

**UNIVERSIDADE DE SÃO PAULO**  
Faculdade de Filosofia Letras e Ciências Humanas  
Departamento de Letras Modernas  
Área de Espanhol

ELISA CRISTINE DO SANTOS

**Corpus multilíngüe especializado em Reciclagem:  
uma análise das faixas de frequência de formas e ocorrências**

São Paulo  
2008

ELISA CRISTINE DO SANTOS

**Corpus multilíngüe especializado em Reciclagem:  
uma análise das faixas de frequência de formas e ocorrências**

Trabalho de Graduação Individual (TGI) apresentado à  
área de Graduação em Língua Espanhola do  
Departamento de Letras Modernas da Faculdade de  
Filosofia, Letras e Ciências Humanas da Universidade  
de São Paulo.

Orientadora: Profa. Dra. Heloísa Pezza Cintrão

São Paulo  
2008

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTES TRABALHOS, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

#### Catálogo na publicação

Santos, Elisa Cristine dos.

Corpus multilíngüe especializado em reciclagem: uma análise das faixas de frequência de formas e ocorrências / Elisa Cristine dos Santos; orientadora Heloísa Pezza Cintrão. -- São Paulo, 2008.

100 f.

Trabalho de Graduação Individual (TGI). Área de Graduação em Língua Espanhola do Departamento de Letras Modernas – Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

## FOLHA DE APROVAÇÃO

Elisa Cristine dos Santos

Trabalho de Graduação Individual (TGI) apresentado à área de Graduação em Língua Espanhola do Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

Aprovado em:

Banca Examinadora

Prof. \_\_\_\_\_

Instituição: \_\_\_\_\_ Assinatura: \_\_\_\_\_

## **DEDICATÓRIA**

À minha mãe e ao meu esposo pelo apoio ao longo do período de elaboração deste trabalho.

## **AGRADECIMENTOS**

À Prof. Dra. Heloísa Pezza Cintrão, pela atenção e apoio durante o período de orientação.

À área de Língua Espanhola do Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, pela oportunidade de realizar o Trabalho de Graduação Individual (TGI) não obrigatório.

## RESUMO

O objetivo inicial deste trabalho é coletar, segundo os princípios da Linguística de corpus, textos contemporâneos em formato eletrônico sobre reciclagem para formar um corpus multilíngüe especializado.

A escolha do tema reciclagem se deve a importância que o assunto adquiriu nos últimos anos por afetar desde os cidadãos e pequenas empresas a grandes governos, além da possibilidade de fornecer um material especializado para profissionais da área de Linguística de Corpus, Tradução e áreas relacionadas ao tema. O processo de reciclagem envolve a separação e destinação adequada de resíduos obsoletos descartados (ex. lixo orgânico, equipamentos eletrônicos, papel, papelão, plástico, pilhas, baterias, vidro, etc.). Os textos escritos sobre o tema, freqüentemente, apresentam as características dos materiais, a descrição do processo de reciclagem de cada um deles, além dos benefícios que podem trazer para o meio ambiente e a economia.

O corpus é formado por três línguas: Português (Brasil), Inglês (Estados Unidos) e Espanhol (Espanha). A partir da criação deste corpus, fazemos a extração de listas de palavras, através do programa AntConc, e a análise das faixas de freqüência de formas e ocorrências, trabalho essencial para o desenvolvimento de outras análises como a padronização e a colocação.

Palavras-chave: reciclagem, corpus, freqüência.

## RESUMEN

El objetivo inicial de este trabajo es compilar, según los principios de la Lingüística de corpus, textos contemporáneos en formato electrónico sobre reciclaje para formar un corpus multilingüe especializado.

La elección del tema reciclaje se debe a la importancia que el tema ha adquirido en los últimos años por afectar directamente desde ciudadanos e pequeñas empresas hasta grandes gobiernos, además de la posibilidad de proveer un material especializado para profesionales del área de Lingüística de Corpus, Traducción y áreas relacionadas al tema. El proceso de reciclaje involucra la separación y destinación adecuada de residuos obsoletos desechados (ej. basura orgánica, computadoras, papel, cartón, plástico, pilas, baterías, vidrio, etc.) Los textos escritos sobre el tema, frecuentemente, presentan las características de los materiales, la descripción del proceso de reciclaje de cada uno, además de los beneficios que pueden traer para el medio ambiente y para la economía.

El corpus es formado por tres lenguas: Portugués (Brasil), Inglés (EE.UU.) y Español (España). A partir de la creación de este corpus, hacemos la extracción de listas de palabras, a través del programa AntConc, y el análisis de los niveles de frecuencia de formas y ocurrencias, trabajo esencial para el desarrollo de otros análisis como la patronización y la colocación.

Palabras-clave: reciclaje, corpus, frecuencia.

## SUMÁRIO

Introdução.....	10
1. Lingüística de corpus: definição e um breve histórico.....	11
1.1. Marcos da Lingüística de Corpus.....	13
1.2. Contextualização.....	16
2. Corpus.....	19
2.1. Definições.....	19
2.2. Tipologia.....	20
2.3. Extensão e Representatividade.....	20
2.4. Especificidade.....	24
2.5. Corpora em diferentes línguas.....	25
3. Organização de corpus.....	27
3.1. Coleta e armazenamento de dados.....	28
3.2. Ferramentas.....	29
3.2.1. WST – WordSmith Tools.....	29
3.2.2. AntConc.....	32
4. Análise de corpus.....	32
4.1. Frequência de palavras.....	32
4.2. Concordâncias e colocados.....	36
4.3. Padronização.....	39
5. Frequência de palavras em corpus multilíngüe especializado em reciclagem.....	41
5.1. Reciclagem.....	41
5.2. Escolha do corpus.....	42
5.3. Contribuição para a Lingüística de corpus.....	43
5.4. Metodologia.....	44
5.5. Listas de frequência de palavras.....	50
5.5.1. Palavras mais e menos frequentes.....	50
6. Considerações finais.....	55
Referências bibliográficas.....	57
Anexo 1. Controle de coleta de textos por website.....	58
Anexo 2. Listas de palavras em Português, Inglês e Espanhol.....	62
Anexo 3. Lista de palavras-chave em Português.....	80

## **Introdução**

A Lingüística de Corpus, com mais de 40 anos de existência, é uma área que trata da criação e utilização de corpora eletrônicos com a finalidade de estudar e analisar fenômenos lingüísticos. A coleta de textos contemporâneos, por exemplo, permite retratar o estágio em que se encontra uma determinada da língua. O objetivo inicial deste trabalho é justamente coletar textos contemporâneos em formato eletrônico sobre reciclagem e formar um corpus comparável multilíngüe especializado.

Antes de iniciar um trabalho como este, é essencial se familiarizar com os conceitos inerentes a área, bem como os estudos desenvolvidos ao longo de sua existência. Este procedimento auxilia na escolha de metodologias e ferramentas adequadas, além de objetos de estudo relevantes para a Lingüística de corpus e áreas afins.

No primeiro capítulo, tratamos de definições, histórico e marcos da Lingüística de Corpus, o que nos permitiu fazer uma breve contextualização da área ao longo de sua história. Os conceitos sobre corpus, tais como definições, tipologia, extensão, representatividade e especificidade foram abordados no segundo capítulo, assim como a produção de corpora em diversas línguas. No terceiro capítulo, buscamos nos familiarizar com a coleta e a organização de um corpus, além do uso de ferramentas durante o processo de análise do material coletado e compilado, como os programas WordSmith Tools e AntConc.

Um corpus multilíngüe especializado nos permite fazer uma série de análises, que foram estudadas no capítulo quatro deste trabalho. A partir de listas de frequências e colocados podemos analisar fenômenos como padronização e estrangeirismos, além do comportamento das palavras em relação à própria frequência e colocação.

O capítulo cinco foi totalmente focado na descrição do corpus que criamos e na utilização de ferramentas voltadas para extração de listas de palavras e palavras-chave, além da análise de faixas de frequência de formas e ocorrências, baseado no estudo de “frequências de palavras da língua portuguesa segundo o Banco do Português”, realizado por Sardinha (2004).

A escolha do tema reciclagem se deve a importância que o assunto adquiriu nos últimos anos por afetar desde os cidadãos e pequenas empresas a grandes governos. O processo de Reciclagem envolve a separação e destinação adequada de resíduos obsoletos descartados (ex. lixo orgânico, equipamentos eletrônicos, papel, papelão, plástico, pilhas, baterias, vidro, etc.). Os textos escritos sobre o tema, frequentemente, apresentam as características dos materiais, a descrição do processo de reciclagem de cada um deles, além dos benefícios que podem trazer para o meio ambiente e a economia.

Após todo este trabalho de pesquisa e prática, apresentamos nossas conclusões e comentários quanto ao cumprimento dos objetivos propostos.

## **1. Lingüística de corpus: definição e um breve histórico**

O lançamento do primeiro corpus lingüístico computadorizado, em 1964, o Brown University Standard Corpus of Present-day American English, com um milhão de palavras, alavancou os estudos em Lingüística de Corpus e permitiu que a observação de padrões lingüísticos, tanto dentro de uma mesma língua quanto comparativamente entre duas ou várias línguas, passasse a ser vista como uma importante fonte de informação para professores, tradutores, lexicógrafos e outros profissionais.

Sardinha define assim essa área de estudos relativamente recente:

A Lingüística de Corpus ocupa-se da coleta e da exploração de corpora, ou conjuntos de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador. (SARDINHA, 2004, p. 3)

Sardinha também nos explica que os *corpora*, como conjunto de documentos, existem desde a Antiguidade, com o Corpus Helenístico, definido por Alexandre, O Grande, e com corpora compostos por citações bíblicas, que se produziram na Antigüidade e na Idade Média. No século XX, uma série de pesquisadores, incluindo educadores como Thorndike e lingüistas como Boas e Fries, também se valeram de corpora para descrever a linguagem. Mas

nesta época, a coleta e a análise de dados eram feitas manualmente, e não através de computadores, e a maioria dos trabalhos enfatizava o ensino de línguas.

Uma importante referência para as dimensões dos grandes corpora computadorizados de língua geral atuais foi um corpus não-computadorizado, o *SEU (Survey of English Usage)*, compilado manualmente em Londres, por uma equipe liderada por Randolph Quirk, a partir de 1959. Este corpus foi planejado inicialmente para comportar a quantidade de um milhão de palavras, proveniente de 200 textos com cinco mil palavras cada. Sua análise foi feita através da organização em fichas, separadas conforme a categoria gramatical, nas quais havia uma palavra do corpus, seguida de sua análise gramatical. “O conjunto de categorias resultante serviu de base para o desenvolvimento dos etiquetadores computadorizados contemporâneos, que fazem a identificação de traços gramaticais automaticamente,” (SARDINHA, p. 3-4)

Este grandioso projeto serviu de base para o Corpus Brown, para a *Comprehensive Grammar of the english language* e para o *London-Lund Corpus*, que constitui a parte falada do *SEU* computadorizada. Apenas em 1989, 30 anos depois de sua criação, o *SEU* pôde ser totalmente computadorizado.

Com o lançamento do *Syntactic Structures*, de Chomsky, no final dos anos 50, a lingüística gerativa e as teorias racionalistas da linguagem tomaram um lugar central nos estudos da linguagem. Estas teorias se fortaleceram em detrimento dos estudos empíricos relacionados a corpora desenvolvidos até então. Uma mudança como esta promoveu e estimulou as críticas ao processo manual de criação e desenvolvimento de corpora gigantescos. Alegaram principalmente que o ser humano não era capaz de controlar de forma confiável uma quantidade de dados dessa grandeza.

Por volta de 1960, com a invenção do computador, os corpora passaram a ser controlados eletronicamente em centros de pesquisas de universidades, o que contribuiu para a retomada dos estudos com mais eficiência e credibilidade. Mas as mudanças foram além. Com o progresso da ciência, nos anos 80, os computadores se estenderam aos lares e permitiram a popularização de corpora e ferramentas de processamento, fortalecendo assim as pesquisas lingüísticas relacionadas a este tema.

Entre os centros de pesquisa, destacam-se os de instituições britânicas e escandinavas com ativa participação e liderança nas pesquisas lingüísticas baseadas em corpus. Fora da Europa, esses estudos estão menos desenvolvidos, mesmo nas instituições norte-americanas, geralmente, muito bem equipadas para pesquisa, inclusive quanto a recursos informáticos. Isto é atribuído, por Sardinha (p. 5) à força da lingüística gerativa-transformacional. Nos EUA o foco é outro: desenvolver a pesquisa em Processamento de Linguagem Natural (PLN) com fins acadêmicos e industriais (comerciais).

Sardinha (2004, p. 6), baseado em seu artigo “Processamento Computacional do Português”, de 1999, afirma que “No Brasil, a Lingüística de Corpus ainda está em estágio inicial. A pesquisa em corpus se dá em centros mais voltados ao Processamento de Linguagem Natural, à Lexicografia e à Lingüística Computacional.”

Parte do fortalecimento da Lingüística de Corpus se deve ao interesse comercial em suas produções. A demanda por dicionários diversos, gramáticas e livros didáticos voltados ao ensino de línguas estrangeiras faz com que editoras busquem parcerias com centros de pesquisa a fim de produzir materiais baseados na observação de corpus. Um exemplo conhecido é o dicionário *Cobuild*, produzido em parceria entre a Universidade de Birmingham (Grã-Bretanha) e a Editora Collins.

Cresce também, o desenvolvimento de centros de pesquisa financiados por empresas de telecomunicações e fabricantes de produtos de informática. Elas se utilizam de pesquisas baseadas em corpus para fins comerciais como: “processamento automático de textos, informatização de grandes bases de dados e a montagem de sistemas inteligentes de reconhecimento de voz e gerenciamento de informação”. (SARDINHA, 2004, p. 6)

### **1.1. Marcos da Lingüística de Corpus**

Sardinha (2004, p. 12-16) considera que, além da significativa produção de corpora eletrônicos em diversas línguas (ver 2.5), a Lingüística de Corpus é marcada por vários acontecimentos que vão desde lançamentos de dicionários até discussões sobre a aplicação de corpora computadorizados no ensino de língua.

1966: Sinclair lança um trabalho pioneiro, que passa a ser utilizado como referência para as pesquisas desenvolvidas em Lingüística de Corpus até os dias de hoje. Leech, “um dos maiores lingüistas de corpus de todos os tempos” publica seu primeiro trabalho em análise de corpus, apontando “a necessidade de análises detalhadas de corpora via computador”.

1982: Francis & Kucera descrevem, de forma computadorizada, as frequências do corpus Brown.

1987: Sinclair et al. lançam o dicionário *Cobuild*, “o primeiro a ser compilado a partir de um corpus computadorizado”.

1988: Biber, baseado em dois grandes corpora da língua inglesa (*LOB* e *London-Lund*), descreveu a “composição lingüística de gêneros” desta língua, apontando assim a “necessidade de investigação do texto”.

1991: Aijmer & Altenberg lançam a primeira obra que usa a nomenclatura *Corpus Linguistics* no título. Sinclair, principal nome na história da Lingüística de Corpus, publica *Corpus, concordance, collocation*, uma reunião de seus principais trabalhos.

1992: “A academia de ciências da Suécia dedicou um de seus célebres seminários Nobel à Lingüística de Corpus. Os mais renomados lingüistas da época compareceram para apresentar um painel do estado da arte naquele momento”.

1994: Kjellmer publica o primeiro dicionário de colocações, baseado no corpus Brown, cujos padrões recorrentes apontados levaram em conta análises estatísticas.

1996: McEnery & Wilson criam um manual de Lingüística de Corpus, didático, com conceitos práticos e teóricos para alunos de Lingüística. Francis & Hunston lançam a primeira gramática do léxico, que “descreve de modo amplo e profundo os padrões verbais da língua inglesa a partir de um corpus, seguindo o princípio básico de identificação de colocações recorrentes, por computador”. O segundo volume, sobre substantivos e adjetivos, foi publicado na seqüência.

1998: Biber, Conrad & Reppen lançaram um manual de Lingüística de Corpus sob a perspectiva americana da área, que era dominada exclusivamente por trabalhos de centros europeus. Granger publica uma coletânea de trabalhos voltados para o ensino e aprendizagem de línguas, consolidando assim um tipo diferente de corpus, o de aprendizes. Partington volta atenções ao praticante da “Lingüística de Corpus caseira” (ex. professores e tradutores), que utiliza “computadores pessoais e corpora pequenos”.

1999: Biber, Johansson, Leech, Conrad & Finegan lançam uma “Gramática da Língua Inglesa baseada em corpus”.

2001: Simpson & Swales publicam um volume sobre o trabalho de Lingüística de Corpus na América do Norte. Ghadessy, Henry & Roseberry dedicam uma coletânea inteira a “questões, usos e aplicações de corpora pequenos, especialmente no contexto de ensino de línguas”.

Sardinha (2004, p. 14-15) afirma que, além das obras supracitadas, as pesquisas da área são veiculadas através de encontros, seminários e periódicos. Seguem abaixo os pioneiros, bem como seu ano de lançamento.

1979: Conferência ICAME para convidados.

1994: Conferência bienal TALC (*Teaching and Learning Corpora*).

1996: PROPOR - Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada. Periódico - *International Journal of Corpus Linguistics*.

1997: Conferência bienal PALC (*Practical Applications of Language Corpora*).

1998: Série *Studies in Corpus Linguistics*, Editora Benjamins.

Quanto às ferramentas computacionais utilizadas para a análise de Corpus, Sardinha (2004, p. 15-16) destaca os pioneiros segundo sua especialidade:

1970: TAGGIT - primeiro etiquetador morfossintático.

1979: CLAWS - etiquetador mais famoso, em utilização para anotar o *BNC (British National Corpus)*.

1987: TACT - programa pioneiro para microcomputadores, utilizado para gerar listas de palavras e concordâncias.

1988: *OCP, The Oxford Concordance Program* - “um dos principais concordanciadores usados em microcomputadores e estações de trabalho”.

1993: MICROCONCORD - “O mais famoso, simples e robusto programa de concordância para microcomputadores até hoje”.

1995: WORDSMITH TOOLS - “o mais completo e versátil conjunto de ferramentas para Lingüística de Corpus.” “O primeiro a aproveitar os recursos do ambiente Windows.”

1997: BRIL TAGGER para DOS - programa que permite contar palavras, concordâncias e etiquetar o corpus em microcomputadores.

1998: QTAG - etiquetador para Java adaptado para microcomputadores e utilizável em outras línguas além do inglês, representando assim uma ruptura da hegemonia desta língua quanto às ferramentas.

## 1.2. Contextualização

Contextualizar a atuação da Lingüística de Corpus envolve desde a apresentação de discussões sobre seu estatuto, até os tipos de pesquisas que ela privilegia, passando por seu emprego no ensino de língua estrangeira, influência no trabalho tradutório e análise multidimensional.

Há pelo menos 16 anos, segundo as obras citadas em Sardinha (2004, p. 35-38) sobre este tema, a Lingüística de Corpus alimenta uma discussão sobre a possibilidade de ser considerada disciplina, metodologia (MCENERY; WILSON, 1996), base metodológica (LEECH, 1992) ou abordagem (HOEY, 1997), por exemplo, uma abordagem filosófica da linguagem (LEECH, 1992) ou uma abordagem lingüística baseada em corpus (BIBER; CONRAD; REPPEN, 1998). A possibilidade de a Lingüística de Corpus ser considerada uma disciplina é a primeira a ser descartada por Leech (apud SARDINHA, 2004, p. 36), pelo fato de ela não apresentar um objeto de pesquisa delimitado, como, por exemplo, a psicolingüística ou a sociolingüística e não se dedicar a um assunto específico.

Para Sardinha (2004, p. 36), o conceito de metodologia adotado é fundamental para definir tal estatuto para a Lingüística de Corpus. “Entendendo metodologia como *instrumental*, então é possível aplicar o instrumental da Lingüística de Corpus livremente e manter a orientação teórica da disciplina original”. Ainda afirma:

Mas a Lingüística de Corpus não se resume a um conjunto de ferramentas. Assim, se entendermos metodologia como um *modo típico de aplicar um conjunto de pressupostos de caráter teórico*, então a Lingüística de Corpus pode ser vista como uma metodologia, pois traz consigo algo mais do que simplesmente o instrumental computacional. Aqui se encaixam as investigações do comportamento do léxico, típicas de lingüistas de corpus autodefinidos, como John Sinclair. A pesquisa de Sinclair acerca da colocação entre itens lexicais, por exemplo, não encontra espaço em outras disciplinas. Ela possui caráter essencialmente ascendente e tem como doutrina a não-categorização *a priori* (*trust the text* é o seu lema). Por isso, exemplifica com precisão a prática empirista e é o pólo mais distante das abordagens racionalistas, causa do surgimento de uma das maiores correntes de pesquisa em corpus.

