ROBERTO DINIZ COSTA

A CONVERSÃO SOB A LUZ DA LINGÜÍSTICA DE CORPUS

MESTRADO EM

LINGÜÍSTICA APLICADA E ESTUDOS DA LINGUAGEM

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO

2006

ROBERTO DINIZ COSTA

A CONVERSÃO SOB A LUZ DA LINGÜÍSTICA DE CORPUS

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de MESTRE em Lingüística Aplicada e Estudos da Linguagem, sob orientação do Prof. Dr. Antonio Paulo Berber Sardinha

PUC - SP
2006

BANCA EXAMINADORA

_____

_____

_____

Para Rosana, Benjamin e Nilce

AGRADECIMENTOS

Ao meu orientador, professor Doutor Antonio Paulo Berber Sardinha, pela orientação, incentivo, amizade e cobrança quando necessária.

À minha mãe, Rosana Santos Diniz, pelos ideais ensinados e pela perseverança exemplar mostrada ao longo de toda sua vida.

À Ana Elvira Luciano Gebara, por todo auxílio durante a graduação e incentivo à criação do embrião que originou este trabalho.

À Carol, que, nos momentos mais necessários e cruciais, esteve disposta a me ajudar.

Ao Yázigi Internexus, que, por intermédio de excelentes professoras que tive, me fez descobrir e cultivar o amor pela língua inglesa.

A meus colegas de sala, Denise Delegá, Daniela Penharvel, Renata Condi, Renata Picasso e Carlos Kauffmann.

Às professoras Rosinda Guerra Ramos, Sumiko Ikeda e ao professor Bruno Dallari pela ajuda e pelo incentivo.

ABSTRACT

The main purpose of this study was to outline, describe what is widely called conversion and show, in an empirical way, how frequent this process is in order to point out which word classes are more converted in the English language.

In order to achieve this objective, this study has found support in Corpus Linguistics, which is a Linguistics field that is concerned about analyzing any language empirically through corpora, which can be defined as authentic text compilations stored in computer files.

Besides being supported by Corpus Linguistics, this dissertation has focused on the study of conversion, which is also called improper derivation in the prescriptive grammar. There is a great range of studies on the conversion process; however, very few of them has demonstrated empirically how frequent conversion is and they have seldom used authentic texts in their analysis. Therefore, this study has tried to fill this gap by being able to show empirically the limits towards conversion through the use of an authentic corpus, the *British National Corpus*.

In order to achieve the objective of this study, the following questions have been stated:

1. What are the word classes that are most and least likely to be converted?

2. What is the conversion rate, in the selected corpus, between the words that are and are not involved in the conversion process?

3. In the selected corpus, what words show the widest array of grammar classes that can possibly be converted?

4. Are there any differences concerning the occurrence of conversion

between the most frequent words and the words that show the widest array of grammar classes that can possibly be converted? If so, why?

The corpus employed in the study was the *British National Corpus*, compiled from 100 million words. These words have been obtained from a wide array of sources, gathered in a way they could show a wide sample of spoken and written British English. It is obvious that, by dealing with a sample of the target language, the corpus might not represent a hundred percent of the language used. This way, it will always be an incomplete sample; therefore, when analyzing corpora, one should work with the concept of probability, which is really helpful to this study, once it made possible to show how probable conversion is.

The results have indicated that conversion is a frequent process in the English language, once it involves 10 word classes and more than 36 million words. Moreover, it might be suggested that conversion is a living process in the English language and, through an empirical way, it is possible to point out that verbs, prepositions and pronouns are the most intrinsically related to conversion word classes.

To sum up, this dissertation hopes that it has contributed to the study of conversion, since there are few studies that have demonstrated empirically how frequent conversion is. This work also presents and discusses the research limitations and its possible applications in the future, apart from pedagogical applications of the results.

RESUMO

Este trabalho teve como objetivos principais delimitar e descrever o processo comumente chamado de conversão, demonstrar, de maneira empírica, sua freqüência para apontar as classes de palavras mais tipicamente convertidas na língua inglesa.

Para tanto, o trabalho encontrou suporte teórico na Lingüística de Corpus, que se preocupa em analisar qualquer língua de maneira empírica, por intermédio de corpora, ou seja, compilações de textos autênticos armazenadas em formato de arquivo de computador.

Além da Lingüística de Corpus, o projeto fundamentou-se no estudo da conversão, que é também chamada de derivação imprópria na gramática normativa. Há vários trabalhos que lidam com a questão da conversão; entretanto, nenhum deles demonstrou empiricamente quão freqüente a conversão é e quase nunca utilizaram textos autênticos como objeto de análise e exemplos. Desse modo, este trabalho buscou preencher essa lacuna, sendo, portanto, capaz de demonstrar empiricamente a extensão da conversão na língua inglesa por meio de um corpus autêntico, o *British National Corpus*.

No intuito de operacionalizar os objetivos deste trabalho, foram elencadas as seguintes perguntas de pesquisa:

1. Quais são as classes de palavras mais e menos passíveis de serem convertidas na língua inglesa?

2. Qual é a proporção, no corpus selecionado, entre as palavras que estão envolvidas no processo de conversão e as que não são modificadas por esse fenômeno?

3. Quais são as palavras que apresentam maior repertório de classes de palavras passíveis de serem convertidas no corpus selecionado?

4. Há diferenças na freqüência do processo de conversão entre as palavras mais freqüentes e as que apresentam maior repertório de classes de palavras passíveis de serem convertidas? Em caso afirmativo, o que explica essa diferença?

O corpus empregado na pesquisa foi o *British National Corpus*, (BNC), compilado a partir de 100 milhões de palavras obtidas da língua inglesa escrita e falada. As amostras de língua foram obtidas de uma grande variedade de fontes, reunidas de modo a mostrar um recorte abrangente do inglês britânico falado e escrito do final do século XX. É evidente que, por se tratar de uma amostra, o corpus pode não representar cem por cento da língua utilizada. Nesse sentido, será sempre uma amostra incompleta; dessa forma, ao estudar um corpus, devemos trabalhar com a questão de probabilidade, o que é extremamente útil para este trabalho, uma vez que nos permitiu determinar quão provável a conversão é.

Os resultados indicaram que a conversão é um processo recorrente na língua inglesa, uma vez que envolve 10 classes de palavras e mais de 36 milhões de palavras. Além disso, pode-se sugerir que a conversão é um processo vivo na língua inglesa e, de maneira empírica, é possível apontar que as classes de palavras que estão mais intrinsecamente relacionadas à conversão são os verbos, as preposições e os pronomes.

Por conseguinte, a pesquisa pretende ter contribuído para o estudo da conversão, uma vez que há poucos trabalhos que demonstrem, de maneira empírica, quão freqüente a conversão é. Desse modo, a pesquisa espera ter feito uma contribuição original para a área. O trabalho ainda apresenta e discute as limitações da pesquisa realizada, aponta possíveis desdobramentos deste estudo em pesquisas futuras e oferece sugestões dos resultados da pesquisa.

SUMÁRIO

Introdução 1

Capítulo 1: Fundamentação Teórica 6

1.1 A Lingüística de Corpus 6
1.1.1 Visão probabilística da linguagem 7
1.1.2 Definição de Corpus 10
1.1.3 Breve histórico da Lingüística de Corpus 11
1.2 Conversão 12
1.2.1 Breve panorama histórico sobre a conversão 12
1.2.1.1 A conversão no latim 12
1.2.1.2 A conversão no *Old English* 13
1.2.1.3 A relação da conversão entre o inglês e as línguas românicas 14
1.2.2 A problemática em definir e delimitar o fenômeno da conversão 15
1.2.3 Conversão e homonímia 21

Capítulo 2: Metodologia de Pesquisa 23

2.1 Objetivos e questões de pesquisa 23
2.2 Descrição do corpus utilizado 24
2.2.1 O *British National Corpus* em números 25
2.2.2 O processo de etiquetagem do corpus 26
2.3 Procedimentos metodológicos 31

Capítulo 3: Apresentação e Análise dos Resultados 58

3.1 O enfoque da probabilidade 59
3.2 O enfoque da possibilidade com maior repertório de etiquetas 61
3.2.1 A conversão na palavra *like* 63
3.2.2 A conversão na palavra *set* 64
3.2.3 A conversão na palavra *forecast* 64
3.3 Análise contrastiva entre as 1000 palavras mais freqüentes e as 1000 que apresentam maior repertório de etiquetas no *British National Corpus* 65

Considerações Finais 68

Referências Bibliográficas 74

Anexos 78

LISTA DOS QUADROS E FIGURAS

Figuras

Figura 01	Página inicial do identificador da conversão	35
Figura 02	Tela de apresentação dos resultados gerais	37
Figura 03	Tela de apresentação da lista de palavras simples	38
Figura 04	Tela de apresentação da lista de palavras completa	38
Figura 05	Tela de apresentação da lista de palavras completa, filtrada	39
Figura 06	Tela de apresentação dos conjuntos de etiquetas originais	40
Figura 07	Tela de apresentação dos conjuntos de etiquetas desmembradas	40
Figura 08	Tela de apresentação dos conjuntos de etiquetas duplas	41
Figura 09	Tela de apresentação dos conjuntos de etiquetas no formato de seqüência	41
Figura 10	Tela de apresentação da somatória de etiquetas	42
Figura 11	Agrupamento de etiquetas iguais	43
Figura 12	Separação das etiquetas e freqüência - tela inicial	44
Figura 13	Separação das etiquetas e freqüência - etapa 1 de 3	45
Figura 14	Separação das etiquetas e freqüência - etapa 2 de 3	46
Figura 15	Separação das etiquetas e freqüência - etapa 3 de 3	47
Figura 16	Etiquetas agrupadas - tela com os resultados	48
Figura 17	Adição de títulos às colunas	49
Figura 18	Tela com os títulos inseridos	50
Figura 19	Utilização do comando *Subtotais*	51
Figura 20	Tela com resultado do comando *Subtotais*	52
Figura 21	Classificação das etiquetas em classes de palavras de acordo com a tabela 02	53
Figura 22	Transferência dos valores das classes de palavras para a planilha 2	54
Figura 23	Filtragem das linhas que não representam subtotais de classes de palavras	55
Figura 24	Totais de freqüência por classe de palavra	56
Figura 25	Tela com resultados finais	57

Gráficos

Gráfico 01	A distribuição da conversão nas classes de palavras a partir das 1000 palavras mais freqüentes no *British National Corpus*	59
Gráfico 02	Número de tokens e a conversão no BNC	60
Gráfico 03	As 1000 palavras que apresentam maior possibilidade de conversão com maior repertório de etiquetas	63

Tabelas

Tabela 01	Relação entre estudiosos e suas respectivas visões acerca da conversão	20
Tabela 02	Composição do *British National Corpus*	26
Tabela 03	Relação entre as etiquetas e as classes de palavras	30
Tabela 04	Palavras do BNC com maior número de etiquetas	31
Tabela 05	1000 palavras mais freqüentes no BNC	32
Tabela 06	Etiquetas desconsideradas neste trabalho	36
Tabela 07	Resultados obtidos a partir das 1000 palavras mais freqüentes No *British National Corpus*	59
Tabela 08	Distribuição da conversão nas 1000 palavras que apresentam maior possibilidade de conversão com maior repertório de etiquetas	62
Tabela 09	Comparação quanto à freqüência entre as 1000 palavras mais freqüentes e as 1000 que apresentam maior repertório de etiquetas no *British National Corpus*	65

INTRODUÇÃO

Sou professor de inglês como língua estrangeira e em uma fatídica noite, ao ler uma redação de uma aluna, vi a seguinte oração *The riches are happy*. Ela me chamou atenção porque me parecia estar gramaticalmente errada: *rich*, sendo adjetivo, não poderia ser posta no plural, para significar *rich people*. *Riches*, eu tinha a impressão, significava *coisas ricas*. Sinceramente não tinha certeza de como corrigir a redação de minha aluna. Poderia, a princípio, colocar a palavra *people* depois da expressão *the rich*. Como estava com tempo, resolvi procurar no dicionário *Cambridge* a palavra *rich* para ver se era oferecida alguma sugestão para o uso dessa palavra.

Ao ler o verbete, vi que havia a opção *the* + *rich* + verbo no plural. Caiu a ficha! Utilizando meu conhecimento sobre a língua portuguesa, lembrei que em português era permitido o uso do adjetivo rico como substantivo. Por que não o seria em inglês? Utilizando meu espírito investigativo, procurei o antônimo dessa palavra – *poor*. O dicionário apresentava a mesma sugestão: *the* + *poor* + verbo no plural. Esses fatos sugeriram que esse fenômeno poderia ser recorrente na língua inglesa, mas nenhum dicionário ou gramática fornecia informação acerca da extensão desse fenômeno chamado de conversão. O que eu poderia dizer a meus alunos sobre a tipicidade do fenômeno? Seria algo comum ou incomum? Em quais classes ou palavras a conversão mais incidiria? Como não sou falante nativo do inglês, não poderia suprir a falta dessa informação com minha intuição, muito embora a intuição do falante nativo não seja confiável para responder questões desse tipo (Sinclair: 1991). Em suma, foram essas inquietações a mola propulsora para o desenvolvimento deste trabalho.

Essa pesquisa no dicionário remeteu-me às aulas que tive no ensino médio, quando havia estudado, em figuras de linguagem, o fenômeno chamado de derivação imprópria ou conversão. Lembrei-me de exemplos um tanto quanto poéticos, como *O despejar do rio* (conversão de verbo para substantivo) e também da minha dificuldade em reconhecê-los na linguagem utilizada por mim no dia a dia. Será que, em inglês, a conversão era um fenômeno produtivo? Será que ocorria somente com exemplos poéticos ou também com palavras que utilizava (e utilizo) nas minhas aulas?

Logo após pensar nessas perguntas, veio a minha cabeça a palavra *water*, que pode ser utilizada como substantivo ou verbo, como nos exemplos *I love to drink water* e *I need to water the plants* respectivamente. O último uso apresenta(va) certa dificuldade para os alunos, uma vez que, em português, muitos verbos formados a partir de substantivos apresentam sufixo, como o par *dicionário* e *dicionarizar*.

Com base em todas as descobertas que fizera, propus-me a estudar este fenômeno em inglês e, logo no início, fui recebido com um *sim* por meu orientador, o professor Doutor Tony Berber Sardinha. Logo nas primeiras orientações surgiram diversos questionamentos, tais como “Por que esse fenômeno existe?”, “Para que ele serve?”, “Quão freqüente ele é?”, “Ele é mesmo freqüente?, Quanto, numericamente falando?” e “Com quais palavras ocorre mais?”. Naquele momento, não tinha as respostas e nem mesmo a certeza de que as conseguiria.

Muitas palavras novas não são tão novas como pensamos. Elas são, de certo modo, apenas ajustes dentro de uma mesma língua, como, por exemplo, acréscimos a itens existentes ou recombinação de elementos (Bartolomé & Cabrera, 2005:10). O fenômeno que me intrigava, surgido a partir do problema da redação que chamara minha atenção, pode ser inserido nesse quadro e é tecnicamente chamado de conversão.

De acordo com Bartolomé & Cabrera (2005:1), a língua inglesa, a exemplo de outras línguas, é extremamente produtiva e devido à sua natureza versátil, pode nela ocorrer diversos processos de formação de palavras a fim de se criar novo léxico. Algum desses processos são extremamente lexicalizados – tais como derivação e composição; entretanto, novas tendências estão obtendo destaque no campo da produtividade, como, por exemplo, abreviação, hibridismo e conversão. Muitos estudiosos dessa área diferem quanto à maneira como esses fenômenos devem ser tratados; todavia, em apenas um ponto há comum acordo: esses novos processos estão se tornando cada vez mais usados.

Cannon (1985:415, apud Bartolomé & Cabrera 2005) afirma que o processo de conversão será cada vez mais ativo e, portanto, criará uma grande parte das palavras novas que surgirão na língua inglesa. Bartolomé & Cabrera (2005:1) comentam que a conversão é provavelmente o processo mais notável no campo de formação de palavras em inglês.

Cannon (1985:429, Bartolomé & Cabrera 2005) ainda afirma que não há uma

maneira precisa de se saber o número de conversões que surgem a cada dia na língua falada, embora se saiba que esse número seja elevado. É exatamente nesse contexto que se insere a Lingüística de Corpus, visto que por intermédio dela é possível demonstrar empiricamente quão freqüente esse fenômeno é. Com embasamento na Lingüística de Corpus é possível ir muito além; ou seja, mostrar quais palavras ou classes de palavras são convertidas e até mesmo os padrões que esse fenômeno segue.

O objetivo da utilização da conversão varia de acordo com o usuário da língua. Aitchison (1989:161) declara que os adultos utilizam a conversão como uma ferramenta de economia de palavras, ao passo que as crianças a utilizam para serem entendidas, apesar de, às vezes, produzirem enunciados não gramaticais. De qualquer forma, a conversão tende a ajudar a tornar a comunicação mais fácil.

Há diferentes visões sobre como a conversão opera, como as unidades convertidas devem ser categorizadas no estudo de morfologia e léxico, e sobre quais são seus limites. Existem autores que delimitam a conversão somente a substantivos e verbos (Karaščuk, 1977, apud Valera 2000), ao passo que outros classificam alguns casos de homonímia como conversão (Smirnickij, 1954, apud Valera 2000).

Alguns estudos diacrônicos sobre a conversão têm como ponto de partida a língua latina, tais como os realizados por Adams (1976) e Hummel (2003). Esses estudos demonstram que a conversão já existia em latim e, conseqüentemente, foi incorporada nas línguas derivadas. De acordo com Postnikova (1999), outros estudos têm como ponto de partida o *Old English* e os resultados obtidos foram muito similares, isto é, a conversão também esteve presente no *Old English.*

Com a evolução das línguas, da tecnologia e com a conseqüente invenção do computador, foi possível compilar grandes números de palavras em corpora digitalizados. Com o advento da utilização do computador como ferramenta para análise de dados, a Lingüística de Corpus vem adquirindo um status de maior importância no cenário acadêmico.

Infelizmente os estudos sobre a conversão não seguiram o mesmo caminho. Atualmente há poucos estudos sobre a conversão que utilizem corpora[1] como objeto de estudo. Muitos estudos ainda utilizam exemplos inventados, fruto de uma

[1] Aqui corpora já entendido como corpora digitalizado.

abordagem chomskyana, que nem sempre representam a língua em uso e quase não há estudos que demonstrem empiricamente quão freqüente e produtivo a conversão é.

Em face ao quadro descrito acima, este trabalho se propõe a delimitar e descrever o processo comumente chamado de conversão, demonstrar, de maneira empírica, sua freqüência para apontar as classes de palavras mais tipicamente convertidas na língua inglesa.

No intuito de operacionalizar os objetivos deste trabalho, foram elencadas as seguintes perguntas de pesquisa:

1. Quais são as classes de palavras mais e menos passíveis de serem convertidas na língua inglesa?

2. Qual é a proporção, no corpus selecionado, entre as palavras que estão envolvidas no processo de conversão e as que não são modificadas por esse fenômeno?

3. Quais são as palavras que apresentam maior repertório de classes de palavras passíveis de serem convertidas no corpus selecionado?

4. Há diferenças na freqüência do processo de conversão entre as palavras mais freqüentes e as que apresentam maior repertório de classes de palavras passíveis de serem convertidas? Em caso afirmativo, o que explica essa diferença?

A fim de responder a essas perguntas, este trabalho está organizado da maneira abaixo.

No capítulo 1, *Fundamentação Teórica*, é fornecido o arcabouço teórico das duas áreas de conhecimento que embasam este trabalho, ou seja, o estudo da Lingüística de Corpus e o estudo da conversão. Para ambas as áreas é traçado um panorama histórico que visa a demonstrar a evolução dos estudos e a inseri-los em

um contexto atual. Esse panorama é fundamentado em Berber Sardinha (2004) e Kennedy (1998) em relação à Lingüística de Corpus e em Adams (1976) e Hummel (2003) em relação à conversão. No mesmo capítulo também é realizado um levantamento sobre a maioria das teorias que definem e delimitam a conversão, onde são discutidas as idéias de Bally (1932, apud Dubois et al 1998), Dokulil (1968), Marchand (1969), Adams (1976), Stein (1977), Görlach (1982), Bauer (1983) e Štekauer (1996) e a Lingüística de Corpus, onde são discutidas as idéias de Berber Sardinha (2004), Hoey (1997), Kennedy (1998) e Sinclair (1991). Nesse capítulo é ainda elaborada uma definição de conversão que balizará o restante deste trabalho.

No capítulo 2, *Metodologia*, são abordados o desenho de uma ferramenta que visa a organizar e classificar os dados analisados e os procedimentos computacionais a fim de filtrar e reorganizar os resultados parciais obtidos.

No capítulo 3, *Apresentação e Discussão dos resultados*, são apresentados e discutidos os resultados decorrentes da análise realizada, de modo a fornecer substrato à elaboração das idéias discutidas nas *Considerações Finais*.

CAPÍTULO 1: FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta as áreas que forneceram embasamento teórico à pesquisa. Primeiramente são apresentados os trabalhos referentes à Lingüística de Corpus, que permeará a visão e abordagem da língua a ser estudada. Essa apresentação ocorrerá por definição dos principais conceitos que a delimitam, tais como corpus, visão probabilística de linguagem e padrões (Berber Sardinha, 2000; Berber Sardinha, 2004; Hoey, 1997; Kennedy, 1998; Sinclair, 1991); elaboração de um panorama histórico (Berber Sardinha, 2004; Kennedy, 1998) e discussão sobre seu status atual (Biber et al, 1998; Kennedy, 1998; McEnery e Wilson, 1996).

A seguir, são elencados os principais temas que englobam a conversão, como, por exemplo, a problemática envolvendo a nomenclatura, a delimitação desse fenômeno e as diferentes definições propostas por estudiosos. E, por fim, é escolhida uma definição que irá permear este trabalho.

1.1 A Lingüística de Corpus

Diferentemente da provável inexistência de uma definição completa sobre Lingüística Aplicada (Celani 1992:15), há um certo consenso e uma delimitação quanto à definição e ao escopo da Lingüística de Corpus. Uma das definições que englobam as principais características da Lingüística de Corpus é proposta por Berber Sardinha (2004: 3):

> "A Lingüística de Corpus ocupa-se da coleta e exploração de corpora, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador".

Ao analisar a definição acima, pode-se questionar, pelo fato da Lingüística de Corpus estar intrinsecamente ligada à coleta e exploração de corpora, se ela não pode ser considerada metodologia[2]. A partir desse ponto, surgem divergências. Uma

[2] Entende-se, neste caso, metodologia como um conjunto de procedimentos a serem seguidos.

corrente, representada por McEnery e Wilson (1996:02), não considera a Lingüística de Corpus uma ramificação da Lingüística, como são a Sociolingüística e a Sintaxe, uma vez que essas disciplinas descrevem/explicam algum aspecto do uso da linguagem, ao passo que a Lingüística de Corpus funciona como uma metodologia que pode ser usada em quase todas as áreas sem se constituir como uma área em si mesma.

Em oposição, Biber et al (1998:9-10) considera que a abordagem (e não metodologia) baseada em corpus pode complementar outras abordagens tradicionais, mas não deve ser vista como a única abordagem correta. Kennedy (1998:07) aponta que a Lingüística de Corpus não deve ser vista como uma teoria de linguagem contrapondo-se a outras como a Gramática Transformacional, mas como uma fonte de evidência que pode servir a qualquer teoria lingüística que aceite a riqueza de dados empíricos, que é uma das características mais marcantes da Lingüística de Corpus.

Este trabalho utilizará como princípio teórico a segunda corrente, formada por Biber et al (1998:9-10) e Kennedy (1998:07), uma vez que não existe uma metodologia única que deve ser adotada a fim de balizar a análise dos dados e, pelo fato de a Lingüística de Corpus estar intrinsecamente ligada a uma visão probabilística de linguagem; ou seja, ao implicar uma visão de linguagem, deixa de ser apenas um conjunto de passos a serem religiosamente seguidos e, desse modo, recebe um status superior. A questão da visão probabilística de linguagem será aprofundada a seguir.

1.1.1 Visão probabilística da linguagem

Para tornar mais clara a concepção de visão probabilística da linguagem é necessário traçar um paralelo entre os conceitos apontados por Halliday e Chomsky. Halliday, lingüista responsável pela criação da Lingüística Sistêmico-Funcional, vê a linguagem como *probabilidade*, ao passo que Chomsky a enxerga essencialmente como *possibilidade* (Kennedy, 1998). Na teoria chomskyiana, o ponto central é a capacidade de um falante nativo elaborar agrupamentos sintáticos que são passíveis de serem utilizados, não importando se façam sentido ou não - para que uma frase seja gramatical, ela tem de respeitar a estruturação sintática da língua. Por outro

lado, a teoria hallidayana procura demonstrar empiricamente quais agrupamentos sintáticos são mais prováveis de ocorrer em um determinado contexto, em uma língua específica. Essa teoria pressupõe que, apesar de muitos traços lingüísticos serem possíveis de serem utilizados, eles não ocorrem com a mesma freqüência e alguns deles não apresentam freqüência superior a zero. Na teoria chomskyiana, a oração *O rato comeu o cachorro* é possível de ser formada; entretanto, do ponto de vista hallidayano, é possível afirmar que essa oração somente ocorreria em um sentido metafórico, o que é corroborado pelo fato de essa oração ter freqüência zero no Banco de Português, maior corpus da língua portuguesa. Portanto, nota-se que o sentido da frase é um fator relevante na determinação da freqüência da palavra ou oração e, como mencionado no início desta seção, a probabilidade de uma palavra ou oração ocorrer é um dos objetos de estudo da Lingüística de Corpus.

Outro exemplo é a diferença de freqüência entre substantivos e as outras classes de palavras. De acordo com Kennedy (1998:103), os substantivos geralmente representam 25% da freqüência de todas as classes de palavras em inglês. Essa diferença é significativa e não aleatória. Assim, a cada quatro palavras, uma provavelmente será um substantivo e outras palavras estarão associadas a esse substantivo para formar agrupamentos maiores. Mais uma vez essas palavras estarão associadas não aleatoriamente, mas de acordo com padrões que são estudados pela Lingüística de Corpus, conforme Berber Sardinha (2000:352):

> "A padronização se evidencia pela recorrência, isto é, uma colocação, coligação ou estrutura, que se repete significativamente, mostra sinais de ser na verdade um *padrão* lexical ou léxico-gramatical. A linguagem forma padrões que apresentam regularidade (se mostram estáveis em momentos distintos, isto é, tem freqüência comparável em corpora distintos) e variação sistemática (correlacionam-se com variedades textuais, genéricas, dialetais etc). Exemplos notáveis da descrição da linguagem por meio da indução de padrões recorrentes são a gramática de verbos (Francis, G. e Hunston, 1996) e de substantivos e adjetivos (Francis, G. e Hunston, 1998) lançadas pelo projeto COBUILD, nas quais se descreve exaustivamente todos os padrões lexicais existentes na língua inglesa."

Segundo Sinclair (1991:170), a colocação é a co-ocorrência de duas ou mais palavras minimamente distantes umas das outras em um texto. De acordo com Hoey (1997), diferentes nomenclaturas são estabelecidas conforme o tipo de co-

ocorrência: se houver uma associação entre itens lexicais, conseqüentemente, haverá uma *colocação*; caso haja uma associação entre itens gramaticais, haverá uma *coligação*; por último, se há uma associação entre uma palavra e conotação (positiva, negativa), então tem-se uma prosódia semântica. A expressão *cachorro-quente* é um exemplo de colocação fixa, pois as duas palavras juntas descrevem um sanduíche e a palavra *cachorro* não pode ser substituída por um sinônimo, como por exemplo *cão*, pois a expressão perderia o sentido, mas as colocações também podem ser menos rígidas, estando presentes nas associações comuns de qualquer palavra, como por exemplo, neste mesmo parágrafo, *associação entre itens*, *pode ser substituída*, *caso haja* e tantas outras.

Por sua vez, o conceito de coligação pode ser exemplificado pela associação de um verbo como *depend* e a classe de preposições. Já o verbo *causar* pode ser considerado um bom exemplo no que concerne à prosódia semântica, visto que a grande maioria dos seus colocados – palavras que o acompanha, têm carga semântica negativa – causar *acidente*, causou *morte*, causou *briga* etc.

Por fim, Sinclair (1991:109) enumera dois princípios colocacionais: o princípio de escolha livre, no qual se abre um leque com diversas escolhas para o falante utilizar, tendo como o único critério a gramaticalidade, por exemplo, tendo como base a estrutura frasal *Sujeito-Verbo-Objeto*. É possível escolher livremente quaisquer palavras para preencher essas categorias sintáticas, podendo gerar frases como *O violino subiu a palavra*, que não possui sentido algum aparente, mas é perfeita do ponto de vista gramatical, pois possui um sujeito, um verbo e um objeto condizente com sua valência. Esse princípio condiz às idéias propostas por Chomsky. Já no segundo princípio, o idiomático, o usuário possui um grande número de frases pré-construídas que constituem escolhas únicas, apesar de essas frases poderem ser analisadas em segmentos menores. Os conceitos de colocação e coligação podem ser englobados nesse segundo princípio, visto que tratam de combinações costumeiras da língua, autorizadas pela convenção do uso. Assim, uma frase como *O violino subiu a palavra* não seria gerada, pois as combinações de palavra nela contida são incomuns na língua em uso.

1.1.2 Definição de Corpus

Um dos mais importantes conceitos na Lingüística de Corpus é o próprio conceito de corpus. Em dicionários como, por exemplo, o Aurélio, é possível encontrar definições de corpus como coleção de textos. Essa definição não representa a abordagem e a visão de língua proposta pela Lingüística de Corpus e difere da definição proposta por Berber Sardinha (2004:18):

> "Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da linguagem ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise."

Berber Sardinha (2004:19) enumera quatro pré-requisitos para a formação de um corpus:

1) O corpus deve ser compilado a partir de textos autênticos, não inventados e tampouco produzidos para fins de pesquisa lingüística.

2) Por autenticidade, entende-se que os textos devem ser produzidos por falantes nativos. A única exceção são os corpora de aprendizes, que podem ser produzidos por falantes não nativos, mas devem receber essa qualificação.

3) O conteúdo do corpus deve ser escolhido cuidadosamente e deve respeitar os critérios 1 e 2. Além disso, o corpus tem de ser fiel ao objetivo do seu criador, ou seja, se o propósito de um corpus escrito é refletir a língua portuguesa no Brasil, a coleta deve ser guiada por um conjunto de critérios que garanta, entre outras coisas, que o corpus contenha o maior número possível de tipos textuais existentes no português, que haja um número aceitável de cada tipo de texto e que a seleção seja aleatória a fim de não contaminar a coleta com variáveis indesejáveis.

