefit from the addition of an ora

articipants receiving azithromycin were less **likely** to become colonized with respiratory p

zed with respiratory pathogens but were more **likely** to become colonized with macrolide-res

e. shows the age at the time of the biopsy, **likely** menopausal status on the basis of age,

after the initial biopsy were 2.5 times as **likely** (P=0.02) to have the cancer in the sam

clusion of the 4-1BB signaling domain. It is **likely** that the CART19-cell-mediated eliminat

ts with screening-detected cancers were more **likely** to have early-stage cancer (stage I or

hat were not detected by screening were more **likely** to be proximal and at a later stage th

tion of earlier-stage tumors, which are less **likely** to cause death than later-stage tumors

ry of grade 2 or higher acute GVHD were more **likely** to have the cord colitis syndrome than

e of precision with which such therapies are **likely** to be practiced in the clinical settin

o be involved in phospholipid metabolism and **likely** to participate in cell-membrane homeos

lood group B antigens are expressed are less **likely** to become ill if infected with Norwalk

H type 1 histo-blood group antigen are less **likely** to become ill if infected with Norwalk

vaccine and 6.7 with placebo; P=0.72), most **likely** owing to the small numbers of particip

ect by enrolling persons in whom illness was **likely** to develop after Norwalk virus exposur

, vaccine recipients were significantly less **likely** to have illness or infection; among th

ve immunity, a multivalent vaccine will most **likely** be required to provide protection agai

solates from the study that yielded 55989 is **likely** to illuminate the evolution of this li

xploit alternative adhesion mechanisms, very **likely** including aggregative adherence fimbri

cipated, suggesting that both protocols were **likely** to have had efficacy, but the finding

ousness and amnesia, some patients are still **likely** to have this complication. Despite its

as new, the recipients may have been more **likely** to report adverse events, especially s

not mandatory; therefore, the rate was most **likely** underreported. Nonserious vaccine r

viders' vigilance making underreporting **likely**, as is shown by the higher rate of rep

kely. Contamination of low-moisture foods is **likely** to lead to prolonged, dispersed outbre

s. Questions about peanut butter consumption **likely** elicited information about peanut butt

mplicate other individual products that most **likely** accounted for only a few cases. This o

treatment on the rate of decline in FEV1 are **likely** to be confounded as a result of bias b

treatment on the rate of decline in FEV1 are **likely** to be confounded as a result of bias b

o received amoxicillin clavulanate were less **likely** than children receiving placebo to

ted mortality. Given this consistency, it is **likely** that these hospitals differ somehow f

ro to 7 minutes. Therefore, it does not seem **likely** that the much lower frequency of ve

between the arrest and the initial ECG is **likely** to be prolonged.The results of this st

lity of CPR by bystanders in the home, is **likely** to be small. The rate of survival afte

t in the home is typically older and more **likely** to ha

Anexo C - Linhas de concordância de *mortality* no *NEJM*:

showing a clinically relevant difference in **mortality** would require thousands of patients
f infants. The Kaplan Meier estimate of the **mortality** rate at 12 months for all four grou
ar to affect the safety of the infants. The **mortality** rate was low among the infants (16.