Ao abordar os principais tipos de pesquisa na Lingüística de Corpus, Sardinha (2004, p. 38), baseado em McEnery e Wilson (1996), nos informa sobre a quantidade de trabalhos produzidos ao longo de 25 anos de pesquisas. Seriam 620 trabalhos no período de 1965 a 1991, sendo que aproximadamente a metade deles foi produzida em apenas cinco anos (os últimos deste período). Da época da publicação desse trabalho (1996) até hoje, passaram-se 12 anos, o que certamente nos permite crer que tenha havido um aumento muito significativo neste número até este momento, devido à necessidade de estudantes, pesquisadores e profissionais de progredirem com seus trabalhos relacionados ao emprego da língua e à tradução, por exemplo.

Sardinha (2004, p. 38) aponta que os trabalhos que se encaixam na Lingüística de Corpus são diversos, porém apresentam características em comum. Conforme Biber, Conrad e Reppen (1998 apud SARDINHA, 2004, p. 3), estas características são:

- São empíricos e analisam os padrões reais de uso em textos naturais.
- Utilizam coletâneas grandes e criteriosas de textos naturais, conhecidas por corpus, como a base da análise.
- Fazem uso extensivo de computadores na análise, empregando técnicas automáticas e interativas.
- Dependem de técnicas quantitativas e qualitativas.

Além destas características, os trabalhos se concentram em algumas áreas específicas. Segundo Sardinha (2004, p. 39), suas principais áreas de pesquisa são a compilação de corpus, a construção de ferramentas que permitam a compilação de dados e sua análise, a

descrição da linguagem e a aplicação dos corpora, o que pode ser exemplificado pelo ensino de línguas e pela tradução.

Segundo Sardinha, a descrição da linguagem é hoje a aplicação mais recorrente e enfoca principalmente o léxico. Através da descrição baseada em um corpus, é possível observar o comportamento de itens lexicais e de estruturas gramaticais em seu ambiente natural, o que Kennedy (1991, apud SARDINHA) chama de ecologia lingüística. Pelo estudo de corpora também tem-se buscado evidências de uma possível padronização do léxico. Para que se identifique um padrão, é necessário que a combinação das palavras ou estruturas de uma língua contribua para o significado de uma determinada palavra e ocorra com certa frequência.

Quanto ao ensino de língua estrangeira, observa-se uma atuação da Lingüística de Corpus, aliada à Lingüística Aplicada, no sentido de permitir a alunos e professores que utilizem os corpora para buscar alternativas que complementem o aprendizado e ensino. Além de usar as aplicações “destinadas especificamente ao ensino, como o *Lexical Syllabus*, o *Lexical Approach* e o *Data Driven Learning*”, atualmente já se usa corpora de aprendizes de língua estrangeira como fonte de informações para o “mapeamento da aquisição de línguas estrangeiras”, ou seja, sobre o uso real que é feito da língua em diferentes fases do aprendizado. (SARDINHA, 2004, p. 296)

Além de todas estas contribuições, a Lingüística de Corpus ainda conta com a abordagem da Análise multidimensional. Esta análise se vale de procedimentos automáticos, semi-automáticos (como a etiquetagem morfosintática) e estatísticos para mapear, em um corpus, as possíveis associações entre uma variedade de traços lingüísticos.

Por meio desta abordagem, afirma Biber (apud SARDINHA, p. 300) que é possível analisar a variação entre registros a partir do cruzamento de dezenas de variáveis lingüísticas correlacionando-as com variáveis situacionais tais como “formalidade, impessoalidade, oralidade, período histórico, estilo”.

## 2. Corpus

A escolha adequada de um corpus é fator determinante para o sucesso de um trabalho em Lingüística de Corpus. Para optar bem é necessário entender as definições que a literatura apresenta, bem como sua tipologia, extensão adequada à finalidade da pesquisa, especificidade e atuação em diferentes línguas.

### 2.1. Definições

Como mostra Sardinha (2004, p. 16-20), a literatura atual apresenta uma série de definições para corpus. Entre elas, está a de Sinclair, que considera corpus “uma coletânea de textos naturais (NATURALLY OCCURRING), escolhidos para caracterizar um estado ou variedade de linguagem” (apud SARDINHA, 2004, p. 16). Enquanto que Leech apresenta as seguintes definições: “Um corpo de material lingüístico que existe em formato eletrônico e que pode ser processado por computador para vários propósitos”; “Corpus de material lingüístico natural (textos inteiros, amostra de textos, ou às vezes somente sentenças desconexas), que são armazenadas em formato legível por máquina” (apud SARDINHA, 2004, p. 17).

Sardinha (2004, p. 16-17) nos lembra bem de que um corpus não é formado por qualquer conjunto de dados. Baseado em Atkins, Clear e Ostler (1992), o autor designa quatro diferentes conjuntos de textos: arquivo (conjunto de textos sem organização prévia), biblioteca eletrônica (conjunto de textos selecionados sob alguns critérios), corpus (seleção objetiva de parte da biblioteca eletrônica) e subcorpus (seleção de parte de um corpus, que pode ser trabalhada de forma flexível no momento da análise). Afirma também, que os textos que fazem parte do corpus devem ser coletados de forma natural para manterem sua autenticidade e permitirem a análise do que, de fato, ocorre na língua. O corpus, por sua vez, é produzido de forma artificial, pois a sua produção já objetiva a pesquisa lingüística.

Após analisar as definições de vários teóricos, o pesquisador brasileiro opta pela definição apresentada em Sanchez e Cantos (apud SARDINHA, p. 18):

Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos na 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.

Para Sardinha (2004, p. 18-19), esta definição é a mais completa, pois compreende a origem dos dados, o propósito, a composição, a formatação, a representatividade e a extensão do corpus.

## 2.2. Tipologia

Em seu *Lingüística de Corpus*, Sardinha (2004, p. 20-22) apresenta os principais tipos de corpora existentes na literatura e suas definições. Optamos por apresentar todos os tipos por ele descritos e agrupados, conforme abaixo.

- Modo: falado, escrito.
- Tempo: sincrônico, diacrônico, contemporâneo, histórico.
- Seleção: de amostragem, monitor, dinâmico ou orgânico, estático, equilibrado.
- Conteúdo: especializado, regional ou dialetal, multilíngüe.
- Autoria: de aprendiz, de língua nativa.
- Disposição interna: paralelo, alinhado.
- Finalidade: de estudo, de referência, de treinamento ou teste.

Comenta também que a classificação dos corpora pode ser feita baseada em perguntas relativas à pluralidade de autoria, origem da autoria, meio, integralidade, especificidade, dialeto, equilíbrio, fechamento, renovação, temporalidade, tradução, intercalação.

## 2.3. Extensão e Representatividade

Segundo Sinclair (apud SARDINHA, p. 22-23), a representatividade de um corpus é dada principalmente pela sua extensão, pois é ela quem permite, através de um número definido de textos e palavras, a visualização, estudo e análise de uma determinada gama de fenômenos dentro de um “sistema probabilístico”, que é a linguagem.

No campo lexical, ao observar a frequência das palavras de uma língua, nota-se que a maioria delas ocorre poucas vezes e por isso dependem de um corpus grande para estarem presentes com pelo menos uma ocorrência. Quanto ao sentido das palavras, uma palavra com alta frequência pode apresentar diversos sentidos e para que o corpus seja representativo de fato deve apresentar ocorrências para cada um dos diferentes sentidos. Ex.: a palavra *logo* pode se apresentar como um advérbio de tempo, modo, intensidade ou se referir de forma coloquial a logomarca ou logotipo, além de outras acepções.

Sardinha (2004, p. 23-24) explica que o corpus é apenas uma amostra de algo imensurável, que é a linguagem. Com isto, o ideal é ter um corpus com o maior tamanho possível, pois assim ele estará o mais próximo da realidade lingüística. Como não se sabe o tamanho real da linguagem, Sardinha reflete sobre a definição da representatividade e aponta que não é possível afirmar se um corpus é, de fato, representativo ou não. Para ele, representatividade e probabilidade estão ligadas, à medida que a linguagem apresenta uma série de traços (lexicais, estruturais, pragmáticos e discursivos) que podem ser relacionados entre si e informarem a probabilidade que determinados traços têm de serem mais ou menos comuns em certos contextos. Este conhecimento sobre a probabilidade e sua relação com os estudos lingüísticos ainda está sendo adquirido e desenvolvido.

Pode-se considerar que a extensão de um corpus comporta três dimensões, o que inclui o número de palavras, o número de textos e um número de gêneros, registros ou tipos textuais..A importância do número de palavras se deve a que em corpora maiores há maior possibilidade de ocorrerem palavras de baixa frequência, que representam a maior parte das palavras de uma língua. Uma língua é formada mais por palavras de baixa frequência do que por palavras de alta frequência. Já um grande número de textos em um corpus de um gênero específico contribui para uma melhor representação do gênero, registro ou tipo textual. (SARDINHA, p. 24-25)

Nosso autor (2004, p. 25) afirma também que apesar da extensão de um corpus ser de extrema importância para cumprir o quesito da representatividade, este ainda é um tema pouco explorado pelos estudiosos. Segundo ele, é possível definir três principais abordagens: impressionística, histórica e estatística.

A abordagem impressionística “baseia-se em constatações derivadas da prática da criação e da exploração de corpora, em geral, feita por autoridades da área” (p. 25). Para exemplificar, Sardinha comenta as definições de Aston, Leech e Sinclair. O primeiro considera um corpus de 20 a 200 mil palavras como sendo pequeno e um corpus de 100 milhões de palavras ou mais, como grande. Para Leech (1991), a quantidade mínima de palavras de um corpus deve ser de um milhão de palavras (taxa usual: *going rate*). Já Sinclair vincula a definição do tamanho mínimo ou ideal de um corpus à tecnologia disponível nos centros de pesquisa utilizados na época da coleta dos dados.

Quanto à abordagem histórica, “fundamenta-se na monitoração dos corpora efetivamente usados pela comunidade”. Sardinha (2004, p. 26), em 2002, elaborou uma classificação para os tamanhos de corpora baseado nos corpora explorados e divulgados por pesquisadores durante quatro anos de conferências de lingüística de corpus.

<b>Tamanho em palavras</b>	<b>Classificação</b>
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

A terceira abordagem adotada por Sardinha (2004, p. 27) é a Estatística, que consiste “na aplicação de teorias estatísticas capazes de identificar quantidades mínimas de palavras, gêneros e textos” ideais para uma amostra representativa através de fórmulas matemáticas.

Apesar dos diferentes tipos de abordagem existentes, Sardinha (2004, p. 98) reforça a necessidade de se cumprir alguns critérios básicos para a seleção de corpora de estudo e de referência no que tange à sua extensão e composição.

Em estudo baseado na “metodologia de estimação de amostras lingüísticas significativas” proposta por Biber , Sardinha (2002) pôde definir um quadro de tamanhos mínimos para corpora geral e especializado, usando como referência o quadro de categorias morfossintáticas da língua inglesa elaborado por Biber.

Estes quadros contribuem para a definição do “tamanho mínimo necessário para um corpus de estudo” que “depende: (a) do que se pretende estudar e (b) da especialização do corpus” (p. 98-99).

#### **Tamanhos mínimos para corpora segundo Biber**

<b>Categoria</b>	<b>Corpus geral</b>	<b>Corpus especializado</b>
Verbo	67187	13848
Substantivo	74551	8555
Adjetivo	149694	21234
Advérbio	205206	68953
Pronome	913256	40945
Numeral	1180815	91161

### Tamanhos mínimos para corpora segundo Berber Sardinha

Corpus	Todas as categorias	Excluindo-se a mais rara	Palavras de Conteúdo
Geral	5495048	1180815	205206
Especializado	91161	65432	68953

Sardinha atenta para alguns tópicos importantes:

- Estes quadros foram elaborados com base na língua inglesa e podem variar “de acordo com a língua e a especialização do corpus” (p. 99);
- “O corpus de referência não deve conter o corpus de estudo”, a fim de evitar interferência na lista de palavras-chave e na análise das palavras mais e menos frequentes no corpus (p. 100);
- “A composição do corpus de referência influencia os tipos de palavras que podem se tornar chave”. O ideal, portanto, é optar por corpora de referência que incluam a maior variedade de gêneros e, deste modo, sejam considerados como “escolha não marcada para estudos de palavra-chave” (p.101).

#### 2.4. Especificidade

A representatividade, a extensão e a especificidade de um corpus estão intrinsecamente ligadas. Para ser representativo um corpus necessita de determinada extensão, que deve ser grande, mas não pode ser total, devido à imensidão da linguagem. Uma forma de aumentar a representatividade é tornar o corpus mais específico. Por exemplo: um corpus que reunisse todas as obras de um autor seria certamente representativo desse autor.

A especificidade de um corpus também pode aumentar sua representatividade quando se restringe seu conteúdo, por exemplo, a uma só variedade de texto, já que “uma variedade específica da linguagem demonstra maior padronização e conseqüente menor variação no

nível do léxico, da gramática, do discurso, ou seja, apresenta maior grau de fechamento (*closure*)”. (SARDINHA, p. 28)

Assim, a restrição de um corpus pode ser feita com base em um determinado tipo de texto (ex. redação de prova vestibular), um tema (ex. fabricação de calçados), ou até mesmo um autor (ex. a obra de Guimarães Rosa), o que contribui significativamente para a Lingüística de Corpus e áreas afins, à medida que se criam corpora extremamente especializados. Porém, Sardinha (p.28) aponta uma dificuldade neste sentido: esses corpora mais específicos, por uma série de motivos, não têm sido disponibilizados por seus compiladores para a comunidade científica. “Desse modo, não satisfazem a condição de serem dados verificáveis, o que compromete a pesquisa em replicabilidade e generalidade.”

Corpora especializados podem ser criados também a partir de corpora de referência, o que pode ser vantajoso pelo fato de os grande corpora de referência, no geral, já estarem etiquetados e ordenados, facilitando o trabalho do pesquisador, tradutor ou interessados.

## **2.5. Corpora em diferentes línguas**

Apesar de a criação de corpora lingüísticos computadorizados ter sido iniciada em 1964, nos Estados Unidos, com o corpus Brown, atualmente a Europa se configura como centro de pesquisa de maior destaque e desenvolvimento. A Grã-Bretanha é um dos mais modernos centros de pesquisa do mundo, abrangendo Universidades como as de Birmingham, Brighton, Lancaster, Liverpool e Londres, que “dedicam-se à pesquisa baseada em corpus para a descrição dos mais variados aspectos da linguagem”. (SARDINHA, 2004, p. 5)

Todos estes avanços permitem o progresso da Lingüística de Corpus tanto no que se refere à teorização sobre a prática da criação e manipulação de corpora, quanto na disseminação desta necessidade para os estudos lingüísticos mundialmente. Estes fatos nos auxiliam a compreender a existência de uma quantidade superior de corpora em língua inglesa frente aos demais idiomas.

Sardinha (2004, p. 7-8), em seu *Lingüística de Corpus*, nos apresenta um primeiro quadro com quase 20 corpora de língua inglesa. São corpora escritos e falados que abrangem

desde o inglês americano, britânico, indiano, neozelandês, passando pelo inglês infantil falado (*POW – Polytechnic of Wales Corpus – e CHILDES – Child Language Data Exchange*) e pelo escrito em jornais e livros americanos (*Longman Written American Corpus*<sup>1</sup>), até o inglês escrito especificamente por estrangeiros (*LACLE – Longman Corpus of Learner’s English – “atualmente Longman Learner’s Corpus*<sup>2</sup>).

Destacam-se também e historicamente:

- 1) o pioneiro *Brown Corpus (Brown University Standard Corpus of Present-day American English)*, de 1964, com a incrível quantidade, para a época, de 1 milhão de palavras.
- 2) O *BNC (British National Corpus)*, de 1995, que “foi o primeiro a conter 100 milhões de palavras” (p. 8).
- 3) O *Bank of English*, que desde a sua criação em 1987<sup>3</sup> até 2002, atingiu a marca de 450 milhões de palavras. Caracteriza-se como “um corpus monitor, orgânico e em crescente expansão”, diferentemente dos dois anteriores, considerados “corpora de amostragem, planejados e fechados”. (p.9)

Quanto aos corpora de Língua Portuguesa, Sardinha (2004, p.9-10) nos traz uma gama de 21 corpora referentes ao português brasileiro e de Portugal, elaborados no Brasil, em Portugal, nos Estados Unidos e até mesmo em Nice, na França . Tratam-se de corpora escritos e falados, provenientes de diversas origens e, em alguns casos especificamente de jornais (*CETEM – Corpus de Extractos de Textos Electrónicos MCT/ Público e Modern Newspapers*), romances (*Modern Portuguese*) e documentos antigos (*Historical Portuguese Prose e Tycho Brohe Parsed Corpus of Historical Portuguese*).

Entre todos os corpora disponíveis, citamos quatro exemplos:

- 1) O *Banco de Português (PUC/SP)*, por ser o maior corpus de português brasileiro, escrito e falado, com 233 milhões de palavras.

---

<sup>1</sup> Componente do *Longman Corpus Network*.

<sup>2</sup> Componente do *Longman Corpus Network*.

<sup>3</sup> “Data refere-se ao *Birmingham Corpus*, do qual o *Bank of English* derivou” (p.8)

- 2) O corpus *Usos do Português* (UNESP/Araraquara), com 200 milhões de palavras de português brasileiro escrito.
- 3) O *COMET – Corpus Multilíngüe para Ensino e Tradução* (USP), com 5 milhões de palavras referentes ao português brasileiro escrito, por ser comparável com inglês.
- 4) O *CR-LW – Corpus de Referência Lácio-Web* (USP), que comporta 5 milhões de palavras do português brasileiro escrito.

Além dos 40 corpora em Inglês e Português, Sardinha (2004, p,12) também cita mais 11 corpora em diversas línguas: alemão, bósnio, chinês, croata, espanhol, francês, italiano, polonês, sueco, e tcheco. Chama-nos a atenção, em primeiro lugar, o *COSMAS (Corpus Access Storage and Maintenance System)*, com seus 1,7 bilhão de palavras em alemão, por superar em 278 % o corpus de inglês britânico *Bank of English*, o maior dos corpora em inglês mencionado por Sardinha, com 450 milhões de palavras até 2002.

Em segundo lugar na lista de corpora de outras línguas, por número de palavras, estão o *Corpus de Referencia del Español Actual (CREA)* e o *Polish National Corpus* (polonês), com 130 milhões de palavras. É interessante observar que o corpus francês *Frantext* não apresenta um número de palavras, e sim a marca de 3.500 obras dos séculos XVI ao XX, sendo 80 % das obras sobre literatura.

Os dados apresentados por Sardinha nos mostram a crescente importância da Lingüística de Corpus em diferentes países.

### **3. Organização de corpus**

O corpus é o elemento essencial para a efetivação dos estudos e análises em Lingüística de Corpus. Ele pode ser constituído por textos de origem impressa (jornais, revistas, folhetos, livros, cartas) ou eletrônica (notícias publicadas na Internet, conteúdo de blogs, e-mails, jornais e revistas eletrônicas). O início dos trabalhos com corpora de grandes dimensões foi marcado pela utilização de corpora não computadorizados manipulados através de fichas, as quais eram criadas, organizadas e utilizadas de forma manual, dificultando a pureza dos dados pelo controle humano e a replicação deles.

Desde os anos de 1960, com a invenção do computador, os centros de pesquisa priorizaram a utilização de corpora computadorizados. Alguns corpora manuais tiveram seus dados inseridos nestas máquinas. Atualmente, com a popularização da Internet (*Word Wide Web*) entre os pesquisadores, é muito comum a utilização de textos já disponibilizados ao público eletronicamente. Porém, nada impede que um pesquisador digitalize textos impressos e os transforme num corpus eletrônico. Possuir um corpus eletrônico é o primeiro passo para se analisar um determinado conteúdo utilizando as ferramentas computacionais atuais.