4) O corpus deve ser representativo da variedade lingüística a ser estudada. Para que isso ocorra, deve se perguntar o que o corpus representa e para quem ele é representativo, ou seja, não existe uma definição absoluta de corpus representativo.

Traçada a definição de corpus que permeará este trabalho, faz-se necessário traçar um breve panorama histórico da Lingüística de Corpus a fim de facilitar a compreensão de sua evolução.

1.1.3 Breve histórico da Lingüística de Corpus

Apesar de hoje em dia ser praticamente impossível dissociar a Lingüística de Corpus do computador, há registros da existência de corpora (aqui entendido como coleção de textos) na Antiguidade e na Idade Média, por intermédio de compilação de textos bíblicos. Como se pode perceber, os corpora daquela época não tinham como objetivo a análise lingüística.

Segundo Berber Sardinha (2000), esse quadro começou a ser alterado em 1921, com o trabalho pioneiro de Thorndike, que manualmente identificou as palavras mais freqüentes na língua inglesa. Esse levantamento foi feito a partir de um corpus de 4,5 milhões de palavras e, quando publicado, impulsionou mudanças no ensino de língua materna nos Estados Unidos e Europa. Quase vinte e cinco anos depois, em 1945, Thorndike revisou seu trabalho inicial e, analisando um corpus maior – dessa vez 18 milhões de palavras, publicou uma obra listando as 30 mil palavras mais comuns da língua inglesa.

De acordo com Leech et al (1992, apud Berber Sardinha), desde o trabalho pioneiro de Thorndike, houve grandes mudanças tecnológicas e, por conseguinte, com o surgimento do computador, tornou-se mais fácil o manuseio de grandes quantidades de dados. Em 1967, os primeiros estudiosos a utilizar esse novo tipo de tecnologia foram Francis e Kucera, criadores do então chamado corpus Brown, que era composto por 500 textos de diferentes gêneros redigidos em inglês americano. (Leech et al 1992, apud Berber Sardinha 2000:2)

Em 1978, foi compilado um outro corpus eletrônico, Lancaster-Oslo/Bergen, mas dessa vez foi utilizado como base o inglês britânico. Similarmente ao corpus

Brown, foi produzida uma lista de palavras, o que possibilitou comparações entre as listas de freqüência, de distribuição e diferença de vocabulário nos dois corpora.

Um outro passo importante dado com o corpus Lancaster-Oslo/Bergen foi a sua etiquetagem gramatical, ocorrida em 1983. A partir desse momento, por conseguinte, ter-se-ia como resultado não somente uma lista de palavras cujo reconhecimento era creditado com base em seu status ortográfico, mas também uma lista que demonstrava a identidade gramatical das palavras.

Como aponta Berber Sardinha (2000:26), em 1982, um outro corpus, o *Bank of English,* foi compilado, tendo como objetivo a formação de um dicionário, o *Cobuild*. A partir desse projeto vários livros e dicionários foram lançados. Atualmente, o *Bank of English* é o maior corpus da língua inglesa, com aproximadamente 450 milhões de palavras.

Por fim, em 1995, ocorre o lançamento de um outro corpus importante na língua inglesa por sua representatividade e diversificação de textos orais e escritos – o *British National Corpus*, que, por ser o corpus utilizado neste trabalho, será descrito na seção *Metodologia*. Por intermédio da análise do corpus utilizado, será apontada a freqüência do fenômeno da conversão; entretanto, a fim de que haja uma maior precisão, é necessário explicar e delimitar o conceito de conversão, o que será realizado abaixo.

1.2 Conversão

Traçadas as principais características da Lingüística de Corpus e o seu respectivo histórico e status atual, faz-se necessário fornecer o embasamento teórico para o outro pilar deste trabalho – o estudo sobre o fenômeno da conversão.

1.2.1 Breve panorama histórico sobre a conversão

1.2.1.1 A conversão no latim

Muitos estudos, cujo escopo é a evolução histórica da conversão com o passar do tempo, têm como ponto de partida a língua latina, especialmente no seu contraste vulgar e erudito, segundo Adams (1976:39) e Hummel (2003).

Hummel (2003) descreve a formação de advérbios a partir de adjetivos em latim. Havia basicamente dois processos para se formar advérbios, diferentes nas duas variantes do latim. Na variante erudita, o processo normal de formação de advérbios ocorria por meio da sufixação, sendo os sufixos mais utilizados o *–e* e *–iter*, conforme os exemplos a seguir:

longus, longa, longum (adj.) → longe (adv.)

fortis (adj.) → fortiter

Entretanto, havia algumas exceções, como o advérbio *facile*:

facilis (adj.) → facile → (adj. Acusativo Singular Neutro) facile (adv.)

O exemplo acima é um típico caso de conversão ocorrida em latim. Apesar de sua comprovada existência, não se sabe sobre a sua freqüência na linguagem informal, presente no latim vulgar. Tem-se conhecimento de que a conversão era considerada um fenômeno marginal no latim clássico. Com a formação de línguas românicas a partir do latim vulgar, houve a transferência desse processo já presente no latim vulgar para as línguas formadas a partir dele, ou por ele influenciadas.

1.2.1.2 A conversão no *Old English*

Além dos casos de conversão em latim, houve também evidências de conversão no *Old English*, que são descritas a partir de Postnikova (1999). De acordo com a autora, no *Old English* havia uma relação de derivação entre o tema presente-infinitivo dos verbos fracos, ou seja, verbos cuja forma passada é formada por adição de um sufixo dental, ou seja, uma flexão que contenha som de /t/ ou /d/ e o tema[3] de substantivos. Para formar o substantivo completo havia outros processos, como, por exemplo, a sufixação; portanto, havia uma certa "disputa" entre os processos de formação de palavras.

[3] Por tema entende-se a combinação entre radical e vogal temática, que também pode ser o morfema-zero. Exemplo: doce = doc (radical) + e (vogal temática nominal)

Já os verbos fortes[4], isto é, verbos cuja flexão é expressa por alomorfia no tema ou pela mudança da vogal ou da consoante no tema, poderiam ter sua vogal ablativa[5] relacionada a um substantivo derivado (*ridan*/*rad* – a partir do verbo *ride* em inglês, *beran*/*bora* a partir do verbo *bear* em inglês). Entretanto, esse tipo de derivação não era muito produtivo. O tema infinitivo dos verbos fortes do inglês atual se tornou representativo de substantivos verbais assim como o tema correspondente dos verbos fracos (*ride* verbo/*ride* substantivo, *look* verbo/*look* substantivo).

Outro processo de conversão que poderia ocorrer era devido às mudanças fonéticas em pares constituídos de um verbo forte e um substantivo que possuíam o mesmo radical. Com o passar do tempo, as diferenças existentes nesses pares no *Old English* desapareceram até que esses pares se tornassem muito semelhantes no inglês moderno. Alguns exemplos são *drinc*/*drinca (drink)* e *helpan*/*help (help)*.

1.2.1.3 A relação da conversão entre o inglês e as línguas românicas

Nesta seção, discorre-se sobre a discussão proposta por Adams (1976). Ela indica que, em inglês, alguns pares dissílabos de substantivos/verbos convertidos e de origem românica diferem quanto à sílaba tônica. O verbo preserva o padrão de tonicidade presente no francês, isto é, a segunda sílaba é a tônica, ao passo que os substantivos apresentam a primeira sílaba como tônica, como por exemplo, *a cómbine*, *to combíne*, *an ínvite*, *to invíte*, *a próduce, to produce*. Para Adams, a tonicidade demonstrada pelos acentos ilustrativos na primeira sílaba dos substantivos pode sugerir que eles sejam derivados dos verbos, o que nem sempre é verdade, visto que o verbo *segmént* é derivado do substantivo *ségment*. Além disso, há casos de pares dissílabos de substantivos/verbos nos quais não há diferenciação quanto à tonicidade – *cómbat*, *cómment*, *cóntact*, *prócess* têm como sílaba tônica a primeira, ao passo que *debáte*, *concérn*, *rebúke*, *regrét*, *retúrn*

[4] Exemplo de verbo forte: *to bring – brought*

[5] Entende-se por ablativo um caso de declinação que assume a função de vários outros casos; assim, o ablativo latino é, ao mesmo tempo, um ablativo, um instrumental, um comitativo, um agentivo e, muitas vezes, um locativo.

apresentam a segunda sílaba como tônica. Portanto, essa classificação intuitiva para qual palavra surgiu primeiro parece não ser a mais correta.

Visto que a intuição e as indicações formais não conseguem elicitar a palavra que surgiu primeiro em um par convertido, a autora assinala que um caminho possível para determinar a palavra original seriam as evidências históricas. De acordo com ela, mais uma vez ocorrem problemas – o substantivo *worship*, por exemplo, existia no *Old English*, ao passo que os primeiros registros do verbo ocorreram por volta de 1200; entretanto, é possível que ele tenha existido muito antes, de acordo com evidências históricas. Outros exemplos são: o substantivo e o verbo *chase* foram emprestados quase que simultaneamente do francês arcaico (*chace*, *chasser*); palavras como *sting* e *support* verbo e substantivo apresentam um problema que descarta as evidências históricas como fator elucidador de ordem de origem – ambos pares possuem a mesma data de origem: *sting*, *Old English* e *support*, século XIV.

Em face à margem de dúvida quanto a determinação de qual palavra surgiu primeiro em muitos pares convertidos e levando em conta que o corpus utilizado representa a língua utilizada em um período específico, neste trabalho a abordagem diacrônica da conversão será preterida juntamente com qualquer abordagem que determine de qual classe de palavra se originou o termo convertido. Na apresentação dos resultados, o agrupamento e a classificação das classes das palavras convertidas serão colocados em ordem alfabética. Assim, por exemplo, o par *doce*$_{\text{(adjetivo-substantivo)}}$ não significa que a palavra *doce*$_{\text{(adjetivo)}}$ se originou primeiro e tampouco o contrário.

1.2.2 A problemática em definir e delimitar o fenômeno da conversão

Ao fazer um levantamento na literatura que aborda esse fenômeno, é possível verificar que há muita discordância no que concerne a nomenclatura e o escopo do fenômeno comumente chamado de conversão – cada estudioso do assunto propõe uma nomenclatura e, como conseqüência, uma definição diferente. As diferenças em termos de nomenclatura devem-se, em grande parte, ao fato de que cada autor define o processo de conversão em um ou mais níveis de linguagem, isto é, a

conversão pode ser um fenômeno de natureza morfossintática, sintática, semântica ou pode mesclar dois ou até mesmo três desses níveis, conforme será demonstrado abaixo.

O primeiro grupo de autores classifica a conversão como um fenômeno de cunho predominantemente morfossintático. Adams (1976:37), em vez de *conversão*, prefere utilizar *derivação por sufixo-zero*. Ela estabelece essa nomenclatura ao comparar pares contrastivos como (1) *to victimize* (tema[6] formado a partir de substantivo *victim* + sufixo verbal *–ize*: fazer alguém de ou tratar alguém como vítima) e (2) *to beggar* (tema formado a partir de substantivo + nenhum sufixo: fazer com que alguém se torne um mendigo, ou seja, empobreça); (1) *to stabilize* (tema formado a partir de adjetivo + sufixo verbal: fazer com que algo se torne estável) e (2) *to firm* (tema formado a partir de adjetivo + nenhum sufixo: fazer com que algo se torne firme).

Em cada par acima, percebe-se a formação de verbos de duas maneiras (1) por um radical com a adição de um sufixo verbal e (2) por um radical, sem a presença do sufixo verbal visível, ou seja, com um sufixo-zero.

Já o segundo grupo de autores, formado por Görlach, Bauer e Štekauer, define conversão como um processo fortemente relacionado à sintaxe. Görlach (1982, apud Sandmann 1996:91) aponta que:

> "Nos manuais ambos os termos são usados às vezes, sem distinção precisa, para fenômenos semelhantes ou até bem diversos. Aqui faz-se a seguinte distinção: derivação zero: De um signo lingüístico é derivado, sem marca expressional (um morfema derivacional), um novo signo, pertencente a outra classe de palavras. A palavra derivada distingue-se da base por sua distribuição **e** seu paradigma diverso: *the clean shoes*; *cleaner*, *cleanest*, *cleanly – to clean shoes*; (*he*) *cleans*, *cleaned*, (*is*) *cleaning*. Conversão: Um signo é empregado em distribuições típicas de outras classes de palavras, mas ele não compartilha todos os empregos da outra classe de palavras nem a flexão."

Görlach também ressalta que a derivação com morfema-zero é um fenômeno da formação de palavras ao passo que a conversão é um fenômeno da sintaxe e que a distinção entre derivação com morfema-zero e conversão é gradual, podendo

[6] Por tema entende-se a combinação entre radical e vogal temática, que também pode ser o morfema-zero. Exemplo: doce = doc (radical) + e (vogal temática nominal)

a conversão desenvolver-se para a derivação-zero. Para ele, se um signo é empregado em distribuições típicas de outras classes de palavras, mas não compartilha todos os empregos nem a flexão da outra classe de palavras, por exemplo, *government* em *government official*, então existe um fenômeno de sintaxe ou conversão.

Bauer (1983:226) sugere que todas as classes de palavras podem passar pelo processo de conversão, e conseqüentemente, produzir palavras de qualquer classe, particularmente das chamadas abertas (substantivo, verbo, adjetivo e advérbio). Para ela, esse fato pode sugerir que não haja regras para a conversão em inglês, ou seja, a conversão é um processo totalmente livre e qualquer lexema pode passar por esse processo em qualquer das classes de palavras abertas à medida que haja uma necessidade. Bauer provavelmente se refere à necessidade de se criar uma palavra nova. A única restrição apontada por ela é que substantivos derivados raramente são convertidos, particularmente, em verbos, visto que há um bloqueio pelo fato de já existir uma palavra que desempenha essa função na língua. Ela cita como exemplo o substantivo derivado *arrival*, que não será convertido em verbo pelo fato de já existir em inglês o verbo *arrive*.

Bauer (1983:227) afirma que a recorrência da conversão pode ser vista como uma ruptura na concepção de que as classes de palavra são "fechadas", isto é, há sempre uma classificação clara e única para todas as palavras:

> "O fato de a conversão ser um fenômeno tão comum pode ser visto como uma ruptura na distinção entre as classes de palavras em inglês e uma tendência a um sistema no qual haja um conjunto de unidades fechadas como os pronomes e um conjunto aberto de unidades lexicais que possam ser utilizadas de acordo com a necessidade[7]."

Štekauer (1996, apud Grzega 2002:4) vê o processo tradicionalmente chamado de conversão como uma simples reestruturação no nível conceitual da palavra e adota uma posição contrária à teoria do sufixo-zero. Como exemplo, tem-

[7] Em inglês – "The commonness of conversion can possibly be seen as breaking down the distinction between form classes in English and leading to a system where there are closed sets such as pronouns and a single open set of lexical items that can be used as required."

se *e-mail* e *to e-mail*. – o que diferencia *e-mail*$_{(verbo)}$ de *e-mail*$_{(substantivo)}$ é a presença da palavra *to* e não a adição de algum sufixo-zero. De acordo com sua teoria, não é possível explicar esse exemplo no nível formal, uma vez que a teoria do sufixo-zero somente faz sentido quando há sufixos "verdadeiros", como por exemplo *-ize* em *stabilize* e *-ify* em *clarify* com a mesma função. Caso contrário, o sufixo-zero seria considerado um morfema singular e casos como *sheep* no plural deveriam ser interpretados como casos de duplo sufixo-zero, ou seja, (*sheep* + sufixo-zero, representando singular + sufixo-zero representando plural) ou então o sufixo-zero representante do singular seria substituído pelo plural, isto é (*sheep* + sufixo-zero representando plural – nesse caso o sufixo-zero representando o singular desapareceria).

Štekauer aponta outro argumento contra a teoria do sufixo-zero: morfemas derivacionais podem ocorrer como alomorfes[8] (*-er*, *-or*, *-ar* ao formar substantivos que denotam agentes, como por exemplo, *teacher*) ou como morfemas homônimos cujo significado seja diferente (*-er*$_1$ significando "agente" e *–er*$_2$ "instrumento"). Para ele, a primeira opção deve ser rejeitada uma vez que um morfema-zero não pode ser alomorfe do sufixo –er, pois, se fosse, funcionaria como uma unidade significativa paralela em relação a um grande número de outros sufixos.

O terceiro grupo apregoa que o fenômeno da conversão é, ao mesmo tempo, sintático e semântico. Dokulil (1968, apud Valera 2004:21) diferencia conversão e mudança de classe de palavra ao afirmar que a mudança de classe de palavras pode ocorrer de diversas maneiras: nominalização, verbalização, adjetivação e adverbialização. Esses processos ilustram fenômenos sintáticos ou semânticos diferentes de conversão. Segundo o autor, quando há mudança na classe de palavra, as palavras não são convertidas totalmente em uma outra palavra pertencente a uma classe gramatical diferente, mas uma forma de uma palavra é interpretada como se pertencesse a uma outra classe de palavras, ou seja, na definição apresentada por Dokulil, o fenômeno da conversão está associado à formação de uma palavra nova e não à utilização de uma forma de uma palavra em outra classe gramatical.

No quarto grupo, composto por Marchand, Sandmann e Stein, classifica-se o

[8] Entende-se por alomorfia a variação ortográfica de um morfema sem mudança no seu significado. Em *infeliz* e *imutável*, por exemplo, tanto *in-* quanto *i-* indicam negação e são, portanto, alomorfes.

processo de conversão como um fenômeno morfossintático e sintático. Marchand (1969, apud Sandmann 1996:90-91) diferencia conversão de derivação-zero e de derivação por morfema-zero. Para ele, derivação por morfema-zero significa a utilização de uma palavra como determinante em um sintagma, cujo determinado não está expresso foneticamente, mas subentende-se que esteja presente na palavra, devido a uma associação com outros sintagmas nos quais o elemento de conteúdo está presente foneticamente no par contrastivo. Como exemplos, cita os verbos formados a partir de adjetivos *legalize* (*legal* + *ize*), *nationalize* (*national* + *ize*) e *sterilize* (*steri* + *ize*) contrastando com os verbos *clean*, *dirty*, *tidy*, formados a partir da adição do adjetivo ao morfema-zero, que não é expresso foneticamente.

Com relação à derivação-zero e conversão, Marchand (90-91) aponta que muitos estudiosos consideram os dois termos sinônimos, entretanto, para ele, essa afirmação somente é verdadeira se *conversão* denotar a transposição sintática da palavra, como a que ocorre com a palavra *government* em *government job*. Para Marchand, a utilização de um substantivo como um pré-adjunto representa um padrão regular sintático e não tem relação alguma com derivação ou tampouco com formação de palavras. Ele considera derivação somente quando uma palavra muda sua classe de palavra ou classe lexical.

Stein (1977, apud Sandmann 1996:90-91) emprega para os fenômenos que Marchand denomina derivação com morfema-zero o termo conversão e para os termos que Marchand denomina conversão, o termo conversão parcial.

Sandmann (1996) defende a utilização do termo derivação imprópria em vez de conversão, visto que, em sua opinião, bem como na opinião dos gramáticos que cunharam essa nomenclatura, é estranha a obtenção de uma palavra nova sem modificar o significante, como acontece em geral, de um signo lingüístico. Sandmann considera derivação própria a derivação combinatória, ou seja, a que produz mudanças formais na palavra.

Por fim, no quinto grupo, afirma-se que o fenômeno da conversão perpassa os três níveis de linguagem. O principal representante desse grupo é Bally (1965, apud Dubois et al, 1998:178), que prefere deslizamento (ou mudança) de sentido em lugar de conversão e define esse fenômeno como toda a derivação em que o sufixo não for aparente, isto é, o deslizamento de sentido consiste em fazer passar uma palavra para outra categoria sem mudança de forma. Como exemplo, cita a palavra

artista, que pode ser utilizada tanto como substantivo e adjetivo nas frases *O artista é famoso* e *Paulo é muito artista*. Desse modo, a classe gramatical da nova palavra será definida por signos exteriores à mesma. Bally também apregoa que o deslizamento (ou mudança) de sentido é um processo de derivação implícita e formador de palavras, cuja produtividade é alta.

A fim de tornar mais claro os níveis de linguagem abordados por cada autor, foi elaborado a tabela-resumo a seguir.

	Morfossintático	**Sintático**	**Semântico**
Bally	**X**	**X**	**X**
Dokulil		**X**	**X**
Marchand	**X**	**X**[9]	
Adams	**X**		
Stein	**X**	**X**	
Sandmann	**X**	**X**	
Görlach		**X**	
Bauer		**X**[10]	
Štekauer		**X**	

Tabela 01 – Relação entre estudiosos e suas respectivas visões acerca da conversão

Após a apresentação de diversas definições sobre o termo conversão, para fins de estudo adotou-se a seguinte definição de trabalho:

A conversão é um fenômeno lingüístico, freqüente na língua inglesa, que reflete o uso de uma mesma forma ortográfica ou variante muito próxima dela em duas ou mais classes/sub-classes gramaticais diferentes, ou seja, os exemplos *Ele precisa beber*$_{\text{(verbo transitivo)}}$ *água*. e *Aquela mulher adora beber*$_{\text{(verbo intransitivo)}}$. ilustram casos de conversão na mesma classe gramatical, havendo mudança de sub-classe. A conversão é um fenômeno cujos padrões léxico-gramaticais podem ser definidos claramente por intermédio de análise de corpus. Desse modo, operacionalizamos os seguintes conceitos: (1) A classe gramatical é indicada por uma etiqueta

[9] Caso conversão signifique transposição sintática.

morfossintática, que é um código inserido por um etiquetador automático junto a cada forma ortográfica do corpus. Devido à profusão de etiquetas, foi preciso reduzir seu repertório segundo os critérios apresentados no capítulo de metodologia; (2) conversão é toda a ocorrência de mais de uma etiqueta morfossintática para uma mesma forma ortográfica encontrada no corpus de estudo, o *British National Corpus*.

A análise dos padrões léxico-gramaticais é de vital importância, visto que a conversão extrapola o limite da palavra, como, por exemplo, *The rich are happy* – não é possível definir se a palavra *rich* é um adjetivo ou substantivo caso seja analisada isoladamente. A fim de refutar a concepção de que há elipse da palavra *people* em conversões de adjetivo para substantivo, alguns adjetivos gentilícios são convertidos em substantivos pela adição do morfema *s*, tais como American$_{(adjetivo)}$ e Americans$_{(substantivo)}$, Brazilian$_{(adjetivo)}$ e Brazilians$_{(substantivo)}$. É de extrema importância ressaltar que a conversão perpassa três níveis de linguagem – morfossintático, sintático e semântico e que, pelo fato de perpassar o nível semântico, pode haver mudanças no sentido da palavra.

A conversão não deve ser confundida com a mudança sintática de uma palavra, como no exemplo *government job*. Nesse caso, *government* não pode ser considerado adjetivo, pois já há na língua inglesa um adjetivo *governmental* correspondente ao substantivo *government*.

1.2.3 Conversão e homonímia

Um dos problemas enfrentados neste trabalho é a incerteza de ocorrência de conversão ocasionada por homonímia. Valera (2004:23) afirma que algumas palavras apresentam a mesma forma, entretanto significados tão distantes que são classificadas como palavras diferentes e, por esse motivo, são homônimas, como por exemplo, a palavra *leaves* em *All the leaves are brown* e *He usually leaves at 8*. O contraste quanto ao significado é estudado na área de homonímia e polissemia, que explicam palavras diferentes que possuem a mesma forma e uma palavra que apresenta significados diferentes, respectivamente.

[10] Bauer questiona a concepção de que as classes de palavras são fechadas.

Em inglês, bem como em português, a homonímia é um fenômeno presente e não representa conversão. Infelizmente, por lidar com quantidades muito grande de palavras no recorte deste trabalho, não foi possível descartar os casos de homonímia nas ocorrências do corpus.

CAPÍTULO 2: METODOLOGIA

O intuito deste capítulo é expor os principais procedimentos metodológicos adotados na classificação e no agrupamento dos dados coletados a partir do corpus utilizado na pesquisa, o *British National Corpus*. Inicialmente, porém, serão reiterados o objetivo deste trabalho e as perguntas de pesquisa, bem como a forma pela qual ambos orientam os critérios adotados para a elaboração da ferramenta utilizada na organização e classificação dos dados analisados e a metodologia adotada.

2.1 Objetivos e questões de pesquisa

O trabalho teve como objetivos principais delimitar e descrever o processo comumente chamado de conversão, demonstrar, de maneira empírica, sua freqüência para apontar as classes de palavras mais tipicamente convertidas na língua inglesa.

No intuito de operacionalizar os objetivos deste trabalho, são elencadas as seguintes perguntas de pesquisa:

1. Quais são as classes de palavras mais e menos passíveis de serem convertidas na língua inglesa?

2. Qual é a proporção, no corpus selecionado, entre as palavras que estão envolvidas no processo de conversão e as que não são modificadas por esse fenômeno?

3. Quais são as palavras que apresentam maior repertório de classes de palavras passíveis de serem convertidas no corpus selecionado?

4. Há diferenças na freqüência do processo de conversão entre as palavras mais freqüentes e as que apresentam maior repertório de classes de palavras passíveis de serem convertidas? Em caso afirmativo, o que explica essa diferença?

2.2 Descrição do corpus utilizado[11]

Uma vez que os objetivos deste trabalho são delimitar e descrever o processo comumente chamado de conversão, demonstrar, de maneira empírica, sua freqüência e apontar quais classes de palavras e palavras são mais passíveis de serem convertidas na língua inglesa, o corpus escolhido foi o *British National Corpus* pelos seguintes motivos:

1) É um corpus representativo da língua inglesa, (Berber Sardinha, 2004:26), pois possui 97.619.934 milhões de palavras de muitos registros e gêneros, da fala e da escrita.

2) É um corpus etiquetado e sua etiquetagem possui aproximadamente 96% de precisão, o que assegura um alto grau de confiabilidade na utilização e análise. Ressalta-se que o processo de etiquetagem é extremamente importante, pois, uma vez etiquetado, o corpus pode ser estudado morfo e sintaticamente, o que é primordial para que os objetivos deste trabalho sejam atingidos.

3) É um corpus acessível, ou seja, qualquer pessoa pode adquiri-lo por 50 libras esterlinas .

O *British National Corpus* (BNC) é a compilação de 97.619.934 milhões de palavras obtidas a partir da língua escrita e falada. As amostras de língua foram obtidas de uma grande variedade de fontes, reunidas de modo a mostrar um recorte abrangente do inglês britânico falado e escrito do final do século XX.

A parte escrita do BNC (90%) inclui, por exemplo, trechos de jornais regionais e nacionais, periódicos técnicos, livros acadêmicos e de ficção, cartas publicadas ou não, memorandos, ensaios de escolas de universidades, entre muitos outros tipos de texto.

A parte falada (10%) inclui uma grande quantidade de conversas informais, gravadas por voluntários selecionados de diferentes idades, localidades, classes

[11] Informações retiradas do site www.comp.lancs.ac.uk

sociais de maneira demograficamente balanceada, juntamente com a língua coletada em diferentes contextos, desde encontros governamentais ou formais até programas de rádio e telefonemas.

A compilação do corpus começou em 1991 e terminou em 1994. Nenhum texto foi adicionado após a finalização do projeto; entretanto, o corpus foi minimamente revisado antes do lançamento da segunda edição: *BNC World*. Desde a finalização do projeto, dois sub-corpora com material do *BNC* foram lançados separadamente: o *BNC Sampler* (uma compilação a partir de um milhão de palavras escritas e um milhão de palavras faladas do corpus) e o *BNC Baby* (quatro amostras de um milhão de palavras de quatro diferentes gêneros).

Algumas características relevantes do BNC:

Monolíngue – O corpus engloba o inglês britânico moderno e não outras línguas usadas na Grã-Bretanha, escritas ou ditas por pessoas reconhecidas como falantes nativos do inglês, embora haja algumas poucas exceções.

Sincrônico – A amostragem que constitui o corpus inclui textos e transcrições de fala obtidas no final dos anos de 1980 e início dos anos de 1990.

Amplo – Inclui muitos estilos e variedades diferentes e não é limitado a uma área particular de assunto, gênero ou registro. Em particular, contém exemplos da língua falada bem como da escrita.

2.2.1 O *British National Corpus* em números

O BNC contém 4054 textos e, em termos de espaço em disco rígido, ocupa aproximadamente 1.5 Gigabytes. No total, engloba 100.476.090 palavras, sendo que o número de palavras etiquetadas é ligeiramente menor – 97.619.934. A distribuição da modalidade dos textos (oral, escrita) está explicada na tabela a seguir:

Classificação dos textos	Número de textos	Tamanho (Kbytes)	Porcentagem em relação ao corpus
Falados em contextos não - governamentais	153	4206058	10.08
Falados em contextos governamentais	757	6135671	7.07
Total falado	910	10341729	17.78
Livros escritos e periódicos	2688	78580018	72.75
Escritos para serem falados	35	1324480	1.98
Escritos diversos	421	7373707	8.09
Total escrito	3144	87278205	82.82

Tabela 02 – Composição do *British National Corpus*

2.2.2 O processo de etiquetagem do corpus

O software de etiquetagem utilizado no BNC, o CLAWS (*the Constituent Likelihood Automatic Word-tagging System*), vem sendo aprimorado desde 1980. O CLAWS atingiu 96-97% de precisão (o grau exato varia de acordo com o tipo de texto), o que é padrão para etiquetadores morfossintáticos. Considerando as maiores categorias, o sistema tem uma taxa de erro de somente 1.5%.[12]

Diversos conjuntos de etiquetas foram utilizados ao longo dos anos. O CLAWS1 tem basicamente 132 etiquetas, muitas idênticas em forma e aplicação no corpus Brown, o corpus pioneiro de inglês americano (cf. seção 1.1.3). Uma revisão do CLAWS em Lancaster em 1983 resultou em um conjunto novo e totalmente revisado de 166 etiquetas, conhecido como CLAWS2.

O conjunto de etiquetas utilizado no BNC, o CLAWS5, apresenta somente cerca de 60 etiquetas. O número pequeno em relação aos outros CLAWS se justifica pelo fato de o BNC lidar com um número muito mais elevado de palavras. Para o BNC Sampler (2 milhões de palavras), o conjunto de etiquetas, o CLAWS6, foi enriquecido e apresenta mais de 160 etiquetas.