Long Interdialytic Interval and **Mortality** among Patients Receiving Hemodialys
rdialytic interval is associated with excess **mortality** among U.S. patients receiving hemod
termine if there were differences in overall **mortality**; cause-specific mortality, with cau
erences in overall mortality; cause-specific **mortality**, with causes grouped as cardiac, va
cular, infectious, and other; cause-specific **mortality** for the five most commonly cited in
ncluded the other events. For cause-specific **mortality**, the groupings listed on the 2004 D
diovascular events (Table 2Table 2Annualized **Mortality** and Cardiovascular-Hospitalization
lytic interval than on other days: all-cause **mortality** (22.1 deaths vs. 18.0 per 100 perso
(22.1 deaths vs. 18.0 per 100 person-years), **mortality** from cardiac causes (10.2 vs. 7.5),
iac causes (10.2 vs. 7.5), infection-related **mortality** (2.5 vs. 2.1), mortality from cardi
, infection-related **m**ortality (2.5 vs. 2.1), **mortality** from cardiac arrest (1.3 vs. 1.0),
mortality from cardiac arrest (1.3 vs. 1.0), **mortality** from myocardial infarction (6.3 vs.
sawtooth pattern was apparent for all-cause **mortality**, mortality from cardiac causes, and
attern was apparent for all-cause mortality, **mortality** from cardiac causes, and cardiovasc
dialysis session (Figure 1Figure 1Annualized **Mortality** and Cardiovascular-Admission Rates
Dialysis Week.). Figure 2Figure 2Annualized **Mortality** and Cardiovascular-Admission Rates
ther Days, According to Subgroup. summarizes **mortality** and cardiovascular-admission rates
rval than on other days, including all-cause **mortality**, mortality from cardiac causes, inf
n other days, including all-cause mortality, **mortality** from cardiac causes, infection-rela
ality from cardiac causes, infection-related **mortality**, mortality from cardiac arrest, and
cardiac causes, infection-related mortality, **mortality** from cardiac arrest, and mortality
ortality, mortality from cardiac arrest, and **mortality** from myocardial infarction. Similar
re frequently on Mondays, and the associated **mortality** was 60% during the first 48 hours a
an occurrence that is associated with higher **mortality**.15,16 Regarding the timing of dialy
ysis between 1990 and 1993. Shift-associated **mortality** differences were present, but only

ts 60 years of age or older, with the lowest **mortality** during the morning shift, intermedi
ality during the morning shift, intermediate **mortality** during the afternoon shift, and the
during the afternoon shift, and the highest **mortality** during the evening shift.17 Morning
ift.17 Morning-shift patients also had lower **mortality** in the study by Bliwise et al.18 B
n. For example, one could argue that because **mortality** and cardiovascular-admission rates
opulation, did not have higher-than-expected **mortality** after the long interdialytic interv
n subgroups and across outcomes, with higher **mortality**, especially for deaths from cardiac
dditional reductions in ischemic events2 and **mortality**3 but at the cost of an increase in
causes is strongly associated with increased **mortality**,21 directly reflects the well-being
isk in terms of cardiovascular morbidity and **mortality**, quality of life, exercise toleranc
iated with more cardiovascular morbidity and **mortality**? First, the incidence of heart fail
ths in the United States.1 Colorectal-cancer **mortality**2-4 and incidence5,6 are reduced wit
of colorectal cancer (by 23%) and associated **mortality** (by 31%).11 In Italy, an 18% reduct
idence and a nonsignificant 22% reduction in **mortality** were observed,12 whereas in Norway,
l and proximal colorectal cancer and related **mortality**. Methods Study Design A total of 15
survival, harms of screening, and all-cause **mortality**. All cancers and deaths were ascert
on-to-screen comparison of colorectal-cancer **mortality** between the study groups. Event rat
cers or deaths) divided by person-years. For **mortality**, person-years were measured from ra
ater relative reduction in colorectal-cancer **mortality** in the intervention group, as compa
confidence intervals for incidence-rate and **mortality** ratios assuming a Poisson distribut
onfidence interval for the colorectal-cancer **mortality** ratio were derived in accordance wi
We assessed the between-group difference in **mortality** with the use of a weighted log-rank
weights that were proportional to the pooled **mortality**. The weighted statistic was chosen
the effect of screening on colorectal-cancer **mortality**. The monitoring design stipulated a
y was 21.9%. Colorectal-Cancer Incidence and **Mortality** According to Study Group Table 2Tab
able 2Table 2Colorectal-Cancer Incidence and **Mortality**. shows colorectal-cancer incidence
ality. shows colorectal-cancer incidence and **mortality** in the intervention group as compar
orectal cancer was 282 (95% CI, 210 to 427). **Mortality** related to colorectal cancer was 2.
risk, 0.74; 95% CI, 0.63 to 0.87; P<0.001). **Mortality** related to distal colorectal cancer
k, 0.50; 95% CI, 0.38 to 0.64; P<0.001), but **mortality** related to proximal colorectal canc
Men had a 34% reduction in colorectal cancer **mortality** (139 vs. 211 deaths; relative risk,