### 3.1. Coleta e armazenamento de dados

A coleta de um corpus eletrônico pode ser feita de diferentes modos:

1. Segundo Sardinha (2004, p. 46-47), é possível, através de uma ferramenta chamada *offline browser*, coletar todo o conteúdo de um *website* e salvá-lo no disco rígido de um computador. O exemplo dado pelo autor se trata de baixar o conteúdo completo de edições eletrônicas do jornal *O Estado de São Paulo*. Para obter mais detalhes sobre a utilização desta ferramenta, consultar o capítulo 2 “Coleta, armazenamento e pré-processamento de corpora” na obra supracitada.
2. Um corpus de estudo especializado também pode ser proveniente de um corpus maior. O corpus de Referência *Lácio-Web*, por exemplo, contém textos relacionados às ciências agrárias, biológicas, humanas, exatas, etc., o que permite que o pesquisador copie os textos apenas do tema ciências humanas e crie assim um subcorpus especializado, que poderá ser utilizado em seus estudos ou como ferramenta de apoio para traduções relacionadas ao tema. Apesar de este ser um corpus de referência, é importante levar em consideração a sugestão de Sardinha (2004, p. 100): não utilizar como corpus de referência um corpus que contenha o corpus de estudo.
3. Outra opção é coletar os textos diretamente em determinadas *webpages*. Por exemplo: um estudo focado na terminologia relacionada ao comportamento humano frente a situações corporativas pode encontrar material em *webpages* de empresas de qualquer ramo de atividade ou, mais especificamente, em *webpages* de consultorias de Recursos Humanos e até mesmo nas de agências de emprego.

Quanto ao armazenamento, dependerá das exigências do programa utilizado para a análise. Veremos a seguir mais detalhes sobre dois programas computacionais utilizados em Lingüística de Corpus.

### **3.2. Ferramentas**

Para realizar trabalhos em Lingüística de Corpus com o objetivo de descrever determinados fenômenos lingüísticos, é necessária a utilização de programas cujas ferramentas permitem essencialmente gerar listas de palavras, palavras-chave e concordâncias.

A seguir são apresentadas algumas funcionalidades do programa *WordSmith Tools* e comentários sobre o programa *AntConc*. O primeiro tem seu uso já consagrado na área, tanto pela qualidade da organização dos dados gerados, quanto pela variedade de funções que disponibiliza. Já o segundo, além de oferecer as principais funcionalidades do *WST*, se destaca ainda por ser um programa de acesso livre, permitindo que as listas de palavras sejam geradas na íntegra, sem a necessidade de compra de uma licença de uso.

#### **3.2.1. WST – *WordSmith Tools***

O *WST – WordSmith Tools* é um programa de computador que permite descrever a linguagem empregada em um determinado corpus, através da identificação e comparação de freqüências e listagem de palavras contextualizadas. O autor desta robusta ferramenta, que pode ser obtida via Internet, é Mike Scott. Baseados em Sardinha (2004, cap. 3), apresentamos a seguir, de forma sucinta, as funcionalidades deste programa em sua versão 3.0.

O programa depende de três princípios básicos para ser capaz de efetuar análises lingüísticas.

- 1) Ocorrência: o item deve estar presente no corpus pelo menos uma única vez, para que possa ser observado.
- 2) Recorrência: o item deve estar presente ao menos duas vezes no corpus.
- 3) Coocorrência: um item aparece em companhia de outro item, de forma sequencial ou não.

O *WST* é composto por três ferramentas (*WordList*, *KeyWords*, *Concord*), quatro utilitários (*Renamer*, *Text Converter*, *Splitter*, *Viewer*), instrumentos e funções.

**Descrição da ferramenta *WordList*:** permite que sejam geradas listas das palavras presentes no corpus. A ferramenta gera sempre duas listas, uma em ordem alfabética e outra por ordem decrescente de frequência, além de uma terceira janela com estatísticas relativas aos dados usados na produção das listas.

**Instrumentos de análise:** *Wordlist*, lista de palavras individuais; *Wordlist, cluster activated*, lista de multipalavras; *Detailed consistency*, lista de palavras de consistência individuais; *Detailed consistency, clusters activated*, lista de multipalavras de consistência; *Statistics*, lista de dimensões e densidade lexical.

**Descrição da ferramenta *KeyWords*:** “permite a seleção de itens de uma lista de palavras (ou mais) por meio da comparação de suas frequências com uma lista de referências. O resultado do contraste é uma lista de palavras-chave, ou palavras estatisticamente diferentes no corpus de estudo e no corpus de referência” (SARDINHA, 2004, p. 96). Para se efetuar esta análise de palavras-chave é imprescindível ter um corpus de estudo e um corpus de referência, selecionados segundo critérios de extensão, composição, especialização e objetivos a serem alcançados.

**Instrumentos de análise:** *Keywords*, lista de palavras-chave; *Database*, banco de dados de listas de palavras-chave; *Key keywords*, lista de palavras-chave-chave; *Associates*, lista de palavras-chave associadas; *Clumps*, lista de agrupamentos textuais; *Keyword plot*, gráfico de distribuição de palavras-chave; *Keyword plot links*, listagem de elos entre palavras-chave.

**Descrição da ferramenta Concord:** “essa ferramenta produz concordâncias ou listagem das ocorrências de um item específico (chamado palavra de busca ou nóculo, que pode ser formado por uma ou mais palavras) acompanhado do texto ao seu redor (o contexto)” (SARDINHA, 2004, p. 105). Dependendo da posição do item na listagem, é possível ter uma variedade de tipos de concordância. Estas concordâncias são essenciais para os estudos de colocação e padronização lexical. “O *Concord* pode ser usado separadamente, para concordâncias avulsas, ou em conjunto com as ferramentas *Wordlist* e *Keywords*, chamado a partir desses programas” (p. 107).

**Instrumentos de análise:** *Concordance*, concordância; *Collocates*, lista de colocados; *Clusters*, lista de agrupamentos lexicais; *Patterns*, lista de padrões de colocados; *Plot*, gráfico de distribuição da palavra de busca.

As funções principais das três ferramentas descritas acima são:

- Lematizar: agrupar em um mesmo item duas ou mais formas diferentes.
- Classificar: ordenar as listas e concordância de acordo com o alfabeto, a frequência ou a posição.
- Delimitar: escolher as partes do corpus que devem ser lidas pelo programa.

Sardinha (2004, p. 89) aponta também alguns ajustes que podem ser feitos pelo usuário para aperfeiçoar o funcionamento destas três ferramentas:

- Horizonte de concordância: é quantidade de palavras ao redor do nóculo definida pelo usuário do programa para ser a faixa dentro da qual se calcularão os colocados;
- Tamanho da multipalavra: determina o tamanho de cada multipalavra, ou seja, quantas palavras tem cada multipalavra;
- Tamanho do agrupamento lexical: “influencia os agrupamentos encontrados na concordância e na produção de listas de palavras”.
- Prova estatística para identificação das palavras-chave: “teste executado durante a comparação das frequências do corpus de estudo com o de referência”.
- Frequência mínima: “número mínimo que um item deve possuir para se tornar elegível para um instrumento”.

- Extensão máxima: “número máximo de ocorrências de uma concordância ou de uma lista de palavras-chave”.

### **3.2.2. AntConc**

O AntConc é um programa que funciona basicamente nos moldes do WordSmith Tools. Ele permite a extração de listas de palavras (*WordList*), listas de concordâncias (*Concordance*) e palavras-chave (*KeyWord*). Uma diferença importante entre estes dois programas é que o WST, em sua versão gratuita, permite que sejam extraídas listas de palavras com um número restrito de itens. No caso do AntConc, o número de itens é livre.

Instruções sobre como instalar e utilizar o programa *AntConc 3.2.1 - A Freeware Concordance Program*, produzidas por Aauri Brezolin, podem ser facilmente encontradas no site do Projeto COMET – Corpus Multilíngüe para Ensino e Tradução (<http://www.fflch.usp.br/dlm/comet/>).

## **4. Análise de corpus**

A análise de um corpus eletrônico pode ser feita a partir das listagens de frequência de palavras, concordâncias e colocados geradas por programas como o *WST* e o *AntConc*. Este procedimento visa obter dados sobre padrões e fenômenos lingüísticos com o intuito de contribuir com os estudos em Lingüística de Corpus, bem como auxiliar, quando possível, o trabalho de tradutores, professores, lexicógrafos e outros profissionais.

### **4.1. Frequência de palavras**

A listagem de frequência de uso das palavras, que pode ser prontamente gerada, tão logo se tenha o corpus digitalizado, fornece base para o entendimento da língua como um todo, seja esta base um corpus de língua geral, como o Banco de Português, ou um corpus especializado em algum tema. A lista de frequências permite observar quais palavras ocorrem

em maior número no corpus e as que ocorrem menos vezes, sendo assim chamadas raras. A relevância desta informação se deve também ao fato de o ser humano não ter habilidades para prover estes dados intuitivamente. É preciso observar empiricamente a frequência de emprego das palavras por diferentes pessoas em contextos definidos para então obter a probabilidade de ocorrência de traços ou estruturas. (SARDINHA, 2004, p. 160)

As listas de frequência, que em alguns casos são denominadas dicionários, podem ser publicadas em diferentes formatos e nas mais diversas línguas. Podem ser parte integrante de um volume especializado, no qual serão publicadas na íntegra. Podem ser publicadas juntamente com outros trabalhos de forma parcial e até mesmo “eletronicamente na Internet, com comentários ou apenas com o conteúdo lexical e numérico”. (Sardinha, 2004, p. 160)

Para exemplificar as publicações destes trabalhos, nos baseamos em algumas línguas de referência: inglês, espanhol, francês, romeno, português, chinês. Com o intuito de colaborar com o ensino de leitura nos Estados Unidos, Edward Thorndike (1921) efetuou manualmente um levantamento frequencial passível de ser utilizado na elaboração de materiais didáticos na época. Irving Lorge (1944) também contribuiu com a elaboração de uma nova versão desta lista e Michael West (1953) publicou a *General Service List of English Words*, ainda produzida de forma manual.

Em 1967, Kucera e Francis, valendo-se do advento do computador, realizaram novas pesquisas e publicaram, além do *Computacional analysis of present-day American English*, uma nova análise frequencial do inglês norte-americano, tendo como base o Corpus Brown. Além desta, houve outras publicações como o *American heritage word frequency book*, de Carroll, Davies e Richman, de 1971, e o *Frequency analysis of English vocabulary and Grammar* (2 vols.), de Johansson e Hofland, publicado em 1989 pela Oxford. O trabalho mais recente apontado por Sardinha (2004, p. 161) é a *Word frequencies in written and spoken English based on British National Corpus*, de Leech, Rayson e Wilson, publicado pela Longman em 2001.

Este levantamento nos permite observar a velocidade com que estes tipos de trabalho são publicados. Em 34 anos tivemos nesta língua, basicamente, quatro grandes trabalhos. Excluindo aqui, as muitas listas que foram divulgadas através da Internet, porém sem serem de fato publicadas. Um exemplo são as listas de Adam Kilgarriff, baseadas no *BNC*.

Alphonse Juilland publicou, sozinho ou em parceria com outros estudiosos, uma série de *frequency dictionaries* de línguas neolatinas. O referente ao espanhol foi lançado em meados dos anos de 1960. Logo a seguir veio o do romeno, seguido anos mais tarde pelo francês, terminando com o *Frequency dictionary of Italian words*, lançado em 1973. Juilland ainda foi editor da série da editora Mouton, encarregada da publicação de um dicionário de frequências do chinês. Infelizmente, não foi incluído na série um dicionário referente ao português, embora Juilland pareça ter tido essa idéia em seus planos. (SARDINHA, 2004, p. 161)

Quanto ao português, há algumas publicações:

- a) “A computacional frequency and range list of five hundred Brazilian Portuguese words”, publicada por Kelly, através da editora da Universidade de Wisconsin, em 1970;
- b) *Frequency dictionary of Portuguese words*, elaborada por Duncan em 1972, “como parte de sua tese de doutoramento em Stanford, a partir de um corpus de português europeu com 500 mil palavras”;
- c) “A face quantitativa da linguagem: um dicionário de frequências do português”, listas de frequências publicadas por Biderman, em 1998, no Brasil. A referência para a produção destas listas foi o Corpus do Português Brasileiro Contemporâneo.

Há outros trabalhos também, como o projeto *Português Fundamental*, coordenado por Bacelar do Nascimento. Projetos como estes visam também prover informações para a elaboração de materiais didáticos, tanto para o ensino de língua materna, quanto para o ensino de língua estrangeira, através da publicação de listas de frequências baseadas em determinados corpora. (SARDINHA, 2004, p. 162). O *Dicionário UNESP do Português Contemporâneo*, publicado, em 2004, por Francisco S. Borba, é um dicionário de português brasileiro elaborado com base nas frequências do vocabulário extraídas de um corpus de 90 milhões de palavras.

A análise de frequência das palavras constitui uma das mais relevantes vertentes de estudo da linguagem tanto para a Lingüística de Corpus, quanto para a Estatística Léxica e a Lingüística Quantitativa. Considerando a linguagem como um sistema probabilístico, o estudo de frequências visando o entendimento do léxico é importante para apontar quais palavras são mais ou menos utilizadas em textos escritos e orais. Este dado contribui também, segundo Biderman (1998 apud SARDINHA, p. 163), para o estabelecimento de normas lingüísticas.

Sinclair, Halliday e Sampson compartilham de uma mesma visão probabilística, reforçada por Guiraud, que afirma em 1954 (apud SARDINHA, 2004, p. 162-163): “as palavras não se dispõem em um plano uniforme no léxico: algumas têm mais do que outras a oportunidade de ser empregadas com frequência”. As palavras podem apresentar frequência de uso alta, intermediária ou baixa. Segundo Guiraud, a palavra é não somente uma criação individual, como também uma criação coletiva, e assume seu valor à “medida em que é aceita, retomada, repetida; por isso ela se define, afinal, pela soma de seus empregos.” Para Biderman (1998 apud SARDINHA, 2004, p. 163):

A frequência é uma característica típica da palavra. Aliás, a norma lingüística se baseia na frequência dos usos lingüísticos. Assim, a norma lingüística nada mais é do que a média dos usos frequentes das palavras que são aceitas pelas comunidades dos falantes. E não é só isso. Também as mudanças lingüísticas que, no decorrer da história, levam de um estado da língua a outro, advêm das frequências de certos usos em detrimento de outros.

Para analisar as frequências de uma língua de modo eficiente, é necessário um corpus eletrônico. Este tipo de análise já foi realizado em diversas línguas, incluindo o Português brasileiro, porém neste caso, em menor quantidade. Além disso, é necessário ter claras as definições de alguns termos-chave. Segue abaixo a lista de definições proposta por Sardinha (2004, p. 165-166).

- Palavras, ocorrências ou itens (*tokens*): contagem de palavras corridas (*running words*); cada palavra conta como uma ocorrência, mesmo que seja repetida. Por exemplo, a frase *o gato e o cão são animais* possui sete palavras *o* (1), *gato* (2), *e* (3), *o* (4), *cão* (5), *são* (6), *animais* (7).

- Palavras, formas, vocábulos ou tipos (*types*): número de palavras diferentes; a frase acima possui seis formas *o* (1), *gato* (2), *e* (3), *cão* (4), *são* (5), *animais* (6), pois a forma *o* apareceu duas vezes.
- Palavras de conteúdo (*content words*), lexicais, plenas ou de classe aberta (*open class words*): palavras que são das categorias morfossintáticas substantivo, adjetivo, verbo, advérbio, interjeição e numeral. Algumas classificações somente incluem aqui os verbos principais.
- Palavras gramaticais (*grammatical words*), funcionais (*function words*), de classe fechada (*closed class*) ou instrumentais: palavras de categorias fechadas (pronome, artigo, preposição, conjunção; algumas classificações incluem aqui verbos auxiliares e modais).
- Lema (*lemma*, *lemmas* ou *lemmata* no plural em inglês) ou lexema: palavras que resumem as suas diversas variantes. Por exemplo, o lema *querer* inclui as várias conjugações desse verbo: quero, queres, queria etc.
- *Hapax legomena*: itens que ocorrem apenas uma vez (isto é, de frequência 1) no corpus.
- *Hapax dislegomena*: itens que ocorrem mais de uma vez no corpus.

Sardinha atenta para a ambigüidade do termo *palavra*, que segundo as definições acima pode se encaixar tanto em *tokens* quanto *types*. Esta é uma discussão que persiste na língua portuguesa e que requer um esclarecimento no momento de uso ou referência. Neste trabalho, trataremos, basicamente, *tokens* como ocorrências e *types* como formas, conforme Biderman (apud SARDINHA, 2004, p. 167).

De modo geral, o estudo de frequência de palavras nos permite analisar palavras mais e menos frequentes, frequência das palavras gramaticais e de conteúdo, além da detecção de empréstimos de outras línguas, o que chamamos de estrangeirismos. Estes aspectos serão abordados de forma mais detalhada ao longo deste trabalho.

## 4.2. Concordâncias e colocados

Concordância é a lista de ocorrências de um determinado item, que permite a análise do seu emprego segundo um contexto de uso. Um modo de iniciar esta análise é a criação de

uma lista eletrônica de frequências, utilizando-se programas como o WST e o AntConc, citados anteriormente. Com base nesta lista são selecionadas as palavras com as quais se quer trabalhar. Estas palavras são utilizadas para gerar a lista de concordância, na qual serão apresentadas de forma centralizada e destacada, acompanhadas pelos outros termos que constituem seu contexto original no corpus. Após esta fase, é possível também definir, utilizando medidas estatísticas de associação, se uma relação entre a palavra selecionada (nódulo) e estes outros termos é aleatória ou não. No caso de uma relação não aleatória, o termo que co-ocorre é chamado de ‘colocado’.

Há vários cálculos estatísticos que são aplicados para verificar o grau de associação entre palavras. Um dos mais comuns é a razão O/E, isto é, Observado/Esperado. Essa medida leva em conta quantas vezes de fato duas palavras ocorreram juntas dentro do horizonte delimitado (o valor observado) e quantas vezes seria esperado que elas coocorressem (o valor esperado), dado o tamanho do corpus e a frequência de cada palavra. (SARDINHA, 2004, p. 201)

Além da razão O/E, há outras formas de medição: a Informação Mútua (*Mutual Information*) e o Escore T (*T-Score*), que são apresentadas detalhadamente por Sardinha (2004, p. 203-217). Ele explica também como fazer cálculo de estatísticas de associação via Internet, WST e em uma planilha eletrônica como a do aplicativo Excel.

Sardinha (2004, p. 187-188) afirma que para desenvolver análises baseadas em concordância, é necessário seguir alguns procedimentos que, normalmente, não são explicados por estudiosos em seus relatos de pesquisa. No intuito de contribuir fortemente com os estudos de Lingüística de Corpus, ele relata com detalhes, no capítulo “Análise de palavras em contexto: concordâncias” (2004) os passos para desenvolver este tipo de trabalho utilizando o WST. Ressalta que além de conhecer o funcionamento desta ferramenta, é preciso entender alguns termos-chave (listados abaixo) para seguir com o processo organização e análise dos dados.

- Nódulo (*node* ou *node word*): palavra (ou grupo de palavras) da qual o usuário tem interesse em obter uma concordância.
- Palavra(s) de busca (*search word(s)*): o mesmo que nódulo.