[12] Análises mais detalhadas sobre as taxas de erros estão disponíveis em www.comp.lancs.ac.uk/ucrel/bnc2/bnc2error.htm

Como se percebe, há uma grande variação no número de etiquetas inserida pelo etiquetador de acordo com a versão do programa. Qualquer um desses conjuntos extrapola em muito o que a maioria das pessoas conhece por classe gramatical, pois para a maioria, haveria cerca de 10 classes gramaticais no inglês (substantivo, adjetivo, verbo, advérbio, numeral, interjeição, artigo, preposição, pronome e conjunção). Os conjuntos de etiquetas do CLAWS ultrapassam em muito esse numero porque (1) etiquetam outros elementos além da classe gramatical, como pontuação e (2) especificam cada uma das possibilidades das classes, atribuindo uma etiqueta diferente para cada caso. Por exemplo, para adjetivo, o CLAWS5 usa três etiquetas *ajo*, *ajc*, *ajs*. Como queríamos trabalhar com as grandes classes gramaticais, tivemos de condensar essa profusão de classes, de tal modo que reuníssemos *aj0*, *ajc*, *ajs* em *adjetivo*. Assim, foi importante que trabalhássemos com um conjunto de etiquetas que fosse o menos extravagante possível, dentre os disponíveis pela convenção do CLAWS. No caso, o conjunto menos extravagante era CLAWS5, com 60 etiquetas, que felizmente era aquele já incorporado ao BNC.

O atual padrão de conjuntos de etiquetas é o CLAWS7, que é similar ao CLAWS6, diferindo apenas nas etiquetas de pontuação. No CLAWS6, todas essas etiquetas de pontuação começam com a letra “Y”.

O conjunto de etiquetas utilizado no BNC e, por conseguinte, neste trabalho, o CLAWS5, pode ser visualizado abaixo:

ETIQUETA	DEFINIÇÃO	EXEMPLOS	CLASSIFICAÇÃO
AJ0	adjetivo não flexionado	good, old	adjetivo
AJC	adjetivo comparativo	better, older	adjetivo
AJS	adjetivo superlativo	best, oldest	adjetivo
AT0	artigo	the, a, an	artigo
AV0	advérbio	often, well, longer, furthest	advérbio
AVP	partícula adverbial	up, off, out	partícula adverbial
AVQ	advérbio iniciado pro wh- ou *how*	when, how, why	advérbio

ETIQUETA	DEFINIÇÃO	EXEMPLOS	CLASSIFICAÇÃO
CJC	conjunção coordenativa	and, or	conjunção
CJS	conjunção subordinativa	although, when	conjunção
CJT	conjunção *que*	that	conjunção
CRD	numeral cardinal exceto *one*	3, fifty-five, 6609	numeral
DPS	pronome possessivo	your, their	pronome
DT0	pronome	these, some	pronome
DTQ	pronome relativo	whose, which	pronome
EX0	pronome existencial *there*		pronome
ITJ	interjeição	oh, yes, mhm	interjeição
NN0	substantivo neutro em relação ao número	aircraft, data	substantivo
NN1	Substantivo singular	pencil, goose	substantivo
NN2	substantivo plural	pencils, geese	substantivo
NNN	substantivo numeral, neutro em relação ao número	dozen, hundred	substantivo
NNN	substantivo numeral no plural	hundreds, thousands	substantivo
NNS	substantivo de tratamento	president, governments, Messrs.	substantivo
NP0	substantivo próprio	London, Michael, Mars	substantivo
NUL	etiqueta nula para itens que não devem ser etiquetados		-
ORD	numeral ordinal	sixth, 77th, last	numeral
PNI	pronome indefinido	none, everything	pronome
PNP	pronome pessoal	you, them, ours	pronome
PNQ	pronome iniciado por -wh	who, whoever	pronome

ETIQUETA	DEFINIÇÃO	EXEMPLOS	CLASSIFICAÇÃO
PNX	pronome reflexivo	itself, ourselves	pronome
POS	morfema possessivo *'s* ou *'*		-
PRF	preposição *of*		preposição
PRP	preposição exceto *of*	for, above, to	preposição
PUL	pontuação, parênteses ou colchetes esquerdo	(, [	-
PUQ	pontuação relativa a citações	" '	-
PUR	pontuação, parênteses ou colchetes direito	),]	-
TO0	marcardor de infinitivo *to*		marcador de infinitivo
UNC	itens não classificados que não pertencem ao léxico inglês		-
VBB	formas bases do verbo *to be* exceto o infinitivo	am, are	verbo
VBD	formas do passado do verbo *to be*	was, were	verbo
VBG	forma do verbo *to be* no gerúndio	being	verbo
VBI	forma infinitiva do verbo *to be*		verbo
VBN	particípio passado do verbo *to be*	been	verbo
VBZ	formas terminadas em -s do verbo *to be*	is ou 's	verbo
VDB	formas bases do verbo *do* exceto o infinitivo	do	verbo
VDD	formas do passado do verbo *do*	did	verbo
VDG	forma do verbo *do* no gerúndio	doing	verbo
VDI	forma infinitiva do verbo *do*		verbo
VDN	particípio passado do verbo *do*	done	verbo

ETIQUETA	DEFINIÇÃO	EXEMPLOS	CLASSIFICAÇÃO
VDZ	forma terminada em -s do verbo *do*	does	verbo
VHD	formas do passado do verbo *have*	had ou 'd	verbo
VHG	forma do verbo *have* no gerúndio	having	verbo
VHI	forma infinitiva do verbo *have*		verbo
VHN	particípio passado do verbo *have*	had	verbo
VHZ	formas terminadas em -s do verbo *have*	has ou 's	verbo
VM0	verbo modal auxiliar	Can, could, will ou 'll	verbo
VVB	forma base de verbos lexicais	take, live	verbo
VVD	formas do passado de verbos lexicais	took, lived	verbo
VVG	forma de verbos lexicais no gerúndio	taking, living	verbo
VVI	forma infinitiva de verbos lexicais	take, live	verbo
VVN	particípio passado de verbos lexicais	taken, lived	verbo
VVZ	forma terminada em -s de verbos lexicais	takes, lives	verbo
XX0	Advérbio negativo *not* ou *n't*		advérbio
ZZ0	símbolo alfabético	A, B, C, D	-

Tabela 03 – Relação entre as etiquetas e classes de palavras

2.3 Procedimentos metodológicos

A fim de obter, a partir do *British National Corpus*, os dois recortes necessários para análise, foi preciso adotar dois conjuntos de procedimentos metodológicos. O primeiro tem como escopo as 1000 palavras com maior número de etiquetas no corpus estudado. Já o segundo tem como objeto as 1000 palavras mais freqüentes no BNC.

Em primeiro lugar, utilizando um programa escrito em Shell Script pelo professor Tony Berber Sardinha, obteve-se um recorte com as 1000 palavras com maior número de etiquetas. Para ilustração, no quadro abaixo, reproduzimos as primeiras 151 palavras com mais etiquetas no *British National Corpus*, obtidas a partir desse recorte.

Nº de etiquetas	Palavra	Nº de etiquetas	Palavra	Nº de etiquetas	Palavra
14	LIKE	9	UPFRONT	9	PLUMB
13	SET	9	TYPECAST	9	PLACING
13	FORECAST	9	TREKKING	9	PET
13	CUT	9	TRADING	9	PASSING
13	A	9	TOURING	9	PARALLEL
12	SPLIT	9	TOP	9	PANTING
12	ROUND	9	THRUST	9	PACKAGING
12	BURST	9	THAT	9	OW
12	BOUND	9	TESTING	9	OUT
11	UPSET	9	THRUST	9	OL
11	UP	9	THAT	9	OFFERING
11	SQUARE	9	TESTING	9	OBTAINING
11	POP	9	TA	9	NT
11	OPENCAST	9	SWIMMING	9	NO
11	LEFT	9	SUPPORTING	9	NESTING
11	LAY	9	SUFFERING	9	MOTORING
11	HURT	9	STYLING	9	MONITORING
11	HIT	9	STRAIN	9	MODELLING
11	FELL	9	STOPPING	9	MESS
11	CROSS	9	STOCK	9	MATCHING
11	CAST	9	STAGING	9	MAT
11	AI	9	SPORTING	9	MANUFACTURING
10	WET	9	SPLITTING	9	MAILING
10	SOUND	9	SPIT	9	LOVING
10	SHED	9	SPENDING	9	LOGGING
10	S	9	SMOKING	9	LISTING
10	OVER	9	SLIT	9	LISTENING
10	OUTCAST	9	SKIING	9	LINKING
10	OPEN	9	SKETCHING	9	LING
10	OFFSET	9	SIGNING	9	LEVEL
10	LIGHT	9	SHOOTING	9	LET
10	GROUND	9	SAMPLING	9	LAPWING

10	FUCKING	9	SAILING	9	LABOURING
10	DO	9	RIDING	9	KILLING
10	DAT	9	REVIEWING	9	JOGGING
10	CRAP	9	RETURNING	9	INTRODUCING
10	CLOSE	9	RESTRUCTURING	9	IMAGING
10	CHANGING	9	RESTING	9	I'
10	BROADCAST	9	REPORTING	9	HEALING
10	BET	9	RECYCLING	9	HEAD
10	BACK	9	READ	9	HANDLING
9	VE	9	PRESENT		

Tabela 04 – Palavras do BNC com maior número de etiquetas

Foi também necessário selecionar, a partir do British National Corpus, as 100 palavras mais freqüentes, que podem ser visualizadas abaixo.

Posição	Palavra	Freq.	%	Posição	Palavra	Freq.	%
1	THE	6.197.320	6,16	**51**	MORE	214.884	0,21
2	OF	3.112.520	3,09	**52**	CAN	214.511	0,21
3	AND	2.692.822	2,68	**53**	OUT	210.479	0,21
4	TO	2.671.334	2,66	**54**	SAID	208.283	0,21
5	A	2.238.218	2,22	**55**	WHO	198.542	0,2
6	IN	1.993.249	1,98	**56**	ABOUT	198.402	0,2
7	THAT	1.078.710	1,07	**57**	DO	183.249	0,18
8	IS	986.173	0,98	**58**	THEM	172.979	0,17
9	IT	951.433	0,95	**59**	SOME	171.309	0,17
10	WAS	901.242	0,9	**60**	TIME	169.053	0,17
11	FOR	900.148	0,89	**61**	HIM	164.711	0,16
12	I	757.695	0,75	**62**	TWO	164.645	0,16
13	ON	753.574	0,75	**63**	INTO	163.635	0,16
14	WITH	677.928	0,67	**64**	ITS	163.353	0,16
15	AS	673.181	0,67	**65**	THEN	160.391	0,16
16	BE	665.935	0,66	**66**	LIKE	157.514	0,16
17	HE	629.427	0,63	**67**	OTHER	156.181	0,16
18	YOU	610.551	0,61	**68**	WELL	154.654	0,15
19	AT	539.395	0,54	**69**	ONLY	153.227	0,15
20	BY	524.610	0,52	**70**	MY	152.310	0,15
21	ARE	464.595	0,46	**71**	COULD	150.368	0,15
22	THIS	463.066	0,46	**72**	THAN	148.329	0,15
23	BUT	460.749	0,46	**73**	NOW	143.971	0,14
24	HAVE	460.223	0,46	**74**	OVER	140.175	0,14
25	NOT	444.186	0,44	**75**	ME	138.856	0,14
26	FROM	436.106	0,43	**76**	YOUR	138.401	0,14
27	HAD	435.867	0,43	**77**	JUST	130.610	0,13

28	HIS	433.515	0,43	**78**	IT'S	129.469	0,13
29	THEY	387.146	0,38	**79**	MAY	129.163	0,13
30	OR	376.088	0,37	**80**	NEW	126.929	0,13
31	WHICH	372.249	0,37	**81**	FIRST	126.923	0,13
32	SHE	349.087	0,35	**82**	THESE	125.508	0,12
33	AN	345.645	0,34	**83**	ALSO	125.019	0,12
34	HER	326.655	0,32	**84**	ANY	124.838	0,12
35	WERE	317.988	0,32	**85**	KNOW	123.006	0,12
36	ONE	308.207	0,31	**86**	VERY	122.970	0,12
37	WE	306.572	0,3	**87**	SEE	118.853	0,12
38	THERE	294.236	0,29	**88**	PEOPLE	118.526	0,12
39	ALL	291.124	0,29	**89**	AFTER	117.792	0,12
40	BEEN	268.515	0,27	**90**	SUCH	109.333	0,11
41	IF	261.423	0,26	**91**	SHOULD	106.959	0,11
42	THEIR	260.946	0,26	**92**	WHERE	106.015	0,11
43	HAS	255.287	0,25	**93**	BACK	105.105	0,1
44	WILL	254.967	0,25	**94**	BECAUSE	102.830	0,1
45	SO	250.384	0,25	**95**	HOW	101.772	0,1
46	NO	239.348	0,24	**96**	WAY	100.719	0,1
47	WOULD	238.777	0,24	**97**	MOST	100.086	0,1
48	WHAT	233.577	0,23	**98**	GET	99.082	0,1
49	UP	229.514	0,23	**99**	DOWN	98.231	0,1
50	WHEN	216.363	0,22	**100**	DON'T	96.963	0,1

Tabela 05 - 1000 palavras mais freqüentes no BNC

De posse das listas com as 1000 palavras mais freqüentes e as 1000 palavras que possuem duas ou mais etiquetas, era necessário agrupar e somar a freqüência de todas etiquetas. Devido à grande quantidade dados, não foi possível somar as freqüências à mão (com calculadora). Parecia inevitável usar um programa de planilha eletrônica, como o Excel ®. Entretanto, a planilha não deu conta dessa tarefa, pois nesse momento, surgiram quatro pontos a serem analisados:

1. O programa Excel não soma números que estão agrupados com letras, ou seja, a simples soma das etiquetas *2pnp* + *10pnp* = *12pnp* não era possível de ser realizada pelo Excel.

2. O programa Excel possui limite de aproximadamente 70.000 linhas. O corpus, se copiado para esse programa, extrapolaria o limite de linhas em muitas vezes.

3. A abordagem manual de muitos dados provenientes do corpus, isto é, manualmente copiando e colando dados do corpus para o Excel, poderia aumentar a margem de erro.

4. O uso do programa Excel possivelmente propiciaria uma análise de cunho predominante ou unicamente quantitativo.

Em face ao quadro descrito acima, em conjunto com o meu professor orientador, passei a desenvolver uma ferramenta específica para identificar possíveis casos de conversão no BNC. Depois de seis meses de desenvolvimento e três meses de teste, essa ferramenta ficou conhecida por *Identificador de conversão no BNC* e está disponível online em www2.laelpucsp.br/corpora/conversão/index.html. Com isso, pretendemos também dar uma contribuição a outros pesquisadores que desejem replicar nosso estudo, para verificar sua exatidão, como também auxiliar a comunidade de pesquisadores e interessados no assunto da conversão e na Lingüística de Corpus.

1.1.1. Lista de itens a serem buscados (em inglês)

Digitar uma palavra ou etiqueta por linha (quantas quiser).

Etiquetas para desconsiderar (quantas quiser).

Freqüência mínima (somatória das duas colunas; digite 0 para não desconsiderar nenhuma):

Fazer (clique uma vez apenas) Reset

Conceito e testagem: Roberto Diniz; Programação: Tony Berber Sardinha (2005).

Figura 01 - Página inicial do identificador da conversão, disponível em http://www2.lael.pucsp.br/corpora/conversao/index.html

No primeiro campo, devem ser digitadas as palavras ou etiquetas a serem buscadas pela ferramenta no corpus a ser analisado, o *BNC*. Cada palavra deve ser digitada embaixo da outra e não ao lado. Não há limite quanto ao número de palavras a serem colocadas nesse campo.

O segundo campo, cujo preenchimento é opcional, tem a função de filtrar etiquetas não pertinentes à análise. Por exemplo, se uma palavra for classificada como apenas letra do alfabeto (etiqueta *zz0*), ela não será pertinente à análise dos casos de conversão. Esse campo é útil também para restringir os casos de conversão existentes em apenas duas ou mais classes de palavras; por exemplo, se o objetivo do trabalho for estudar as conversões existentes apenas em adjetivos e substantivos, basta digitar os códigos das outras etiquetas que não correspondem a substantivos e adjetivos, que elas serão ignoradas.

As etiquetas desconsideradas neste trabalho e as respectivas justificativas de exclusão estão representadas na tabela abaixo:

nul	etiqueta nula para itens não considerados válidos pelo etiquetador como, por exemplo, símbolos
pos	Morfema possessivo *'s* ou *'*
pul	pontuação, parênteses ou colchetes esquerdo
pun	pontuação geral
puq	pontuação relativa a citações
pur	pontuação, parênteses ou colchetes direito
to0	marcardor de infinitivo *to*
unc	itens não classificados que não pertencem ao léxico inglês
zz0	símbolo alfabético

Tabela 06 – Etiquetas desconsideradas neste trabalho

Por fim, o campo três tem a função de filtrar as etiquetas pela freqüência, ou seja, preenchendo esse campo com um valor, o programa não exibirá as etiquetas que possuam freqüência menor do que o valor estipulado. Esse campo tem também como objetivo filtrar as palavras que tenham sido etiquetadas erroneamente, uma vez que, em alguns casos, uma freqüência baixa pode ser ocasionada por um erro de etiquetagem. Como exemplo, a palavra *window*$_{(verbo)}$ possui uma ocorrência no

corpus, mas o dicionário não apresenta *window* como verbo. Caso o campo três fosse preenchido, essa ocorrência não apareceria nos resultados. De qualquer modo, é interessante ressaltar que a etiquetagem possui 96% de precisão.

O valor digitado filtrará a somatória total das etiquetas. Algumas palavras apresentam grau tão elevado de ambigüidade quanto à classificação de classe gramatical, que o etiquetador as classifica como pertencente a duas classes de palavras, isto é, essas palavras apresentam etiquetas duplas ou *portmanteau tags*. Por exemplo, a palavra *like* pode ser preposição ou conjunção (Swan 1980: 73).

No exemplo abaixo, utilizei a palavra *like*, desprezei as etiquetas *unc* e *zz0* e estabeleci como mínimo aceitável as ocorrências iguais ou superiores a 200.

2. Identificador de conversão -- CEPRIL, LAEL, PUC/SP

Tue Jan 3 00:13:58 BRST 2006

Resultados

1. 1 palavra(s) foram pesquisadas
    1. Lista de palavras simples.
    2. Lista de palavras completa.
    3. Lista de palavras completa, filtrada.
2. O valor de corte foi 200.
3. A(s) etiqueta(s) ignorada(s) foi/foram: zz0 unc.
4. A soma da primeira coluna numérica é 11840.
5. A soma da segunda coluna numérica é 151590.
6. O total das duas colunas juntas é 163430.
7. Conjuntos de etiquetas originais
8. Conjuntos de etiquetas desmembradas
9. Conjunto de etiquetas duplas
10. Conjunto de etiquetas no formato de seqüência
11. Somatória de etiquetas

Figura 02 - Tela de apresentação dos resultados gerais

A figura 02 é a tela de apresentação dos resultados. Ela é formada por onze itens que serão explicados detalhadamente a seguir.

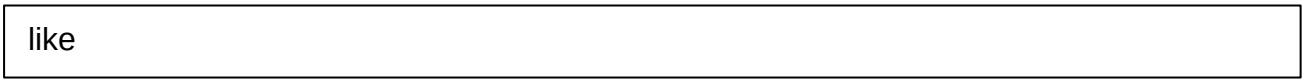


Figura 03 - Tela de apresentação da lista de palavras simples

A figura 03 tem como função apresentar as palavras que foram digitadas e encontradas no corpus de acordo com os critérios estabelecidos na figura 01. No exemplo utilizado, a palavra *like* correspondeu aos critérios estabelecidos (freqüência superior a 200 e etiquetas diferentes de *unc* e *zz0*) e por isso foi exibida nessa tela.

Palavra	Número de etiquetas diferentes	Etiquetas	Número de ocorrências no corpus falado	Número de ocorrências no corpus escrito
Like	14	aj0	1342	2635
Like	14	aj0-av0	667	1345
Like	14	aj0-nn1	543	810
Like	14	av0	532	3318
Like	14	cjs	327	690
Like	14	cjs-prp	984	2204
Like	14	nn1	44	45
Like	14	nn1-vvb	8	9
Like	14	np0	16	17
Like	14	pnp	1	1
Like	14	prp	3680	108988
Like	14	unc	4	6
Like	14	vvb	2074	10115
Like	14	vvi	2788	23664

Figura 04 - Tela de apresentação da lista de palavras completa

Na figura 04 são exibidos respectivamente a palavra, o número de etiquetas, as etiquetas, o número de ocorrências no corpus falado e no escrito. A tela acima, diferentemente da tela seguinte, não é apresentada a listra filtrada, ou seja, nesse

caso, a ferramenta exibe as etiquetas *zz0* e *unc* e as que possuem frequência inferior a 200. Observando os resultados, pela etiquetagem é possível ter uma visão geral de todas as classes de palavras e também notar casos de conversão e homonímia, quando houver.

Palavra	Número de etiquetas diferentes	Etiquetas	Número de ocorrências no corpus falado	Número de ocorrências no corpus escrito
Like	14	aj0	1342	2635
Like	14	av0	532	3318
Like	14	cjs	327	690
Like	14	prp	3680	108988
Like	14	vvb	2074	10115
Like	14	vvi	2788	23664
Like	14	aj0-av0	333.5	672.5
Like	14	aj0-nn1	271.5	405.0
Like	14	cjs-prp	492.0	1102.0

Figura 05 - Tela de apresentação da lista de palavras completa, filtrada

Diferentemente da tela de apresentação da lista de palavras completas, na figura 05 são exibidas as etiquetas que passaram pelo filtro determinado previamente, neste caso, freqüência de 200 e eliminação das etiquetas *zz0* e *unc*. Como se pode perceber, houve uma alteração na freqüência das etiquetas duplas – comparando a etiqueta dupla *like 14 cjs-prp 984 2204*, retirada da sexta linha da figura 4, e a mesma etiqueta dupla na figura 5, *like 14 cjs-prp 492.0 1102.0*; nota-se que os valores da freqüência relativas às variantes escritas e faladas foram divididas aritmeticamente na figura 5. A existência de etiquetas duplas suscitou providências especiais. Não se pode simplesmente desprezá-las, pelo risco de se perder dados. Uma etiqueta dupla pode representar 100% de freqüência da primeira classe de palavras envolvida, ou 100% da segunda, ou um valor intermediário entre as duas. Diante dessa ambigüidade entre classes, adotou-se a divisão aritmética entre a freqüência total das duas classes e conseqüentemente creditou-se metade da freqüência para cada uma.

No exemplo apresentado, *like 14 cjs-prp 984 2204*, retirado da figura 04, é possível que todas as 3188 ocorrências (984 faladas + 2204 escritas) da palavra sejam de adjetivos, ou todas as 3188 de preposições, ou um valor intermediário

entre os dois, o que é mais plausível. Como não se sabe de fato a freqüência de adjetivos e de preposições da forma *like* a partir do que o etiquetador informa, foi utilizada a média aritmética dos dois valores, ou seja, *like* classificado como conjunção apresenta 1594 ocorrências, 482 provenientes do corpus escrito e 1102 do falado; *like* classificado como preposição apresenta 1594 ocorrências, 482 provenientes do corpus escrito e 1102 do falado.

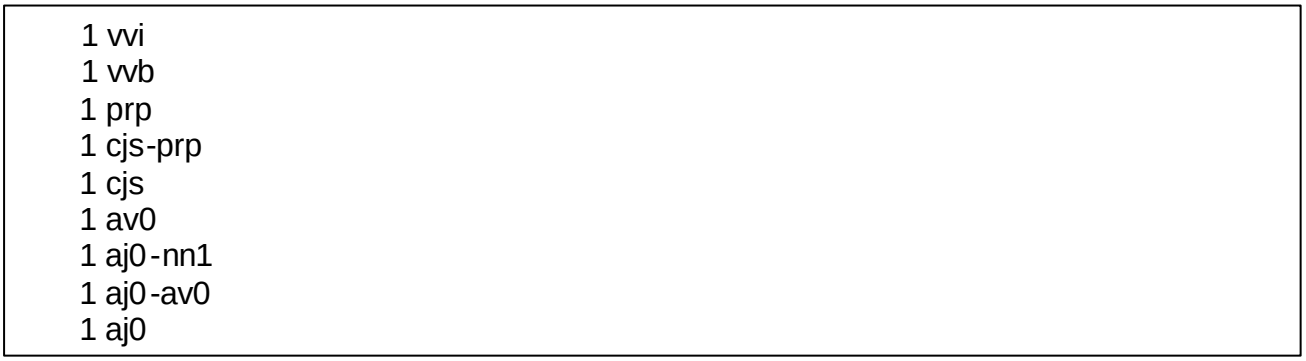


Figura 06 - Tela de apresentação dos conjuntos de etiquetas originais

Na figura 06 são exibidas somente as etiquetas encontradas que satisfazem os critérios escolhidos anteriormente, na tela 01. Observando a figura 06, é possível, visualizar as etiquetas e conseqüentemente as classes de palavras envolvidas no processo de conversão de uma maneira mais resumida e direta, uma vez que são exibidas diferentes classes de palavras que uma mesma palavra possui. Na tela de apresentação de etiquetas originais são exibidas as *portmanteau tags*, ou seja, as etiquetas duplas.

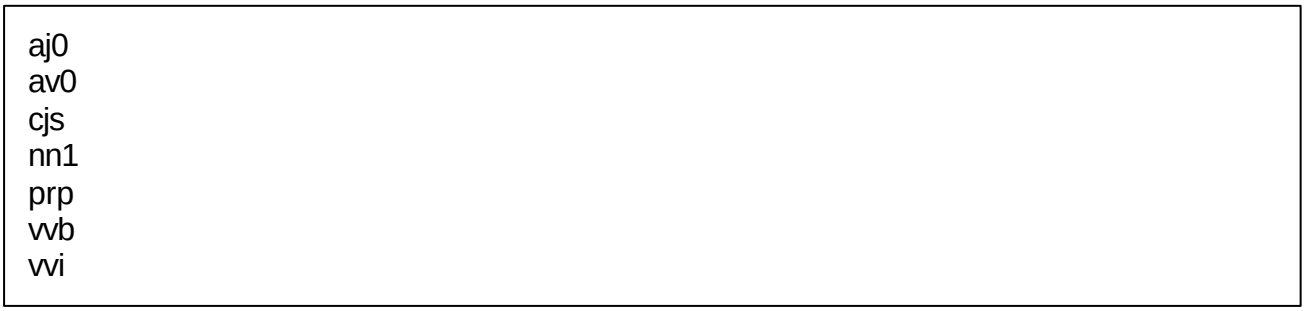


Figura 07 - Tela de apresentação dos conjuntos de etiquetas desmembradas

Diferente da figura 06, na figura 07 são somente exibidas as classes de

palavras envolvidas no processo de conversão na(s) palavra(s) pesquisada(s). Sua função é tornar ainda mais fácil a visualização das classes envolvidas no processo de conversão, independentemente das classes serem oriundas de etiquetas duplas.

Figura 08 - Tela de apresentação dos conjuntos de etiquetas duplas

A figura 08 permite, de uma forma direta, a visualização das etiquetas duplas. A partir dessa primeira observação, é possível apontar quais classes de palavras apresentam ambigüidade em relação à sua etiquetagem e conseqüente classificação morfológica.

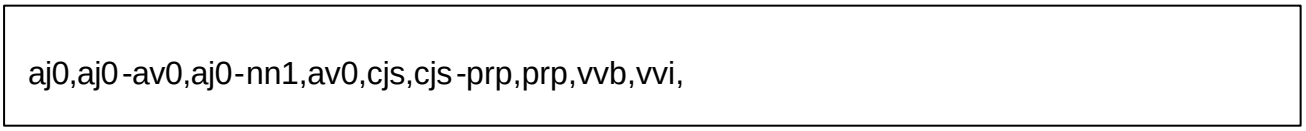


Figura 09 - Tela de apresentação dos conjuntos de etiquetas no formato de seqüência

A figura 09 tem como objetivo exibir, de uma maneira holística e seqüencial, todas as etiquetas que podem classificar morfologicamente uma determinada palavra. Ela pode ser utilizada para verificar se duas palavras com a mesma classificação morfológica apresentam a possibilidade de serem convertidas pelas mesmas classes de palavras. Como exemplo, as palavras *book*[13] e *report* apresentam os mesmos conjuntos de etiquetas no formato de seqüência, isto é, elas são convertidas pelas mesmas classes de palavras. Para verificar se isso ocorre, basta comparar os resultados obtidos.

[13] A palavra *book*, bem como a palavra *report*, apresenta *nn1,nn1-vvb,np0,vvb,vvi*, como conjunto de etiquetas no formato de seqüência.

aj0 5659.5
av0 4856
cjs 2611
nn1 676.5
prp 114262
vvb 12189
vvi 26452

Figura 10 - Tela de apresentação da somatória de etiquetas

Nessa tela é exibida a somatória total de todas as etiquetas que satisfizeram os critérios estabelecidos na figura 01. Essa somatória já contabiliza a divisão das etiquetas duplas. Por meio dos resultados finais obtidos, pode-se apontar a freqüência relativa às classes de palavras convertidas em uma determinada palavra.

Obtidos a partir da ferramenta, os resultados foram colados no Excel a fim de que se obtivesse a freqüência total de cada classe de palavra. Para realizar essa operação foi necessário adotar os procedimentos a seguir.

Em primeiro lugar, clicar em *classificar* para que as etiquetas iguais permanecessem próximas.