0). The relative risks for colorectal-cancer **mortality** among participants 55 to 64 years o and Proximal Colorectal-Cancer Incidence and **Mortality**.). The cumulative incidences of pro ntervention group. Overall colorectal-cancer **mortality** and mortality related to distal and oup. Overall colorectal-cancer mortality and **mortality** related to distal and to proximal c 1B, 1D, and 1F. Between-group differences in **mortality** emerged within a few years and pers stal colorectal cancer, but no difference in **mortality** related to proximal colorectal canc P<0.001 for both comparisons). Incidence and **Mortality** According to Location and Stage of able 4Table 4Colorectal-Cancer Incidence and **Mortality** According to Location and Stage. sh Stage. shows colorectal-cancer incidence and **mortality** according to the location in the co 61.7% for stage IV cancers (66 fewer cases). **Mortality** related to distal colorectal cancer igmoid had little effect on the incidence or **mortality** results. Cancer Treatment, Screenin a 26% reduction in overall colorectal-cancer **mortality** and a 21% reduction in the incidenc ction in the incidence of colorectal cancer. **Mortality** related to distal colorectal cancer d, but there was no significant reduction in **mortality** related to proximal cancer. The bas ng. The observed reductions in incidence and **mortality** are similar to the results of the U xamination on colorectal-cancer incidence or **mortality**. In the United States, endoscopic s s testing probably reduced the difference in **mortality** and incidence between the two group ally reduces colorectal-cancer incidence and **mortality**, especially with regard to distal c benefit of endoscopic screening in reducing **mortality**31,32 from and the incidence33,34 of th regard to colorectal-cancer incidence and **mortality** was limited to the distal colon, bu proximal colorectal cancer (9%) and related **mortality** (15%) were not significant. In the ng period. We did not observe a reduction in **mortality** related to proximal colorectal canc of the benefit in reducing colorectal-cancer **mortality** from screening derives from a reduc n stage IV disease,4 which has a much higher **mortality** than lower stages. In the PLCO tria he incidence of stage IV disease and related **mortality** were observed. In contrast, for can the incidence of stage IV disease or related **mortality** were identified. Furthermore, in th ce of proximal colorectal cancer and related **mortality** of the additional polyp removal wit veness of flexible sigmoidoscopy in reducing **mortality** related to distal colorectal cancer igmoidoscopic screening on colorectal-cancer **mortality** and incidence in subgroups, the res ced by 12% (a nonsignificant difference) and **mortality** by 33% (a significant difference).4 oth groups, incidence was reduced by 21% and **mortality** by 26%. Endoscopic testing appears

decrease in colorectal-cancer incidence and **mortality**. The incidence of colorectal cancer
d proximal colon. A significant reduction in **mortality** was observed only for cancer in the
e syndrome was not associated with increased **mortality** among patients undergoing cord-
which has resulted in significant increases in **mortality**, length of hospital stay, and hospi
anesthesia administered or in postoperative **mortality**.11,21,22 Both protocols in the B-Un
n St. Louis, and 43 in Winnipeg). The 30-day **mortality** was 1.96% (57 of 2907 patients) in
associated with an increase in postoperative **mortality**21,22 or in the amount of anesthetic
unstable phenomenon20 that does not predict **mortality** when post-bronchodilator FEV1 is
unstable phenomenon20 that does not predict **mortality** when post-bronchodilator FEV1 is t
iability in the Measurement of Hospital-wide **Mortality** Rates Abstract Several countries us
Abstract Several countries use hospital-wide **mortality** rates to evaluate the quality of ho
sment of methods to calculate this aggregate **mortality** metric for use as a measure of hosp
nd for hospital-level observed and expected **mortality** rates. We compared the numbers and
ds. We also compared hospitals' standardized **mortality** ratios and classification of hospi
ratios and classification of hospitals with **mortality** rates that were higher or lower tha
riteria, two methods calculated in-hospital **mortality** rates (4.0% and 5.9%) that were twi
n coefficients) of discharge-level predicted **mortality** probabilities ranged from 0.46 to
that had higher-than- expected hospital-wide **mortality** when classified by one method had l
sified by one method had lower-than-expected **mortality** when classified by one or more of
common methods for calculating hospital-wide **mortality** produced substantially different re
othesized association between hospital-wide **mortality** and quality of care. (Funded by the
ious approaches to estimating hospital- wide **mortality** rates as performance metrics. Medic
nce metrics. Medicare released hospital-wide **mortality** rates for its beneficiaries beginn
rious methods for calculating hospital-wide **mortality** rates.7 At the DHCFP's request, we
s to calculate risk-adjusted, hospital-wide **mortality** rates with the use of data from sta
on outcomes after discharge, such as 30-day **mortality**. We asked vendors to apply their ri
evel estimates (e.g., observed and expected **mortality** rates) for each hospital. We perfor
t inclusion and exclusion criteria in their **mortality** algorithms. We therefore calculated
charges and hospitals, including in-hospital **mortality** rates; the age, sex, and race or e
ach method, we used the predicted and actual **mortality** rates that were based on the disch
od. Some methods produced risk- standardized **mortality** ratios, whereas others produced ris