- Palavra(s) de contexto (*context word(s)*): palavras que devem estar próximas do nóculo durante a busca. São opcionais, ou seja, é possível fazer uma concordância sem especificá-las. Por exemplo: para buscar as expressões *tem interesse*, *tem algum interesse*, *tem nenhum interesse* e outras variações, deve-se especificar *interesse* como palavra de contexto e *tem* como nóculo (ou vice-versa).
- Horizonte(s) (*horizon(s)*): distância máxima entre o nóculo e a palavra de contexto. No exemplo, a horizonte entre *tem* e *interesse* é zero para a esquerda (ou seja, *interesse* não deve ocorrer à esquerda de *tem*) e dois à direita (ou seja, *interesse* pode ser a primeira ou a segunda palavra à direita de *tem*). Normalmente, usa-se um horizonte de 2:2 ou 3:3, isto é, de duas ou três palavras de cada lado do nóculo (Stubbs, 1995, p. 33), mas não mais que quatro, pois segundo Sinclair (1995), as relações colocacionais de interesse não estão além desta faixa.
- Janela (*window*): o mesmo que horizonte.
- Caractere curinga (*wildcard*): um caractere que pode significar muitos outros. Tipicamente, é o asterisco. Por exemplo, se o usuário tiver interesse em buscar as diversas formas do verbo *anunciar* (anuncio, anuncias, anuncia, etc.), pode digitar como palavra de busca a seguinte expressão: *anunc\**.
- Colocado(s) (*collocate(s)*): palavra(s) que ocorre(m) ao redor do nóculo, em posições relativas (primeira à esquerda, segunda à esquerda). Difere de palavra de contexto porque esta é opcional, definida pelo usuário no momento da busca. Os colocados, contudo, são todas as palavras que ocorrem perto do nóculo, dentro do horizonte especificado, incluindo as palavras de busca que existirem.
- Classificação (*sorting*): ordenação das linhas da concordância por critério alfabético de acordo com uma posição selecionada da concordância, como por exemplo o primeiro colocado à esquerda do nóculo ou o segundo colocado à direita. Pode ser feita de modo cumulativo, isto é, por mais de uma posição. Por exemplo, as linhas podem ser classificadas de acordo com a ordem alfabética da primeira palavra à esquerda (como critério principal) e pela segunda palavra à esquerda (como critério secundário). (SARDINHA, 2004, p. 188)

Tagnin (2007), em seu artigo “A identificação de equivalentes tradutórios em corpora comparáveis”, mostra a importância e utilidade da análise baseada em listas de concordância através da análise de dois corpora, um em inglês e um português, sobre “Instrumentos Contratuais”. Após extrair listas de frequências em cada uma destas línguas, verificou que a palavra de conteúdo *contrato* era a de maior frequência, com 1832 ocorrências. Isso a levou a buscar na lista de frequências em inglês seu cognato *contract*. Observou que esta palavra ocorria apenas 186 vezes, um número relativamente baixo para um corpus especializado sobre

o mesmo tema, enquanto que a palavra *agreement* ocorria 1724 vezes. Através desta análise de concordância, cujos corpora se caracterizavam por apresentarem contextos similares, foi possível confirmar a equivalência entre os termos *contrato* e *agreement* e não entre *contrato* e *contract*.

Com a busca da palavra *contract*, Tagnin observou que ela co-ocorria com a palavra *adhesion*. Desta forma, encontrou no corpus ocorrências para *adhesion contract*, mas não para *adhesion agreement*. Confirmou estas ocorrências com uma busca no Google e concluiu que identificar os termos com maior número de ocorrência não é o bastante para considerá-los equivalentes em duas línguas, é preciso analisar o contexto em que ocorrem, assim como seus colocados.

### 4.3. Padronização

A padronização é o conjunto de “regularidades formadas pela recorrência de itens coocorrentes” e se aplica a todas as palavras de uma língua. (SARDINHA, 2004, p. 221) Para realizar análises neste campo é necessário trabalhar com a concordância e a lista de colocados, visando a observação dos contextos em que estes itens ocorrem. A partir deste ponto é possível listar quais palavras, de conteúdo ou não, acompanham este termo e que sentidos lhe agregam, dependendo também da posição em que se apresentam.

A quantidade de pesquisas sobre a padronização do português brasileiro ainda é pequena em relação aos trabalhos sobre a língua inglesa, que representam a maioria. Volpe Nunes et al., em seu artigo “Desafios na construção de recursos lingüísticos para o processamento do português do Brasil”, publicado em 2005 (apud SARDINHA, 2004, p. 222), consideram como os principais desafios no país a “criação de recursos lingüísticos e a formação de equipes interdisciplinares”.

Apesar disto, há publicações acerca de frequências, marcadores conversacionais, colocações referentes a cores, unidades polilexicais, como *greve de fome* e *cair do céu*,

padrões identificados em corpora de português brasileiro e europeu, além de trabalhos voltados à tradução entre o português e o inglês.

Sardinha (2004, p. 233-234), em seu *Linguística de Corpus*, apresentou um estudo sobre a padronização da partícula *só*, baseada em um corpus de português brasileiro. Iniciou o processo de análise extraindo 16 mil ocorrências da palavra *só* num corpus de 160.280 ocorrências, utilizando o WordSmith Concord. Selecionou os 150 colocados mais frequentes e transportou estes dados para um programa de cálculos estatísticos, o SPSS. Classificou então, através de uma Análise de Agrupamento (*Cluster Analysis*), estes colocados em grupos de acordo com a sua frequência em relação às posições que apresentavam em torno do termo principal. Diferenciando-se da maioria dos estudos, que realizam a análise tendo como base apenas a lista de concordâncias.

Interpretou as concordâncias de modo qualitativo e observou que a) os verbos apresentavam uma padronização consistente; b) algumas formas do mesmo verbo possuíam padronização distinta; c) ocorria a intercolocação, de onde se originam padrões mais complexos; d) padrões diferentes realizam sentidos também diferentes; e e) os sentidos apurados não puderam ser dedutíveis de explicações dicionarizadas do português. Após esta análise, apontou os fatores positivos e negativos decorrentes das ferramentas disponíveis e os pontos que “corroboram os princípios da descrição léxico-gramatical constante na literatura da Linguística de Corpus”. São eles:

1. as escolhas léxico-gramaticais não são feitas em isolamento, mas sim em cadeia, segundo o princípio idiomático;
  2. as palavras possuem padronização, observável a partir da exploração de um corpus computadorizado;
  3. formas derivadas de uma mesma raiz devem ser consideradas itens lexicais distintos; e
  4. diferenças entre padrões colocacionais indicam diferenças de sentidos.
- (SARDINHA, 2004, p. 234)

## 5. Frequência de palavras em corpus multilíngüe especializado em reciclagem

Para iniciar estudos em Lingüística de Corpus, algumas das ações essenciais que precisam ser tomadas são a coleta e a organização de um corpus eletrônico. Para este trabalho, optamos por criar um corpus multilíngüe especializado em reciclagem. A escolha do tema se deve à sua contemporaneidade e à possibilidade de fornecer um material especializado para profissionais da área de Lingüística de Corpus, Tradução e áreas relacionadas ao tema.

Inicialmente, planejamos trabalhar apenas com textos em português brasileiro e espanhol, porém cogitamos a hipótese de este corpus apresentar influências da língua inglesa, pelo fato de os Estados Unidos estarem bem desenvolvidos em relação ao tema da reciclagem. Desta forma, optamos por criar um corpus multilíngüe formado por três línguas: Português (Brasil), Inglês (Estados Unidos) e Espanhol (Espanha). A partir da criação deste corpus, propomos a extração de listas de palavras, através do programa AntConc, e a análise das faixas de frequência de formas e ocorrências, trabalho essencial para o desenvolvimento de outras análises como a padronização e a colocação.

### 5.1. Reciclagem

Reciclagem, segundo umas das definições do *Dicionário Houaiss da língua portuguesa*, em formato eletrônico, é a “recuperação da parte reutilizável dos dejetos do sistema de produção ou de consumo, para reintroduzi-los no ciclo de produção de que provêm Ex.: r. do papel, do vidro, da água”. (<http://houaiss.uol.com.br>)

A preocupação com a reciclagem vem se tornando cada vez mais freqüente em uma série de países, mas com proporções e abordagens diferentes. Isto se deve à necessidade de empresas e pessoas de reaproveitarem materiais com o intuito de reduzir o impacto do homem sobre o meio ambiente (reutilizando vasilhames, sacos plásticos e papel, por exemplo) e também obterem benefícios econômicos (reduzindo custos de produção através da reutilização de materiais obsoletos como matéria-prima ou através da reutilização doméstica de utensílios e materiais, ao invés de comprar novos) e sociais (gerando empregos principalmente para comunidades carentes, que vêem neste processo uma oportunidade de se integrarem ao mercado de trabalho).

O ato de reciclar é um dos pilares da Sustentabilidade Ambiental, que envolve também a educação ambiental, a elaboração de leis sobre o tema e o desenvolvimento de produtos e tecnologias que impactem o mínimo possível o ambiente e possam garantir ao máximo a conservação do planeta para as gerações futuras.

## 5.2. Escolha do corpus

Segundo Sardinha (2004, p.18-19), para definir as características de um corpus é necessário considerar alguns pontos importantes, como a origem dos dados, o propósito, a composição, a formatação, a representatividade e a extensão do corpus. O corpus escolhido apresenta as seguintes características:

- Origem dos dados: textos coletados em websites de organizações que prestam serviços relacionados ao tema da reciclagem no Brasil, Estados Unidos e Espanha.
- Propósito: estudar as faixas de frequência de formas e ocorrências.
- Composição: textos relacionados ao tema da reciclagem, o que abrange informações sobre a destinação de resíduos e o meio ambiente.
- Formatação: textos em formato eletrônico, salvos na extensão .txt.
- Representatividade: o corpus multilíngüe é especializado no tema reciclagem.
- Extensão: mínimo de 80.000 palavras em cada língua.

Para caracterizar o corpus foi necessário também considerar a tipologia proposta por Sardinha (2004, p. 20-21):

- Modo: escrito.
- Tempo: contemporâneo.
- Seleção: de amostragem e estático.
- Conteúdo: especializado e multilíngüe.
- Autoria: de língua nativa.
- Finalidade: de estudo.

Segundo Sardinha (2004, p. 98-99), um corpus especializado deve ser constituído de no mínimo 68.953 palavras de conteúdo. As palavras de conteúdo, segundo Biber (apud SARDINHA, 2004, p. 98-99), são verbo, substantivo, adjetivo e advérbio. Para o corpus em questão, optamos por considerar as palavras de conteúdo e os pronomes, excluindo apenas os numerais, que se fossem incluídos levariam o corpus a um mínimo de 91.161. Deste modo, adicionando uma margem de 16%, definimos o corpus mínimo representativo de uma linguagem especializada com 80.000 palavras, sendo um corpus pequeno médio, segundo a abordagem histórica apresentada por Sardinha em 1.9 (p. 26).

Em seu corpus especializado na linguagem profissional, Sardinha (2004, p. 147) definiu o número mínimo de palavras como sendo 100.000, o que inclui a categoria dos numerais. Neste caso, seria necessária também uma análise de sentido para diferenciar termos como *um* numeral e *um* artigo indefinido masculino, em português, por exemplo. Não consideramos relevante para o desenvolvimento do trabalho a inclusão dos numerais, o que pôde ser corroborado quando observamos na lista de frequências a presença de numerais da seguinte forma:

<b>Língua</b>	<b>Ranque</b>	<b>Frequência</b>	<b>Forma</b>
Português	15	582	um
Português	75	110	mil
Espanhol	13	868	un
Espanhol	99	77	dos
Inglês	73	165	one
Inglês	82	143	million

A representatividade de nosso corpus pode ser confirmada à medida que apresenta a extensão proposta por Sardinha (2004, p. 98-99) e um alto grau de especificidade, que se deve à seleção criteriosa de textos eletrônicos sobre o tema reciclagem. Os textos tratam mais especificamente dos resíduos gerados pela indústria e pela comunidade, suas características, históricos, processos de reciclagem e reutilização pré e pós-processamento.

### **5.3. Contribuição para a Lingüística de corpus**

As contribuições que uma coleta e um estudo de corpus como os que realizamos podem trazer para a Lingüística de Corpus são várias. Estas contribuições englobam: a corroboração de teorias lingüísticas através do emprego de metodologias de coleta; a comprovação do adequado funcionamento de programas de computador não tão consagrados, como o AntConc, que permite a descrição da linguagem de forma organizada, para a criação de listas de palavras; a produção de um corpus especializado em um tema contemporâneo, como o da reciclagem; o fornecimento de dados para a possível elaboração de dicionários e materiais didáticos, além de, se publicadas na internet, servirem como material de consulta para tradutores, especialistas na área de reciclagem, lingüistas e outros profissionais.

#### **5.4. Metodologia**

Para produzir o corpus multilíngüe especializado em reciclagem, foi necessário, primeiramente, estudar a obra de um dos maiores especialistas em Lingüística de Corpus no Brasil: *Lingüística de Corpus*, de Tony Berber Sardinha (2004). A partir da incorporação de conceitos básicos foi possível definir características essenciais para a coleta do corpus (ver 5.1.2), fazer um levantamento de dados que se dividiu em três fases, inserir dados no programa AntConc e, com esta ferramenta, gerar listas de palavras, listas de palavras-chave e realizar uma análise das faixas de frequência de formas e ocorrências.

##### **Levantamento de dados – fase I**

Após breve pesquisa e observação de websites que apresentam conteúdo relacionado à Reciclagem e Meio Ambiente, optou-se por iniciar o levantamento de dados nas páginas dos Ministérios do Meio Ambiente do Brasil, Espanha e Estados Unidos. Esta escolha se deveu à possibilidade de obter dados comparáveis, oriundos de instituições confiáveis e produzidos, supostamente, em sua língua nativa.

Este piloto permitiu observar as diferenças de forma e conteúdo existentes entre websites relacionados ao mesmo tema. O primeiro levantamento foi feito com o website do Ministério do Meio Ambiente do Brasil ([www.mma.gov.br](http://www.mma.gov.br)), no qual foram coletados dados sobre consumo sustentável, economia e gestão de resíduos sólidos. O website seguinte foi o

do Ministerio de Medio Ambiente da Espanha ([www.mma.es](http://www.mma.es)), cujo conteúdo coletado trata da descrição do Ministério e do tratamento de águas.

Observou-se certa dificuldade em coletar dados comparáveis e estritamente ligados ao tema Reciclagem, visto que os Ministérios apresentam, em geral, informações relacionadas às agências governamentais de meio ambiente e estatísticas. Deste modo, optou-se por encerrar a coleta de dados dos Ministérios neste ponto, devido ao fato de o conteúdo não atender aos objetivos inicialmente definidos para a pesquisa.

O website do Ministério do Meio Ambiente do Brasil apresenta dados sobre a Instituição e depois se divide em três partes denominadas Temas, Colegiados e Serviços. Os dados sobre a Instituição tratam-se, na verdade, da localização do Ministério em Brasília e de uma lista de outros órgãos, colegiados e entidades relacionadas. Cada uma das três partes trata de assuntos diversos:

- Temas: Agenda 21; Amazônia; assuntos internacionais; biodiversidade e florestas; conferência nacional; desenvolvimento sustentável; educação ambiental; fortalecimento do SISNAMA (Sistema Nacional de Meio Ambiente); mudanças climáticas e qualidade ambiental; recursos hídricos e ambiente urbano; revitalização do São Francisco.
- Colegiados apresenta os vários conselhos formados para apoiar o governo: CGEN – Conselho de Gestão do Patrimônio Genético; CNRH – Conselho Nacional de Recursos Hídricos; CONABIO – Comissão Nacional de Biodiversidade; CONACER – Comissão Nacional do Programa Cerrado Sustentável; CONAMA – Conselho Nacional do Meio Ambiente; FNMA – Fundo Nacional do Meio Ambiente.
- Serviços: biblioteca virtual; consulta pública e editais; eventos; nosso ambiente; sala de imprensa; Sibeia (Sistema Brasileiro de Informação em Educação Ambiental); SINIMA (Sistema Nacional de Informação sobre Meio Ambiente).

O tema desenvolvimento sustentável, através do “Guia de boas práticas para o consumo sustentável”, foi o maior provedor de informações relacionadas com o tema da reciclagem. Este guia apresenta orientações sobre como consumir de forma sustentável e dados sobre os recursos disponíveis no planeta.

Os oito textos coletados neste website tratam de consumo sustentável; comércio e meio ambiente; ICMS (imposto) ecológico; gestão de águas; gestão de meio ambiente; gestão de resíduos sólidos e rotulagem, totalizando mais de 100.000 palavras. Poderíamos ter coletado mais textos sobre território e mudanças climáticas, porém este não é o foco de nosso trabalho. Com isto, observamos que os textos em geral são sobre meio ambiente e não especificamente sobre reciclagem de materiais.

O website do Ministerio de Medio Ambiente da Espanha apresenta dados sobre: o ministério; legislação; águas continentais, marítimas e litorâneas; biodiversidade; mudanças climáticas; qualidade e contaminação; avaliação ambiental; autoridades ambientais; informação estatística e ambiental; desenvolvimento territorial; ajudas e subvenções; formação e educação ambiental e entidades ambientais.

Neste website foi possível coletar apenas dois textos que tratam da apresentação do ministério e da gestão de águas, totalizando mais de 10.000 palavras. Optou-se por não coletar mais dados, pois este ministério não apresentava textos relativamente comparáveis com os textos que obtivemos em português. Nem todos os temas coletados em português foram encontrados em espanhol: consumo sustentável; comercio e meio ambiente; economia; gestão de resíduos sólidos; rotulagem. Com isto, partimos para uma nova fase de pesquisa e definição de websites possivelmente comparáveis e capazes de fornecerem os dados necessários para o início do projeto.

## **Levantamento de dados – fase II**

A segunda fase de levantamento de dados teve como foco a coleta de artigos científicos sobre reciclagem. A busca por artigos publicados eletronicamente nos permitiu observar que a coleta deste tipo de texto poderia restringir e interferir nos dados quanto à faixa de frequência de formas e ocorrências. Isto porque os artigos encontrados tratavam de temas muito específicos, como a reciclagem de barcos, óleos, papéis, eletrônicos, pet e plásticos, o que poderia fazer com que o número de formas empregadas se reduzisse. Optou-se então por não construir um corpus de reciclagem baseado em artigos científicos.

## **Levantamento de dados – fase III**

A terceira fase de levantamento de dados foi mais certa. Optamos por coletar textos de websites de organizações ligadas ao tema da reciclagem no Brasil, Estados Unidos e Espanha. Para isto utilizamos a ferramenta “pesquisa avançada” do Google (www.google.com.br), na qual definimos: a extensão .org; os idiomas português, inglês e espanhol; as regiões Brasil, Estados Unidos e Espanha; e as palavras-chave *reciclagem*, para o português brasileiro, *recycling*, para o inglês, e *reciclaje*, para a língua espanhola.

A pesquisa começou bem, porém tivemos que combinar mais algumas palavras às palavras-chave na língua portuguesa e espanhola para otimizar a busca, que estava trazendo muitos sites que apresentavam conteúdo relacionado à reciclagem, mas não o perfil que definimos, conforme abaixo. As palavras utilizadas foram, para o português, *borracha*, *papel*, *eletrônico*, *resíduo*, *pet*, *vidro*, e para o espanhol, *alumínio*, *metal* e *vida sostenible*.

Como os sites relacionados à reciclagem abordam o tema considerando diversos aspectos, que abrangem desde a venda de produtos verdes até a indicação de empresas que vendem máquinas e equipamentos utilizados no processo de reciclagem, definimos alguns critérios para a seleção do conteúdo.

Foram excluídos da coleta: página inicial (quando apresentava informações muito truncadas), informações sobre eventos, links, máquinas e equipamentos, tabelas de cotação de material no mercado, listas de empresas recicladoras, cadastro no site, agenda, histórico de notícias e publicações, imprensa (notícias na mídia), arquivos (em extensão .pdf ou .doc, por exemplo) sobre educação ambiental, cartilhas e manuais, acervos multimídia, pesquisas e enquetes, “fale conosco”, prêmios, memória ou histórico da organização, equipe que trabalha na organização, parceiros, “ecoprodutos”, jogos, políticas públicas e leis, projetos ambientais, cursos e instruções sobre como fazer arte e artesanato com materiais recicláveis.

Foram coletados textos sobre: reciclagem de materiais, descarte de resíduos, coleta seletiva, institucional e missão da organização, serviços destinados a cooperativas e sucateiros, fichas técnicas, dicas e curiosidades sobre a reciclagem de materiais, histórico e origem dos resíduos, respostas a dúvidas frequentes e instalações próprias para reciclagem.

A coleta de textos em inglês foi a mais rápida, pois em um único website encontramos muitas informações, o que já totalizou 51.203 palavras. No caso do português, coletamos

textos em 20 websites, sendo que alguns deles pertenciam a associações de indústrias brasileiras que divulgam a sua preocupação com o destino de resíduos pós-consumo (em geral, embalagens) e dos resíduos gerados em suas operações (como aparas de papel, por exemplo). A coleta realizada nos sites da Espanha contou com a inclusão de alguns textos relacionados aos temas eficiência energética, consumo responsável e sustentável e mudança climática. Este procedimento foi necessário devido à dificuldade de encontrar sites que tratam especificamente sobre o assunto, como no Brasil e Estados Unidos. Observamos também, nos sites espanhóis, a publicação de uma série de listas de discussões, *blogs* e artigos não científicos sobre o assunto.

Ao realizar a busca por sites .org encontramos organizações voltadas a trabalhos de consultoria, conscientização e educação sobre o assunto e também associações relacionadas a determinadas classes de trabalhadores ou destinadas a assuntos específicos, como associação de químicos e centros de reciclagem.