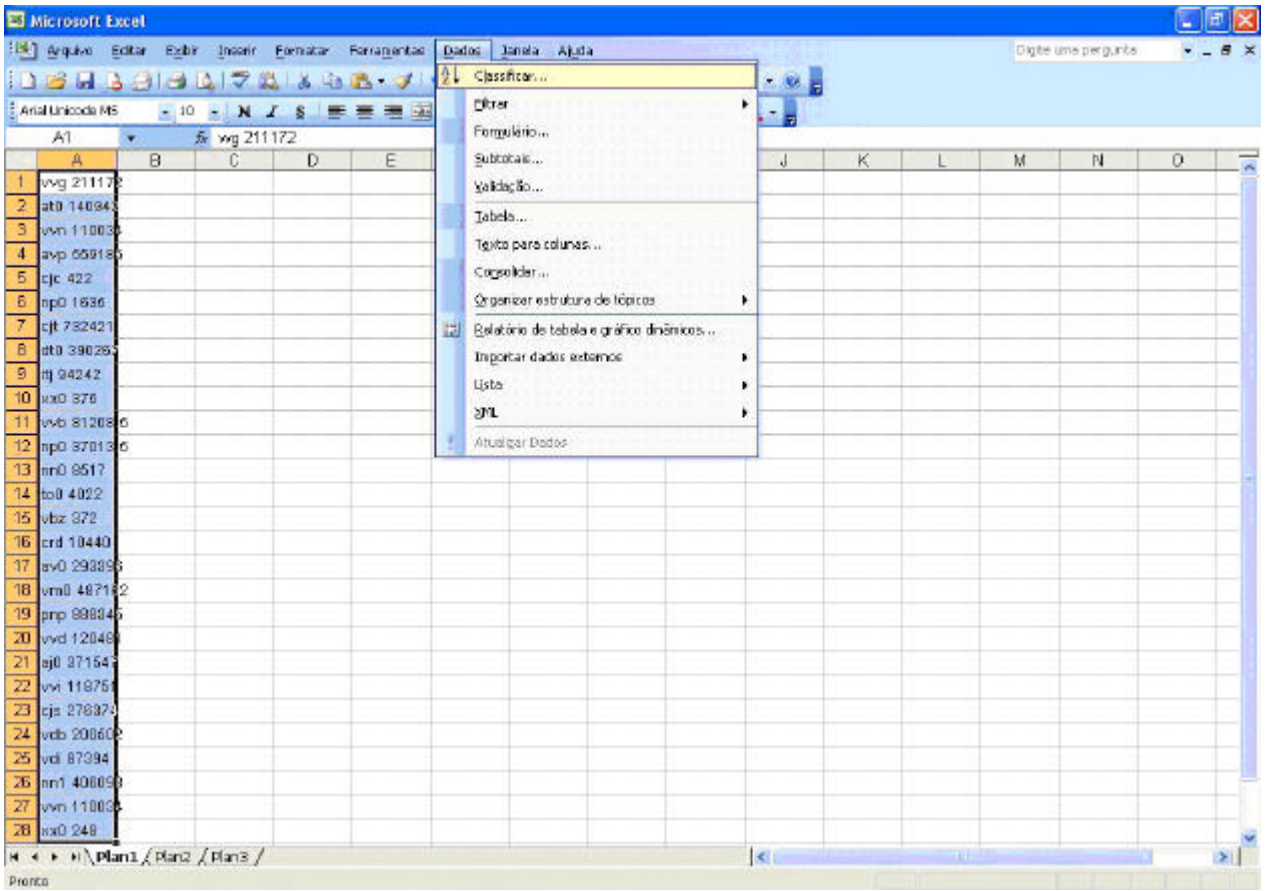


Figura 11 – Agrupamento de etiquetas iguais

Em segundo lugar, selecionar e separar, em duas colunas, as etiquetas e a freqüência.

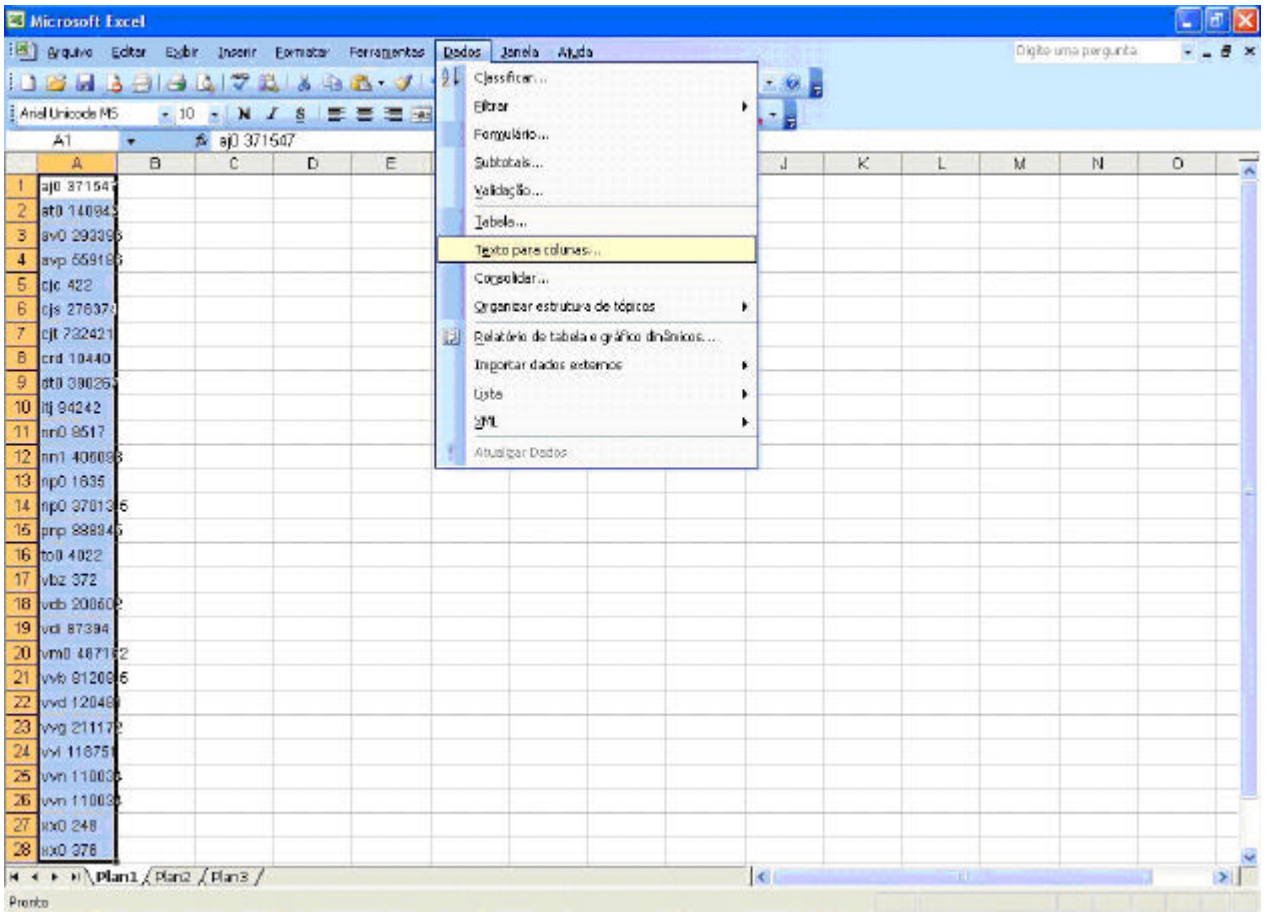


Figura 12 – Separação das etiquetas e da freqüência – tela inicial

As figuras 13, 14 e 15 demonstram as etapas seguidas para a seleção e separação das duas colunas. Na figura 16 as colunas já estão separadas.

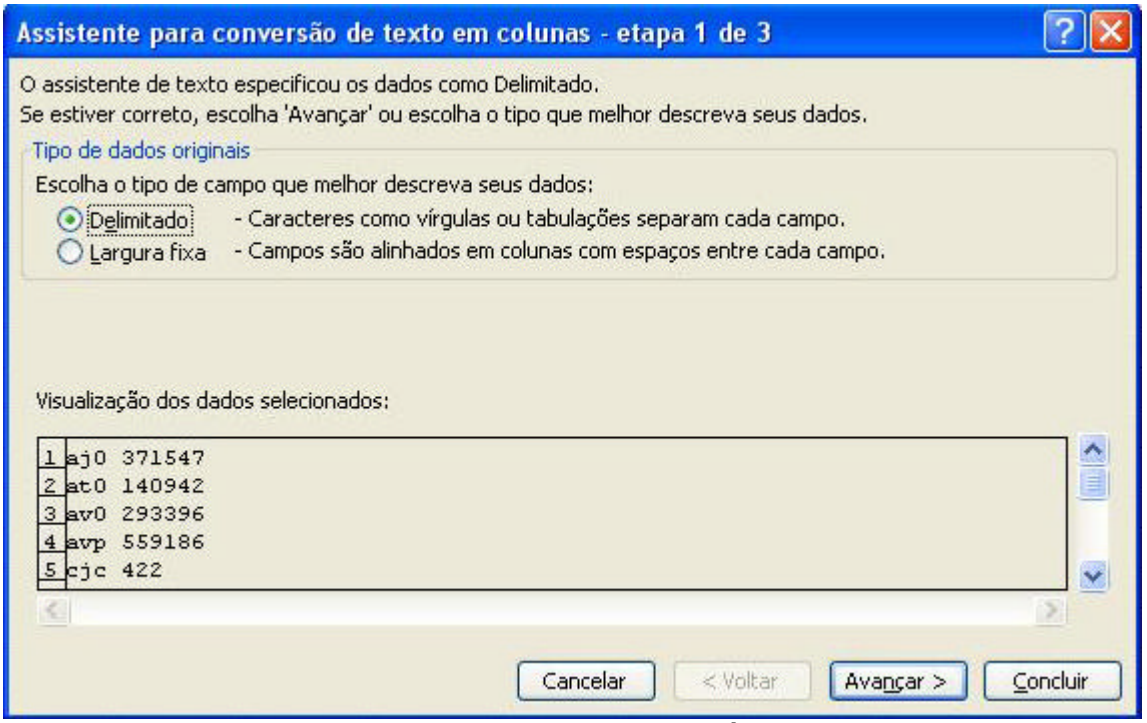


Figura 13 - Separação das etiquetas e da freqüência - etapa 1 de 3

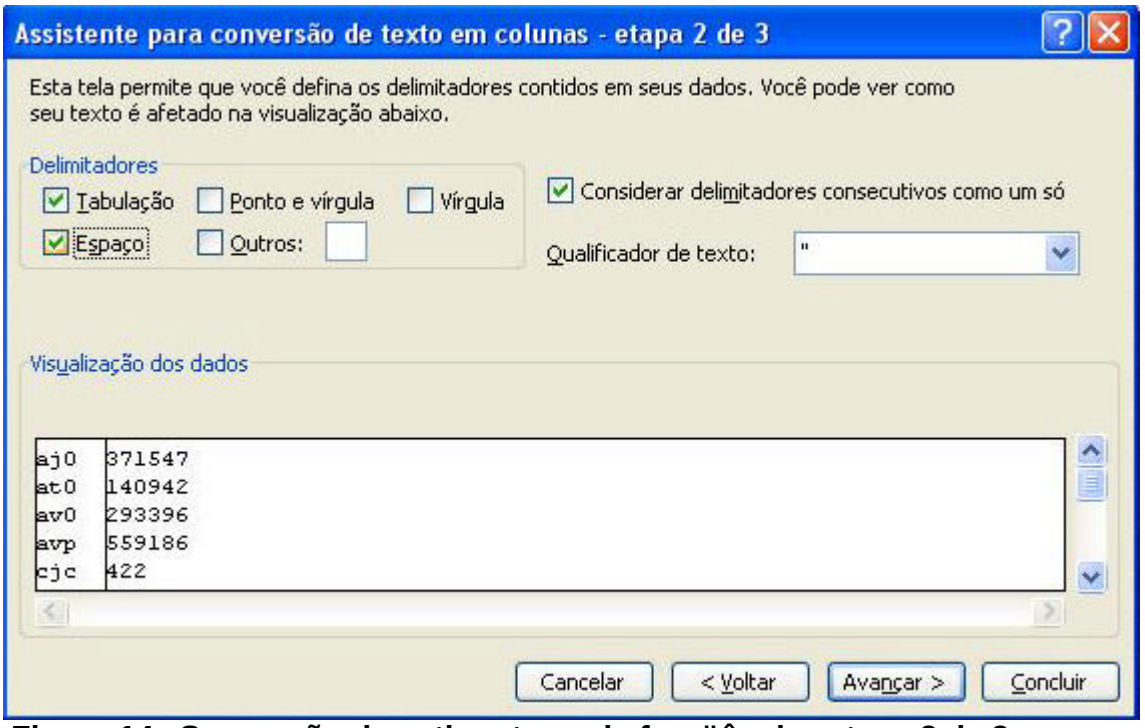


Figura 14 - Separação das etiquetas e da freqüência - etapa 2 de 3

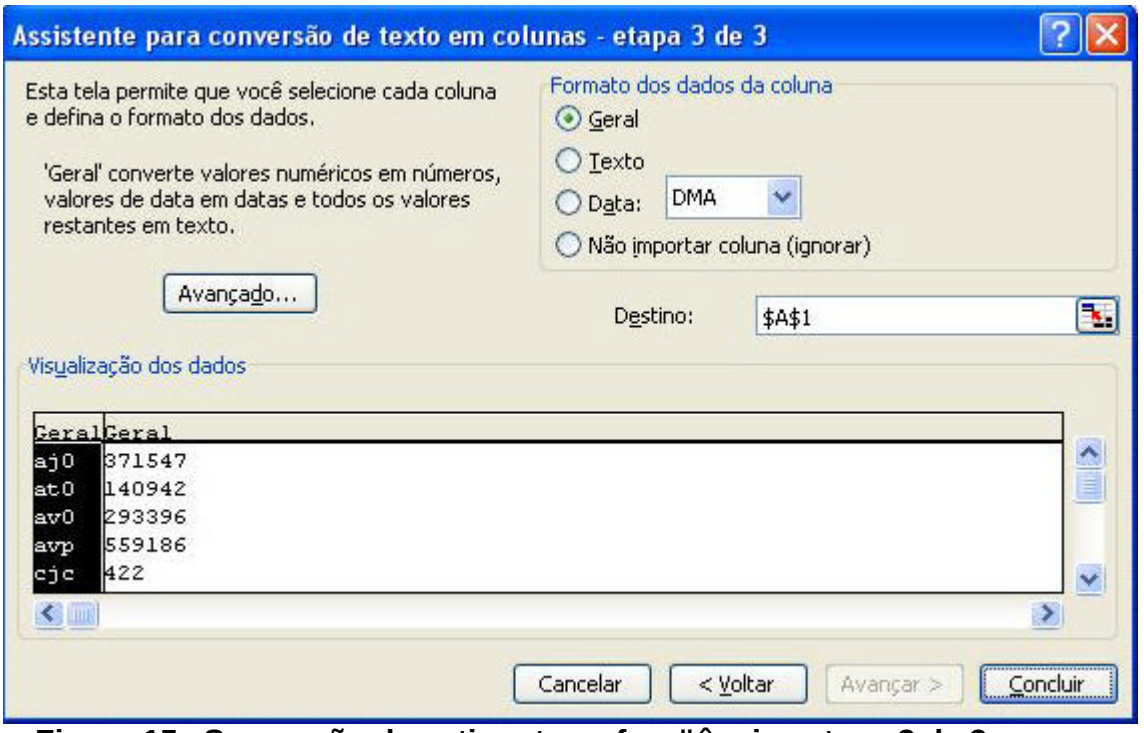
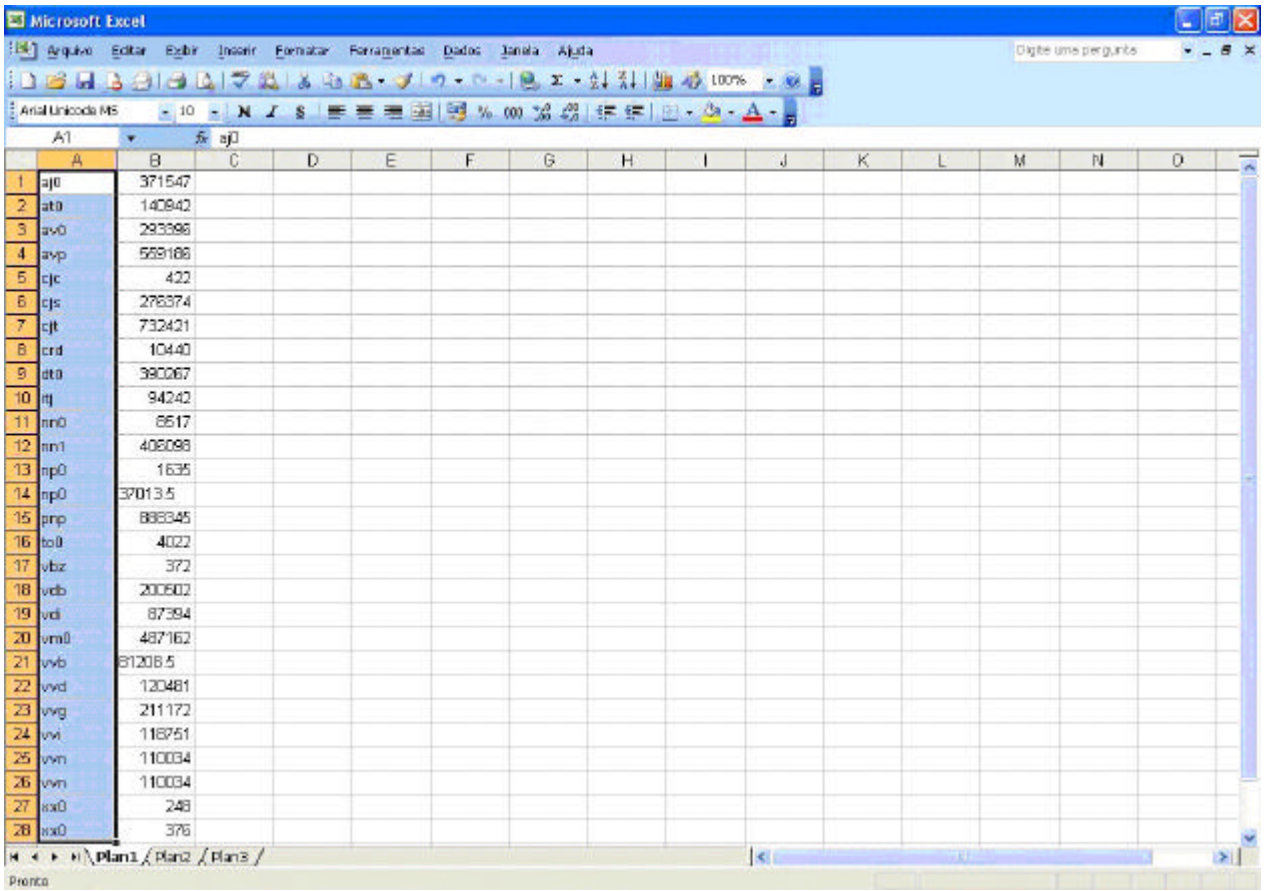


Figura 15 - Separação das etiquetas e freqüência - etapa 3 de 3



	A	B
1	aj0	371547
2	at0	140842
3	av0	293398
4	avp	569186
5	cjc	422
6	cjs	276374
7	cjt	732421
8	crd	10440
9	dt0	390267
10	ftj	94242
11	nn0	6617
12	nn1	408098
13	np0	1635
14	np0	370135
15	prp	888345
16	to0	4022
17	vbz	372
18	vdb	200602
19	vdi	87394
20	vm0	487162
21	vvb	312085
22	vvd	120481
23	vvg	211172
24	vvi	116751
25	vvn	110034
26	vvn	110034
27	xx0	248
28	xx0	376

Figura 16 - Etiquetas agrupadas - tela com os resultados

Em terceiro lugar, colocar título nas colunas.

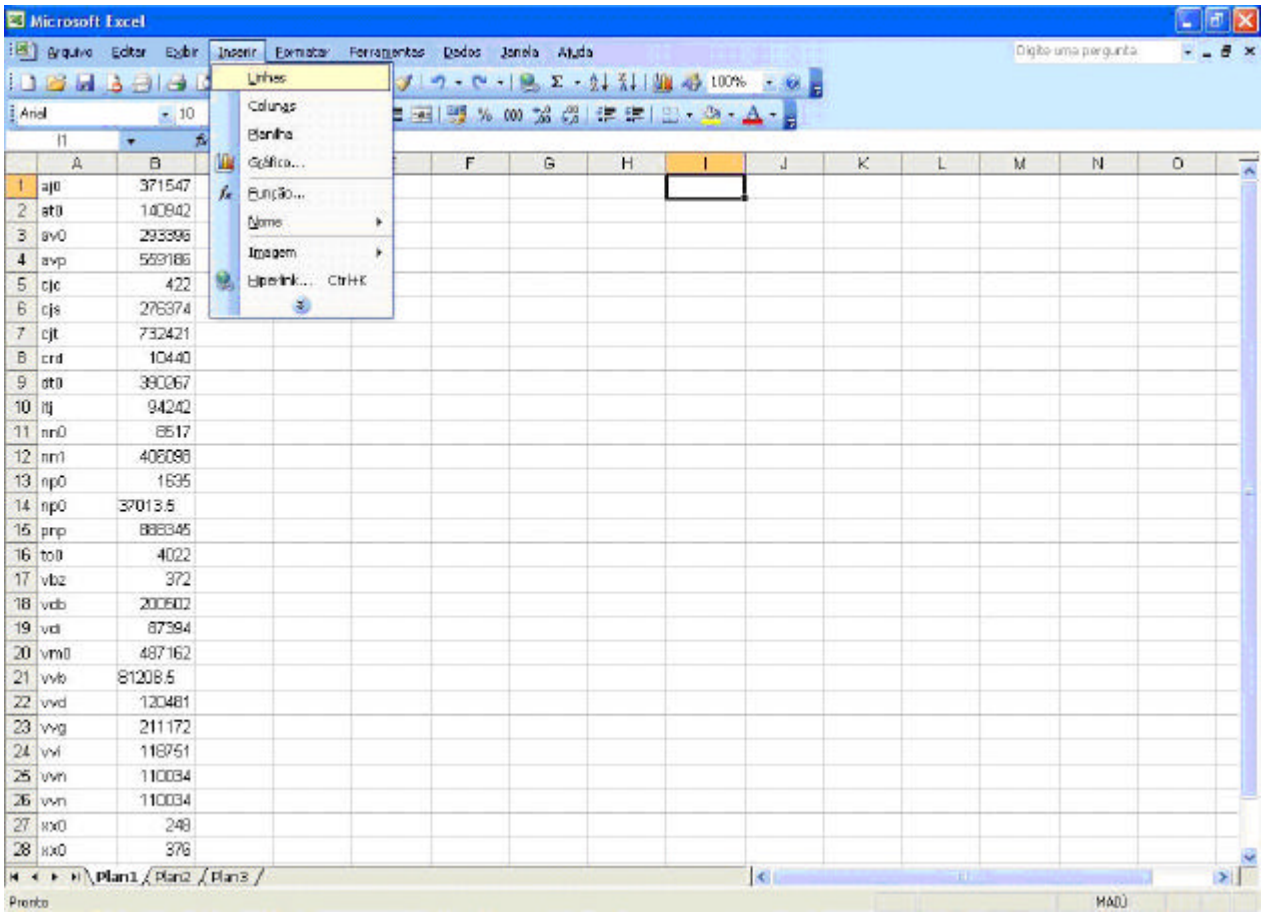
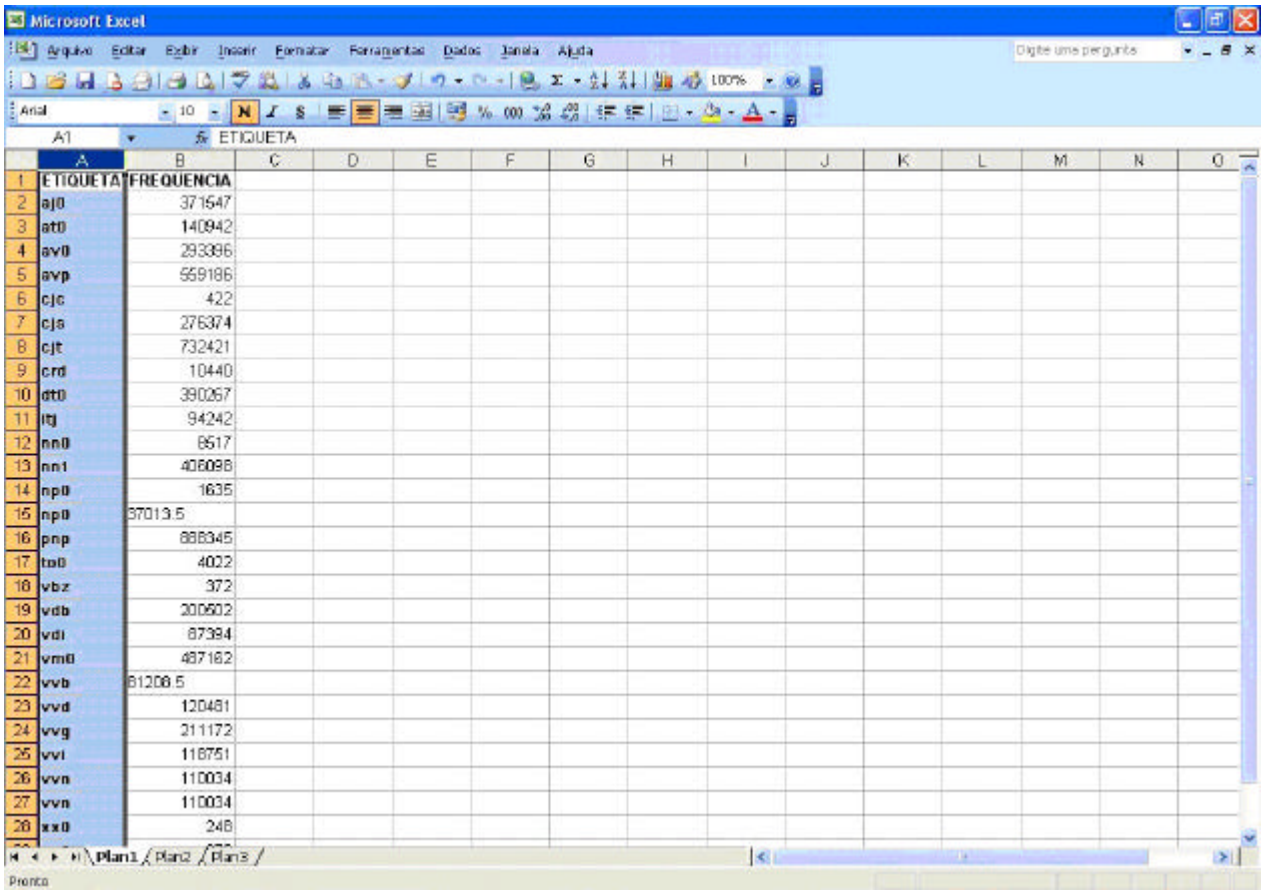


Figura 17 - Adição de títulos às colunas



ETIQUETA	FREQUENCIA
aj0	371547
at0	140942
av0	293396
avp	559186
cjc	422
cjs	276374
cjt	732421
crd	10440
dt0	390267
itj	94242
nn0	8617
nn1	406098
np0	1635
np0	37013.5
pnp	888345
to0	4022
vbz	372
vdb	200602
vdi	87394
vm0	487162
vvb	81208.5
vvd	120481
vvg	211172
vvi	118751
vvn	110034
vvn	110034
xx0	248

Figura 18 - Tela com os títulos inseridos

Por fim, utilizar o comando *subtotais* a fim de ter a somatória dos totais de cada etiqueta.