s, whereas others produced risk-standardized **mortality** rates. We converted all measures t
variance procedure that modeled standardized **mortality** ratios as a function of method fix
ch method indicated whether a hospital was a **mortality** outlier in each of the 3 years. We
ually exclusive groups: higher-than-expected **mortality**, as-expected mortality, and lower-
higher-than-expected mortality, as-expected **mortality**, and lower- than-expected mortality
expected mortality, and lower- than-expected **mortality**. Typically, outlier designations we
ars of data, along with annual standardized **mortality** ratios (including standard errors)
pected, as-expected, or higher-than-expected **mortality**, with a P value of less than 0.05
e study period, with an overall in-hospital **mortality** rate of 2.1% (Table 1TABLE 1 Hospi
to Four Methods of Measuring Hospital-wide **Mortality**.). Two methods excluded some
methods excluded some hospitals from their **mortality** calculations. Percentages of discha
Dr. Foster methods had overall in- hospital **mortality** rates of 5.9% and 4.0%, respectivel
s who were older and had higher in-hospital **mortality**, longer lengths of stay, and more r
oduced different estimates of hospital-level **mortality** and assessments of whether individ
higher-than-expected or lower-than-expected **mortality**. Every method provided hospital-wi
e correlations of hospital-wide standardized **mortality** ratios depended on the weighting o
elation among 78 Hospital-Level Standardized **Mortality** Ratios for Fiscal Year 2007, Accord
1 Comparison of Hospital-Level Standardized **Mortality** Ratios (SMRs) for Fiscal Year 2007,
e Thomson Reuters cohort for Hospital C. The **mortality** rate among these 3% of discharges
stantially higher than Hospital C's overall **mortality** rate (2.2%). Intraclass correlation
lower- than-expected or higher-than-expected **mortality** were high and varied among
ure 2FIGURE 2 Percentages of Hospitals with **Mortality** Rates Higher or Lower Than Expected
s having higher- than-expected hospital-wide **mortality**, as compared with 38 of 77 hospital
t11 between methods in classifying hospital **mortality** performance, depending on the year
s having higher- than-expected hospital-wide **mortality** by one method, 12 were
ly classified as having lower-than-expected **mortality** by other methods (6 by one method,
ethods). The goal of assessing hospital-wide **mortality** rates is to make inferences about t
itals. Proponents believe that hospital-wide **mortality** metrics provide useful warning fla
ed, and tabulated accurately. Hospital-wide **mortality** encompasses wide-ranging diagnoses,
the four methods for assessing hospital-wide **mortality** that we studied are currently mark
g. We found that estimates of hospital-wide **mortality** could vary, sometimes widely, among