### **Coleta e organização do corpus**

A coleta e organização do corpus passaram pelos seguintes passos:

- acessamos o Google;
- utilizamos a ferramenta “pesquisa avançada”, conforme descrito anteriormente;
- selecionamos e acessamos o *website* sobre reciclagem;
- selecionamos o texto adequado aos critérios pré-estabelecidos;
- copiamos o texto escolhido;
- colamos o texto copiado no Bloco de Notas;
- criamos o cabeçalho, conforme exemplo a seguir;
- salvamos o arquivo com a extensão .txt.

Este procedimento foi seguido para cada um dos websites, pois somente arquivos salvos na extensão .txt podem ser utilizados no programa AntConc. Isto significa que estes arquivos contêm somente caracteres do teclado (letras, números e símbolos ortográficos), sem códigos de formatação específicos para certos programas (como o Microsoft Word ®).” (Sardinha, 2004, p.51).

Ao coletar os textos, já tínhamos definido o cabeçalho que foi utilizado em cada um dos arquivos que formam o corpus, conforme exemplo abaixo. O cabeçalho é importante porque “contém informações sobre o texto, tais como a origem, a data de coleta, o grupo de pesquisa responsável, o tamanho do texto, sistema de transcrição, detalhes do *copyright*, a autoria, os participantes.” (Sardinha, 2004, p. 74). Isto facilita o processamento através de ferramentas eletrônicas, além de deixar o corpus organizado e preparado para ser utilizado como corpus de estudo para outros pesquisadores. Para informações mais detalhadas sobre etiquetadores ou cabeçalhos consultamos “2.4. Criação de cabeçalhos” em Sardinha (2004).

Exemplo de cabeçalho:

```
<pg web>http://www.iar.org.br</pg web>
<title>instituto ambiental reciclar</title>
<data>27set08</data>
<tema>serviços</tema>
<tema>meio ambiente</tema>
<país>brasil</país>
```

Ao coletar os textos e salvá-los em arquivos, imediatamente já os organizamos em pastas, seguindo o critério abaixo.

- criamos uma pasta chamada “Corpus Multilíngüe especializado em Reciclagem”;
- dentro desta pasta criamos três outras pastas chamadas “ORG Português”, “ORG Inglês” e “ORG Espanhol”;
- salvamos os arquivos conforme a língua empregada no texto, o país de origem, o tipo de *website* (neste caso, temos apenas o tipo “organização”, porém optamos por deixar o corpus preparado para receber arquivos de outros tipos) e a ordem em que foram coletados. Ex.: PBORG010.

<b>PB</b>	<b>ORG</b>	<b>010</b>
Língua e país	Tipo	Ordem de coleta

Sendo:

<b>Legenda</b>	<b>Língua</b>	<b>País</b>
EE	Espanhol	Espanha
IE	Inglês	Estados Unidos
PB	Português	Brasil

O controle de coleta de textos por *website* se encontra no anexo 1, ao final deste trabalho.

### **Utilização do programa AntConc**

Após a coleta e organização do corpus, seguimos as instruções de Brezolin, publicadas no site do Projeto COMET – Corpus Multilíngüe para Ensino e Tradução (<http://www.fflch.usp.br/dlm/comet/>), para trabalhar os dados no programa AntConc. Extraímos listas de palavras em cada uma das línguas, separadamente, e uma lista de palavras-chave do português. A criação desta última lista só foi possível porque inserimos também no programa o corpus de referência do português brasileiro Lácio-Web. Estas listas se encontram nos anexos 2 e 3 deste trabalho.

## **5.5. Listas de frequência de palavras**

As palavras de uma língua podem apresentar frequência de uso alta, intermediária ou baixa, o que pode ser observado através das listas de frequência. As palavras que ocorrem menos vezes são consideradas raras, porém não menos importantes, afinal, a maioria das palavras que formam uma língua apresentam esta condição. Além disto, é possível analisar a frequência de palavras gramaticais e de conteúdo, observar o emprego de estrangeirismos e utilizá-las como fonte para seleção de palavras que farão parte de um estudo de colocados. As análises produzidas a partir destas listas fornecem dados que podem contribuir para a descrição dos padrões lingüísticos atuais e nos estudos lingüísticos diacrônicos.

### **5.5.1. Palavras mais e menos frequentes**

Para analisar as palavras mais e menos freqüentes do corpus multilíngüe, iniciamos com a elaboração de quadros de faixas de freqüência para cada uma das três línguas. Estes quadros foram elaborados com base no quadro apresentado por Sardinha (2004, p. 168), no capítulo 6 de seu *Lingüística de Corpus*, denominado “Freqüências de palavras da língua portuguesa segundo o Banco de Português”.

Estes quadros apresentam cinco faixas de freqüência, sendo que a última faixa abarca as palavras com mil ocorrências ou mais. O critério utilizado para a definição desta faixa foi o número máximo de freqüência de uma única palavra encontrado em cada língua. Observamos que entre as palavras do inglês a que mais ocorre é a palavra *the*, 3713 vezes. Em ordem crescente, temos a palavra *de*, que ocorre 5822 vezes no português brasileiro, e finalmente a palavra *de* que ocorre 7349 vezes no espanhol, o que faz com que as faixas acima deste número se tornem desnecessárias. A última faixa utilizada por Sardinha comporta palavras com mais de um milhão de ocorrências dentro de um corpus de mais de 120 milhões de formas (em 2000).

#### Faixas de freqüência no corpus sobre Reciclagem em Português Brasileiro

Faixa	Formas ( <i>types</i> )	% formas	Ocorrências ( <i>tokens</i> )	% ocorrências	Freqüência média
1	4230	45,7%	4230	5,1%	1
2 a 9	3914	42,3%	14511	17,4%	4
10 a 99	1029	11,1%	25191	30,2%	24
100 a 999	73	0,8%	19442	23,3%	266
1000 em diante	9	0,1%	19962	24,0%	2218
<b>Total</b>	<b>9255</b>	<b>100%</b>	<b>83336</b>	<b>100%</b>	<b>9</b>

No português brasileiro podemos observar que dentro das 83.336 ocorrências apresentadas temos apenas 9.255 formas distintas, o que inclui palavras gramaticais (como pronomes, artigos e preposições) e palavras de conteúdo (como substantivos, adjetivos e numerais). Observa-se que a maioria das formas está na primeira faixa de freqüência, ou seja, que ocorrem uma única vez no corpus, e, portanto, são raras. Esta faixa é apenas 8,07% maior do que a segunda faixa. Com isto, temos 88% das formas do corpus ocorrendo de uma a nove vezes cada.

A frequência de uso das formas pode ser baixa, intermediária, alta ou altíssima. Sardinha (2004, p. 169) considera que esta última frequência se refere a palavras com um milhão ou mais de ocorrências, o que não cabe a este corpus. As formas que se enquadram na segunda faixa de frequência intermediária (100 a 999) e na faixa alta (de 1.000 em diante) representam uma minoria, que neste caso não soma 1%. São apenas nove as palavras de alta frequência neste corpus. Estes dados corroboram o que a léxico-estatística e a lingüística quantitativa já apontam: “as frequências de palavras não são distribuídas equitativamente; algumas são mais comuns que outras” (SARDINHA, 2004, p. 168). Isto significa que a maioria das formas de um corpus ocorre apenas uma vez e, portanto, apresentam um baixo número de ocorrências. Neste caso, os *hapax legomena* representam 45,7% do total das formas e 5,1% do total de ocorrências.

**Formas de alta frequência no corpus  
sobre Reciclagem em Português Brasileiro**

<b>Ranque</b>	<b>Frequência</b>	<b>Forma</b>
1	5822	de
2	2874	e
3	2667	a
4	2382	o
5	1355	que
6	1304	para
7	1292	do
8	1250	em
9	1016	da

As nove formas de alta frequência listadas acima ocorrem 19.962 vezes no corpus, o que representa 24% do total de ocorrências e 372% mais ocorrências do que as formas de baixa frequência. Apesar disso, estas formas não correspondem, neste corpus, à faixa com maior número de ocorrências. O maior número de ocorrências se refere às palavras da faixa de 10 a 99, que correspondem a 30, 2% de um total de 83.336 ocorrências.

**Faixas de frequência no corpus sobre Reciclagem em Espanhol**

<b>Faixa</b>	<b>Formas (types)</b>	<b>% formas</b>	<b>Ocorrências (tokens)</b>	<b>% ocorrências</b>	<b>Frequência média</b>
1	5159	51,6%	5159	6,1%	1
2 a 9	3748	37,5%	13488	15,9%	4
10 a 99	1025	10,2%	24332	28,6%	24
100 a 999	59	0,6%	16186	19,0%	274
1000 em diante	10	0,1%	25851	30,4%	2585
<b>Total</b>	<b>10001</b>	<b>100%</b>	<b>85016</b>	<b>100%</b>	<b>9</b>

Ao analisar as faixas de frequência do espanhol, observamos que em um total de 85.016 ocorrências, 10.001 são formas distintas. O corpus nesta língua apresenta 8% mais formas do que o português brasileiro. Além disso, a quantidade de *hapax legomena* encontrada ultrapassa 51% do número total de formas, que somada à quantidade da segunda faixa, representam 89,1% do total. O restante das formas se divide entre as faixas intermediária e alta, que correspondem a 10,9% do total. Assim como no português, as palavras de alta frequência são apenas 0,1% do total de formas.

**Formas de alta frequência no corpus  
sobre Reciclagem em Espanhol**

<b>Ranque</b>	<b>Frequência</b>	<b>Forma</b>
1	7349	de
2	3054	la
3	2662	y
4	2475	en
5	2340	el
6	2189	que
7	1787	los
8	1396	a
9	1345	se
10	1254	las

As dez palavras que ocorrem mais de mil vezes cada, conforme acima, correspondem a 30,4% do total de ocorrências, ou seja, 25.851 em um universo de 85.016. As mais de

25.000 ocorrências tornam estas palavras 401,09% mais frequentes do que as *hapax legomena*.

**Faixas de frequência no corpus sobre Reciclagem em Inglês**

<b>Faixa</b>	<b>Formas (types)</b>	<b>% formas</b>	<b>Ocorrências (tokens)</b>	<b>% ocorrências</b>	<b>Frequência média</b>
1	2902	41,4%	2902	3,3%	1
2 a 9	2893	41,3%	11042	12,6%	4
10 a 99	1096	15,7%	29851	33,9%	27
100 a 999	102	1,5%	25137	28,6%	246
1000 em diante	10	0,1%	19015	21,6%	1902
<b>Total</b>	<b>7003</b>	<b>100%</b>	<b>87947</b>	<b>100%</b>	<b>13</b>

O corpus em língua inglesa, representado por 87.947 ocorrências, a maior quantidade do corpus, se diferencia do português brasileiro e do espanhol por apresentar também a menor quantidade de formas: 7.003. Isto significa que o corpus é formado por uma quantidade menor de formas que se repetem mais vezes. Já é possível observar que a faixa de frequência de 100 a 999 apresenta um número maior de formas em relação às outras línguas e, se a somarmos com a faixa de frequência alta, teremos 1,6%, o dobro da média entre o português brasileiro e o espanhol.

O número de *hapax legomena* em inglês é o menor entre os três índices, com 2902 formas (41,4%), e as duas faixas de baixa frequência também apresentam a menor soma: 5795 formas (82,8%). Ainda quanto às formas, a faixa de frequência alta corresponde a 0,1% do total, assim como as demais línguas.

**Formas de alta frequência no corpus  
sobre Reciclagem em Inglês**

<b>Ranque</b>	<b>Frequência</b>	<b>Forma</b>
1	3713	the
2	2722	and
3	2520	of
4	2289	to
5	1656	a
6	1525	in
7	1278	is
8	1155	recycling
9	1136	for
10	1021	are

A faixa de frequência alta, no caso das ocorrências, representa 21,6% do total, ficando atrás das faixas intermediárias de 100 a 999 (28,6%) e de 10 a 99 (33,9%). Esta relação foi distinta em cada uma das línguas.

Observando as faixas de frequência de formas e ocorrências nas três línguas e relacionando-as entre si, verificamos que o português brasileiro e o espanhol se apresentam relativamente de forma equilibrada em comparação ao inglês, conforme indica a frequência média. Segundo este índice, o inglês apresenta uma frequência média (13) 44% maior do que a frequência média do português brasileiro (9) e do espanhol (9). Além disso, a quantidade das formas que ocorrem 100 vezes ou mais em inglês representa o dobro da média deste índice nas duas outras línguas. Um ponto em comum entre elas é a soma das duas faixas de baixa frequência de formas que representa entre 80 e 90% do total. Já uma diferença se observa quanto à ordem em que as faixas se apresentam quanto ao número de ocorrências.

## **6. Considerações finais**

O objetivo deste trabalho foi criar um corpus multilíngüe especializado em reciclagem e analisar a faixa de frequência de formas e ocorrências, segundo um estudo realizado por Sardinha (2004) com base no Banco de Português. Para cumprir estes objetivos buscamos

informações sobre o estatuto, o histórico e os conceitos propostos pela Lingüística de corpus, bem como as metodologias e ferramentas disponíveis para a realização de estudos na área.

Observamos que as referências teóricas e os exemplos encontrados na obra *Lingüística de corpus* (SARDINHA, 2004) foram realmente essenciais no decorrer de todo o trabalho. Através deles, foi possível tornar nosso corpus específico e representativo, além de corroborar teorias já consagradas e comprovar a funcionalidade de programas que operacionalizam a descrição lingüística baseada em corpus. Aprendemos a utilizar o programa WordSmith Tools e nos sentimos motivados a utilizar um programa semelhante, o AntConc, que não oferece restrições como a versão demonstrativa do WST.

A partir da criação do corpus, pudemos extrair não somente listas de palavras em português, inglês e espanhol, como também uma lista de palavras-chave do português, tendo como base o corpus de referência Lácio-Web, constituído por cinco milhões de palavras do português brasileiro escrito. A compilação das listas de palavras permitiu uma análise das faixas de freqüência de formas e ocorrências em cada uma das línguas, além de uma análise comparativa entre elas.

Concluimos que este estudo inicial nos permitiu assimilar conceitos essenciais para a prática da Lingüística de corpus, bem como nos habilitou para a produção e ampliação de corpora multilíngües, para a análise de faixas de freqüência e para dar continuidade ao trabalho realizando estudos sobre colocados, padronização e uso de estrangeirismos.

## Referências bibliográficas

BENEDUZI, R. “Extração e classificação das colocações a partir de um corpus de língua espanhola”. In: VI Encontro de Lingüística de Corpus, 2007, São Paulo. Anais do VI Encontro de Lingüística de Corpus, 2007. p. 1-19. Disponível em: <[http://www.4shared.com/file/43833368/7d950ac/Extrao\\_e\\_Classificao\\_das\\_Colocaes\\_a\\_partir\\_de\\_um\\_Corpus\\_de\\_Lngua.html?dirPwdVerified=593c9804](http://www.4shared.com/file/43833368/7d950ac/Extrao_e_Classificao_das_Colocaes_a_partir_de_um_Corpus_de_Lngua.html?dirPwdVerified=593c9804)>.

*Dicionário Houaiss da língua portuguesa*. Disponível em: <<http://houaiss.uol.com.br>>

ENVIRONMENTAL PROTECTION AGENCY OF UNITED STATES. Disponível em: <<http://www.epa.gov>>

MINISTERIO DE MEDIO AMBIENTE DE ESPANHA. Disponível em: <<http://www.mma.es>>

MINISTÉRIO DO MEIO AMBIENTE DO BRASIL. Disponível em: <<http://www.mma.gov.br/>>

Projeto COMET – Corpus Multilíngüe para Ensino e Tradução. Disponível em: <<http://www.fflch.usp.br/dlm/comet>>.

SARDINHA, T. B. *Lingüística de Corpus*. Barueri, SP: Manole, 2004.

TAGNIN, S. E. O. “A identificação de equivalentes tradutórios em corpora comparáveis”. In: Anais do I Congresso Internacional da ABRAPUI: Belo Horizonte, 3 a 6 de junho de 2007. Disponível em: <[http://www.fflch.usp.br/dlm/comet/Novo/Stella\\_Abrapui%202007\\_artigo.pdf](http://www.fflch.usp.br/dlm/comet/Novo/Stella_Abrapui%202007_artigo.pdf)>

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

## Anexo 1

**Controle de coleta de textos por website**

<b>Língua País</b>	<b>Tipo de website</b>	<b>Palavra- chave</b>	<b>Número coleta</b>	<b>Data</b>	<b>Nome da ORG</b>
EE	ORG	Reciclaje	1	28/9/2008	Org, de func. Química del cloro
EE	ORG	Reciclaje	2	28/9/2008	Centro de Reciclaje Zaragoza
EE	ORG	Reciclaje	3	28/9/2008	Pequeño Green Peace
EE	ORG	Reciclaje	4	28/9/2008	Terra Ecologia Práctica
EE	ORG	Reciclaje	5	28/9/2008	Fundación Tierra
EE	ORG	Reciclaje	6	28/9/2008	Educación Ambiental Corazón Verde
EE	ORG	Reciclaje	7	28/9/2008	Fenacor
EE	ORG	Reciclaje	8	28/9/2008	Recicla Papel
EE	ORG	Reciclaje	9	28/9/2008	Ecología y desarrollo
EE	ORG	Reciclaje	10	28/9/2008	Consumo responsable
EE	ORG	Reciclaje	11	28/9/2008	Oviedo Ecológico
EE	ORG	Reciclaje	12	28/9/2008	Arpal Aluminio

EE	ORG	Reciclaje	13	28/9/2008	Consello de Santiago
EE	ORG	Reciclaje	14	28/9/2008	Vida Sostenible
EE	ORG	Reciclaje	15	12/10/2008	Asociación Cultural y Deportiva el Cacahuete Viajero
EE	ORG	Reciclaje	16	12/10/2008	Asociación de Empresas de Recogida, Tratamiento y Reciclaje de Aceite y Grasas
EE	ORG	Reciclaje	17	12/10/2008	Asociación Espanola de Recogedores de Pilas, Acumuladores y Móviles
EE	ORG	Reciclaje	18	12/10/2008	Asociación Tenerifeña de Amigos de la Naturaleza
EE	ORG	Reciclaje	19	12/10/2008	Asociación de Latas de Bebidas
EE	ORG	Reciclaje	20	12/10/2008	Fórum de comércio
EE	ORG	Reciclaje	21	12/10/2008	Facua
IE	ORG	Recycling	1	28/9/2008	Earth 911
IE	ORG	Recycling	2	12/10/2008	Lane County
IE	ORG	Recycling	3	12/10/2008	Steel Recycling Institute
IE	ORG	Recycling	4	12/10/2008	Recycle room
IE	ORG	Recycling	5	12/10/2008	Lamp recycle

IE	ORG	Recycling	6	12/10/2008	Reach recycling
IE	ORG	Recycling	7	12/10/2008	Wastecap
IE	ORG	Recycling	8	13/10/2008	Recycle Ann Arbor
IE	ORG	Recycling	9	13/10/2008	Recycle Hawaii
IE	ORG	Recycling	10	13/10/2008	American Forest & Paper Association
IE	ORG	Recycling	11	13/10/2008	Glass Packaging Institute
PB	ORG	Reciclagem	1	24/8/2008	CEMPRE
PB	ORG	Reciclagem	2	24/8/2008	AKATU
PB	ORG	Reciclagem	3	24/8/2008	5 Elementos
PB	ORG	Reciclagem	4	24/8/2008	Ajuda Brasil
PB	ORG	Reciclagem	5	24/8/2008	Recicloteca
PB	ORG	Reciclagem	6	24/8/2008	Reviverde
PB	ORG	Reciclagem	7	27/9/2008	Instituto Ambiental Reciclar
PB	ORG	Reciclagem	8	27/9/2008	Portal do Meio Ambiente

PB	ORG	Reciclagem	9	27/9/2008	Pueras
PB	ORG	Reciclagem	10	27/9/2008	Instituto Triângulo
PB	ORG	Reciclagem	11	28/9/2008	Cia Eco Educação Ambiental
PB	ORG	Reciclagem	12	8/10/2008	Instituto do PVC
PB	ORG	Reciclagem	13	8/10/2008	Associação Brasileira de Embalagem
PB	ORG	Reciclagem	14	8/10/2008	Associação Brasileira da Indústria do pet
PB	ORG	Reciclagem	15	8/10/2008	Associação Brasileira das Indústrias de Refrigerantes e de Bebidas Não Alcoólicas
PB	ORG	Reciclagem	16	10/10/2008	Associação Brasileira do Alumínio
PB	ORG	Reciclagem	17	10/10/2008	Associação Técnica Brasileira das Indústrias Automáticas de Vidro
PB	ORG	Reciclagem	18	12/10/2008	Associação Brasileira das Indústrias Recicladoras de Papel
PB	ORG	Reciclagem	19	12/10/2008	Plastivida
PB	ORG	Reciclagem	20	12/10/2008	Federação das Indústrias do Estado do Paraná