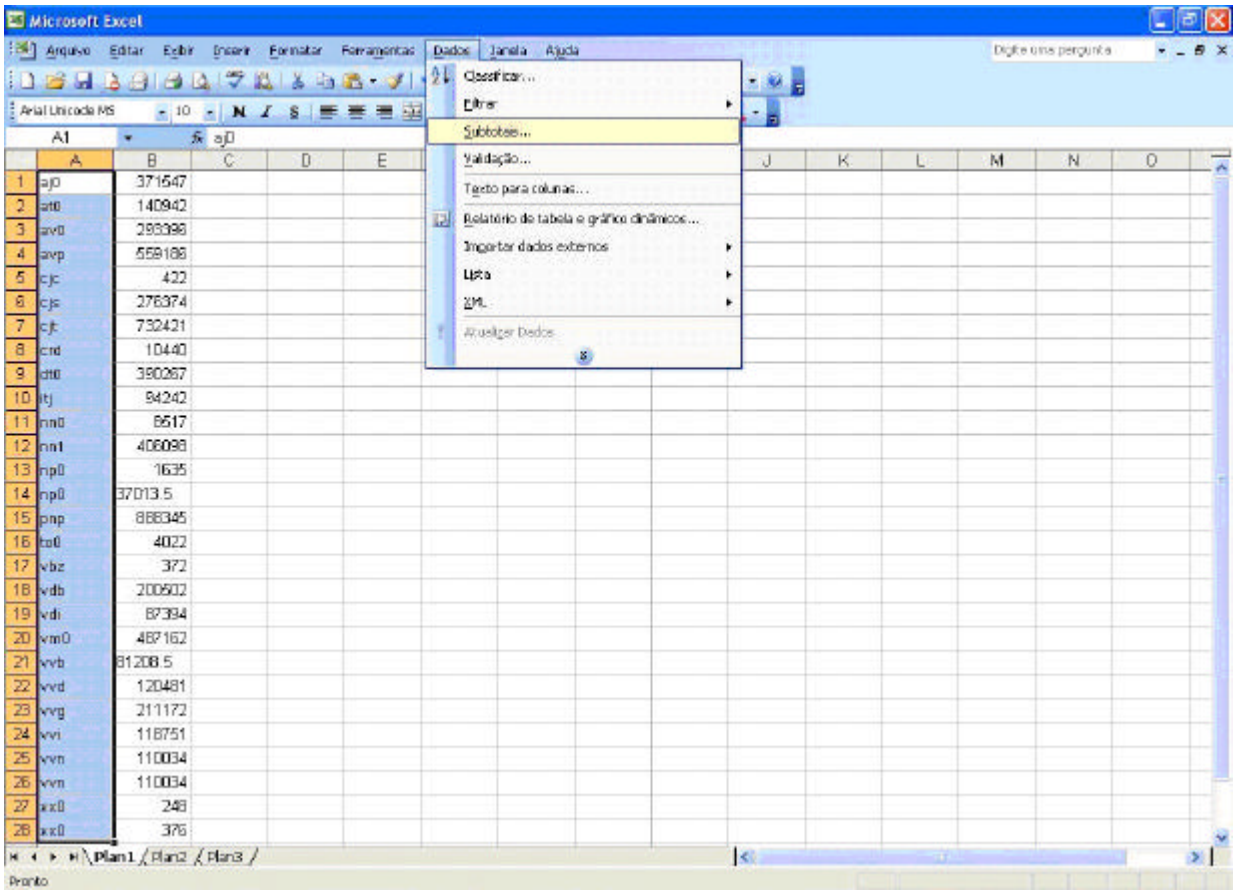
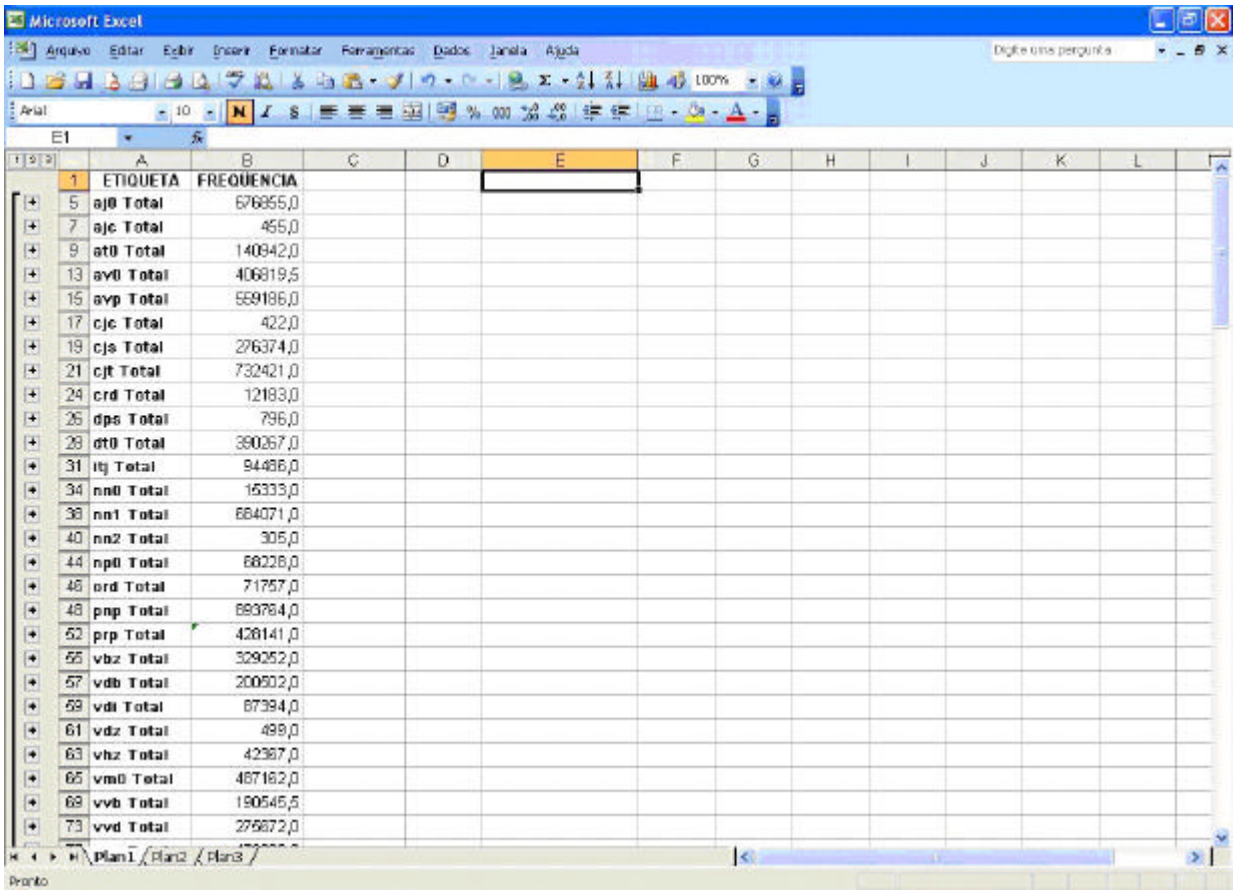


Figura 19 - Utilização do comando *Subtotais*

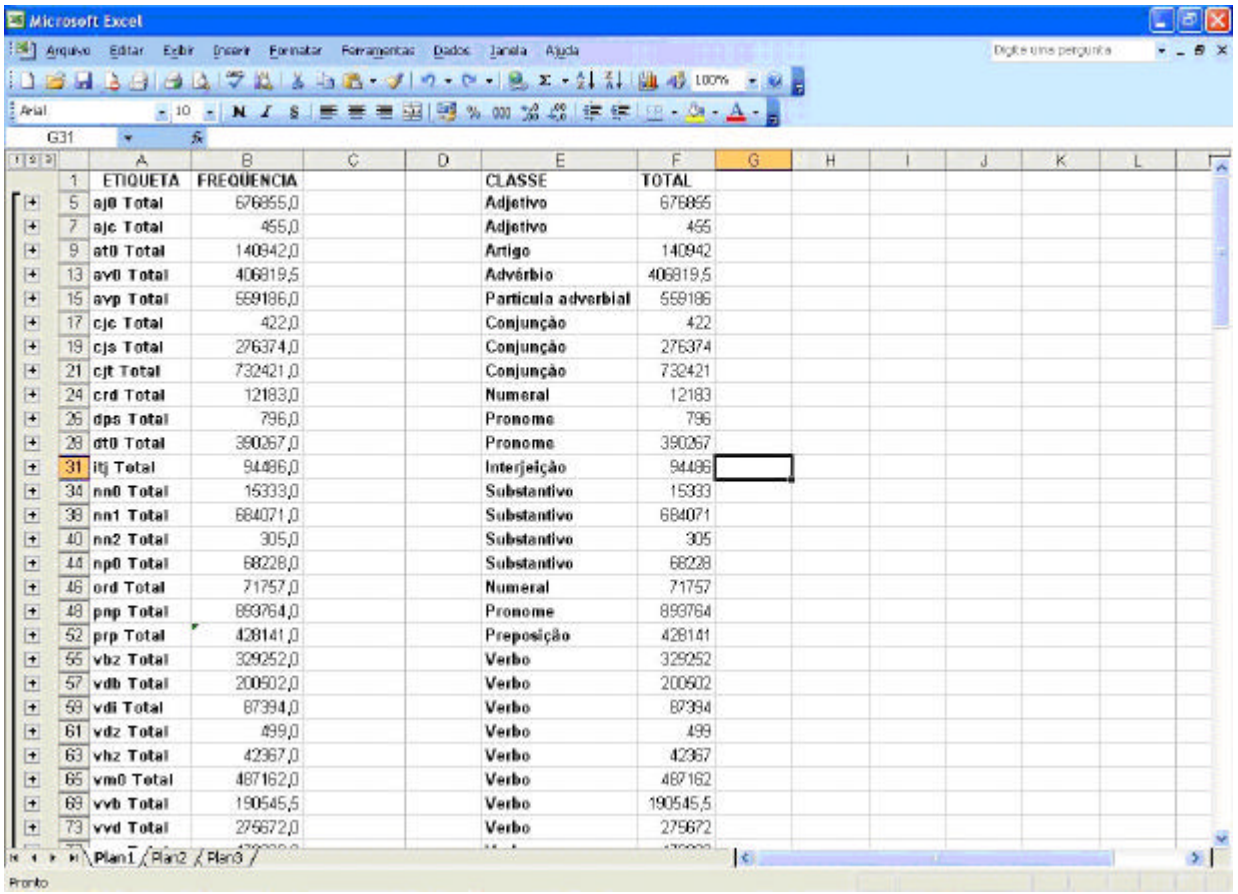


	A	B
1	ETIQUETA	FREQUENCIA
5	aj0 Total	676855,0
7	ajc Total	455,0
9	at0 Total	140942,0
13	av0 Total	406819,5
15	avp Total	559186,0
17	cjc Total	422,0
19	cjs Total	276374,0
21	cjt Total	732421,0
24	crd Total	12183,0
26	dps Total	796,0
28	dt0 Total	390267,0
31	itj Total	94486,0
34	nn0 Total	15333,0
36	nn1 Total	664071,0
40	nn2 Total	305,0
44	np0 Total	68226,0
46	ord Total	71757,0
48	pnp Total	693784,0
52	prp Total	428141,0
55	vbz Total	329252,0
57	vdb Total	200502,0
59	vdi Total	87394,0
61	vdz Total	499,0
63	vhz Total	42367,0
65	vm0 Total	467162,0
69	vvb Total	190545,5
73	vvd Total	276672,0

Figura 20 - Tela com resultado do comando *Subtotais*

Após esses passos foi possível obter os subtotais de cada etiqueta, mas não o de cada classe de palavra, uma vez que uma classe de palavra pode ser representada por mais do que uma etiqueta. Para obter os totais de cada classe de palavra, foram adotados os seguintes passos:

- ✓ Classificação semi-automática interativa das etiquetas em classes de palavras de acordo com a tabela 02 (cf. figura 21).
- ✓ Cálculo dos subtotais das classes de palavras a serem feitos em uma outra planilha (*plan 2*) , uma vez que o Excel não aceita dois cálculos de subtotais na mesma planilha (cf. figura 22).



	A	B	C	D	E	F	G	H	I	J	K	L
1	ETIQUETA	FREQUENCIA			CLASSE	TOTAL						
5	aj0 Total	676855,0			Adjetivo	676855						
7	ajc Total	455,0			Adjetivo	455						
9	at0 Total	140942,0			Artigo	140942						
13	av0 Total	406819,5			Advérbio	406819,5						
15	avp Total	559186,0			Particula adverbial	559186						
17	cjc Total	422,0			Conjunção	422						
19	cjs Total	276374,0			Conjunção	276374						
21	cjt Total	732421,0			Conjunção	732421						
24	crd Total	12183,0			Numeral	12183						
26	dps Total	796,0			Pronome	796						
28	dt0 Total	390267,0			Pronome	390267						
31	itj Total	94486,0			Interjeição	94486						
34	nn0 Total	15333,0			Substantivo	15333						
38	nn1 Total	684071,0			Substantivo	684071						
40	nn2 Total	305,0			Substantivo	305						
44	np0 Total	68228,0			Substantivo	68228						
46	ord Total	71757,0			Numeral	71757						
48	pnp Total	869764,0			Pronome	869764						
52	prp Total	428141,0			Preposição	428141						
55	vbz Total	329252,0			Verbo	329252						
57	vdb Total	200502,0			Verbo	200502						
59	vdi Total	87394,0			Verbo	87394						
61	vdz Total	499,0			Verbo	499						
63	vhz Total	42367,0			Verbo	42367						
65	vm0 Total	487162,0			Verbo	487162						
69	vvb Total	190545,5			Verbo	190545,5						
73	vvd Total	275672,0			Verbo	275672						

Figura 21 - Classificação das etiquetas em classes de palavras de acordo com a tabela 02

Para transferir os dados da primeira tabela de subtotais para a segunda planilha, não foi possível utilizar o comando de copiar célula, digitando o sinal ‘=’ na célula em branco e clicando em *enter* na célula a ser copiada. A solução encontrada foi utilizar o comando *colar especial – valores*.

No entanto, ao colar os dados em uma nova planilha, o programa Excel exibe todas as linhas do comando subtotais e não somente os totais de cada classe, conforme mostra a figura 22. É possível observar que na coluna das *Classes*, há espaços em branco, que não representam os subtotais das classes e, por conseguinte, essas linhas com espaços em branco devem ser eliminadas.

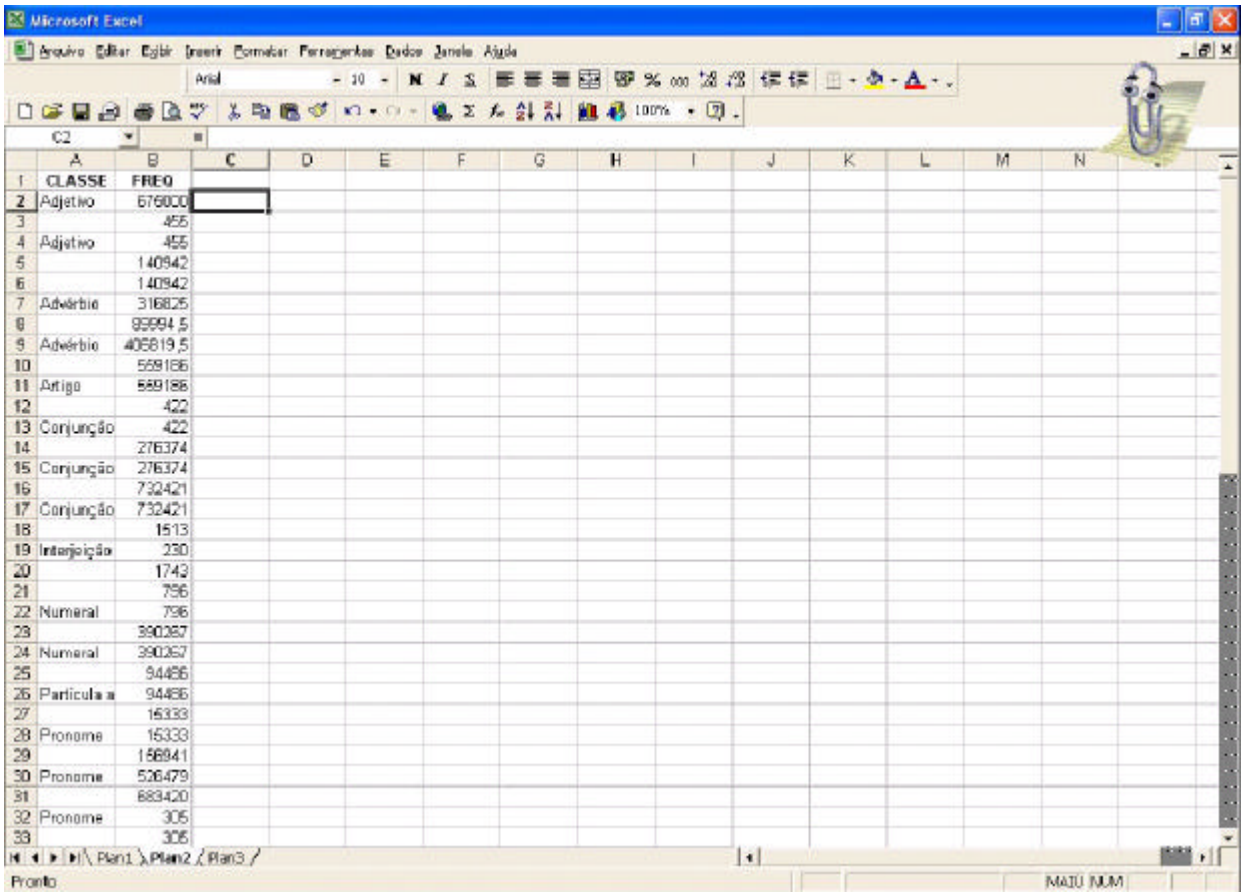
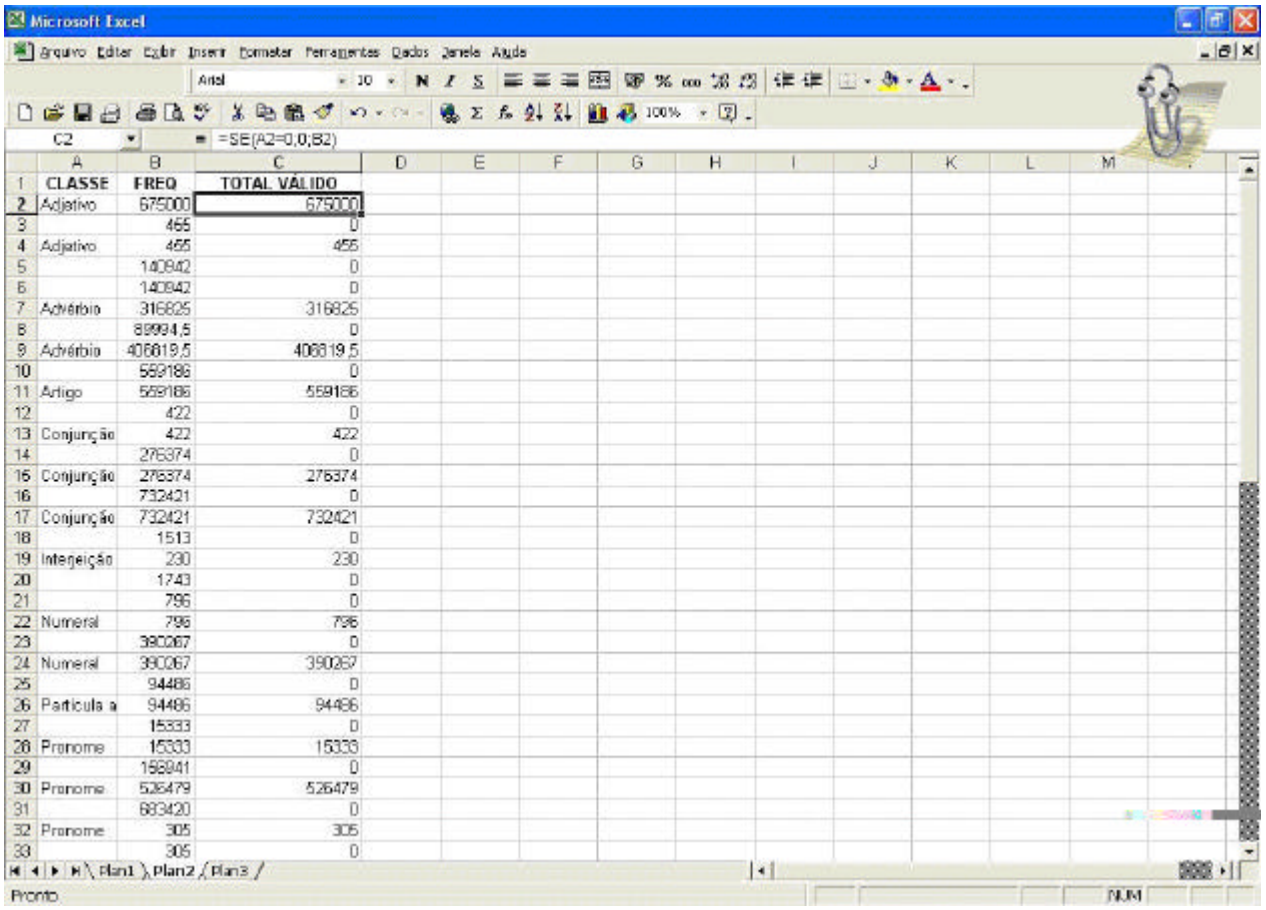


Figura 22 –Transferência dos valores das classes de palavras para a planilha 2

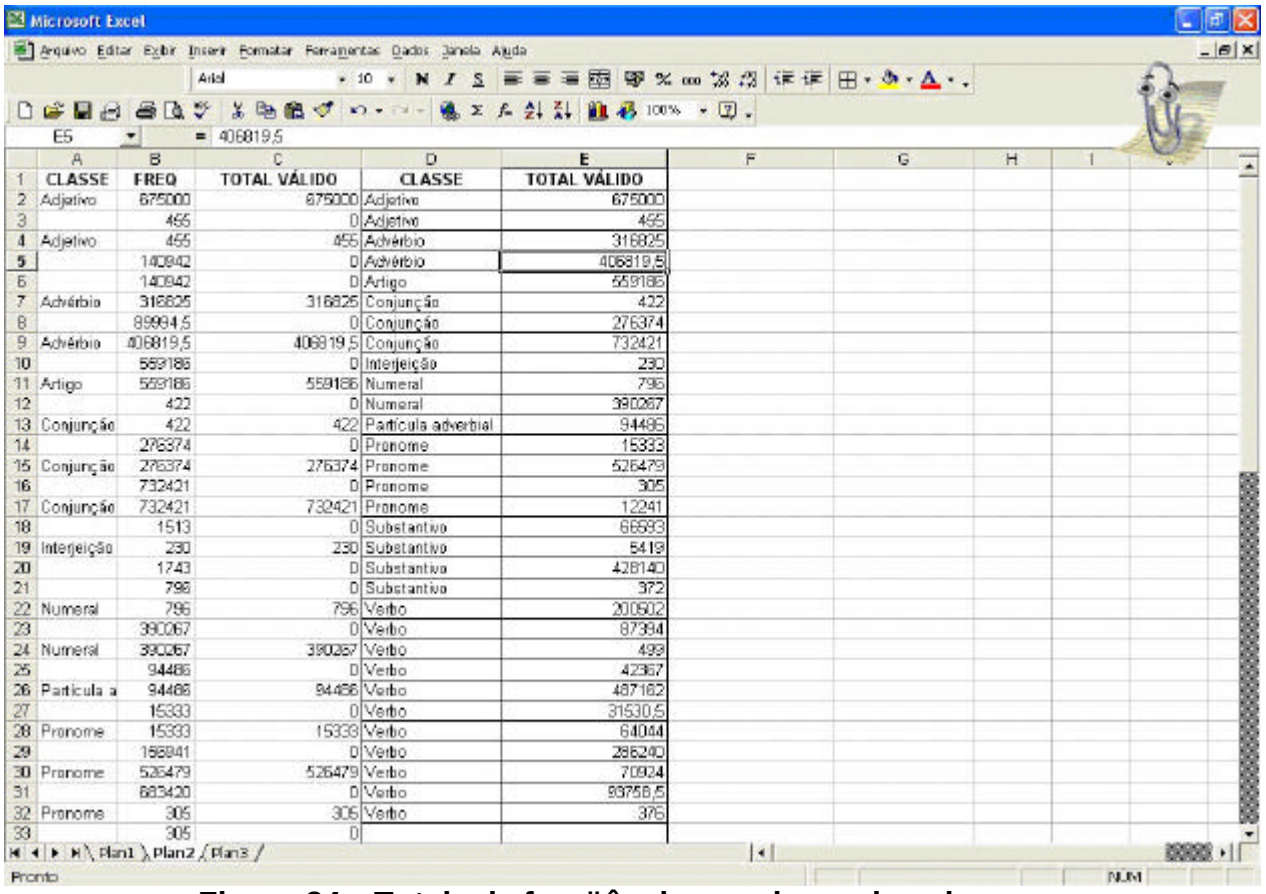
Para obter os subtotais finais de cada classe, é necessário filtrar os valores, isto é, deletar os valores que não representem o total de cada classe e, a seguir, calcular os subtotais dos valores restantes, conforme a figura 23. Para poder filtrar os valores que não representavam os subtotais, foi criada uma coluna - *Total válido*, com a seguinte fórmula =SE(A2=0;0;B2). A partir dessa fórmula, somente aparecerão os valores em cujas colunas representem totais das classes de palavras.



	A	B	C
1	CLASSE	FREQ	TOTAL VÁLIDO
2	Adjativo	675000	675000
3		455	0
4	Adjativo	455	455
5		140842	0
6		140842	0
7	Advérbio	316825	316825
8		89994,5	0
9	Advérbio	406819,5	406819,5
10		559186	0
11	Artigo	559186	559186
12		422	0
13	Conjunção	422	422
14		276374	0
15	Conjunção	276374	276374
16		732421	0
17	Conjunção	732421	732421
18		1513	0
19	Interjeição	230	230
20		1743	0
21		796	0
22	Numeral	796	796
23		390267	0
24	Numeral	390267	390267
25		94486	0
26	Particula a	94486	94486
27		15333	0
28	Pronome	15333	15333
29		156941	0
30	Pronome	526479	526479
31		683420	0
32	Pronome	305	305
33		305	0

Figura 23 – Filtragem das linhas que não representam subtotais de classes de palavras

Por fim, foram replicados os mesmos passos demonstrados nas figuras 19 e 20. Caso uma outra ferramenta tivesse sido desenvolvida para fazer a soma dos totais, o uso do Excel seria evitado e o processo não precisaria ser repetido diversas vezes. Porém, devido à pressão do tempo, não foi possível o desenvolvimento dessa ferramenta. Os resultados (cf. figura 24) demonstram as freqüências totais por classe de palavra e não mais as freqüências parciais. Com base nesses resultados foi calculada a porcentagem (cf. figura 25) de cada classe de palavra a fim de que fosse realizada a análise dos dados, que será demonstrada no capítulo 3, *apresentação e discussão de resultados*.



Microsoft Excel

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda

Arial 10

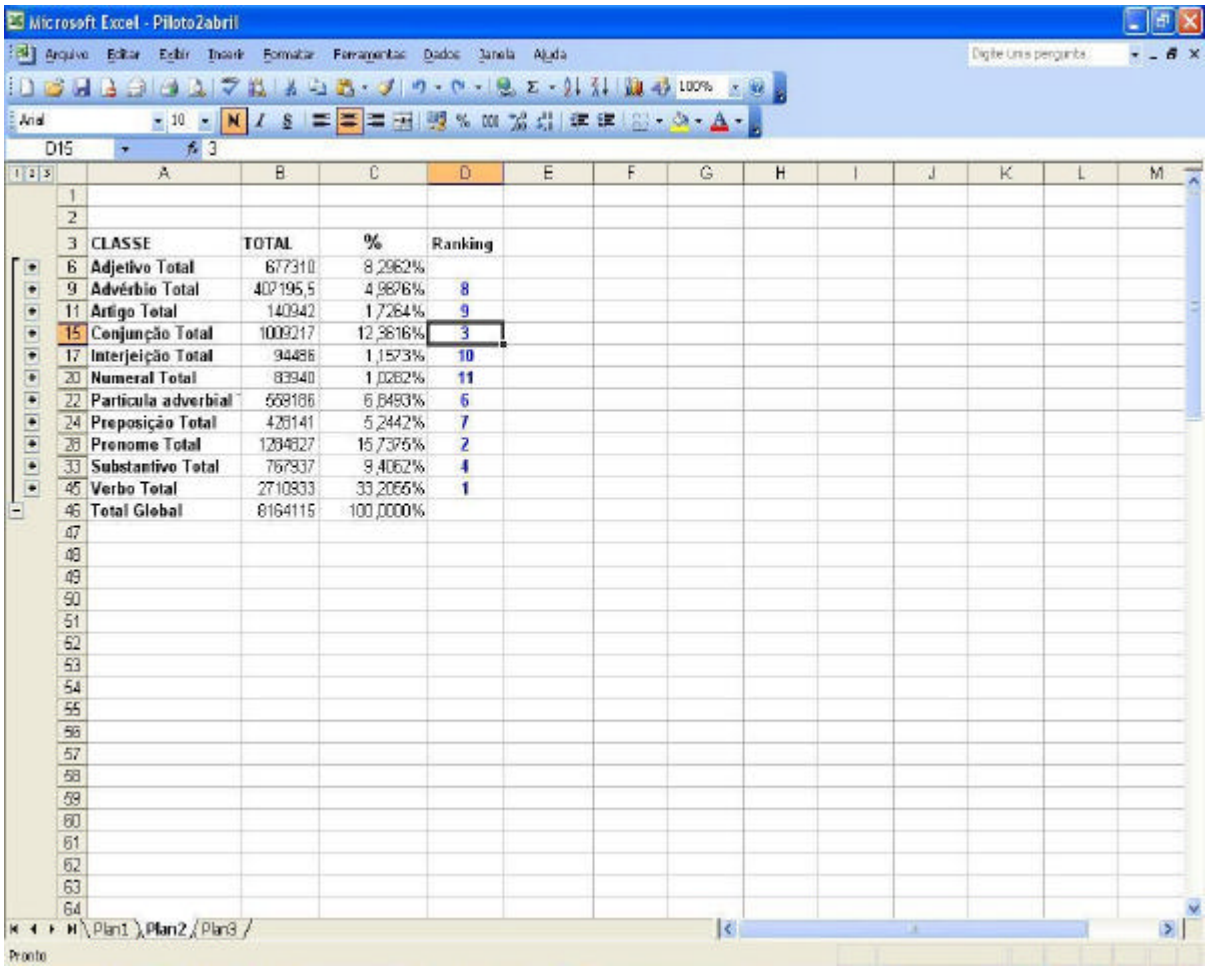
E5 = 406819,5

	A	B	C	D	E	F	G	H	I
1	CLASSE	FREQ	TOTAL VÁLIDO	CLASSE	TOTAL VÁLIDO				
2	Adjetivo	675000	675000	Adjetivo	675000				
3		455	0	Adjetivo	455				
4	Adjetivo	455	455	Advérbio	316825				
5		140842	0	Advérbio	406819,5				
6		140842	0	Artigo	559186				
7	Advérbio	316825	316825	Conjunção	422				
8		89994,5	0	Conjunção	276374				
9	Advérbio	406819,5	406819,5	Conjunção	732421				
10		559186	0	Interjeição	230				
11	Artigo	559186	559186	Numeral	796				
12		422	0	Numeral	390267				
13	Conjunção	422	422	Partícula adverbial	94486				
14		276374	0	Pronome	15333				
15	Conjunção	276374	276374	Pronome	526479				
16		732421	0	Pronome	305				
17	Conjunção	732421	732421	Pronome	12241				
18		1513	0	Substantivo	66593				
19	Interjeição	230	230	Substantivo	5419				
20		1743	0	Substantivo	428140				
21		796	0	Substantivo	372				
22	Numeral	796	796	Verbo	200502				
23		390267	0	Verbo	87394				
24	Numeral	390267	390267	Verbo	499				
25		94486	0	Verbo	42367				
26	Partícula a	94486	94486	Verbo	487162				
27		15333	0	Verbo	31530,5				
28	Pronome	15333	15333	Verbo	64044				
29		156941	0	Verbo	286240				
30	Pronome	526479	526479	Verbo	70924				
31		683420	0	Verbo	93758,5				
32	Pronome	305	305	Verbo	376				
33		305	0						

Plan1 Plan2 Plan3

Pronto NUM

Figura 24 – Totais de freqüência por classe de palavra



	A	B	C	D
3	CLASSE	TOTAL	%	Ranking
6	Adjetivo Total	677310	8,2862%	
9	Advérbio Total	402195,5	4,9876%	8
11	Artigo Total	140942	1,7264%	9
15	Conjunção Total	1009217	12,3616%	3
17	Interjeição Total	94486	1,1573%	10
20	Numeral Total	83940	1,0282%	11
22	Particula adverbial	559186	6,8493%	6
24	Preposição Total	428141	5,2442%	7
28	Pronome Total	1284827	15,7375%	2
33	Substantivo Total	767937	9,4062%	4
45	Verbo Total	2710933	33,2055%	1
46	Total Global	8164115	100,0000%	

Figura 25 - Tela com resultados finais

3. Apresentação e discussão dos resultados

Neste capitulo são apresentados os resultados da análise dos dados a partir de dois enfoques que se complementam. Esses dois enfoques partem do principio de que era impossível trabalhar com todas as 97.619.934 formas (*types*) do inglês. Foi preciso trabalhar com uma amostra de 1000 palavras para viabilizar a pesquisa. Essa amostra de 1000 palavras foi colhida de duas maneiras: uma, entre as 1000 palavras mais freqüentes do BNC; outra, entre as 1000 palavras com mais etiquetas. Cada um desses enfoques representou uma análise em separado, porém seguindo os mesmos passos ilustrados na metodologia. Devido à necessidade de nomear cada um dos enfoques, decidimos chamar o primeiro de enfoque de probabilidade. Esse enfoque nos permitiu descobrir qual a probabilidade de conversão, em termos gerais, no BNC. Neste, observamos a conversão entre as ocorrências (*tokens*) da língua, mais especificamente, entre as 1000 palavras correspondem a 69.257.801 *tokens*, ou aproximadamente 69% do total de ocorrências do corpus. O outro enfoque foi chamado de *a perspectiva da possibilidade*. Esse enfoque nos permitiu descobrir qual a possibilidade de palavras se converterem no corpus, independente de serem muito ou pouco freqüentes. Nele, verifica-se a ocorrência de conversão entre as 1000 palavras com maior número de etiquetas segundo o *BNC*. Por exemplo, a palavra *sweet* pode ter 1000 ocorrências como nn1 (substantivo singular), 500 como nn2 (substantivo plural) e 10000 como aj0 (adjetivo não flexionado), que representam 3 maneiras potenciais de ela se converter.