hospitals at either end of the spectrum of **mortality** categories (higher than expected vs
ertainty in the presumed association between **mortality** and hospital quality. With more li
for which the linkages between in-hospital **mortality** and quality are problematic. In add
ng performance on the basis of hospital-wide **mortality** rates raise the inevitable questio
among methods for estimating hospital-wide **mortality** are similar to the findings of a s
findings of a study22of condition-specific **mortality** rates and hospital performance. The
s. A broader question is whether in-hospital **mortality** rates are valid quality indicators.
examined the relationships between hospital **mortality** rates for individual conditions an
bsence of an association between in-hospital **mortality** and quality or, given the study de
ore worrisome when it comes to hospital-wide **mortality** measures that encompass a diverse
ma or advanced cancer), linking in-hospital **mortality** with hospital quality is clearly pr
ing whether to include them in hospital-wide **mortality** estimates. Differences in hospital
ortality estimates. Differences in hospital **mortality** findings among methods also probabl
ntial for gaming exists when risk-adjusted **mortality** rates are used for public reportin
pparent temporal reduction in hospital-wide **mortality** rates in the United Kingdom.5 The u
mask the very quality problems that hospital-**mortality** metrics aim to identify. One poten
ee hospitals as having higher-than-expected **mortality**. In fiscal year 2007, all four meth
one hospital as having lower-than-expected **mortality** and another hospital as having high
her hospital as having higher-than- expected **mortality**. Given this consistency, it is like
are unlikely to use more than one method of **mortality** measurement. Since they typically
e fundamental hypothesis that hospital-wide **mortality** is a valid metric for the quality o
bility that the estimation of hospital-wide **mortality** rates on the basis of nationally st
ts suggest that efforts to use hospital-wide **mortality** rates to evaluate the quality of c
tial alternative to the use of hospital-wide **mortality** rates as a metric would be to esti
oup of diagnoses for which the link between **mortality** and quality is most plausible, samp

ANEXO D - Linhas de concordância com *cardiac arrest(s)*:

tachycardia as the first recorded rhythm after **out-of-hospital cardiac arrest** has

his may be the more common initial rhythm when **out-of-hospital cardiac arrest** occurs in

we conducted a prospective cohort study of **out-of-hospital cardiac arrest** in adults in 10 Nor

ar tachycardia as the first recorded rhythm in **out-of-hospital cardiac arrest** has declined dramat

rained laypersons improves survival after an **out-of-hospital cardiac arrest**.6 In contrast, l

d emergency-medical-services (EMS) registry of **out-of-hospital cardiac arrest**.3 We carried out a

ons 19 years of age or older with nontraumatic **out-of-hospital cardiac arrest** for whom external d

a) are a relatively infrequent presentation of **out-of-hospital cardiac arrest** (with an overall

ies outside the home. Thus, the location of an **out-of-hospital cardiac arrest** may be a surrogate

A Trial of an Impedance Threshold Device in **Out-of-hospital cardiac arrest** Standard cardiopulm

charge with satisfactory function for adults with **out-of-hospital cardiac arrest**. Methods Study Sett

Patient Population All adults with nontraumatic, **out-of-hospital cardiac arrest** were eligible for e

uring standard CPR in patients with nontraumatic, **out-of-hospital cardiac arrest**. One possible expla

uring standard CPR in patients with nontraumatic, **out-of-hospital cardiac arrest**. Use of the active

rly versus Later Rhythm Analysis in Patients with **Out-of-hospital cardiac arrest** Out-of-hospital car

s in Patients with Out-of-hospital cardiac arrest **Out-of-hospital cardiac arrest** is a common and let

ival to hospital discharge among patients with an **out-of-hospital cardiac arrest** who are treated by

ates in recent years. The traditional approach to **out-of-hospital cardiac arrest** has been to emphasi

d all persons 18 years of age or older who had an **out-of-hospital cardiac arrest** that was not the re

l, we tested the hypothesis that patients with an **out-of-hospital cardiac arrest** might benefit from

ival. Given the complex clinical circumstances of **out-of-hospital cardiac arrest**, precise control of likely to be

practiced in the clinical setting of **out-of-hospital cardiac arrest**. Furthermore,

cardiac rhythm during CPR in patients who had an **out-of-hospital cardiac arrest** that was

07 cases (25%), as compared with 61 of 161 **EMS-witnessed cardiac arrests** (38%) that

ublic location (Table 2 and Figure 2). For **EMS-witnessed cardiac arrests**, the odds ratio

shockable arrhythmias was higher for **bystander-witnessed cardiac arrests** in a public

EMS arrival was modestly longer for **bystander-witnessed cardiac arrests** at home than

to EMS response time in the case of **bystander-witnessed cardiac arrests** in public

on successful bystander-applied AED shocks and **witnessed cardiac arrests** in both public

(usually a rural farmland location). **Bystander-witnessed cardiac arrest** was defined as

harge after adjustment for age, sex, **bystander-witnessed cardiac arrest**, the delivery of

Of the 3451 patients with **bystander-witnessed cardiac arrest** that occurred in the home