## Anexo 2

## Listas de palavras em Português, Inglês e Espanhol (as 1000 primeiras formas)

Português			Inglês			Espanhol		
Freq.	Forma		Freq.	Forma		Freq.	Forma	
1	5822	de	1	3713	the	1	7349	de
2	2874	e	2	2722	and	2	3054	la
3	2667	a	3	2520	of	3	2662	y
4	2382	o	4	2289	to	4	2475	en
5	1355	que	5	1656	a	5	2340	el
6	1304	para	6	1525	in	6	2189	que
7	1292	do	7	1278	is	7	1787	los
8	1250	em	8	1155	recycling	8	1396	a
9	1016	da	9	1136	for	9	1345	se
10	987	é	10	1021	are	10	1254	las
11	707	com	11	856	or	11	990	para
12	688	no	12	745	can	12	876	del
13	687	os	13	690	be	13	868	un
14	616	por	14	665	as	14	793	papel
15	582	um	15	665	that	15	712	por
16	568	as	16	604	it	16	684	una
17	557	na	17	549	paper	17	667	con
18	548	reciclagem	18	541	recycled	18	622	es
19	508	uma	19	532	your	19	594	o
20	505	se	20	502	s	20	547	no
21	474	como	21	488	on	21	447	como
22	452	dos	22	481	from	22	441	al
23	449	ou	23	478	you	23	413	más
24	433	são	24	461	with	24	383	su
25	424	ser	25	395	by	25	328	resíduos
26	409	não	26	395	products	26	264	son
27	390	papel	27	395	used	27	235	reciclado
28	353	mais	28	380	materials	28	228	este
29	315	material	29	380	steel	29	228	lo
30	309	das	30	379	waste	30	217	productos
31	299	lixo	31	369	glass	31	216	recogida
32	292	alumínio	32	339	plastic	32	192	consumo
33	289	ao	33	314	this	33	192	latas
34	271	brasil	34	304	use	34	188	sin
35	261	embalagens	35	298	into	35	185	envases
36	255	coleta	36	298	more	36	181	sobre
37	249	sua	37	273	at	37	180	puede
38	240	produtos	38	271	new	38	177	reciclaje
39	239	à	39	263	not	39	171	si
40	226	vidro	40	251	these	40	164	medio
41	205	pode	41	249	other	41	162	también
42	190	materiais	42	244	about	42	162	ya
43	190	plástico	43	243	have	43	156	proceso
44	189	produção	44	241	recycle	44	151	pueden
45	186	meio	45	238	an	45	151	sus
46	183	plásticos	46	235	they	46	148	cartón
47	173	processo	47	230	mercury	47	147	bebidas
48	169	água	48	229	all	48	146	pilas

49	166	resíduos	49	225	oil	49	144	energía
50	164	outros	50	219	percent	50	143	uso
51	164	também	51	219	what	51	138	ambiente
52	162	reciclado	52	214	containers	52	138	materiales
53	161	nos	53	210	than	53	137	esta
54	160	ambiente	54	206	material	54	137	ha
55	158	consumo	55	200	if	55	135	ser
56	147	ambiental	56	198	up	56	133	cada
57	142	indústria	57	197	will	57	131	entre
58	141	seu	58	196	when	58	128	e
59	139	pela	59	192	cans	59	122	todo
60	135	toneladas	60	190	also	60	121	pasta
61	130	seletiva	61	189	most	61	117	estos
62	129	entre	62	182	may	62	115	forma
63	129	vida	63	182	program	63	114	otros
64	127	pet	64	181	such	64	108	vez
65	125	foi	65	174	energy	65	106	muy
66	123	forma	66	174	make	66	106	producción
67	121	energia	67	173	many	67	105	material
68	120	pvc	68	171	has	68	102	etc
69	117	até	69	169	how	69	100	información
70	115	embalagem	70	167	information	70	99	además
71	115	latas	71	166	water	71	98	aluminio
72	115	pelo	72	165	lead	72	96	está
73	114	podem	73	165	one	73	95	vidrio
74	113	já	74	165	process	74	94	menos
75	110	mil	75	165	some	75	93	españa
76	107	além	76	161	their	76	93	papeles
77	107	nas	77	158	which	77	92	parte
78	106	produto	78	156	made	78	91	gran
79	105	tem	79	152	car	79	91	mayor
80	104	matéria	80	148	out	80	88	agua
81	102	anos	81	144	aluminum	81	88	calidad
82	102	mercado	82	143	million	82	88	selectiva
83	96	quando	83	142	bottles	83	88	tipo
84	96	suas	84	141	them	84	85	cuando
85	95	aço	85	138	disposal	85	85	han
86	95	etc	86	138	there	86	85	todos
87	95	sem	87	131	collection	87	84	desde
88	94	cerca	88	131	food	88	83	fabricación
89	93	desenvolvimento	89	131	paint	89	83	tiene
90	93	fabricação	90	128	industry	90	81	así
91	93	uso	91	127	batteries	91	81	sistema
92	91	papéis	92	126	our	92	80	basura
93	90	ainda	93	124	do	93	80	empresas
94	89	ano	94	122	massachusetts	94	80	hasta
95	89	maior	95	117	benefits	95	80	pero
96	89	muito	96	117	u	96	79	sólo
97	87	onde	97	116	but	97	78	plástico
98	87	pneus	98	115	product	98	77	ambiental
99	85	recicláveis	99	114	local	99	77	dos
100	84	grande	100	114	programs	100	77	fibras
101	83	empresas	101	111	compost	101	77	países
102	83	partir	102	111	recovered	102	77	tanto
103	82	sobre	103	110	find	103	76	gestión
104	81	baterias	104	109	because	104	76	sistemas
105	81	cada	105	109	we	105	74	contenedores
106	79	prima	106	107	tires	106	74	hay
107	78	aos	107	107	year	107	73	madera

108	78	através
109	78	mesmo
110	78	país
111	77	indústrias
112	75	metais
113	74	seus
114	74	tipos
115	72	deve
116	72	pilhas
117	72	sucata
118	71	óleo
119	70	garrafas
120	70	programa
121	69	alimentos
122	69	importante
123	69	rebia
124	69	todos
125	68	pois
126	67	quantidade
127	67	redução
128	66	peso
129	65	este
130	65	há
131	65	parte
132	64	apenas
133	64	mas
134	64	qualidade
135	64	tipo
136	62	desde
137	62	está
138	62	às
139	61	foram
140	61	metal
141	61	milhões
142	60	ambientais
143	60	nacional
144	60	seja
145	59	reciclar
146	59	trabalho
147	58	ciclo
148	58	exemplo
149	58	você
150	57	assim
151	57	brasileira
152	57	fibras
153	57	volume
154	56	após
155	56	aterros
156	56	hoje
157	56	reciclável
158	55	recursos
159	54	construção
160	54	dados
161	54	diferentes
162	54	industrial
163	54	outras
164	54	sociedade
165	54	usado
166	54	vidros

108	105	environmental
109	105	hazardous
110	104	reuse
111	101	only
112	100	container
113	99	over
114	97	recyclable
115	97	where
116	96	each
117	96	facts
118	96	tips
119	94	help
120	94	high
121	94	its
122	93	so
123	93	using
124	92	cardboard
125	91	amount
126	91	through
127	90	tons
128	89	items
129	89	no
130	89	office
131	88	bags
132	87	earth
133	87	old
134	87	packaging
135	86	environment
136	86	lamps
137	86	landfills
138	85	less
139	85	non
140	85	should
141	84	composting
142	83	any
143	83	motor
144	83	off
145	83	take
146	82	been
147	82	every
148	82	scrap
149	81	include
150	81	like
151	81	quality
152	80	plastics
153	78	cost
154	78	recovery
155	78	t
156	77	metal
157	77	pvc
158	77	resources
159	77	services
160	77	while
161	76	costs
162	76	state
163	76	then
164	75	computer
165	75	e
166	75	pounds

108	72	vida
109	71	mercado
110	70	materia
111	69	mediante
112	67	lata
113	66	desarrollo
114	65	asociación
115	65	años
116	64	cantidad
117	64	según
118	63	año
119	63	cambio
120	63	estas
121	62	nos
122	61	problemas
123	61	reciclar
124	60	donde
125	60	fsc
126	59	están
127	59	importante
128	59	pvc
129	59	recuperación
130	59	recursos
131	57	reducir
132	57	utilizar
133	55	debe
134	55	nuevo
135	55	sea
136	54	ambientales
137	54	ejemplo
138	54	envase
139	54	oficina
140	54	otro
141	54	qué
142	54	tierra
143	53	actividades
144	53	grandes
145	53	impresión
146	53	tienen
147	53	través
148	52	acero
149	52	caso
150	52	cloro
151	52	metales
152	52	mundo
153	52	necesario
154	51	hacer
155	51	materias
156	51	partir
157	51	tipos
158	51	toneladas
159	51	tu
160	50	ahorro
161	50	durante
162	50	existen
163	50	mismo
164	49	industria
165	49	otras
166	49	podemos

167	53	isso	167	75	wastecap	167	49	producto
168	53	pessoas	168	74	contain	168	49	utiliza
169	53	projeto	169	74	reduce	169	48	contaminación
170	52	informação	170	74	toxic	170	48	diferentes
171	52	lata	171	73	curbside	171	48	general
172	52	sendo	172	73	end	172	48	personas
173	52	será	173	72	recyclables	173	48	usados
174	51	lâmpadas	174	72	types	174	47	consumidores
175	51	mundo	175	71	household	175	47	deben
176	51	valor	176	71	including	176	47	siempre
177	50	atividade	177	71	keep	177	46	contenedor
178	50	borracha	178	70	community	178	46	cualquier
179	50	final	179	70	furniture	179	46	cómo
180	50	geração	180	70	often	180	46	plantas
181	50	instituto	181	69	businesses	181	45	algunos
182	50	todo	182	68	available	182	45	cara
183	50	vez	183	68	parts	183	45	muchas
184	50	vezes	184	68	soil	184	45	nuestro
185	49	composto	185	68	trash	185	45	recortes
186	49	setor	186	68	well	186	45	sociedad
187	48	depois	187	67	even	187	45	uno
188	48	economia	188	67	manufacturing	188	45	virgen
189	48	outro	189	67	need	189	44	todas
190	48	quanto	190	67	was	190	43	baterías
191	48	transporte	191	66	pesticides	191	43	nuestra
192	47	bebidas	192	66	source	192	43	plásticos
193	47	características	193	66	www	193	43	primas
194	47	chumbo	194	65	before	194	42	certificación
195	47	esse	195	65	saves	195	42	color
196	47	esta	196	63	air	196	42	tener
197	47	grupo	197	63	different	197	42	usado
198	47	longa	198	63	here	198	41	algunas
199	46	bem	199	63	much	199	41	consumidor
200	46	caso	200	62	cleaning	200	41	ello
201	46	diversos	201	62	years	201	41	objetivo
202	46	índice	202	61	corrugated	202	41	vertederos
203	45	essa	203	60	landfill	203	40	climático
204	45	geral	204	59	large	204	40	datos
205	45	média	205	59	must	205	40	emisiones
206	45	naturais	206	59	org	206	40	puedes
207	45	qualquer	207	59	were	207	40	puntos
208	44	associação	208	58	based	208	40	tratamiento
209	44	principais	209	58	being	209	39	aunque
210	44	processos	210	58	clean	210	39	fue
211	44	sistema	211	58	collected	211	39	hace
212	44	utilizado	212	58	construction	212	39	imprimir
213	43	alguns	213	58	down	213	39	muchos
214	43	caixas	214	58	page	214	39	naturales
215	43	día	215	58	re	215	39	sector
216	43	mercúrio	216	58	regulations	216	38	embargo
217	43	novas	217	58	space	217	38	menor
218	43	tempo	218	57	boxes	218	38	porque
219	43	total	219	57	business	219	38	sino
220	43	usados	220	57	color	220	38	tiempo
221	42	estão	221	57	important	221	37	bien
222	42	grandes	222	57	processing	222	37	cantidades
223	42	madeira	223	56	after	223	37	debemos
224	42	possível	224	56	c	224	37	día
225	42	qual	225	56	just	225	37	esto

226	42	social	226	56	next	226	37	nuevos
227	41	atividades	227	56	time	227	37	objetivos
228	41	benefícios	228	56	way	228	37	permite
229	41	civil	229	55	buy	229	37	posible
230	41	deste	230	55	equipment	230	37	punto
231	41	informações	231	55	metals	231	37	separación
232	41	muitos	232	55	people	232	36	coste
233	41	ondulado	233	55	states	233	36	embalajes
234	41	países	234	55	stream	234	36	fundación
235	41	reaproveitamento	235	54	found	235	36	m
236	41	recuperação	236	54	properly	236	36	ni
237	40	aparas	237	54	remove	237	36	prima
238	40	cor	238	54	two	238	36	veces
239	40	fazer	239	53	according	239	35	blanco
240	40	folhas	240	53	content	240	35	capacidad
241	40	fonte	241	53	get	241	35	conocer
242	40	programas	242	53	three	242	35	mecánica
243	39	educação	243	52	home	243	35	servicio
244	39	têm	244	52	life	244	35	servicios
245	39	utilização	245	52	making	245	35	será
246	38	ações	246	52	save	246	35	verde
247	38	compostagem	247	51	area	247	34	aire
248	38	devem	248	51	bin	248	34	depositar
249	38	durante	249	51	pile	249	34	fabricar
250	38	mundial	250	50	proper	250	34	fin
251	38	outra	251	50	sure	251	34	impacto
252	38	solo	252	50	tire	252	34	sostenible
253	38	só	253	50	today	253	34	toda
254	37	aplicações	254	49	automotive	254	34	tres
255	37	celulose	255	49	companies	255	34	unos
256	37	ele	256	49	health	256	34	utilización
257	37	fabricantes	257	49	i	257	34	utilizan
258	37	limpeza	258	49	provide	258	33	antes
259	37	matérias	259	49	value	259	33	casi
260	37	novos	260	49	work	260	33	ciudad
261	37	pelos	261	48	back	261	33	compra
262	37	petróleo	262	48	cars	262	33	cuenta
263	37	sustentabilidade	263	48	consumer	263	33	estudio
264	37	sólidos	264	48	however	264	33	lugar
265	37	ter	265	48	organic	265	33	metal
266	37	vários	266	48	separated	266	33	mucho
267	36	catadores	267	48	service	267	33	ordenador
268	36	industriais	268	47	d	268	33	otra
269	36	melhor	269	47	go	269	33	sido
270	36	natureza	270	47	natural	270	33	urbanos
271	36	paulo	271	47	paperboard	271	32	alimentos
272	36	primas	272	46	green	272	32	elementos
273	36	útil	273	46	light	273	32	reducción
274	35	custo	274	46	produce	274	32	resto
275	35	eua	275	46	reduction	275	31	bosques
276	35	história	276	46	solid	276	31	calidades
277	35	população	277	46	system	277	31	embalaje
278	35	recicladadas	278	46	type	278	31	evitar
279	35	representa	279	45	books	279	31	gases
280	35	resistência	280	45	center	280	31	hojas
281	35	separação	281	45	check	281	31	responsable
282	35	sustentável	282	45	contact	282	31	social
283	35	tratamento	283	45	containing	283	31	tecnología
284	35	°	284	45	cullet	284	30	educación

285	34	br
286	34	consumidor
287	34	equipamentos
288	34	menos
289	34	mistura
290	34	química
291	34	recicladados
292	34	reduzir
293	34	relação
294	34	segundo
295	34	serão
296	34	sociais
297	34	tonelada
298	34	usadas
299	34	vantagens
300	33	alta
301	33	ar
302	33	duas
303	33	estados
304	33	feita
305	33	impacto
306	33	papelão
307	33	pelas
308	33	polietileno
309	33	rio
310	33	sanitários
311	33	serviços
312	33	todas
313	33	utilizados
314	32	desta
315	32	formas
316	32	produzido
317	32	serem
318	31	algumas
319	31	brasileiro
320	31	comum
321	31	contaminantes
322	31	elétrica
323	31	ferro
324	31	leve
325	31	menor
326	31	principalmente
327	31	rede
328	31	três
329	31	vem
330	31	www
331	30	aproximadamente
332	30	cempre
333	30	chapas
334	30	composição
335	30	especiais
336	30	fábricas
337	30	local
338	30	natural
339	30	pós
340	30	somente
341	30	técnica
342	29	cidade
343	29	desse

285	45	deposit
286	45	mixed
287	45	money
288	45	once
289	45	place
290	45	site
291	44	could
292	44	electronics
293	44	facility
294	44	few
295	44	first
296	44	good
297	44	organization
298	44	since
299	44	toys
300	44	usually
301	43	bottle
302	43	drop
303	43	example
304	43	facilities
305	43	market
306	43	markets
307	43	number
308	43	raw
309	43	reused
310	43	who
311	42	bins
312	42	both
313	42	electronic
314	42	increase
315	42	list
316	42	management
317	42	small
318	41	beverage
319	41	during
320	41	epa
321	41	fiber
322	41	part
323	41	recyclers
324	41	uses
325	41	valuable
326	40	america
327	40	battery
328	40	communities
329	40	fluids
330	40	occ
331	40	per
332	40	produced
333	40	production
334	40	public
335	40	rubber
336	40	vehicle
337	40	wood
338	39	another
339	39	average
340	39	considered
341	39	directory
342	39	enough
343	39	reducing

285	30	europea
286	30	maioria
287	30	medioambiental
288	30	móveis
289	30	nueva
290	30	nuevas
291	30	realizar
292	30	siglo
293	30	trabajo
294	30	transporte
295	29	bajo
296	29	comercio
297	29	doble
298	29	energética
299	29	entidades
300	29	fibra
301	29	hoy
302	29	mercurio
303	29	misma
304	29	peso
305	29	problema
306	29	pues
307	29	químicos
308	29	sustancias
309	28	debido
310	28	ecológico
311	28	ellos
312	28	entidad
313	28	europa
314	28	incineración
315	28	incluso
316	28	le
317	28	libre
318	28	manera
319	28	medidas
320	28	nivel
321	28	origen
322	28	poco
323	28	supone
324	28	usar
325	28	vertedero
326	27	aparatos
327	27	ayuntamiento
328	27	características
329	27	efecto
330	27	equipos
331	27	espacio
332	27	existe
333	27	forestal
334	27	grupo
335	27	kg
336	27	mejor
337	27	número
338	27	población
339	27	principales
340	27	realiza
341	27	salud
342	27	tan
343	27	útil