Esses dois enfoques são distintos porque, de certo modo, o enfoque da probabilidade parte do principio da ocorrência na linguagem em uso, o que pressupõe um corpus eletrônico e ferramentas de processamento de corpus – esse enfoque é oferecido naturalmente pela Lingüística de Corpus. Já o enfoque da possibilidade é aquele que tende a ser preferido em abordagens teóricas cujos dados podem ser obtidos pela intuição ou pela coleta de dados sem apoio de um corpus eletrônico.

3.1 O enfoque da probabilidade

Após os processos de inserção, filtragem e contagem no identificador de conversão e no programa Excel das 1000 palavras mais freqüentes do corpus (cf. Capítulo 2, Metodologia), obteve-se como resultado os dados a seguir:

CLASSE	TOTAL	PORCENTAGEM	POSIÇÃO
Verbo Total	9.396.915	25,4%	**1**
Preposição Total	9.063.688	24,5%	**2**
Pronome Total	5.234.972	14,2%	**3**
Substantivo Total	4.070.274	11,0%	**4**
Advérbio Total	2.631.899	7,1%	**5**
Adjetivo Total	2.585.653	7,0%	**6**
Conjunção Total	2.068.711	5,6%	**7**
Partícula adverbial Total	843.059	2,2%	**8**
Artigo Total	488.311	1,3%	**9**
Numeral Total	349.761	0,9%	**10**
Interjeição Total	128.694	0,3%	**11**
Total Global	36.861.937	100,0%	

Tabela 07 - Resultados obtidos a partir das 1000 palavras mais freqüentes no *British National Corpus*

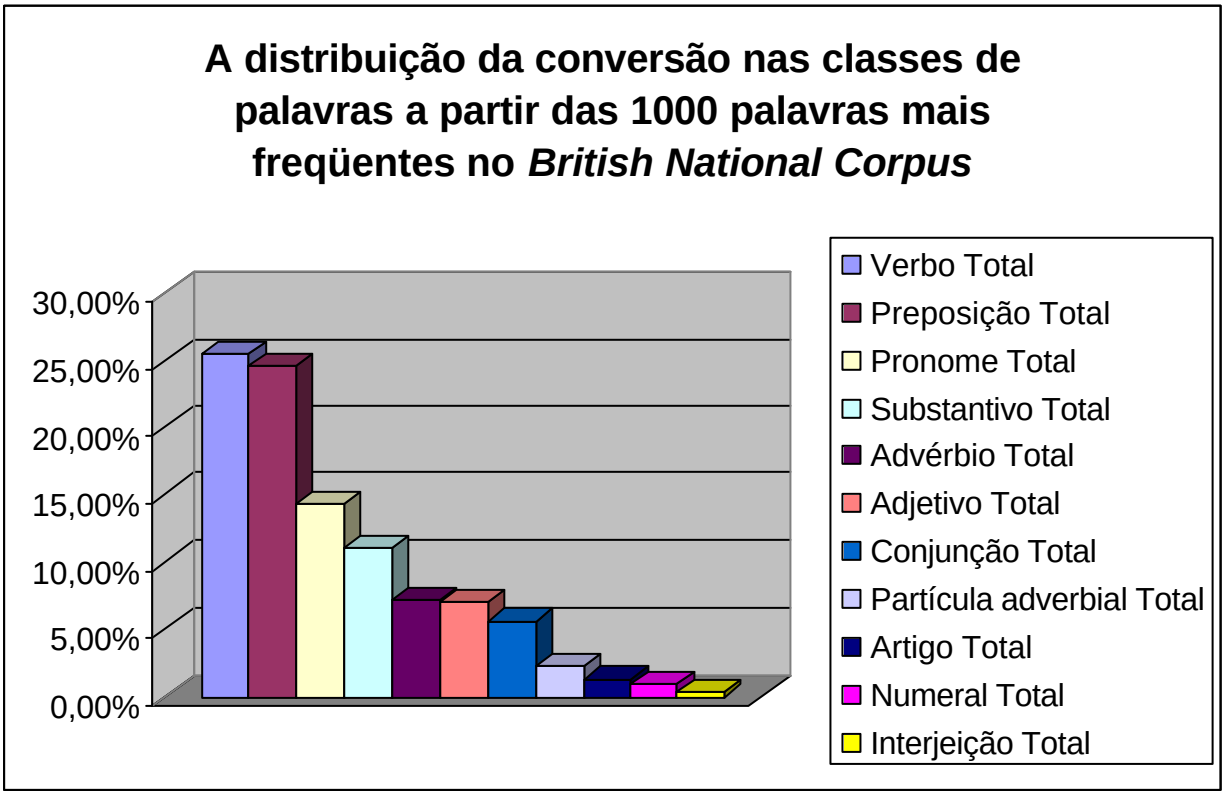


Gráfico 01 – A distribuição da conversão nas classes de palavras a partir das

1000 palavras mais freqüentes no *British National Corpus*

Observando a tabela 07, é possível afirmar que os verbos, as preposições e os pronomes apresentam maior freqüência entre as palavras mais convertidas e, portanto, são mais passíveis de conversão, ao passo que os artigos, os numerais e as interjeições são as classes que apresentam menor freqüência e, portanto, são as menos passíveis de serem convertidas. Esse conjunto de observações responde à primeira pergunta de pesquisa, isto é, quais são as classes de palavras mais e menos passíveis de serem convertidas na língua inglesa.

Levando-se em conta a freqüência global de todas as classes, tem-se como possível conclusão que, partindo de um universo de 69.257.801 *tokens*, que representam as 1000 palavras mais freqüentes no BNC, 36.861.937 *tokens* (53% do total) (cf. tabela 07) estejam potencialmente envolvidos no processo de conversão. De acordo com Halliday (1993), essa observação retrata um exemplo de *equiprobable distribuition*, uma vez que a proporção entre a probabilidade de uma palavra estar relacionada e não estar relacionada ao fenômeno da conversão é de aproximadamente 1 para 1. Com base nessa conclusão, é possível demonstrar empiricamente quão freqüente a conversão é.

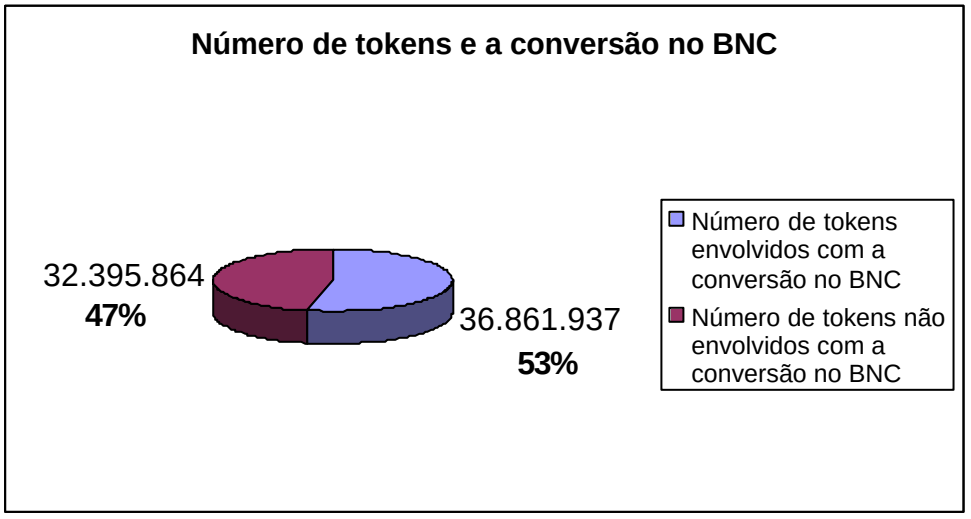


Gráfico 02 – Número de tokens e a conversão no BNC

Observando o gráfico 02, é possível concluir que a cada duas palavras da língua inglesa, uma provavelmente estará relacionada ao fenômeno da conversão.

3.2 O enfoque da possibilidade com maior repertório de etiquetas

No intuito de retomar as palavras que apresentam maior número de etiquetas e conseqüentemente responder à terceira pergunta de pesquisa - quais são as palavras que apresentam maior repertório de classes de palavras passíveis de serem convertidas, faz-se necessário copiar a tabela 04 – *Palavras do BNC com maior número de etiquetas*.

Nº de etiquetas	Palavra
14	LIKE
13	SET
13	FORECAST
13	CUT
13	A
12	SPLIT
12	ROUND
12	BURST
12	BOUND
11	UPSET
11	UP
11	SQUARE
11	POP
11	OPENCAST
11	LEFT
11	LAY
11	HURT
11	HIT
11	FELL
11	CROSS
11	CAST
11	AI
10	WET
10	SOUND
10	SHED
10	S
10	OVER
10	OUTCAST
10	OPEN
10	OFFSET
10	LIGHT
10	GROUND
10	FUCKING
10	DO
10	DAT
10	CRAP
10	CLOSE
10	CHANGING

Nº de etiquetas	Palavra
9	UPFRONT
9	TYPECAST
9	TREKKING
9	TRADING
9	TOURING
9	TOP
9	THRUST
9	THAT
9	TESTING
9	THRUST
9	THAT
9	TESTING
9	TA
9	SWIMMING
9	SUPPORTING
9	SUFFERING
9	STYLING
9	STRAIN
9	STOPPING
9	STOCK
9	STAGING
9	SPORTING
9	SPLITTING
9	SPIT
9	SPENDING
9	SMOKING
9	SLIT
9	SKIING
9	SKETCHING
9	SIGNING
9	SHOOTING
9	SAMPLING
9	SAILING
9	RIDING
9	REVIEWING
9	RETURNING
9	RESTRUCTURING
9	RESTING

Nº de etiquetas	Palavra
9	PLUMB
9	PLACING
9	PET
9	PASSING
9	PARALLEL
9	PANTING
9	PACKAGING
9	OW
9	OUT
9	OL
9	OFFERING
9	OBTAINING
9	NT
9	NO
9	NESTING
9	MOTORING
9	MONITORING
9	MODELLING
9	MESS
9	MATCHING
9	MAT
9	MANUFACTURING
9	MAILING
9	LOVING
9	LOGGING
9	LISTING
9	LISTENING
9	LINKING
9	LING
9	LEVEL
9	LET
9	LAPWING
9	LABOURING
9	KILLING
9	JOGGING
9	INTRODUCING
9	IMAGING
9	I'

10	BROADCAST	9	REPORTING	9	HEALING
10	BET	9	RECYCLING	9	HEAD
10	BACK	9	READ	9	HANDLING
9	VE	9	PRESENT		

Cópia da tabela 04 – Palavras do BNC com maior número de etiquetas

Se por um lado, partindo das 1000 palavras mais freqüentes do BNC, os verbos, as preposições e os pronomes são as classes de palavras com maior probabilidade de conversão (cf. tabela 07), os verbos, as conjunções e os substantivos são as classes mais passíveis de conversão, partindo das 1000 palavras que possuem maior repertório de classes de palavras associadas à conversão. Essa conclusão pode ser observada na tabela abaixo:

CLASSE	TOTAL	Porcentagem	Posição
Verbo Total	2.705.173	37,2%	**1**
Conjunção Total	1.009.217	13,9%	**2**
Substantivo Total	765.652	10,5%	**3**
Adjetivo Total	675.455	9,3%	**4**
Partícula adverbial Total	559.186	7,7%	**5**
Preposição Total	428.141	5,9%	**6**
Advérbio Total	407.195,5	5,6%	**7**
Pronome Total	396.482	5,4%	**8**
Artigo Total	140.942	1,9%	**9**
Interjeição Total	94.486	1,3%	**10**
Numeral Total	73.500	1,0%	**11**
Total Global	7.255.429,5	100,0%	

Tabela 08 - Distribuição da conversão nas 1000 palavras que apresentam maior possibilidade de conversão com maior repertório de etiquetas

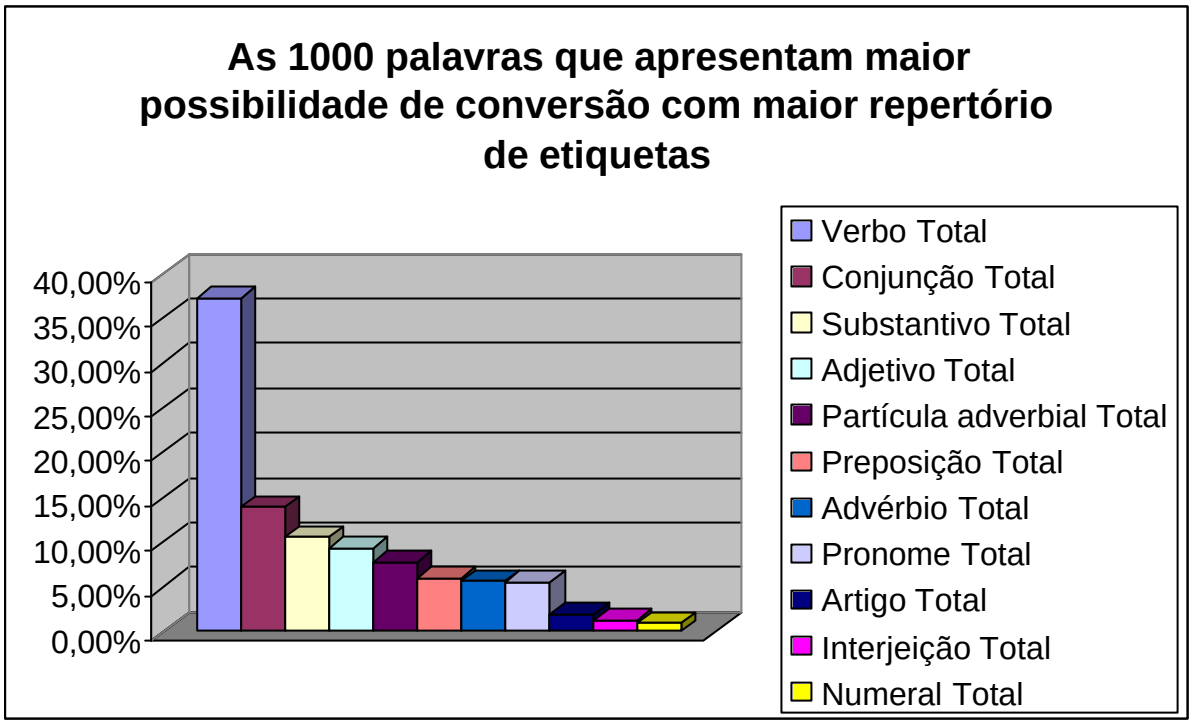


Gráfico 03 – As 1000 palavras que apresentam maior possibilidade de conversão com maior repertório de etiquetas

De modo similar à tabela 07, os artigos, as interjeições e os numerais são as classes de palavras menos passíveis de conversão, apenas invertendo-se a posição da freqüência das interjeições e dos numerais. A fim de exemplificar empiricamente quão freqüente a conversão é, faz-se necessário copiar algumas linhas de concordância encontradas no corpus e suas respectivas etiquetas gramaticais.

3.2.1 A conversão na palavra *like*

Na concordância abaixo, percebe-se que a palavra *like* ocorre como verbo nas linhas 1 e 2, substantivo na linha 3, preposição na linha 4, advérbio na linha 5, conjunção na 6 e adjetivo na linha 7. Nas linhas 8, 9 e 11, o etiquetador, devido à ambigüidade, não foi capaz de diferenciar se a palavra é preposição ou conjunção na linha 8, advérbio ou adjetivo na linha 9, adjetivo ou substantivo na linha 10 e conjunção ou preposição na linha 11.

nursing support through our 24 hour on call facility. I would **like** to be an ACET volunteer **(VVI)**
art and work are nevertheless very often the same. I **like** to think that he would have accepted **(VVB)**

: by Acts of Parliament, voted by landlords to benefit their **like**. In France they had loosed **(NN1)**
This includes nurses and volunteers. At the moment it looks **like** my usual stint of ironing **(PRP)**
goddawful evening. She wouldn't have supper for a start. **Like**, it was work not socialising **(AV0)**
to sell yourself --; and if you're a character actor **like** I am I can not show myself **(CJS)**
the same time and place, the appellant was committing a **like** act with the co-defendant. **(AJ0)**
I could no longer rely on friends for help with everyday chores **like** shopping and housework **(PRP-CJS)**
Nobody gives a shit about you and your research. **Like**, why did you feel the need to do on **(AV0-AJ0)**
, St John Ambulance men, special constables, and the **like** were all used to define **(AJ0-NN1)**
police officers but before you? It's quick **like** quite **like**ly that C I D officers almost certainly, C I **(UNC)**
photograph. The main thing is that it needs to look **like** you! **(CJS-PRP)**

3.2.2 A conversão na palavra *set*

Na concordância abaixo, percebe-se que a palavra *set* ocorre como verbo nas linhas 1, 3, 4, 5 e 6, substantivo na linha 2 e adjetivo na linha 7. Nas linhas 8, 9 e 10 o etiquetador, devido à ambigüidade, não foi capaz de diferenciar se a palavra é substantivo ou adjetivo na linha 8, verbo ou substantivo na linha 9 e verbo ou adjetivo na linha 10.

a Presidential Task Force on human rights was **set** up, consisting of senior figures close to **(VVN)**
In her mind there was a **set** of events which includes reading about a work of art and looking **(NN1)**
what art criticism is, or to **set** out a theory of criticism, beyond the suggestions of some **(VVI)**
these demolitions: "If we pulled one down, he **set** up three." The wrecker explains that, **(VVD)**
Thus if you **set** aside £25 per month it would take two years to accumulate the **(VVB)**
culture which appeared among the promoters of the Pro-Life Campaign, **set** up to achieve **(VVD-VVN)**
special feature with some great tips of her own. Get **set** for a bulb bonanza too. **(AJ0)**
school finals. No. I'm quite hopeless at the **set** auditions, I think. I know it's silly but I **(NN1-AJ0)**
the world for the two of us, he wrote. **Set** it on its feet, he wrote, and then let it **(VVB-NN1)**
all know the face and hands of the figure, **set** in its marble chair, in that circle of fantastic **(VVN-AJ0)**

3.2.3 A conversão na palavra *forecast*

Na concordância abaixo, percebe-se que a palavra *forecast* ocorre como substantivo na linha 1, verbo nas linhas 3, 4, 5 e 6. Nas linhas 2, o etiquetador, devido à ambigüidade, não foi capaz de diferenciar se a palavra é substantivo ou verbo.

found by flying out of the lift. 2 Check the **forecast** for risk of storms and do not enter large clouds **(NN1)**
privatisation offering at the time of the crash, its profit **forecast** for the year ending 31 December, **(NN1-VVB)**
available, was 7.3 per cent, and it is **forecast** to be unchanged when September figures are released **(VVN)**
the shares they wanted," one analyst said. Analysts **forecast** an extra fillip to early water **(VVB)**
and wants to start a family. "We can not **forecast** the success of a reversal operation in an t **(VVI)**
Consultant Peter Thompson destroyed a hesitant rally when he **forecast** the FT-SE share index ending **(VVD)**

3.3 Análise contrastiva entre as 1000 palavras mais freqüentes e as 1000 que apresentam maior repertório de etiquetas no *British National Corpus*

A fim de facilitar a visualização e conseqüente comparação entre as duas perspectivas de análises, foi elaborada esta tabela-resumo:

	1000 mais freqüentes		1000 com mais classes	
	Porcentagem	**Posição**	**Porcentagem**	**Posição**
Verbo	25,49%	1	37,28%	1
Preposição	24,58%	2	5,90%	6
Pronome	14,20%	3	5,46%	8
Substantivo	11,04%	4	10,55%	3
Advérbio	7,13%	5	5,61%	7
Adjetivo	7,01%	6	9,30%	4
Conjunção	5,61%	7	13,90%	2
Partícula adverbial	2,28%	8	7,70%	5
Artigo	1,32%	9	1,94%	9
Numeral	0,94%	10	1,01%	11
Interjeição	0,34%	11	1,30%	10

Tabela 09 - Comparação quanto à freqüência entre as 1000 palavras mais freqüentes e as 1000 que apresentam maior repertório de etiquetas no *British National Corpus*

Apesar de, nas duas tabelas, os verbos figurarem como classe gramatical mais freqüente, é notável a diferença entre a primeira e a segunda classe mais freqüente em cada tabela. Nas 1000 palavras mais frequentes (cf. tabela 07)**,** essa diferença é de 0,91%, ao passo que na tabela 08 essa diferença é de 23,38%. Isso significa que, à medida que o número de palavras diferentes (tokens) aumenta, são maiores as chances de haver melhor distribuição da freqüência da conversão entre

as diferentes classes das palavras. Essa observação, de certo modo, contraria a tendência das 1000 palavras com mais etiquetas, pelo fato de apresentarem maior diversidade de etiquetas, terem a freqüência melhor distribuída entre esse maior número de classes e sub-classes de palavras. Com a conclusão de que quanto maior a quantidade de *tokens*, melhor é a distribuição entre as classes de palavras, a quarta pergunta de pesquisa é respondida. Podemos interpretar essa diferença pelo fato de certas classes muito freqüentes (como as preposições) terem um espectro limitado de conversão. A preposição *about*, por exemplo, embora seja muito freqüente no inglês, somente ocorre como preposição e advérbio. Já os verbos, que são muito freqüentes na língua, também possuem um repertório amplo de possibilidades de conversão. Como vimos, *set*, que está entre as 1000 palavras do inglês, possui 10 possibilidades de conversão. As diferenças entre as segundas e terceiras posições não são tão grandes quanto a diferença entre as primeiras e segundas posições. As segundas posições são representadas por preposições e conjunções, que são classes compostas de palavras gramaticais e não de conteúdo, o que ocorre com as terceiras posições, uma vez que são ocupadas por pronomes e substantivos. Apontadas e discutidas as três primeiras posições nas tabelas, é possível, de maneira holística, traçar similaridades e diferenças entre os resultados das duas análises. Esse processo pode ser feito ao dividir os resultados em 3 grupos.

(a) Classes que apresentam grandes diferenças na freqüência (representada pela porcentagem), mas que mantêm a mesma posição nas duas análises.

(b) Classes que apresentam grandes diferenças na freqüência e possuam posição diferente nas duas análises.

(c) Classes que apresentam a mesma freqüência e posições parecidas nas duas análises.

Com base nesses três grupos, tem-se como elementos dos grupos (a) os verbos; (b) as preposições, os pronomes, as conjunções, as partículas adverbiais; (c) os substantivos, os advérbios, os adjetivos, os artigos, os numerais e as interjeições.

Observando as 10 classes de palavras que compõem os grupos (a), (b) e (c), nota-se que elas representam 100% das classes de palavras que o etiquetador possui (cf. tabela 03 – *Relação entre as etiquetas e classes de palavras*). Essa

conclusão corrobora empiricamente a teoria de Bauer (1983) que a conversão pode ocorrer em qualquer classe de palavra.

Além da freqüência e posição das classes de palavras, é importante ressaltar um outro ponto sobre a análise da tabela 09, que mostra a freqüência absoluta de cada classe de palavra e não exatamente a direção da conversão. Por esse motivo, não se pode afirmar que a conversão entre verbo e preposição, ou entre preposição e verbo, é a mais freqüente.

Considerações Finais

Neste ponto fazemos um fechamento do trabalho, retomando os pontos principais da pesquisa, apontando limitações e fazendo sugestões de pesquisa futura e possíveis aplicações pedagógicas dos resultados.

Conforme apontado na Introdução, a conversão é um fenômeno extremamente produtivo no que concerne a formação de palavras e provavelmente essa produtividade será maior, uma vez que favorece a economia de palavras, isto é, utiliza-se uma palavra já conhecida pelo falante e a emprega de uma maneira diferente. À medida que esse fenômeno se torna cada vez mais freqüente, surgem algumas questões relacionadas a quão freqüente esse fenômeno é e quais os limites de sua utilização e produtividade. Devido à profusão de perspectivas teóricas sobre o assunto, foi preciso compilar uma definição de trabalho, que fosse compatível com a perspectiva da análise de um corpus grande por meios automáticos ou semi-automaticos. Nossa definição foi a seguinte:

A conversão é um fenômeno lingüístico, freqüente na língua inglesa, que reflete o uso de uma mesma forma ortográfica ou variante muito próxima dela em duas ou mais classes/sub-classes gramaticais diferentes, ou seja, os exemplos *Ele precisa beber*$_{\text{(verbo transitivo)}}$ *água* e *Aquela mulher adora beber*$_{\text{(verbo intransitivo)}}$. ilustram casos de conversão na mesma classe gramatical, havendo mudança de sub-classe. A conversão é um fenômeno cujos padrões léxico-gramaticais podem ser definidos claramente por intermédio de análise de corpus. Desse modo, operacionalizamos os seguintes conceitos: (1) A classe gramatical é indicada por uma etiqueta morfossintática, que é um código inserido por um etiquetador automático junto a cada forma ortográfica do corpus. Devido à profusão de etiquetas, foi preciso reduzir seu repertorio segundo os critérios apresentados no capítulo de metodologia; (2) conversão é toda a ocorrência de mais de uma etiqueta morfossintática para uma mesma forma ortográfica encontrada no corpus de estudo, o *British National Corpus*.

O estudo dos padrões é de vital importância, visto que a conversão extrapola o limite da palavra, como, por exemplo, *The rich are happy* – não é possível definir se a palavra *rich* é um adjetivo ou substantivo caso seja analisada isoladamente. A fim de refutar a concepção de que há elipse da palavra *people* em conversões de adjetivo para substantivo, alguns adjetivos gentilícios são convertidos em

substantivos pela adição do morfema *s*, tais como *American*$_{(adjetivo)}$ e *Americans*$_{(substantivo)}$, *Brazilian*$_{(adjetivo)}$ e *Brazilians*$_{(substantivo)}$. É de extrema importância ressaltar que a conversão perpassa três níveis de linguagem – morfossintático, sintático e semântico e que, pelo fato de perpassar o nível semântico, pode haver mudanças no sentido da palavra.

Este trabalho teve como objetivos justamente demonstrar empiricamente e utilizando corpus autêntico, o *British National Corpus*, de que forma a conversão ocorre e quão freqüente a conversão é na língua inglesa. Para tanto, o projeto encontrou suporte teórico na Lingüística de Corpus, que é uma área que se preocupa em analisar qualquer língua de maneira empírica, por intermédio de corpora, ou seja, compilações de textos autênticos.

Apesar de não figurar entre os objetivos iniciais deste trabalho, foram criados mecanismos mais sofisticados para análise de corpus, como a ferramenta apresentada no capítulo dedicado à metodologia. Por intermédio da criação de novas ferramentas computacionais, é possível contribuir para as pesquisas que envolvem corpora, uma vez que se torna mais fácil, rápido e confiável o manuseio e a análise de grandes quantidades de textos, o que foi essencial neste trabalho.

Por intermédio da ferramenta criada, os dados provenientes do *British National Corpus* foram selecionados, filtrados e agrupados. Em seguida, os dados obtidos foram colados no programa *Excel* no intuito de serem contabilizados e fornecerem substrato à análise.

O processo de análise consistiu na demonstração empírica e autêntica do processo de conversão em 10 classes de palavras e a análise da distribuição de sua freqüência nessas 10 classes. Tomou-se como elementos balizadores para a análise as seguintes perguntas de pesquisa:

1. Quais são as classes de palavras mais e menos passíveis de serem convertidas na língua inglesa?

2. Qual é a proporção, no corpus selecionado, entre as palavras que estão envolvidas no processo de conversão e as que não são modificadas por esse fenômeno?

3. Quais são as palavras que apresentam maior repertório de classes de palavras passíveis de serem convertidas na língua inglesa?

4. Há diferenças na freqüência do processo de conversão entre as palavras mais freqüentes e as que apresentam maior repertório de classes de palavras passíveis de serem convertidas? Em caso afirmativo, o que explica essa diferença?

Os resultados apontaram que:

(1) Os verbos, as preposições e os pronomes apresentam maior freqüência entre as palavras mais convertidas e, portanto, são mais passíveis de conversão, ao passo que os artigos, os numerais e as interjeições são as classes que apresentam menor freqüência e, portanto, são as menos passíveis de serem convertidas.

(2) Partindo de um universo de 69.257.801 *tokens*, que representam as 1000 palavras mais freqüentes no BNC, 36.861.937 *tokens* (cf. tabela 07) estejam envolvidos diretamente no processo de conversão, ou seja, aproximadamente 53% dos *tokens*.

(3) *Like*, *set* e *forecast* são as palavras pertencentes ao léxico inglês que apresentam maior repertório de classes de palavras associadas ao fenômeno da conversão.

(4) À medida que o número de palavras diferentes (tokens) aumenta, são maiores as chances de haver melhor distribuição da freqüência da conversão entre as diferentes classes das palavras. Essa observação, de certo modo, contraria a tendência das 1000 palavras com mais etiquetas, pelo fato de apresentarem maior diversidade de etiquetas, de terem a freqüência melhor distribuída entre esse maior número de classes e sub-classes de palavras.

Tomando por base os resultados encontrados, o conceito de derivação imprópria conforme entendido na nossa definição de trabalho pode ser questionado,

uma vez que a cada duas palavras na língua inglesa, uma provavelmente estará relacionada a esse fenômeno. A alta freqüência da conversão na língua inglesa pode ser justificada pelas concepções de mutabilidade da língua e arbitrariedade do signo, propostas por Saussure (1970:90/92):

> "Uma língua é radicalmente incapaz de se defender dos fatores que deslocam, de minuto a minuto, a relação entre significado e significante. É uma das conseqüências da arbitrariedade do signo...
>
> Como o signo lingüístico é arbitrário, pareceria que a língua, assim definida, é um sistema livre organizável à vontade, dependendo unicamente de um princípio racional."

Partindo do princípio que a língua é um sistema livre, a nomenclatura *derivação imprópria* não faz sentido, visto que a idéia de liberdade do sistema é contrária a qualquer juízo de valor em relação aos processos de criação de palavras. Por intermédio dos resultados obtidos, é também possível questionar a aplicabilidade da utilização da nomenclatura *derivação imprópria*, uma vez que metade das palavras encontradas no corpus estaria relacionada a um fenômeno marginal e *impróprio*, ou seja, mais de 36 milhões de palavras teriam sido formadas por um fenômeno impróprio.

Os resultados também demonstraram empiricamente que as palavras nem sempre podem ser classificadas com 100% de certeza quanto às suas classes gramaticais. Essa conclusão pode conduzir ao questionamento sobre a própria questão da existência de tantas classes e subclasses de palavras e de conseqüentemente de como e para que a língua e a gramática são estudadas. Muitas vezes, em livros didáticos, os alunos são induzidos a acreditar em uma língua que é imutável, na qual não existem ambigüidades e que há sempre uma única resposta clara para tudo. Este trabalho procurou refutar essa tendência, ao demonstrar que muitas palavras na língua inglesa perpassam por várias classes gramaticais, sem sofrer alteração aparente em sua forma.

O presente trabalho possui algumas limitações, ocasionadas pela necessidade da analise automática e semi-automática de grande quantidade de dados, o que nos distanciou do que seria aceito do ponto de vista teórico caso tivéssemos analisado manualmente um número menor de ocorrências. A principal

delas é que não foi possível separar os casos de homonímia, como a palavra *leaves* que pode significar *folha* em *All the leaves are brown* ou *sai* em *He usually leaves at 8*. A segunda limitação é que não foi possível determinar os agrupamentos de classes de palavras que são convertidas, ou seja, não foi possível determinar quais pares de classes são convertidas. Essa descoberta, realizada de forma empírica, poderia apontar que, por exemplo, 10% das conversões envolvem exclusivamente conjunções e preposições.

A terceira limitação diz respeito à não separação da ocorrência entre classes e subclasses de palavras. Caso mudanças entre subclasses de palavras, como substantivo simples e próprio, não fossem levadas em conta, provavelmente os resultados seriam diferentes. De modo similar, caso as palavras homógrafas fossem retiradas, os resultados poderiam ter sido diferentes.

Por fim, a última limitação foi a não elaboração de uma outra nomenclatura para o fenômeno estudado que não fosse conversão. Foi possível refutar algumas nomenclaturas, como derivação imprópria; no entanto, não foi possível, mas seria importante, encontrar uma outra denominação que englobasse todas as características do fenômeno estudado, segundo nossa metodologia.

Há vários aspectos que podem ser explorados em pesquisa futura. O primeiro deles é uma comparação, realizada de forma empírica entre a freqüência da conversão nas línguas inglesa e portuguesa. Esse estudo poderá mostrar as diferenças e similaridades desse processo nas duas línguas.

O segundo aspecto tem como escopo a análise das palavras formadas a partir de conversão em inglês. É possível indagar se a palavra correspondente em português é formada também por conversão. Em caso negativo, é interessante estudar e demonstrar quais outros processos de formação de palavras exercem essa função que a conversão desempenha na língua inglesa.

Por fim, é possível também traçar um paralelo na língua inglesa entre o processos de anteposição de substantivos, a formação de adjetivos e a conversão na língua inglesa, como, por exemplo, *government* em *government* job em vez de *governmental job*. Por intermédio dessa pesquisa, seria possível determinar se algum caso de conversão pode bloquear a utilização de um adjetivo já existente na língua inglesa.

Pode-se pensar em algumas possíveis aplicações pedagógicas para os

resultados desta pesquisa. A primeira delas é demonstrar nos livros de gramática, mostrar que uma determinada palavra em inglês pode assumir diversas classes gramaticais sem sofrer mudança em sua forma. A segunda possível aplicação seria, por intermédio de exercícios contrastivos, demonstrar diferenças de padronização léxico-gramatical entre a língua materna e a alvo.

O trabalho aqui apresentado espera, assim, ter preenchido uma lacuna importante na literatura ao demonstrar empiricamente e utilizar corpus autêntico quão freqüente a conversão é na língua inglesa.

Referências bibliográficas

Adams, V. (1976). *An Introduction to Modern English Word Formation*. Londres: Longman.

Aitchison, J. (1989). *Words in the Mind - An Introduction to the Mental Lexicon*. Oxford: Basil Blackwell.

Bally, C. (1965). *Linguistique générale et linguistique française*. Suisse: Editions Francke Berne.

Bartolomé, A. & Cabrera, G. (2005). *Grammatical Conversion in English: Some new trends in lexical evolution*. Disponível no endereço: http://accurapid.com/journal/31conversion.htm. Acessado em 21/03/2006.

Bauer, L. (1983). *English Word-formation*. Cambridge: University Press.

Berber Sardinha, A. P. (2000). *Lingüística de Corpus: Histórico e Problemática*. **D.E.L.T.A., 16** (2), 323-367.

Berber Sardinha, A. P. (2004). *Lingüística de Corpus*. São Paulo: Editora Manole.

Biber, D., Conrad, S. e Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure And Use*. Cambridge: Cambridge University Press.

Celani, M.A.A. (1992) 'Afinal, o que é Lingüística Aplicada?'. In: Paschoal, M. S. Z. de e M.A.A.Celani (orgs.) (1992) *Lingüística Aplicada: da Aplicação da Lingüística à Lingüística Transdisciplinar*. São Paulo: Educ.

Cambridge Dictionaries Online. Disponível na Internet no endereço http://dictionary.cambridge.org. Acessado em 07/06/2006.

Cannon, G. (1985). *Functional Shift in English*. [S.1]: Linguistics.

Dokulil, M. (1968). Sobre a questão da assim chamada derivação-zero. In Brekle H.E. & l. Lipka (eds.). *Formação de Palavras, Sintaxe e Morfologia*. Publicação em homenagem a Hans Marchand.

Dubois, J. et al (1998). *Dicionário de Lingüística*. São Paulo: Cultrix.

Francis, G. & S. Hunston (1996). *Grammar Patterns 1: Verbs*. London: HarperCollins, COBUILD.

Francis, G. & S. Hunston (1998) *Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins, COBUILD.

Francis, G. & S. Hunston (1999). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins Publishing Company

Görlach, M. (1982). *Introdução à História da Língua Inglesa*. Heidelberg: Quelle & Meyer.

Grzega, J. (2002). *Some thoughts on a cognitive onomasiological approach to word-formation with special reference to Engish*. Disponível em http://www1.ku-eichstaett.de/SLF/EngluVglSW/grzega1023.pdf. Acessado em 10/04/2006.

Halliday, M. A. K. (1993). Quantitative studies and probabilities in grammar. In Michael Hoey (Ed.), *Data, description, discourse* (1-25). London: HarperCollins.

Hoey, M. (1997). From concordance to text structure: new uses for computer corpora. In B. Lewandoswka-Tomaszczyk e P. J. Melia (org.). *PALC'97 – Practical Applications in Language Corpora*. Lódz: Lódz University Press.

Hummel, M. (2003). A conversão do adjetivo em advérbio em perspectiva sincrônica e diacrônica. Disponível em http://www.geocities.com/ail_br/aconversaodoadjectivoem.htm. Acessado em 10/04/2006.

Karaščuk, P. (1977). *Vysšaja škola*. Moscou: Krajina.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Londres: Longman.

Leech (1992). Corpora and theories of linguistic performance. In: J. SVARTVIK (org.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin, New York: De Gruyter.

Marchand, H. (1969). *The Categories and Types of Present-Day English Word-Formation.* München: C. H. Becksche Verlagsbuchhandlung.

McEnery, T. e Wilson, A. (1996). *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Postnikova, A. (1999). *Word-formation of the English language. Conversion*. Disponível em http://referatw.ru/cgi-bin/main.cgi?level=5&p1=89&p2=8&p3=17181. Acessado em 10/04/2006.

Sandmann, A. (1996). *Formação de palavras no português brasileiro contemporâneo*. Curitiba: UFPR.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smirnickij, A. (1954). Po povodu konversiji v anglijskom jazyke. *Inostrannyje jazyky v ✶kole 3*.

Stein, G. (1977). The Place of Word-Formation in Linguistic Description", in H. E. Brekle und D. Kastovsky (eds.): *Perspektiven der Wortbildungsforschung. Beiträge zum Wuppertaler Wortbildungskolloquium.* Bonn: Bouvier Verlag Herbert Grundmann.

Štekauer, P. (1996). *A Theory of Conversion in English*. Frankfurt am Main: Peter Lang.

Swan, M. (1980). *Practical English usage*. Oxford: Oxford U.P.

Valera, S. (2000). Conversion and onomasiological theory. Disponível em http://journals.cambridge.org/article_S0022226799007951. Acessado em 21/03/2006.

Valera, S. (2004). *Conversion vs. unmarked word-class change*. Disponível em www.skase.sk/Volumes/JTL01/valera.pdf. Acessado em 21/03/2006.

Anexos

Anexo 1: As mil palavras mais freqüentes no *British National Corpus*

Posição	Palavra	Freq.	%	Posição	Palavra	Freq.	%
1	THE	6.197.320	6,16	**501**	LOW	20.531	0,02
2	OF	3.112.520	3,09	**502**	SOMETIMES	20.519	0,02
3	AND	2.692.822	2,68	**503**	THUS	20.510	0,02
4	TO	2.671.334	2,66	**504**	COMING	20.473	0,02
5	A	2.238.218	2,22	**505**	COMMON	20.454	0,02
6	IN	1.993.249	1,98	**506**	MOVE	20.449	0,02
7	THAT	1.078.710	1,07	**507**	GETTING	20.398	0,02
8	IS	986.173	0,98	**508**	SINGLE	20.378	0,02
9	IT	951.433	0,95	**509**	V	20.368	0,02
10	WAS	901.242	0,9	**510**	BROUGHT	20.361	0,02
11	FOR	900.148	0,89	**511**	CHURCH	20.323	0,02
12	I	757.695	0,75	**512**	WORD	20.316	0,02
13	ON	753.574	0,75	**513**	NEEDS	20.229	0,02
14	WITH	677.928	0,67	**514**	HEARD	20.198	0,02
15	AS	673.181	0,67	**515**	SHALL	20.191	0,02
16	BE	665.935	0,66	**516**	BRITAIN	20.114	0,02
17	HE	629.427	0,63	**517**	ELSE	20.079	0,02
18	YOU	610.551	0,61	**518**	INVOLVED	20.016	0,02
19	AT	539.395	0,54	**519**	STREET	20.014	0,02
20	BY	524.610	0,52	**520**	RATE	19.997	0,02
21	ARE	464.595	0,46	**521**	HUMAN	19.976	0,02
22	THIS	463.066	0,46	**522**	CALL	19.900	0,02
23	BUT	460.749	0,46	**523**	AGO	19.863	0,02
24	HAVE	460.223	0,46	**524**	WHOSE	19.848	0,02
25	NOT	444.186	0,44	**525**	I'D	19.781	0,02
26	FROM	436.106	0,43	**526**	BUILDING	19.651	0,02
27	HAD	435.867	0,43	**527**	INDUSTRY	19.622	0,02
28	HIS	433.515	0,43	**528**	GROUPS	19.567	0,02
29	THEY	387.146	0,38	**529**	CENTRAL	19.515	0,02
30	OR	376.088	0,37	**530**	EIGHT	19.507	0,02
31	WHICH	372.249	0,37	**531**	LOST	19.507	0,02
32	SHE	349.087	0,35	**532**	GONE	19.491	0,02
33	AN	345.645	0,34	**533**	ASK	19.412	0,02
34	HER	326.655	0,32	**534**	HISTORY	19.409	0,02
35	WERE	317.988	0,32	**535**	TERM	19.382	0,02
36	ONE	308.207	0,31	**536**	SELF	19.364	0,02
37	WE	306.572	0,3	**537**	FOOD	19.282	0,02
38	THERE	294.236	0,29	**538**	UNITED	19.232	0,02
39	ALL	291.124	0,29	**539**	SEVEN	19.228	0,02
40	BEEN	268.515	0,27	**540**	LANGUAGE	19.226	0,02
41	IF	261.423	0,26	**541**	PRICE	19.202	0,02
42	THEIR	260.946	0,26	**542**	HUNDRED	19.201	0,02
43	HAS	255.287	0,25	**543**	CHANGES	19.196	0,02
44	WILL	254.967	0,25	**544**	SECTION	19.196	0,02

45	SO	250.384	0,25	**545**	HANDS	19.178	0,02
46	NO	239.348	0,24	**546**	TEAM	19.171	0,02
47	WOULD	238.777	0,24	**547**	USUALLY	19.168	0,02
48	WHAT	233.577	0,23	**548**	NEAR	19.137	0,02
49	UP	229.514	0,23	**549**	WE'RE	19.091	0,02
50	WHEN	216.363	0,22	**550**	REMEMBER	19.074	0,02
51	MORE	214.884	0,21	**551**	SIR	19.062	0,02
52	CAN	214.511	0,21	**552**	PROGRAMME	19.048	0,02
53	OUT	210.479	0,21	**553**	CUT	19.029	0,02
54	SAID	208.283	0,21	**554**	HOURS	19.029	0,02
55	WHO	198.542	0,2	**555**	COMMITTEE	18.945	0,02
56	ABOUT	198.402	0,2	**556**	INDEED	18.891	0,02
57	DO	183.249	0,18	**557**	TYPE	18.826	0,02
58	THEM	172.979	0,17	**558**	REASON	18.754	0,02
59	SOME	171.309	0,17	**559**	CERTAINLY	18.651	0,02
60	TIME	169.053	0,17	**560**	TRYING	18.624	0,02
61	HIM	164.711	0,16	**561**	VALUE	18.614	0,02
62	TWO	164.645	0,16	**562**	DECISION	18.605	0,02
63	INTO	163.635	0,16	**563**	DATA	18.602	0,02
64	ITS	163.353	0,16	**564**	SIMILAR	18.521	0,02
65	THEN	160.391	0,16	**565**	ROLE	18.503	0,02
66	LIKE	157.514	0,16	**566**	MEMBER	18.448	0,02
67	OTHER	156.181	0,16	**567**	SAYING	18.411	0,02
68	WELL	154.654	0,15	**568**	INDIVIDUAL	18.400	0,02
69	ONLY	153.227	0,15	**569**	MINUTES	18.382	0,02
70	MY	152.310	0,15	**570**	CASES	18.357	0,02
71	COULD	150.368	0,15	**571**	YESTERDAY	18.326	0,02
72	THAN	148.329	0,15	**572**	TRUE	18.297	0,02
73	NOW	143.971	0,14	**573**	PRIVATE	18.273	0,02
74	OVER	140.175	0,14	**574**	STATES	18.212	0,02
75	ME	138.856	0,14	**575**	EVERYTHING	18.165	0,02
76	YOUR	138.401	0,14	**576**	NECESSARY	18.129	0,02
77	JUST	130.610	0,13	**577**	DATE	18.098	0,02
78	IT'S	129.469	0,13	**578**	ISN'T	18.053	0,02
79	MAY	129.163	0,13	**579**	TAX	18.010	0,02
80	NEW	126.929	0,13	**580**	M	18.009	0,02
81	FIRST	126.923	0,13	**581**	FIGURE	17.982	0,02
82	THESE	125.508	0,12	**582**	NATURE	17.967	0,02
83	ALSO	125.019	0,12	**583**	AUTHORITY	17.934	0,02
84	ANY	124.838	0,12	**584**	COMPANIES	17.886	0,02
85	KNOW	123.006	0,12	**585**	SOMEONE	17.817	0,02
86	VERY	122.970	0,12	**586**	FRENCH	17.807	0,02
87	SEE	118.853	0,12	**587**	TOWN	17.799	0,02
88	PEOPLE	118.526	0,12	**588**	RETURN	17.771	0,02
89	AFTER	117.792	0,12	**589**	SIMPLY	17.771	0,02
90	SUCH	109.333	0,11	**590**	ESPECIALLY	17.767	0,02
91	SHOULD	106.959	0,11	**591**	PERSONAL	17.739	0,02
92	WHERE	106.015	0,11	**592**	SHE'S	17.733	0,02
93	BACK	105.105	0,1	**593**	HOPE	17.670	0,02
94	BECAUSE	102.830	0,1	**594**	TOTAL	17.640	0,02

95	HOW	101.772	0,1	**595**	UNION	17.597	0,02
96	WAY	100.719	0,1	**596**	DOESN'T	17.592	0,02
97	MOST	100.086	0,1	**597**	PAPER	17.531	0,02
98	GET	99.082	0,1	**598**	BANK	17.495	0,02
99	DOWN	98.231	0,1	**599**	STARTED	17.484	0,02
100	DON'T	96.963	0,1	**600**	PATIENTS	17.473	0,02
101	DID	96.300	0,1	**601**	SYSTEMS	17.458	0,02
102	MADE	96.268	0,1	**602**	LIVE	17.436	0,02
103	RIGHT	95.999	0,1	**603**	CO	17.408	0,02
104	OUR	94.997	0,09	**604**	DEAL	17.398	0,02
105	MUCH	93.585	0,09	**605**	AMERICAN	17.396	0,02
106	WORK	92.643	0,09	**606**	COULDN'T	17.347	0,02
107	BETWEEN	92.225	0,09	**607**	DEPARTMENT	17.299	0,02
108	THINK	91.471	0,09	**608**	HERSELF	17.211	0,02
109	GO	91.073	0,09	**609**	LONGER	17.211	0,02
110	EVEN	90.987	0,09	**610**	EUROPE	17.168	0,02
111	MANY	90.608	0,09	**611**	PRACTICE	17.155	0,02
112	BEING	90.558	0,09	**612**	RED	17.155	0,02
113	YEARS	90.527	0,09	**613**	STAGE	17.132	0,02
114	ER	90.333	0,09	**614**	PRESIDENT	17.122	0,02
115	THOSE	88.801	0,09	**615**	PROVIDED	17.084	0,02
116	GOT	88.604	0,09	**616**	TALK	17.053	0,02
117	BEFORE	88.304	0,09	**617**	FORMER	17.052	0,02
118	THREE	86.519	0,09	**618**	SEEM	16.994	0,02
119	YEAR	84.631	0,08	**619**	STRONG	16.969	0,02
120	GOOD	84.600	0,08	**620**	INCREASE	16.918	0,02
121	THROUGH	84.361	0,08	**621**	QUALITY	16.910	0,02
122	US	82.790	0,08	**622**	UK	16.908	0,02
123	YEAH	82.406	0,08	**623**	GROUND	16.858	0,02
124	MAKE	81.145	0,08	**624**	EXPECTED	16.842	0,02
125	OFF	77.337	0,08	**625**	COUNTRIES	16.817	0,02
126	LAST	77.105	0,08	**626**	NEEDED	16.765	0,02
127	STILL	74.976	0,07	**627**	MAKES	16.750	0,02
128	TAKE	72.802	0,07	**628**	FINANCIAL	16.726	0,02
129	MUST	71.408	0,07	**629**	EARLIER	16.712	0,02
130	OWN	71.275	0,07	**630**	ACCOUNT	16.699	0,02
131	TOO	70.396	0,07	**631**	FINAL	16.690	0,02
132	COME	69.473	0,07	**632**	ISSUE	16.663	0,02
133	SAY	69.058	0,07	**633**	R	16.659	0,02
134	HERE	68.914	0,07	**634**	HIGHER	16.651	0,02
135	BOTH	68.791	0,07	**635**	LIVING	16.651	0,02
136	OH	67.804	0,07	**636**	SECRETARY	16.628	0,02
137	THAT'S	67.683	0,07	**637**	WOULDN'T	16.609	0,02
138	USED	67.381	0,07	**638**	WE'VE	16.582	0,02
139	LONG	67.344	0,07	**639**	FOREIGN	16.547	0,02
140	MR	67.181	0,07	**640**	BED	16.499	0,02
141	DAY	67.163	0,07	**641**	COS	16.488	0,02
142	GOING	66.644	0,07	**642**	PARENTS	16.445	0,02
143	OLD	65.624	0,07	**643**	REQUIRED	16.443	0,02
144	LITTLE	65.500	0,07	**644**	CONCERNED	16.311	0,02

145	I'M	65.271	0,06	**645**	MARCH	16.307	0,02
146	UNDER	64.748	0,06	**646**	WIFE	16.267	0,02
147	USE	63.554	0,06	**647**	REST	16.264	0,02
148	ERM	62.781	0,06	**648**	APPROACH	16.260	0,02
149	SAME	62.588	0,06	**649**	LED	16.253	0,02
150	EACH	61.635	0,06	**650**	FORCE	16.242	0,02
151	MIGHT	61.509	0,06	**651**	WRONG	16.162	0,02
152	HOWEVER	60.576	0,06	**652**	UNIVERSITY	16.137	0,02
153	ANOTHER	60.478	0,06	**653**	FRIEND	16.122	0,02
154	AGAIN	60.125	0,06	**654**	ART	16.073	0,02
155	YES	59.968	0,06	**655**	WON'T	16.013	0,02
156	MAN	59.848	0,06	**656**	SITUATION	16.009	0,02
157	PUT	59.509	0,06	**657**	PRODUCTION	15.943	0,02
158	LIFE	58.795	0,06	**658**	COMES	15.928	0,02
159	GOVERNMENT	57.908	0,06	**659**	OFFER	15.922	0,02
160	PART	57.459	0,06	**660**	STOP	15.919	0,02
161	WANT	57.321	0,06	**661**	PAID	15.914	0,02
162	THOUGHT	56.941	0,06	**662**	ACCORDING	15.879	0,02
163	WORLD	56.876	0,06	**663**	D	15.879	0,02
164	AGAINST	56.687	0,06	**664**	RECENT	15.875	0,02
165	WHILE	56.673	0,06	**665**	FIELD	15.851	0,02
166	NEVER	56.090	0,06	**666**	LORD	15.741	0,02
167	NEED	56.056	0,06	**667**	CLUB	15.739	0,02
168	LOOK	54.645	0,05	**668**	MIDDLE	15.738	0,02
169	HOME	54.189	0,05	**669**	G	15.710	0,02
170	PER	54.143	0,05	**670**	BAD	15.677	0,02
171	HOUSE	52.777	0,05	**671**	FORWARD	15.613	0,02
172	SOMETHING	52.180	0,05	**672**	ADDED	15.597	0,02
173	END	51.872	0,05	**673**	DUE	15.578	0,02
174	FOUR	51.234	0,05	**674**	BOARD	15.544	0,02
175	COURSE	50.950	0,05	**675**	MONTH	15.543	0,02
176	DOES	50.933	0,05	**676**	SCHOOLS	15.542	0,02
177	AWAY	50.764	0,05	**677**	CONDITIONS	15.523	0,02
178	WHY	50.457	0,05	**678**	VARIOUS	15.523	0,02
179	SINCE	50.231	0,05	**679**	UNDERSTAND	15.502	0,02
180	PLACE	49.958	0,05	**680**	GREEN	15.501	0,02
181	NUMBER	49.531	0,05	**681**	GREATER	15.476	0,02
182	LEFT	49.396	0,05	**682**	BRING	15.456	0,02
183	FOUND	49.147	0,05	**683**	RESULTS	15.444	0,02
184	GREAT	48.516	0,05	**684**	WEEKS	15.429	0,02
185	DIFFERENT	48.452	0,05	**685**	AMOUNT	15.427	0,02
186	WENT	48.165	0,05	**686**	CLEARLY	15.405	0,02
187	REALLY	47.912	0,05	**687**	RECORD	15.394	0,02
188	CAME	47.062	0,05	**688**	WHAT'S	15.376	0,02
189	LOCAL	46.666	0,05	**689**	DE	15.313	0,02
190	SET	46.652	0,05	**690**	AWARD	15.304	0,02
191	CASE	46.544	0,05	**691**	COSTS	15.266	0,02
192	DIDN'T	46.474	0,05	**692**	POOR	15.242	0,02
193	FIVE	46.428	0,05	**693**	INCLUDE	15.233	0,02
194	WITHIN	46.353	0,05	**694**	MUSIC	15.232	0,02

195	NEXT	46.328	0,05	**695**	POST	15.203	0,02
196	ALWAYS	46.179	0,05	**696**	PROJECT	15.197	0,02
197	WITHOUT	46.178	0,05	**697**	TRIED	15.195	0,02
198	THOUGH	45.597	0,05	**698**	PLAN	15.178	0,02
199	SMALL	45.432	0,05	**699**	DAVID	15.168	0,02
200	AROUND	45.267	0,04	**700**	EASY	15.164	0,02
201	GIVE	45.147	0,04	**701**	DECIDED	15.158	0,02
202	HIGH	45.114	0,04	**702**	FRIENDS	15.149	0,02
203	FEW	45.029	0,04	**703**	NINE	15.091	0,02
204	SYSTEM	44.955	0,04	**704**	SHOWN	15.038	0,01
205	DURING	44.049	0,04	**705**	GAME	15.024	0,01
206	SECOND	43.707	0,04	**706**	HOSPITAL	15.015	0,01
207	ALTHOUGH	43.688	0,04	**707**	WORKERS	15.009	0,01
208	MEAN	43.595	0,04	**708**	GIRL	15.007	0,01
209	RATHER	42.878	0,04	**709**	WAYS	14.983	0,01
210	SOCIAL	42.860	0,04	**710**	APRIL	14.935	0,01
211	THINGS	42.355	0,04	**711**	FOLLOWED	14.905	0,01
212	FIND	42.341	0,04	**712**	CHAPTER	14.903	0,01
213	CHILDREN	42.196	0,04	**713**	ROYAL	14.869	0,01
214	GIVEN	41.985	0,04	**714**	DESCRIBED	14.859	0,01
215	POINT	41.879	0,04	**715**	JUNE	14.854	0,01
216	FAR	41.783	0,04	**716**	LEAD	14.843	0,01
217	UNTIL	41.486	0,04	**717**	GOES	14.806	0,01
218	QUITE	41.035	0,04	**718**	SOUND	14.803	0,01
219	GENERAL	40.878	0,04	**719**	PLANNING	14.800	0,01
220	GROUP	40.823	0,04	**720**	KNOWLEDGE	14.794	0,01
221	EVERY	40.340	0,04	**721**	ST	14.782	0,01
222	STATE	40.005	0,04	**722**	FIRE	14.780	0,01
223	PARTY	39.735	0,04	**723**	ANSWER	14.745	0,01
224	LESS	39.734	0,04	**724**	NEWS	14.731	0,01
225	SAYS	39.719	0,04	**725**	AGREED	14.722	0,01
226	IMPORTANT	39.648	0,04	**726**	SERIES	14.711	0,01
227	LATER	39.506	0,04	**727**	POINTS	14.680	0,01
228	PUBLIC	39.447	0,04	**728**	STUDENTS	14.669	0,01
229	HAND	39.176	0,04	**729**	HAIR	14.655	0,01
230	TOOK	39.116	0,04	**730**	CURRENT	14.632	0,01
231	INFORMATION	39.085	0,04	**731**	ANYONE	14.629	0,01
232	HEAD	38.900	0,04	**732**	DESPITE	14.622	0,01
233	ONCE	38.881	0,04	**733**	INSTEAD	14.622	0,01
234	HALF	38.836	0,04	**734**	COMPUTER	14.598	0,01
235	CENT	38.647	0,04	**735**	WORKS	14.575	0,01
236	HELP	38.409	0,04	**736**	QUESTIONS	14.554	0,01
237	YOU'RE	38.360	0,04	**737**	MOVED	14.531	0,01
238	NATIONAL	38.286	0,04	**738**	BUILT	14.519	0,01
239	SCHOOL	38.274	0,04	**739**	NATURAL	14.503	0,01
240	OFTEN	37.915	0,04	**740**	BASIS	14.487	0,01
241	MONEY	37.872	0,04	**741**	RUNNING	14.454	0,01
242	FACT	37.726	0,04	**742**	BELOW	14.433	0,01
243	I'VE	37.625	0,04	**743**	WIDE	14.429	0,01
244	SEEN	37.625	0,04	**744**	INSIDE	14.408	0,01

245	MEN	37.435	0,04	**745**	SEA	14.394	0,01
246	TOLD	37.168	0,04	**746**	KEPT	14.376	0,01
247	BEST	36.857	0,04	**747**	PLEASE	14.369	0,01
248	NIGHT	36.724	0,04	**748**	ALLOWED	14.332	0,01
249	BRITISH	36.525	0,04	**749**	ALONE	14.281	0,01
250	EXAMPLE	36.494	0,04	**750**	HEART	14.262	0,01
251	COMPANY	36.493	0,04	**751**	MET	14.212	0,01
252	LARGE	36.457	0,04	**752**	DARK	14.178	0,01
253	WATER	36.435	0,04	**753**	TEST	14.170	0,01
254	FURTHER	36.323	0,04	**754**	FEET	14.136	0,01
255	BETTER	36.297	0,04	**755**	MEET	14.125	0,01
256	CALLED	36.136	0,04	**756**	SIMPLE	14.071	0,01
257	BUSINESS	36.096	0,04	**757**	SECURITY	14.005	0,01
258	WHETHER	36.074	0,04	**758**	LETTER	13.980	0,01
259	FACE	35.983	0,04	**759**	WRITTEN	13.976	0,01
260	P	35.690	0,04	**760**	DIED	13.968	0,01
261	TAKEN	35.611	0,04	**761**	STRUCTURE	13.927	0,01
262	DONE	35.385	0,04	**762**	STORY	13.908	0,01
263	HAVING	35.268	0,04	**763**	HAPPENED	13.881	0,01
264	ORDER	35.264	0,04	**764**	HEAR	13.858	0,01
265	YET	35.202	0,03	**765**	CARRIED	13.798	0,01
266	AREA	35.086	0,03	**766**	SHARE	13.793	0,01
267	LOOKED	35.074	0,03	**767**	HOLD	13.790	0,01
268	FORM	35.061	0,03	**768**	ATTENTION	13.758	0,01
269	PERHAPS	35.025	0,03	**769**	OPERATION	13.747	0,01
270	SIDE	34.999	0,03	**770**	SENT	13.735	0,01
271	THING	34.882	0,03	**771**	STUDIES	13.732	0,01
272	WOMEN	34.666	0,03	**772**	CAPITAL	13.723	0,01
273	EARLY	34.406	0,03	**773**	MANAGER	13.661	0,01
274	ALREADY	34.321	0,03	**774**	MATERIAL	13.626	0,01
275	POSSIBLE	34.258	0,03	**775**	EVENING	13.610	0,01
276	LONDON	34.146	0,03	**776**	CONSIDERED	13.604	0,01
277	FAMILY	34.047	0,03	**777**	THINKING	13.571	0,01
278	ROOM	33.955	0,03	**778**	TALKING	13.564	0,01
279	NOTHING	33.845	0,03	**779**	FINE	13.532	0,01
280	MM	33.657	0,03	**780**	MODERN	13.513	0,01
281	YOUNG	33.528	0,03	**781**	BILL	13.466	0,01
282	FULL	33.466	0,03	**782**	PERFORMANCE	13.464	0,01
283	ASKED	33.393	0,03	**783**	SUCCESS	13.461	0,01
284	ROUND	33.294	0,03	**784**	SPACE	13.447	0,01
285	SIX	33.243	0,03	**785**	MOVEMENT	13.413	0,01
286	DAYS	33.151	0,03	**786**	MODEL	13.412	0,01
287	POWER	33.090	0,03	**787**	LEGAL	13.393	0,01
288	THERE'S	33.022	0,03	**788**	ANALYSIS	13.392	0,01
289	HE'S	32.988	0,03	**789**	CHANCE	13.363	0,01
290	DEVELOPMENT	32.587	0,03	**790**	RECEIVED	13.319	0,01
291	ENOUGH	32.587	0,03	**791**	AGREEMENT	13.318	0,01
292	CHANGE	32.253	0,03	**792**	BOOKS	13.318	0,01
293	OPEN	32.235	0,03	**793**	CAUSE	13.293	0,01
294	WORKING	31.769	0,03	**794**	PRESS	13.278	0,01