344	29	era	344	38	accept	344	26	actual
345	29	janelas	345	38	af	345	26	campanha
346	29	lixões	346	38	computers	346	26	compostaje
347	29	responsáveis	347	38	dispose	347	26	empresa
348	29	restos	348	38	does	348	26	evita
349	29	tecnologia	349	38	five	349	26	final
350	29	tubos	350	38	know	350	26	hecho
351	29	várias	351	38	pa	351	26	industriales
352	29	óleos	352	38	power	352	26	instalaciones
353	28	automóveis	353	38	school	353	26	medioambientales
354	28	cooperativas	354	38	see	354	26	millones
355	28	crescimento	355	38	those	355	26	oficinas
356	28	densidade	356	38	virgin	356	26	paso
357	28	disso	357	37	common	357	26	sean
358	28	origem	358	37	county	358	26	sociales
359	28	pasta	359	37	dep	359	25	actualmente
360	28	quais	360	37	don	360	25	botellas
361	28	químicos	361	37	instead	361	25	casos
362	28	resinas	362	37	manufacturers	362	25	contaminantes
363	28	urbano	363	37	residents	363	25	contenido
364	27	antes	364	37	same	364	25	contienen
365	27	cadeia	365	37	would	365	25	encontrar
366	27	combustível	366	36	debris	366	25	fácil
367	27	comunicação	367	36	devices	367	25	hacia
368	27	demais	368	36	fact	368	25	precedentes
369	27	esses	369	36	forest	369	25	residuo
370	27	isto	370	36	free	370	25	siguientes
371	27	novo	371	36	makes	371	25	tecnologías
372	27	objetivos	372	36	polystyrene	372	24	actividad
373	27	passa	373	36	protection	373	24	ahorrar
374	27	polpa	374	36	rate	374	24	alta
375	27	primeiro	375	36	store	375	24	beneficios
376	27	r	376	36	typically	376	24	centro
377	26	destinação	377	36	very	377	24	ciclo
378	26	devido	378	36	vinyl	378	24	comunicación
379	26	eles	379	35	association	379	24	etiquetas
380	26	essas	380	35	best	380	24	mientras
381	26	estrutura	381	35	billion	381	24	máquinas
382	26	existem	382	35	consumption	382	24	nuestros
383	26	frascos	383	35	gas	383	24	periódicos
384	26	importância	384	35	removed	384	24	persona
385	26	janeiro	385	35	set	385	24	primera
386	26	obtenção	386	35	storage	386	24	producen
387	26	orgânico	387	35	white	387	24	protección
388	26	participação	388	34	around	388	24	proyectos
389	26	poder	389	34	building	389	24	renovables
390	26	projetos	390	34	bulbs	390	24	tapa
391	26	reciclada	391	34	collect	391	24	temperatura
392	26	regiões	392	34	commercial	392	24	total
393	26	sempre	393	34	company	393	24	varias
394	26	sistemas	394	34	contaminants	394	24	zonas
395	26	técnicas	395	34	emissions	395	23	aumento
396	26	unidos	396	34	form	396	23	centros
397	25	aumento	397	34	homes	397	23	componentes
398	25	bauxita	398	34	htm	398	23	compuestos
399	25	ciudades	399	34	long	399	23	control
400	25	compostos	400	34	nearly	400	23	cual
401	25	conta	401	34	options	401	23	desechos
402	25	contaminação	402	34	ton	402	23	fabricantes

403	25	escritório	403	33	acid	403	23	historia
404	25	estar	404	33	children	404	23	impresos
405	25	feito	405	33	consider	405	23	mejora
406	25	http	406	33	create	406	23	mundial
407	25	instrumentos	407	33	currently	407	23	orgánica
408	25	jornal	408	33	goal	408	23	país
409	25	necessária	409	33	known	409	23	reciclados
410	25	objetivo	410	33	links	410	23	recuperar
411	25	permite	411	33	organizations	411	23	restos
412	25	pesquisa	412	33	times	412	23	sólidos
413	25	porém	413	33	us	413	23	ventajas
414	25	preservação	414	33	variety	414	23	vivienda
415	25	produzir	415	32	ann	415	23	árboles
416	25	utilizadas	416	32	arbor	416	22	aceites
417	25	ácido	417	32	break	417	22	alto
418	24	brinquedos	418	32	fluorescent	418	22	aspectos
419	24	calor	419	32	fuel	419	22	buen
420	24	consciente	420	32	iron	420	22	calor
421	24	criação	421	32	magazines	421	22	cambios
422	24	demanda	422	32	municipal	422	22	ciudadanos
423	24	desperdício	423	32	plant	423	22	conseguir
424	24	dois	424	32	several	424	22	decir
425	24	empresa	425	32	staff	425	22	después
426	24	entidades	426	32	united	426	22	documentos
427	24	faz	427	32	various	427	22	doméstico
428	24	muitas	428	32	visit	428	22	española
429	24	necessário	429	31	avoid	429	22	forestales
430	24	neste	430	31	city	430	22	fotocopiadoras
431	24	numa	431	31	day	431	22	importantes
432	24	número	432	31	disposed	432	22	iniciativas
433	24	processamento	433	31	filters	433	22	medios
434	24	promover	434	31	generally	434	22	nacional
435	24	quase	435	31	low	435	22	naturaleza
436	24	renda	436	31	newspaper	436	22	pequeños
437	24	resistente	437	31	now	437	22	petróleo
438	24	resíduo	438	31	pollution	438	22	primer
439	24	setores	439	31	styrofoam	439	22	procesos
440	24	solução	440	31	top	440	22	produce
441	24	taxa	441	30	american	441	22	puedan
442	23	ação	442	30	approximately	442	22	puerta
443	23	baixa	443	30	auto	443	22	química
444	23	d	444	30	away	444	22	r
445	23	desses	445	30	bag	445	22	requiere
446	23	escolas	446	30	commonly	446	22	responsabilidad
447	23	estes	447	30	crt	447	22	reutilizar
448	23	europa	448	30	department	448	22	selección
449	23	formação	449	30	g	449	22	situación
450	23	identificação	450	30	locations	450	22	técnicas
451	23	insumos	451	30	locator	451	22	u
452	23	junto	452	30	monitor	452	22	unas
453	23	latinhas	453	30	phone	453	21	aguas
454	23	limitações	454	30	separate	454	21	artículos
455	23	litros	455	30	single	455	21	bolsa
456	23	m	456	30	still	456	21	bolsas
457	23	novamente	457	29	automobile	457	21	cadmio
458	23	orgânicos	458	29	cause	458	21	directamente
459	23	passo	459	29	centers	459	21	distintos
460	23	polímeros	460	29	come	460	21	distribución
461	23	porque	461	29	effective	461	21	eficiente

462	23	primeira	462	29	either	462	21	entorno
463	23	quem	463	29	learn	463	21	generados
464	23	quilo	464	29	mail	464	21	global
465	23	redes	465	29	purchasing	465	21	limpios
466	23	sucateiros	466	29	reusable	466	21	medida
467	23	tanto	467	29	size	467	21	modo
468	23	transformação	468	29	smaller	468	21	necesidad
469	23	venda	469	29	systems	469	21	objetos
470	23	áreas	470	29	wastes	470	21	personal
471	23	árvores	471	28	additional	471	21	planta
472	22	aplicação	472	28	appliances	472	21	reciclan
473	22	aspectos	473	28	below	473	21	recuperado
474	22	base	474	28	box	474	21	respecto
475	22	c	475	28	call	475	21	sobres
476	22	conselho	476	28	cannot	476	21	totalmente
477	22	cores	477	28	chemicals	477	21	usos
478	22	distribuição	478	28	components	478	21	usuarios
479	22	empregos	479	28	donation	479	21	volumen
480	22	entidade	480	28	fish	480	21	zaragoza
481	22	estimular	481	28	fluid	481	21	últimos
482	22	evitar	482	28	four	482	20	actualidad
483	22	filme	483	28	great	483	20	consecuencia
484	22	impurezas	484	28	laws	484	20	costes
485	22	incineração	485	28	plants	485	20	dar
486	22	kg	486	28	possible	486	20	económico
487	22	nova	487	28	reduces	487	20	electricidad
488	22	orgânica	488	28	required	488	20	electrodomésticos
489	22	refrigerantes	489	28	result	489	20	eliminar
490	22	saúde	490	28	section	490	20	energético
491	22	seguintes	491	28	under	491	20	energías
492	22	tais	492	28	vehicles	492	20	fabrican
493	22	transparente	493	28	web	493	20	hogar
494	22	vai	494	28	whole	494	20	impresoras
495	22	virgem	495	27	accessories	495	20	incremento
496	21	agentes	496	27	demand	496	20	instalación
497	21	análise	497	27	donating	497	20	largo
498	21	atualmente	498	27	drink	498	20	limpieza
499	21	coletado	499	27	efforts	499	20	mismos
500	21	conceito	500	27	gets	500	20	máquina
501	21	custos	501	27	little	501	20	peligrosos
502	21	doméstico	502	27	means	502	20	podría
503	21	domésticos	503	27	national	503	20	principalmente
504	21	então	504	27	pay	504	20	promover
505	21	faça	505	27	pick	505	20	publicidad
506	21	fornos	506	27	purchase	506	20	recoger
507	21	interesse	507	27	put	507	20	reduce
508	21	maneira	508	26	accepted	508	20	reutilización
509	21	minério	509	26	always	509	20	serie
510	21	nosso	510	26	brown	510	20	tapas
511	21	plano	511	26	comes	511	20	te
512	21	portanto	512	26	country	512	20	unión
513	21	possibilidade	513	26	estimated	513	20	va
514	21	principal	514	26	etc	514	19	aceíte
515	21	problemas	515	26	further	515	19	aragón
516	21	queima	516	26	gallons	516	19	basuras
517	21	saber	517	26	improve	517	19	carga
518	21	separado	518	26	listing	518	19	chatarra
519	21	utensílios	519	26	reduced	519	19	ciudades
520	21	veículos	520	26	sent	520	19	colaboración

521	21	volta	521	26	supply	521	19	colectores
522	20	aumentando	522	26	why	522	19	consume
523	20	caco	523	26	without	523	19	dentro
524	20	capacitação	524	26	world	524	19	distintas
525	20	econômicos	525	25	almost	525	19	domésticos
526	20	entanto	526	25	associated	526	19	estados
527	20	escala	527	25	automobiles	527	19	formación
528	20	estado	528	25	donate	528	19	generar
529	20	etapas	529	25	grades	529	19	industrial
530	20	existe	530	25	hhw	530	19	interior
531	20	gestão	531	25	largest	531	19	kraft
532	20	instituições	532	25	ma	532	19	marcha
533	20	moldura	533	25	takes	533	19	mejorar
534	20	perda	534	25	trees	534	19	máximo
535	20	peças	535	25	ways	535	19	obtener
536	20	postos	536	24	americans	536	19	participación
537	20	realização	537	24	chemical	537	19	prensa
538	20	região	538	24	clear	538	19	práctica
539	20	sacos	539	24	colored	539	19	relación
540	20	sul	540	24	depending	540	19	revistas
541	19	adequada	541	24	efficient	541	19	separar
542	19	aterro	542	24	employees	542	19	solo
543	19	cacos	543	24	ensure	543	19	teléfono
544	19	componentes	544	24	examples	544	18	acuerdo
545	19	conforme	545	24	happens	545	18	administraciones
546	19	conhecido	546	24	manufacturer	546	18	bateria
547	19	contribuir	547	24	might	547	18	cabo
548	19	dentro	548	24	needs	548	18	carbono
549	19	descartáveis	549	24	requirements	549	18	clasificación
550	19	diretamente	550	24	resource	550	18	colores
551	19	ela	551	24	simple	551	18	combustión
552	19	elas	552	24	stored	552	18	construcción
553	19	etapa	553	24	together	553	18	critérios
554	19	fim	554	24	too	554	18	departamento
555	19	gera	555	24	total	555	18	dioxinas
556	19	gerado	556	24	view	556	18	días
557	19	gerenciamento	557	23	across	557	18	eficiencia
558	19	implantação	558	23	add	558	18	elemento
559	19	lugar	559	23	chlorine	559	18	estar
560	19	municípios	560	23	consumers	560	18	fuera
561	19	norte	561	23	due	561	18	hora
562	19	oxigênio	562	23	filter	562	18	junto
563	19	passar	563	23	foam	563	18	limpio
564	19	pilha	564	23	generated	564	18	metálicos
565	19	poluição	565	23	ground	565	18	mezclados
566	19	prata	566	23	habitat	566	18	org
567	19	preço	567	23	half	567	18	partes
568	19	primário	568	23	hawai	568	18	planeta
569	19	produz	569	23	heat	569	18	productores
570	19	proteção	570	23	industries	570	18	publicaciones
571	19	resina	571	23	lamp	571	18	solar
572	19	resultados	572	23	least	572	18	trabajadores
573	19	reutilização	573	23	methods	573	18	tóxicos
574	19	rígido	574	23	newspapers	574	17	acumuladores
575	19	seguida	575	23	potential	575	17	ante
576	19	segurança	576	23	provides	576	17	atmósfera
577	19	site	577	23	reach	577	17	aún
578	19	temperatura	578	23	recycler	578	17	cadena
579	19	veja	579	23	reusing	579	17	caras

580	19	área	580	23	schools	580	17	casa
581	18	acordo	581	23	sold	581	17	celulosa
582	18	alumina	582	23	step	582	17	compras
583	18	atuação	583	22	addition	583	17	condiciones
584	18	conama	584	22	become	584	17	consiste
585	18	curso	585	22	customers	585	17	continuación
586	18	degradação	586	22	cut	586	17	dirección
587	18	descarte	587	22	dry	587	17	diversos
588	18	fato	588	22	following	588	17	eliminación
589	18	função	589	22	garden	589	17	eléctricos
590	18	gerados	590	22	guide	590	17	eso
591	18	hora	591	22	haulers	591	17	fabricado
592	18	importantes	592	22	heavy	592	17	fuentes
593	18	isopor	593	22	highly	593	17	función
594	18	luz	594	22	increased	594	17	internacional
595	18	manutenção	595	22	lower	595	17	les
596	18	placas	596	22	mills	596	17	litros
597	18	plantas	597	22	newsprint	597	17	llegar
598	18	preciso	598	22	own	598	17	marca
599	18	recipientes	599	22	post	599	17	momento
600	18	retirada	600	22	quantity	600	17	natural
601	18	segmentos	601	22	rather	601	17	necesidades
602	18	sejam	602	22	receive	602	17	normalmente
603	18	tintas	603	22	significant	603	17	p
604	18	toda	604	22	specific	604	17	pequeñas
605	18	tudo	605	22	throw	605	17	poder
606	18	índices	606	21	actually	606	17	precio
607	17	abividro	607	21	again	607	17	refrescos
608	17	alto	608	21	along	608	17	siendo
609	17	artefatos	609	21	between	609	17	sí
610	17	baixo	610	21	build	610	17	t
611	17	cabos	611	21	carpet	611	17	tales
612	17	carbono	612	21	click	612	17	temperaturas
613	17	comunidade	613	21	contamination	613	17	tras
614	17	condições	614	21	covers	614	17	utilizado
615	17	cursos	615	21	creates	615	17	utilizados
616	17	dessa	616	21	demolition	616	17	valor
617	17	destino	617	21	difficult	617	17	ver
618	17	disposição	618	21	easy	618	17	vista
619	17	domiciliar	619	21	figures	619	16	actuales
620	17	energética	620	21	greenhouse	620	16	almacenamiento
621	17	equivalente	621	21	impact	621	16	alrededor
622	17	filmes	622	21	incinerators	622	16	aquellas
623	17	folha	623	21	inside	623	16	asimismo
624	17	htm	624	21	lumber	624	16	asociaciones
625	17	lado	625	21	matter	625	16	cerca
626	17	legislação	626	21	processed	626	16	co
627	17	maioria	627	21	require	627	16	comunidades
628	17	manganês	628	21	requires	628	16	contiene
629	17	metade	629	21	screen	629	16	cooper
630	17	milhão	630	21	soda	630	16	depuración
631	17	máquina	631	21	supplies	631	16	diseño
632	17	ocorre	632	21	technology	632	16	distribuidores
633	17	pesados	633	20	able	633	16	efectos
634	17	podendo	634	20	better	634	16	escritura
635	17	ponto	635	20	certain	635	16	especialmente
636	17	propriedades	636	20	contains	636	16	facilitar
637	17	prática	637	20	control	637	16	fomentar
638	17	pó	638	20	education	638	16	generación

639	17	recicladores
640	17	responsabilidade
641	17	s
642	17	separados
643	17	significa
644	17	totalmente
645	17	urbanos
646	17	zinco
647	16	aparelhos
648	16	apesar
649	16	aqui
650	16	areia
651	16	aspecto
652	16	camadas
653	16	capacidade
654	16	coloque
655	16	comercialização
656	16	cozinha
657	16	democratização
658	16	diversas
659	16	doenças
660	16	econômico
661	16	fácil
662	16	homem
663	16	impactos
664	16	iniciativas
665	16	lo
666	16	massa
667	16	mecânica
668	16	mesma
669	16	nós
670	16	pouco
671	16	presente
672	16	produzidas
673	16	pré
674	16	recicla
675	16	recicladoras
676	16	resolução
677	16	supermercados
678	16	sólido
679	16	usada
680	16	utilizada
681	16	vantagem
682	16	vegetal
683	15	alemanha
684	15	aquecimento
685	15	brasileiros
686	15	camada
687	15	casa
688	15	conhecendo
689	15	contato
690	15	cádmio
691	15	desempenho
692	15	deverá
693	15	divulgação
694	15	dá
695	15	década
696	15	economiza
697	15	eficiência

639	20	floor
640	20	hard
641	20	longer
642	20	main
643	20	needed
644	20	option
645	20	pieces
646	20	placed
647	20	pressure
648	20	ready
649	20	shipping
650	20	taken
651	20	tin
652	20	universal
653	19	antifreeze
654	19	asphalt
655	19	bring
656	19	called
657	19	carbon
658	19	change
659	19	colors
660	19	crushed
661	19	determine
662	19	empty
663	19	especially
664	19	furnace
665	19	greater
666	19	handling
667	19	hauler
668	19	house
669	19	institute
670	19	location
671	19	meet
672	19	mix
673	19	monitors
674	19	near
675	19	opportunities
676	19	prevent
677	19	sand
678	19	upon
679	19	useful
680	19	weight
681	19	working
682	18	added
683	18	brake
684	18	cartons
685	18	cases
686	18	casing
687	18	catalogs
688	18	cell
689	18	cent
690	18	concrete
691	18	copy
692	18	created
693	18	deposits
694	18	drain
695	18	drinking
696	18	exposure
697	18	ferrous

639	16	igual
640	16	magnética
641	16	mantenimiento
642	16	modelo
643	16	municipios
644	16	nuestras
645	16	obtiene
646	16	organización
647	16	pesados
648	16	pila
649	16	políticas
650	16	posterior
651	16	presenta
652	16	producir
653	16	propio
654	16	proveedores
655	16	proyecto
656	16	pueda
657	16	recicla
658	16	sería
659	16	solución
660	16	superior
661	16	trata
662	16	vegetales
663	16	vertido
664	16	volver
665	16	zinc
666	15	ahora
667	15	alguna
668	15	andalucía
669	15	apenas
670	15	caliente
671	15	carpetas
672	15	clima
673	15	comunidad
674	15	conservación
675	15	constituye
676	15	cuanto
677	15	estrategia
678	15	finales
679	15	fracción
680	15	frente
681	15	futuro
682	15	g
683	15	hábitos
684	15	idea
685	15	justo
686	15	kilos
687	15	kioto
688	15	lado
689	15	libros
690	15	luego
691	15	miembros
692	15	mismas
693	15	mm
694	15	método
695	15	móvil
696	15	neumáticos
697	15	nunca

698	15	encaminhado
699	15	entretanto
700	15	espaço
701	15	evitando
702	15	excesso
703	15	feitos
704	15	fora
705	15	gestor
706	15	governo
707	15	graças
708	15	inclusive
709	15	iniciativa
710	15	lavagem
711	15	liquidificador
712	15	los
713	15	município
714	15	nossa
715	15	num
716	15	org
717	15	pedaços
718	15	período
719	15	planeta
720	15	plásticas
721	15	pneu
722	15	potes
723	15	presença
724	15	própria
725	15	realizada
726	15	sacolas
727	15	secretaria
728	15	selos
729	15	substâncias
730	15	série
731	15	tendo
732	15	transformados
733	15	triângulo
734	15	vapor
735	15	óxido
736	14	abaixo
737	14	animais
738	14	artesanal
739	14	associações
740	14	bateria
741	14	básico
742	14	chegar
743	14	classe
744	14	cobre
745	14	comprar
746	14	consumir
747	14	copos
748	14	diagnóstico
749	14	eletrodomésticos
750	14	equipe
751	14	extração
752	14	fardos
753	14	fase
754	14	for
755	14	impermeável
756	14	interior

698	18	future
699	18	grass
700	18	harmful
701	18	includes
702	18	industrial
703	18	last
704	18	leaves
705	18	loop
706	18	milk
707	18	mulching
708	18	participation
709	18	processes
710	18	protect
711	18	pulp
712	18	quantities
713	18	remaining
714	18	sites
715	18	sustainable
716	18	transportation
717	18	treated
718	18	try
719	17	adding
720	17	animal
721	17	animals
722	17	annually
723	17	cadmium
724	17	cleaners
725	17	contaminate
726	17	contaminated
727	17	contents
728	17	coordinator
729	17	copper
730	17	days
731	17	developed
732	17	easily
733	17	gallon
734	17	generate
735	17	hours
736	17	imposed
737	17	ingredients
738	17	keeping
739	17	likely
740	17	machines
741	17	method
742	17	order
743	17	pet
744	17	release
745	17	residuals
746	17	safe
747	17	shredded
748	17	six
749	17	sources
750	17	therefore
751	17	transfer
752	17	transmission
753	17	tubes
754	17	users
755	17	volume
756	17	wire