295	WEEK	31.618	0,03	**795**	KING	13.261	0,01
296	ALMOST	31.605	0,03	**796**	PRODUCED	13.247	0,01
297	SERVICE	31.595	0,03	**797**	POPULATION	13.244	0,01
298	JOHN	31.567	0,03	**798**	STANDARD	13.239	0,01
299	CAN'T	31.464	0,03	**799**	STOOD	13.226	0,01
300	MAKING	31.379	0,03	**800**	T	13.175	0,01
301	LEAST	31.242	0,03	**801**	THEORY	13.173	0,01
302	MARKET	31.176	0,03	**802**	SIZE	13.162	0,01
303	WAR	31.117	0,03	**803**	FEELING	13.147	0,01
304	HIMSELF	31.062	0,03	**804**	ENVIRONMENT	13.146	0,01
305	SUPPORT	31.048	0,03	**805**	GROWTH	13.121	0,01
306	POLITICAL	31.005	0,03	**806**	WHATEVER	13.115	0,01
307	WHOLE	31.005	0,03	**807**	HOUR	13.097	0,01
308	C	30.926	0,03	**808**	RIGHTS	13.096	0,01
309	TOGETHER	30.921	0,03	**809**	NICE	13.076	0,01
310	TELL	30.886	0,03	**810**	DESIGN	13.067	0,01
311	ABLE	30.622	0,03	**811**	INCREASED	13.047	0,01
312	BECOME	30.400	0,03	**812**	HAVEN'T	13.041	0,01
313	TIMES	30.168	0,03	**813**	FINALLY	13.012	0,01
314	MEMBERS	29.905	0,03	**814**	KEY	13.009	0,01
315	CONTROL	29.802	0,03	**815**	BUY	13.001	0,01
316	EYES	29.727	0,03	**816**	AUTHORITIES	12.987	0,01
317	PROBLEM	29.445	0,03	**817**	LIST	12.983	0,01
318	COUNCIL	29.344	0,03	**818**	RELATIONSHIP	12.959	0,01
319	LAW	29.290	0,03	**819**	WHOM	12.955	0,01
320	PARTICULAR	29.191	0,03	**820**	GIVING	12.917	0,01
321	EVER	29.141	0,03	**821**	LOWER	12.895	0,01
322	LEVEL	29.036	0,03	**822**	COMPLETE	12.855	0,01
323	COUNTRY	29.014	0,03	**823**	SON	12.834	0,01
324	ANYTHING	28.846	0,03	**824**	DEAD	12.831	0,01
325	MAJOR	28.803	0,03	**825**	WORKED	12.831	0,01
326	LOOKING	28.763	0,03	**826**	HE'D	12.809	0,01
327	DOING	28.720	0,03	**827**	ESTABLISHED	12.797	0,01
328	LOT	28.609	0,03	**828**	PARTIES	12.769	0,01
329	TOWARDS	28.601	0,03	**829**	CONTINUED	12.747	0,01
330	MIND	28.528	0,03	**830**	STAY	12.746	0,01
331	COURT	28.474	0,03	**831**	ENERGY	12.734	0,01
332	INTEREST	28.400	0,03	**832**	STYLE	12.731	0,01
333	I'LL	28.375	0,03	**833**	DEVELOPED	12.719	0,01
334	BIT	28.303	0,03	**834**	PROPERTY	12.704	0,01
335	OTHERS	28.282	0,03	**835**	BOY	12.687	0,01
336	HELD	28.217	0,03	**836**	VISIT	12.663	0,01
337	CAR	28.189	0,03	**837**	COLD	12.649	0,01
338	TOP	28.118	0,03	**838**	MISS	12.644	0,01
339	FELT	28.073	0,03	**839**	WORTH	12.601	0,01
340	KNOWN	27.953	0,03	**840**	NORMAL	12.586	0,01
341	POLICE	27.926	0,03	**841**	DR	12.568	0,01
342	KEEP	27.871	0,03	**842**	MYSELF	12.533	0,01
343	ROAD	27.858	0,03	**843**	CUP	12.531	0,01
344	EITHER	27.784	0,03	**844**	INCOME	12.516	0,01

345	REPORT	27.630	0,03	**845**	PRIME	12.498	0,01
346	PROBLEMS	27.544	0,03	**846**	NOR	12.475	0,01
347	ACT	27.529	0,03	**847**	WALL	12.475	0,01
348	SORT	27.410	0,03	**848**	SERIOUS	12.473	0,01
349	AVAILABLE	27.289	0,03	**849**	RISK	12.468	0,01
350	PROBABLY	27.222	0,03	**850**	DIRECTOR	12.435	0,01
351	BASED	27.212	0,03	**851**	THROUGHOUT	12.432	0,01
352	RESEARCH	27.205	0,03	**852**	RECENTLY	12.424	0,01
353	AM	27.088	0,03	**853**	REACHED	12.411	0,01
354	FOLLOWING	27.037	0,03	**854**	TREATMENT	12.393	0,01
355	LET	26.908	0,03	**855**	QUICKLY	12.388	0,01
356	SHOW	26.904	0,03	**856**	EVERYONE	12.386	0,01
357	TERMS	26.896	0,03	**857**	MARK	12.382	0,01
358	MEANS	26.833	0,03	**858**	BEHAVIOUR	12.368	0,01
359	POLICY	26.831	0,03	**859**	INCLUDED	12.360	0,01
360	SAW	26.818	0,03	**860**	LEVELS	12.354	0,01
361	PAST	26.717	0,03	**861**	DOUBT	12.343	0,01
362	LINE	26.711	0,03	**862**	RE	12.321	0,01
363	NAME	26.530	0,03	**863**	BEGINNING	12.316	0,01
364	WHITE	26.449	0,03	**864**	GONNA	12.298	0,01
365	FEEL	26.398	0,03	**865**	PREVIOUS	12.298	0,01
366	B	26.379	0,03	**866**	ISSUES	12.289	0,01
367	QUESTION	26.355	0,03	**867**	YOU'LL	12.265	0,01
368	SOUTH	26.281	0,03	**868**	THANK	12.256	0,01
369	CLEAR	26.185	0,03	**869**	SAT	12.250	0,01
370	EDUCATION	26.159	0,03	**870**	PRESSURE	12.243	0,01
371	BLACK	26.143	0,03	**871**	SOVIET	12.234	0,01
372	LABOUR	26.082	0,03	**872**	CHOICE	12.232	0,01
373	DOOR	26.033	0,03	**873**	DEFENCE	12.232	0,01
374	KNEW	26.005	0,03	**874**	POUNDS	12.224	0,01
375	BIG	25.975	0,03	**875**	FLOOR	12.220	0,01
376	ABOVE	25.929	0,03	**876**	CHARGE	12.208	0,01
377	ACTUALLY	25.926	0,03	**877**	HALL	12.201	0,01
378	OFFICE	25.889	0,03	**878**	OKAY	12.192	0,01
379	PRESENT	25.852	0,03	**879**	SPECIFIC	12.177	0,01
380	VIEW	25.701	0,03	**880**	ANYWAY	12.169	0,01
381	BODY	25.464	0,03	**881**	JULY	12.155	0,01
382	CARE	25.433	0,03	**882**	RESPONSE	12.150	0,01
383	VOICE	25.406	0,03	**883**	SIGNIFICANT	12.125	0,01
384	MAIN	25.354	0,03	**884**	SCOTLAND	12.103	0,01
385	BOOK	25.258	0,03	**885**	COUPLE	12.088	0,01
386	HEALTH	25.233	0,03	**886**	INDUSTRIAL	12.088	0,01
387	ACROSS	25.224	0,03	**887**	SCHEME	12.077	0,01
388	MILLION	25.104	0,02	**888**	REPORTED	12.074	0,01
389	NORTH	25.074	0,02	**889**	CONTRACT	12.064	0,01
390	SERVICES	25.051	0,02	**890**	TECHNOLOGY	12.047	0,01
391	SURE	25.047	0,02	**891**	COLOUR	11.976	0,01
392	THEY'RE	25.002	0,02	**892**	X	11.972	0,01
393	HARD	24.941	0,02	**893**	DIRECT	11.957	0,01
394	MONTHS	24.858	0,02	**894**	GERMAN	11.956	0,01

395	TURNED	24.816	0,02	**895**	LOSS	11.951	0,01
396	USING	24.776	0,02	**896**	WISH	11.884	0,01
397	LIGHT	24.728	0,02	**897**	FRANCE	11.877	0,01
398	E	24.643	0,02	**898**	DETAILS	11.859	0,01
399	CENTURY	24.633	0,02	**899**	TAKES	11.837	0,01
400	WORDS	24.596	0,02	**900**	WON	11.837	0,01
401	KIND	24.495	0,02	**901**	HAPPY	11.833	0,01
402	PERIOD	24.441	0,02	**902**	WRITING	11.828	0,01
403	ECONOMIC	24.437	0,02	**903**	STAND	11.825	0,01
404	ALONG	24.425	0,02	**904**	BEYOND	11.822	0,01
405	FREE	24.382	0,02	**905**	FORCES	11.803	0,01
406	RUN	24.380	0,02	**906**	SUDDENLY	11.800	0,01
407	START	24.241	0,02	**907**	CONTINUE	11.795	0,01
408	INCLUDING	24.184	0,02	**908**	SPENT	11.787	0,01
409	PERSON	24.146	0,02	**909**	CHAIRMAN	11.782	0,01
410	SEVERAL	24.054	0,02	**910**	OPENED	11.773	0,01
411	SHORT	23.984	0,02	**911**	BLUE	11.768	0,01
412	WEST	23.937	0,02	**912**	MILITARY	11.752	0,01
413	ENGLISH	23.896	0,02	**913**	CHIEF	11.747	0,01
414	REAL	23.895	0,02	**914**	CONSIDER	11.742	0,01
415	CENTRE	23.864	0,02	**915**	PLANS	11.740	0,01
416	MOTHER	23.846	0,02	**916**	PETER	11.737	0,01
417	LOVE	23.799	0,02	**917**	ONES	11.736	0,01
418	BEHIND	23.753	0,02	**918**	FORMS	11.730	0,01
419	BEGAN	23.752	0,02	**919**	THIRTY	11.723	0,01
420	ITSELF	23.752	0,02	**920**	PARTS	11.714	0,01
421	SEEMED	23.736	0,02	**921**	SHOWS	11.692	0,01
422	THEMSELVES	23.693	0,02	**922**	PAGE	11.688	0,01
423	CLASS	23.662	0,02	**923**	RATES	11.649	0,01
424	UPON	23.656	0,02	**924**	PRODUCE	11.638	0,01
425	SUBJECT	23.606	0,02	**925**	ACTIVITIES	11.622	0,01
426	EFFECT	23.574	0,02	**926**	ACTIVITY	11.618	0,01
427	WANTED	23.570	0,02	**927**	GENERALLY	11.596	0,01
428	MINISTER	23.437	0,02	**928**	TEACHERS	11.594	0,01
429	TODAY	23.437	0,02	**929**	DEEP	11.568	0,01
430	AREAS	23.426	0,02	**930**	PARK	11.566	0,01
431	JOB	23.353	0,02	**931**	ASSOCIATION	11.563	0,01
432	LIKELY	23.347	0,02	**932**	APPROPRIATE	11.535	0,01
433	POSITION	23.327	0,02	**933**	CHANGED	11.527	0,01
434	THEREFORE	23.239	0,02	**934**	NEARLY	11.524	0,01
435	S	23.219	0,02	**935**	SUN	11.522	0,01
436	COMMUNITY	23.043	0,02	**936**	PRODUCT	11.504	0,01
437	READ	23.036	0,02	**937**	ALLOW	11.494	0,01
438	FRONT	22.978	0,02	**938**	LOOKS	11.493	0,01
439	AGE	22.932	0,02	**939**	FIGURES	11.472	0,01
440	MATTER	22.908	0,02	**940**	L	11.456	0,01
441	GAVE	22.901	0,02	**941**	ROSE	11.440	0,01
442	AMONG	22.880	0,02	**942**	ATTEMPT	11.437	0,01
443	SOCIETY	22.862	0,02	**943**	MALE	11.431	0,01
444	THIRD	22.846	0,02	**944**	FALL	11.430	0,01

445	STAFF	22.845	0,02	**945**	LEADING	11.425	0,01
446	PROCESS	22.795	0,02	**946**	SORRY	11.420	0,01
447	TAKING	22.750	0,02	**947**	ARMS	11.415	0,01
448	CITY	22.718	0,02	**948**	COVER	11.412	0,01
449	FUTURE	22.670	0,02	**949**	ORIGINAL	11.407	0,01
450	PAY	22.529	0,02	**950**	NUMBERS	11.377	0,01
451	CHILD	22.506	0,02	**951**	CLAIM	11.367	0,01
452	NON	22.411	0,02	**952**	PROFESSIONAL	11.366	0,01
453	ACTION	22.360	0,02	**953**	FIRM	11.362	0,01
454	PROVIDE	22.342	0,02	**954**	DIFFERENCE	11.360	0,01
455	STUDY	22.340	0,02	**955**	BASIC	11.326	0,01
456	YOU'VE	22.340	0,02	**956**	RELATIONS	11.322	0,01
457	CERTAIN	22.313	0,02	**957**	MEANT	11.318	0,01
458	BECAME	22.309	0,02	**958**	VILLAGE	11.314	0,01
459	PLAY	22.267	0,02	**959**	PLAYED	11.276	0,01
460	SENSE	22.231	0,02	**960**	WE'LL	11.276	0,01
461	COST	22.229	0,02	**961**	EFFECTS	11.273	0,01
462	SPECIAL	22.200	0,02	**962**	HOTEL	11.265	0,01
463	INTERNATIONAL	22.155	0,02	**963**	CROSS	11.251	0,01
464	MRS	22.155	0,02	**964**	POTENTIAL	11.248	0,01
465	RESULT	22.155	0,02	**965**	WRITE	11.238	0,01
466	DIFFICULT	22.091	0,02	**966**	ARMY	11.227	0,01
467	LAND	22.090	0,02	**967**	GARDEN	11.219	0,01
468	PARTICULARLY	22.056	0,02	**968**	REPORTS	11.184	0,01
469	CLOSE	22.054	0,02	**969**	READING	11.166	0,01
470	OUTSIDE	21.991	0,02	**970**	SUMMER	11.137	0,01
471	MANAGEMENT	21.982	0,02	**971**	DEMAND	11.132	0,01
472	CANNOT	21.807	0,02	**972**	PLAYING	11.129	0,01
473	ENGLAND	21.751	0,02	**973**	SUGGESTED	11.111	0,01
474	MOMENT	21.726	0,02	**974**	IDEAS	11.109	0,01
475	IDEA	21.719	0,02	**975**	FISH	11.101	0,01
476	WOMAN	21.656	0,02	**976**	INDEPENDENT	11.087	0,01
477	TURN	21.653	0,02	**977**	COUNTY	11.079	0,01
478	EXPERIENCE	21.558	0,02	**978**	UNIT	11.063	0,01
479	EVIDENCE	21.537	0,02	**979**	ACCESS	11.060	0,01
480	TRY	21.505	0,02	**980**	PICTURE	11.057	0,01
481	BELIEVE	21.275	0,02	**981**	SHOP	11.035	0,01
482	RANGE	21.258	0,02	**982**	UNLESS	11.029	0,01
483	SEEMS	21.202	0,02	**983**	YOURSELF	11.019	0,01
484	TEN	21.194	0,02	**984**	INVESTMENT	11.017	0,01
485	FATHER	21.189	0,02	**985**	VIDEO	11.011	0,01
486	GOD	21.161	0,02	**986**	RULES	10.991	0,01
487	TWENTY	21.142	0,02	**987**	READY	10.990	0,01
488	EAST	21.077	0,02	**988**	APPEAL	10.984	0,01
489	MORNING	21.017	0,02	**989**	OBVIOUSLY	10.977	0,01
490	TABLE	20.977	0,02	**990**	BENEFIT	10.967	0,01
491	WASN'T	20.952	0,02	**991**	NOTE	10.961	0,01
492	MEETING	20.904	0,02	**992**	WINDOW	10.953	0,01
493	DEATH	20.877	0,02	**993**	STANDING	10.947	0,01
494	TRAINING	20.843	0,02	**994**	APPEAR	10.946	0,01

495	EUROPEAN	20.826	0,02	**995**	CONTACT	10.925	0,01
496	AIR	20.801	0,02	**996**	NORTHERN	10.923	0,01
497	LEAVE	20.745	0,02	**997**	EMPLOYMENT	10.894	0,01
498	SOON	20.735	0,02	**998**	BLOOD	10.888	0,01
499	TRADE	20.720	0,02	**999**	OIL	10.887	0,01
500	LATE	20.600	0,02	**1000**	REASONS	10.880	0,01

Anexo 2: As mil palavras com mais etiquetas no *British National Corpus*

Nº de etiquetas	Palavra	Nº de etiquetas	Palavra	Nº de etiquetas	Palavra
14	like	8	stalking	8	floating
13	set	8	staining	8	fleeting
13	forecast	8	stained	8	fleeing
13	cut	8	stain	8	fixing
13	a	8	staggering	8	fitted
12	split	8	staffing	8	fit
12	round	8	stack	8	firing
12	down	8	squatting	8	finishing
12	burst	8	spruce	8	fine
12	bound	8	spread	8	filming
11	upset	8	spotted	8	fight
11	up	8	sponging	8	feuding
11	square	8	spitting	8	ferreting
11	pop	8	spiralling	8	fermenting
11	opencast	8	spinning	8	fer
11	left	8	sowing	8	fed
11	lay	8	sounding	8	featuring
11	hurt	8	soaring	8	fave
11	hit	8	soaking	8	fattening
11	fell	8	snorting	8	fast
11	cross	8	snoring	8	failing
11	cast	8	smuggled	8	fai
11	ai	8	smoothing	8	fading
10	wet	8	smoked	8	exposed
10	sound	8	smiling	8	exporting
10	shed	8	smelting	8	exploring
10	s	8	smelling	8	exploded
10	over	8	smallholding	8	expanded
10	outcast	8	sleeved	8	exhausted
10	open	8	sizzling	8	excluded
10	offset	8	situated	8	examining
10	light	8	sinking	8	eventing
10	ing	8	silver	8	even
10	ground	8	signified	8	evaporated
10	fucking	8	signalling	8	et
10	following	8	signal	8	establishing
10	do	8	sightseeing	8	ep
10	dat	8	si	8	enterprising
10	crap	8	shut	8	entering
10	close	8	shrinking	8	enlightened
10	changing	8	shouting	8	engineered
10	broadcast	8	short	8	endangered
10	bet	8	shit	8	encouraged

10	back	8	shining	8	encoding
10	an'	8	shifting	8	enclosed
9	wrong	8	shelving	8	enamelled
9	well	8	shaving	8	emi
9	visiting	8	shattered	8	emerging
9	videoconferencing	8	sharing	8	embedded
9	ve	8	shaping	8	embalming
9	upfront	8	shaped	8	electorate
9	typecast	8	shaking	8	electioneering
9	trekking	8	sentencing	8	educated
9	trading	8	sensing	8	editing
9	touring	8	seeking	8	earning
9	top	8	seasoned	8	earn
9	timing	8	sealed	8	dun
9	thrust	8	se	8	duff
9	that	8	screwing	8	dub
9	testing	8	screaming	8	drumming
9	ta	8	scrambled	8	drowning
9	swimming	8	scouting	8	drink
9	supporting	8	scouring	8	dried
9	suffering	8	score	8	dredging
9	styling	8	scheduled	8	dreaded
9	strain	8	scented	8	dram
9	stopping	8	scanning	8	drafting
9	stock	8	scaling	8	dose
9	staging	8	save	8	don
9	sporting	8	savage	8	doe
9	splitting	8	sauce	8	dock
9	spit	8	saturated	8	disturbed
9	spending	8	rush	8	distributed
9	smoking	8	run	8	distressed
9	slit	8	rowing	8	distracted
9	skiing	8	rotting	8	distilling
9	sketching	8	rose	8	dissenting
9	signing	8	rooted	8	disqualified
9	shooting	8	roasted	8	dispossessed
9	sampling	8	rip	8	displaying
9	sailing	8	ringed	8	discussing
9	right	8	revised	8	discounting
9	riding	8	retiring	8	disciplining
9	reviewing	8	retired	8	disappearing
9	returning	8	restricted	8	disappeared
9	restructuring	8	requesting	8	directorate
9	resting	8	requested	8	directing
9	reporting	8	rented	8	dipping
9	recycling	8	renewed	8	dip
9	read	8	renaming	8	digging
9	rating	8	removing	8	dieting
9	raising	8	rem	8	diet
9	purchasing	8	relieved	8	diesel

9	publishing	8	released	8	di
9	pricing	8	relaxed	8	determining
9	present	8	relax	8	destroyed
9	plumb	8	relating	8	designing
9	placing	8	regulated	8	deprived
9	pet	8	regained	8	deposited
9	passing	8	refurbished	8	dented
9	parallel	8	reducing	8	demonstrate
9	panting	8	recycled	8	delegated
9	paging	8	recruiting	8	delayed
9	packaging	8	recovered	8	defining
9	ow	8	receiving	8	defeated
9	out	8	recast	8	dedicated
9	ol	8	reassured	8	decorating
9	offering	8	rave	8	decorated
9	obtaining	8	rap	8	decoding
9	nt	8	rambling	8	deciding
9	no	8	rallying	8	debating
9	nesting	8	raise	8	dating
9	motoring	8	rail	8	darn
9	monitoring	8	rafting	8	damned
9	modelling	8	racketeering	8	damaged
9	mess	8	r'	8	cycling
9	matching	8	quote	8	custom
9	mat	8	qualified	8	cushioning
9	manufacturing	8	put	8	curtain
9	mailing	8	pulling	8	cup
9	loving	8	proving	8	cum
9	logging	8	protecting	8	crying
9	listing	8	prospecting	8	crumbling
9	listening	8	prospect	8	crowning
9	linking	8	prop	8	crossing
9	ling	8	prompted	8	crooked
9	level	8	promoting	8	crippled
9	let	8	promoted	8	creating
9	lapwing	8	pro	8	crash
9	labouring	8	printed	8	crack
9	knit	8	pressed	8	cow
9	killing	8	preserved	8	coupling
9	jogging	8	prescribing	8	coup
9	introducing	8	preparing	8	counting
9	imaging	8	predicting	8	counter
9	i'	8	preaching	8	cot
9	healing	8	praying	8	costing
9	head	8	practising	8	cost
9	handling	8	pounding	8	corrected
9	grey	8	potting	8	cornered
9	green	8	postgraduate	8	corgi
9	grading	8	polling	8	coping
9	funding	8	polled	8	cop

9	functioning	8	polishing	8	coordinating
9	full	8	polish	8	cook
9	frank	8	poisoned	8	convicted
9	founding	8	poaching	8	conveyancing
9	flying	8	pit	8	converting
9	flush	8	pioneering	8	controlling
9	fishkeeping	8	pine	8	contracting
9	fishing	8	photographed	8	continuing
9	filling	8	persecuted	8	contented
9	filing	8	permitting	8	consulting
9	farming	8	pat	8	constrained
9	falling	8	pass	8	consolidated
9	factoring	8	participating	8	conjuring
9	f&mdash	8	parenting	8	confused
9	express	8	paralysed	8	condemned
9	eq	8	panicking	8	conceived
9	entertaining	8	padded	8	computed
9	enduring	8	packed	8	compiled
9	em	8	overcrowded	8	competing
9	el	8	outsourcing	8	committed
9	eating	8	outraged	8	commissioned
9	dying	8	ot	8	combined
9	drunk	8	organizing	8	coloured
9	drilling	8	organising	8	collecting
9	dreaming	8	organised	8	collected
9	drain	8	orbit	8	coded
9	downsizing	8	omitted	8	coated
9	double	8	okay	8	co-ordinating
9	domain	8	occupied	8	closing
9	diving	8	obsessed	8	clip
9	dis	8	observing	8	cleared
9	dining	8	numbering	8	clear
9	dere	8	nu	8	clean
9	der	8	nought	8	classified
9	damn	8	nominate	8	christening
9	cutting	8	negotiating	8	cheered
9	cruising	8	neglected	8	checking
9	cream	8	near	8	chased
9	cool	8	naturalized	8	charred
9	computing	8	nap	8	challenging
9	coding	8	n'	8	censored
9	coaching	8	muse	8	celibate
9	cleansing	8	murdered	8	celebrated
9	cleaning	8	mounted	8	celebrate
9	ching	8	moulded	8	caving
9	champion	8	moss	8	cautioning
9	catering	8	moisturising	8	causing
9	caring	8	moi	8	cataloguing
9	can	8	moderate	8	casting
9	camping	8	model	8	cart

9	buying	8	mixing	8	carp
9	bust	8	missing	8	carol
9	bursting	8	mini	8	carn
9	budgeting	8	mine	8	canvassing
9	briefing	8	mimicking	8	canning
9	brewing	8	mid	8	campaigning
9	breathing	8	metering	8	camp
9	branding	8	met	8	cabling
9	brain	8	messaging	8	cab
9	boxing	8	merchandising	8	bungling
9	borrowing	8	meet	8	bullying
9	bombing	8	medicine	8	bullied
9	boeing	8	mechanised	8	brushed
9	bleeding	8	measuring	8	browning
9	blazing	8	mean	8	brooding
9	black	8	masked	8	brit
9	beat	8	marking	8	brief
9	bay	8	marching	8	breaking
9	balancing	8	mapping	8	brawling
9	avoiding	8	manufacture	8	brass
9	auditing	8	managing	8	braking
9	as	8	malting	8	boxed
9	addressing	8	maintaining	8	bowing
9	a'	8	machining	8	bored
8	zoom	8	m	8	booming
8	y'	8	low	8	boom
8	wrestling	8	love	8	bookselling
8	wrecked	8	loose	8	bonding
8	wooing	8	lodged	8	bomb
8	wonder	8	localised	8	boating
8	wiz	8	loaded	8	blurring
8	witnessing	8	linked	8	blundering
8	wishing	8	lifting	8	blue
8	windsurfing	8	licensing	8	blowing
8	windowing	8	liberating	8	blinking
8	will	8	letting	8	bleedin'
8	wi	8	lent	8	bled
8	whiting	8	lending	8	birdwatching
8	whirling	8	lemming	8	bing
8	whip	8	leasing	8	billing
8	weighting	8	leased	8	betting
8	weighing	8	leaking	8	bent
8	weeping	8	leaked	8	benchmarking
8	weed	8	laying	8	beleaguered
8	wee	8	layered	8	beckoning
8	weaving	8	laundering	8	batting
8	weathering	8	last	8	bathed
8	wearing	8	laminate	8	bath
8	wasting	8	labour	8	bate
8	wasted	8	labelled	8	bat

8	washing	8	knocking	8	base
8	wash	8	knock	8	barred
8	warming	8	kneeling	8	bargaining
8	wanted	8	kit	8	bargain
8	wandering	8	kiss	8	barbed
8	wailing	8	kipling	8	banqueting
8	voting	8	kicking	8	banning
8	vesting	8	key	8	banned
8	veiled	8	kate	8	bang
8	vanishing	8	karate	8	banding
8	vanished	8	jousting	8	ballooning
8	unfortunate	8	jose	8	balanced
8	undergraduate	8	joking	8	baked
8	twisting	8	jobbing	8	bain
8	tuning	8	je	8	backswing
8	tufted	8	itching	8	backpacking
8	troubled	8	it'	8	ax
8	trooping	8	iss	8	awarding
8	trim	8	interviewing	8	auld
8	trilling	8	interpreting	8	audit
8	travelling	8	intensified	8	attracting
8	travelled	8	integrated	8	attended
8	trapped	8	integrate	8	atm
8	transferred	8	int	8	assured
8	tramp	8	insulated	8	assorted
8	trail	8	inspired	8	ass
8	trafficking	8	influencing	8	arriving
8	tracking	8	increase	8	arrested
8	tow	8	incorporate	8	archiving
8	tour	8	improved	8	arch
8	total	8	impoverished	8	apr
8	tortured	8	imp	8	approved
8	tooling	8	im	8	approaching
8	toning	8	illustrated	8	appoint
8	toast	8	ignored	8	ap
8	tl	8	ig	8	ankle
8	tipping	8	identified	8	allied
8	tinted	8	ici	8	alain
8	tilting	8	iain	8	aix
8	till	8	i	8	aim
8	ti	8	hung	8	aided
8	throwing	8	hunched	8	aggregated
8	thrashing	8	hse	8	ageing
8	textured	8	housebuilding	8	afhq
8	tested	8	hooked	8	advising
8	terrified	8	hmi	8	advised
8	terminated	8	hitting	8	adj
8	tender	8	hitched	8	adding
8	tasting	8	hire	8	acquitted
8	tapping	8	high	8	aching

8	tap	8	hide	8	accused
8	taming	8	help	8	acclaimed
8	tab	8	heated	8	abseiling
8	switching	8	haunting	8	ab
8	swinging	8	hate	8	's
8	swearing	8	harm	8	'e
8	swapping	8	hare	7	zoomed
8	suspended	8	handicapped	7	zoning
8	surveying	8	handheld	7	zone
8	superseded	8	guiding	7	zipped
8	supercomputing	8	guided	7	yu
8	sucralfate	8	grouping	7	ytp
8	styled	8	grooming	7	yoke
8	stunned	8	grazing	7	yodelling
8	stuffed	8	governing	7	yi
8	stuff	8	golfing	7	yer
8	struggling	8	going	7	yellowing
8	structuring	8	gill	7	yelling
8	structured	8	gilded	7	yell
8	stripping	8	ger	7	yawning
8	stripped	8	georg	7	yawn
8	striking	8	ge	7	yam
8	stretching	8	gaping	7	wrought
8	straw	8	gambling	7	wrinkled
8	strauss	8	gaining	7	wriggling
8	strat	8	furnishing	7	wrecking
8	stranded	8	fur	7	wrapped
8	stored	8	fundraising	7	wounding
8	stockbroking	8	fundholding	7	wounded
8	stitched	8	funded	7	wound
8	stimulated	8	fringe	7	wot
8	stiffening	8	frayed	7	worshipping
8	stepping	8	franchising	7	worsening
8	step	8	framed	7	worrying
8	steering	8	forewarned	7	worming
8	stealing	8	forecasting	7	worm
8	staying	8	foot	7	workplace
8	star	8	fm	7	working
8	standing	8	fly	7	workforce
8	standardised	8	flowering	7	wondering
				7	wolf