698	15	obstante
699	15	organizaciones
700	15	permiten
701	15	porcentaje
702	15	principal
703	15	procedente
704	15	programa
705	15	prácticas
706	15	recogen
707	15	rsu
708	15	segunda
709	15	suelo
710	15	superficie
711	15	tinta
712	15	tirar
713	15	todavía
714	15	tonelada
715	15	transformación
716	15	áreas
717	14	asegurar
718	14	atacos
719	14	base
720	14	blanqueo
721	14	botón
722	14	campañas
723	14	certificado
724	14	cierre
725	14	clientes
726	14	compromiso
727	14	contribuir
728	14	creación
729	14	da
730	14	determinados
731	14	disminuir
732	14	eco
733	14	económicos
734	14	electrónico
735	14	encuentran
736	14	español
737	14	etiqueta
738	14	evoswitch
739	14	existencia
740	14	exterior
741	14	falta
742	14	generan
743	14	impactos
744	14	imprentas
745	14	incineradora
746	14	incineradoras
747	14	internet
748	14	interés
749	14	ligeros
750	14	llegan
751	14	mayores
752	14	medicamentos
753	14	ocasiones
754	14	orgánicos
755	14	plomo
756	14	política

757	14	locais
758	14	modo
759	14	máximo
760	14	nome
761	14	ong
762	14	padrão
763	14	passou
764	14	precisam
765	14	produzidos
766	14	público
767	14	recolhimento
768	14	seleção
769	14	separar
770	14	somos
771	14	século
772	14	tampas
773	14	tecnologias
774	14	tenham
775	14	transformando
776	14	troca
777	14	umidade
778	14	usar
779	14	voltar
780	13	abipet
781	13	acesso
782	13	alternativa
783	13	armazenamento
784	13	atingir
785	13	aumentar
786	13	busca
787	13	coletados
788	13	computadores
789	13	comunidades
790	13	condomínios
791	13	consciência
792	13	contendo
793	13	contribuição
794	13	cooperativa
795	13	corresponde
796	13	crescente
797	13	decomposição
798	13	deles
799	13	descartados
800	13	desenvolver
801	13	dessas
802	13	dias
803	13	eco
804	13	econômica
805	13	embora
806	13	especial
807	13	especificações
808	13	estudo
809	13	exemplos
810	13	facilitar
811	13	falta
812	13	fazendo
813	13	ferrosos
814	13	fontes
815	13	forem

757	16	already
758	16	although
759	16	areas
760	16	b
761	16	bodies
762	16	broken
763	16	crt
764	16	cups
765	16	damage
766	16	domestic
767	16	fees
768	16	general
769	16	give
770	16	goods
771	16	helps
772	16	identify
773	16	leftover
774	16	levels
775	16	listed
776	16	lives
777	16	located
778	16	look
779	16	manufacture
780	16	mats
781	16	original
782	16	pesticide
783	16	piles
784	16	please
785	16	private
786	16	processors
787	16	released
788	16	results
789	16	return
790	16	second
791	16	separation
792	16	special
793	16	strength
794	16	television
795	16	think
796	16	unit
797	16	until
798	16	want
799	16	whether
800	16	within
801	16	yard
802	15	activities
803	15	additives
804	15	agency
805	15	alternatives
806	15	amounts
807	15	benefit
808	15	child
809	15	coal
810	15	coated
811	15	commodities
812	15	contract
813	15	development
814	15	educate
815	15	fill

757	14	poner
758	14	posibilidades
759	14	potencial
760	14	pública
761	14	públicas
762	14	quieres
763	14	realizan
764	14	ropa
765	14	seguridad
766	14	tasa
767	14	tóxicas
768	14	unidos
769	14	valores
770	14	varios
771	14	verdes
772	13	algo
773	13	alicante
774	13	alternativas
775	13	ambos
776	13	animales
777	13	aplicaciones
778	13	aquellos
779	13	azul
780	13	buena
781	13	cajas
782	13	ciclos
783	13	comerciales
784	13	comprar
785	13	consecuencias
786	13	contacto
787	13	correo
788	13	crecimiento
789	13	demás
790	13	derivados
791	13	diferencia
792	13	difícil
793	13	documento
794	13	ecológica
795	13	elaboración
796	13	eléctrica
797	13	encuadernación
798	13	era
799	13	ese
800	13	espacios
801	13	estado
802	13	extracción
803	13	formas
804	13	fotocopias
805	13	fundamental
806	13	gas
807	13	grado
808	13	había
809	13	hacen
810	13	iluminación
811	13	iniciativa
812	13	invernadero
813	13	ley
814	13	niveles
815	13	nosotros

816	13	formado
817	13	fundição
818	13	fusão
819	13	garrafa
820	13	gerada
821	13	gerando
822	13	infinitas
823	13	iso
824	13	japonês
825	13	maiores
826	13	mercados
827	13	meses
828	13	necessidade
829	13	números
830	13	objetos
831	13	organização
832	13	presentes
833	13	produtiva
834	13	produtivo
835	13	pública
836	13	quilos
837	13	químico
838	13	separadas
839	13	simples
840	13	tal
841	13	tela
842	13	telefone
843	13	terra
844	13	transformado
845	13	tão
846	13	unidades
847	13	usina
848	13	verde
849	13	voltando
850	12	acontece
851	12	adubo
852	12	alimentícios
853	12	art
854	12	articulação
855	12	bacia
856	12	bilhões
857	12	bom
858	12	branco
859	12	brasileiras
860	12	carro
861	12	centro
862	12	classificação
863	12	comerciais
864	12	comissão
865	12	compra
866	12	comuns
867	12	conjunto
868	12	conscientização
869	12	consiste
870	12	criado
871	12	dentre
872	12	depende
873	12	design
874	12	deverão

816	15	housing
817	15	http
818	15	humans
819	15	individual
820	15	island
821	15	jars
822	15	labels
823	15	liquid
824	15	marketed
825	15	necessary
826	15	north
827	15	pages
828	15	paints
829	15	phones
830	15	point
831	15	practices
832	15	prepare
833	15	produces
834	15	referred
835	15	road
836	15	safety
837	15	savings
838	15	sell
839	15	soils
840	15	sort
841	15	sorting
842	15	start
843	15	stewardship
844	15	substances
845	15	support
846	15	thus
847	15	understand
848	14	bales
849	14	burning
850	14	care
851	14	current
852	14	daily
853	14	doing
854	14	donations
855	14	enter
856	14	extremely
857	14	fee
858	14	friendly
859	14	global
860	14	government
861	14	grade
862	14	hdpe
863	14	hold
864	14	human
865	14	incineration
866	14	insulation
867	14	key
868	14	latex
869	14	limited
870	14	machine
871	14	millions
872	14	myth
873	14	naturally
874	14	nature

816	13	plan
817	13	posteriormente
818	13	procedimientos
819	13	producido
820	13	protocolo
821	13	puesto
822	13	públicos
823	13	realizado
824	13	red
825	13	sentido
826	13	solamente
827	13	somos
828	13	soporte
829	13	tamaño
830	13	término
831	13	usadas
832	13	vehículos
833	13	viviendas
834	12	análisis
835	12	aproximadamente
836	12	aquí
837	12	arpal
838	12	blancos
839	12	capa
840	12	carbón
841	12	combustibles
842	12	compost
843	12	común
844	12	conjunto
845	12	debería
846	12	descarga
847	12	dicen
848	12	directiva
849	12	disponer
850	12	diversas
851	12	dióxido
852	12	electrónicos
853	12	envasador
854	12	esa
855	12	estamos
856	12	exentos
857	12	fabricados
858	12	fax
859	12	folletos
860	12	forman
861	12	fácilmente
862	12	genera
863	12	guardar
864	12	habitual
865	12	hablar
866	12	habrá
867	12	hechos
868	12	implantación
869	12	incluyen
870	12	informar
871	12	listado
872	12	llega
873	12	llevar
874	12	locales

875	12	diminuir
876	12	disponíveis
877	12	disponível
878	12	elaboração
879	12	encaminhados
880	12	enquanto
881	12	equipamento
882	12	estas
883	12	existentes
884	12	fabricados
885	12	facilidade
886	12	facilmente
887	12	fins
888	12	fios
889	12	ganhos
890	12	gases
891	12	geralmente
892	12	grupos
893	12	ideal
894	12	institucional
895	12	início
896	12	jornais
897	12	longo
898	12	manter
899	12	mineração
900	12	moléculas
901	12	municipal
902	12	necessidades
903	12	normalmente
904	12	obtem
905	12	paraná
906	12	parceiros
907	12	partes
908	12	peneira
909	12	peneiras
910	12	perfil
911	12	poliéster
912	12	pontos
913	12	possui
914	12	potencial
915	12	recicladora
916	12	reduz
917	12	reduzindo
918	12	resfriamento
919	12	responsável
920	12	resultado
921	12	resultante
922	12	risco
923	12	sucatas
924	12	suficiente
925	12	temos
926	12	teor
927	12	triagem
928	12	técnico
929	12	utilizam
930	12	visando
931	12	visão
932	12	últimos
933	11	aberto

875	14	news
876	14	numbers
877	14	online
878	14	petroleum
879	14	pollutants
880	14	polyethylene
881	14	projects
882	14	refined
883	14	rinse
884	14	run
885	14	saved
886	14	serious
887	14	shipped
888	14	shop
889	14	something
890	14	sometimes
891	14	sorted
892	14	steps
893	14	tax
894	14	toy
895	14	treatment
896	14	typical
897	14	yourself
898	13	access
899	13	ask
900	13	base
901	13	becomes
902	13	big
903	13	bill
904	13	buying
905	13	caps
906	13	cfls
907	13	com
908	13	compact
909	13	consumed
910	13	continue
911	13	council
912	13	cover
913	13	creating
914	13	dangerous
915	13	data
916	13	discards
917	13	doesn
918	13	done
919	13	dust
920	13	electrical
921	13	electricity
922	13	equivalent
923	13	event
924	13	family
925	13	federal
926	13	feed
927	13	fibers
928	13	forms
929	13	fuels
930	13	garbage
931	13	generators
932	13	given
933	13	glossy

875	12	línea
876	12	masa
877	12	media
878	12	medioambiente
879	12	mezcla
880	12	mínimo
881	12	ofrece
882	12	ordenadores
883	12	organismos
884	12	papelera
885	12	presentan
886	12	principios
887	12	procede
888	12	propia
889	12	propiedades
890	12	puesta
891	12	químicas
892	12	recuperadas
893	12	siguiente
894	12	socios
895	12	soluciones
896	12	sur
897	12	tal
898	12	tcf
899	12	tenemos
900	12	ue
901	12	usan
902	12	van
903	12	vía
904	12	web
905	12	ámbito
906	12	área
907	12	único
908	11	acción
909	11	administración
910	11	alternativa
911	11	aprovechamiento
912	11	ayudar
913	11	bienes
914	11	cartones
915	11	carácter
916	11	certificados
917	11	colas
918	11	comercios
919	11	concepto
920	11	conocimiento
921	11	contra
922	11	coordinación
923	11	defensa
924	11	destino
925	11	dicho
926	11	dificultades
927	11	edificios
928	11	elige
929	11	ellas
930	11	empleo
931	11	esos
932	11	especial
933	11	especiales

934	11	ambientalmente	934	13	grocery	934	11	especies
935	11	aqueles	935	13	higher	935	11	estudios
936	11	associados	936	13	hot	936	11	formato
937	11	banho	937	13	households	937	11	fuelle
938	11	blocos	938	13	incandescent	938	11	funcionamiento
939	11	característica	939	13	instructions	939	11	garantizar
940	11	casacas	940	13	involved	940	11	generaciones
941	11	causa	941	13	involves	941	11	grados
942	11	chamado	942	13	label	942	11	haya
943	11	combustíveis	943	13	let	943	11	hemos
944	11	começou	944	13	level	944	11	hogares
945	11	compromisso	945	13	major	945	11	humana
946	11	conservação	946	13	mixing	946	11	imprescindible
947	11	considerando	947	13	occurs	947	11	introducir
948	11	consume	948	13	others	948	11	ir
949	11	consumidas	949	13	packing	949	11	litio
950	11	consumidores	950	13	past	950	11	llamar
951	11	contribuindo	951	13	pennsylvania	951	11	lucro
952	11	cordas	952	13	percentage	952	11	luz
953	11	cujo	953	13	person	953	11	lámparas
954	11	cultura	954	13	picked	954	11	líneas
955	11	descoberto	955	13	promote	955	11	mala
956	11	dez	956	13	providing	956	11	mitad
957	11	dicas	957	13	purposes	957	11	muebles
958	11	difícil	958	13	rates	958	11	necesaria
959	11	disponibilidade	959	13	ray	959	11	necesita
960	11	drenagem	960	13	related	960	11	obtención
961	11	eletricidade	961	13	residential	961	11	ofrecen
962	11	energético	962	13	shape	962	11	opción
963	11	envolvidas	963	13	spent	963	11	pastas
964	11	estudos	964	13	spill	964	11	posibilidad
965	11	evita	965	13	standards	965	11	posibles
966	11	executiva	966	13	thermometers	966	11	presencia
967	11	fabricado	967	13	thermostats	967	11	presión
968	11	fabricar	968	13	tools	968	11	primeros
969	11	fazem	969	13	toxins	969	11	promoción
970	11	feitas	970	13	transport	970	11	quiere
971	11	fibra	971	13	window	971	11	recicladass
972	11	gerar	972	12	allows	972	11	reciclarse
973	11	gerações	973	12	ban	973	11	resistencia
974	11	identificar	974	12	blue	974	11	responsables
975	11	impressão	975	12	body	975	11	resulta
976	11	limites	976	12	boston	976	11	riesgo
977	11	lâminas	977	12	bottom	977	11	segundo
978	11	lâmpada	978	12	cereal	978	11	sensibilización
979	11	metálicas	979	12	charge	979	11	significa
980	11	microorganismos	980	12	choice	980	11	similares
981	11	missão	981	12	cities	981	11	spittelau
982	11	misturado	982	12	combined	982	11	subcarpetas
983	11	máquinas	983	12	commingled	983	11	suelen
984	11	móveis	984	12	cubic	984	11	teléfonos
985	11	nem	985	12	designed	985	11	temas
986	11	origens	986	12	eco	986	11	tiendas
987	11	parcerias	987	12	effort	987	11	tradicionales
988	11	participantes	988	12	engine	988	11	tubos
989	11	perigosos	989	12	exchange	989	11	técnica
990	11	pisos	990	12	full	990	11	utilizadas
991	11	plástica	991	12	gases	991	11	venta
992	11	poliestireno	992	12	going	992	11	zona

993	11	pp
994	11	prensadas
995	11	problema
996	11	proporciona
997	11	recebe
998	11	residências
999	11	saiba
1000	11	sido

993	12	growing
994	12	juice
995	12	layers
996	12	left
997	12	loss
998	12	maintain
999	12	manufactured
1000	12	melted

993	11	última
994	11	último
995	10	adecuada
996	10	adquirir
997	10	agricultura
998	10	alcanzar
999	10	anual
1000	10	asegura

### Anexo 3

#### Lista de palavras-chave em Português

	<b>Freq.</b>	<b>Keyness</b>	<b>Keyword</b>
1	548	4039455	reciclagem
2	292	2075191	alumínio
3	261	2006192	embalagens
4	299	1828842	lixo
5	255	1448511	coleta
6	226	1423251	vidro
7	390	1351483	papel
8	162	1295954	reciclado
9	183	1203512	plásticos
10	190	1155477	plástico
11	120	977608	pvc
12	315	926823	material
13	127	921882	pet
14	115	902980	latas
15	130	901308	seletiva
16	166	867935	resíduos
17	115	768191	embalagem
18	85	695238	recicláveis
19	135	671316	toneladas
20	69	616081	rebia
21	240	603636	produtos
22	87	581169	pneus
23	95	571387	aço
24	81	554042	baterias
25	72	541083	sucata
26	190	494981	materiais
27	158	491546	consumo
28	72	444908	pilhas
29	56	440427	reciclável
30	70	432550	garrafas
31	147	418450	ambiental
32	59	415907	reciclar
33	75	378289	metais
34	5822	373370	de
35	52	370053	lata
36	91	368135	papéis
37	56	363841	aterros
38	142	353135	indústria
39	41	344243	ondulado
40	40	341114	aparas
41	169	332037	água
42	93	318114	fabricação
43	160	316472	ambiente
44	57	314707	fibras
45	79	301008	prima

46	71	298132	óleo
47	51	295509	lâmpadas
48	35	286805	recicladas
49	41	285071	reaproveitamento
50	38	284461	compostagem
51	54	281622	vidros
52	104	278268	matéria
53	61	273371	metal
54	30	267861	cempre
55	47	263625	chumbo
56	50	262648	borracha
57	36	262463	catadores
58	47	252333	bebidas
59	77	237531	indústrias
60	34	229045	reciclados
61	37	228633	sustentabilidade
62	271	225946	brasil
63	34	220443	tonelada
64	33	217612	polietileno
65	31	214785	contaminantes
66	30	206742	chapas
67	23	205360	sucateiros
68	186	203292	meio
69	26	201401	reciclada
70	121	196707	energia
71	25	196300	bauxita
72	33	190499	papelão
73	33	186263	sanitários
74	29	183452	lixões
75	43	183146	caixas
76	2874	181451	e
77	189	180601	produção
78	43	178908	mercúrio
79	69	172853	alimentos
80	27	172614	polpa
81	23	170142	latinhas
82	28	160910	resinas
83	22	159289	incineração
84	37	159162	celulose
85	26	156736	frascos
86	106	155495	produto
87	66	154203	peso
88	37	146558	sólidos
89	21	144519	utensílios
90	16	142859	recicladoras
91	60	140410	ambientais
92	29	140147	óleos
93	19	139518	cacos
94	17	139048	abividro
95	29	135098	restos
96	1304	133897	para
97	17	131430	recicladores

98	20	131118	moldura
99	30	130737	fábricas
100	58	130544	ciclo
101	37	130304	limpeza
102	22	129640	refrigerantes
103	29	128322	tubos
104	40	127041	folhas
105	15	126473	liquidificador
106	95	126228	etc
107	24	126171	desperdício
108	37	125981	fabricantes
109	20	120955	caco
110	26	117543	destinação
111	19	117406	descartáveis
112	21	117070	coletado
113	26	116987	orgânico
114	20	116192	sacos
115	13	116073	abipet
116	42	115350	madeira
117	29	112828	janelas
118	23	112027	quilo
119	15	111553	sacolas
120	28	111231	cooperativas
121	48	110824	transporte
122	47	110371	longa
123	31	110106	leve
124	19	110048	aterro
125	49	109280	composto
126	28	109210	pasta
127	18	108606	isopor
128	35	107677	sustentável
129	18	107369	recipientes
130	12	107145	recicladora
131	24	106897	brinquedos
132	15	106571	economiza
133	22	106227	impurezas
134	14	106025	fardos
135	28	105765	automóveis
136	19	104902	reutilização
137	22	103668	virgem
138	14	103217	tampas
139	15	100512	potes
140	18	99787	descarte
141	21	99182	fornos
142	21	99182	minério
143	11	98216	trapos
144	15	97113	selos
145	24	95995	resistente
146	19	94944	pilha
147	129	94131	vida
148	15	94087	cádmio
149	35	92997	separação

150	37	92484	petróleo
151	67	92262	redução
152	23	91561	orgânicos
153	10	89287	btus
154	10	89287	enfardado
155	10	89287	inquebrável
156	10	89287	rerrefino
157	19	88972	resina
158	13	88348	condomínios
159	74	88269	tipos
160	110	88229	mil
161	14	87910	copos
162	14	86510	impermeável
163	424	86296	ser
164	17	86113	manganês
165	41	85836	benefícios
166	43	85586	usados
167	17	85372	artefatos
168	54	84418	industrial
169	54	83976	usado
170	36	82778	industriais
171	22	82499	transparente
172	31	82456	ferro
173	23	82325	litros
174	12	82298	sucatas
175	67	82018	quantidade
176	13	81734	ferrosos
177	25	81473	escritório
178	164	80411	outros
179	9	80358	anip
180	9	80358	pebd
181	9	80358	plastivida
182	34	80022	usadas
183	19	78967	prata
184	21	78865	doméstico
185	21	78865	domésticos
186	16	78796	coloque
187	36	78712	útil
188	17	78194	cabos
189	15	78138	triângulo
190	21	77785	separado
191	13	77747	fundição
192	18	77337	conama
193	987	77295	é
194	12	77060	poliéster
195	102	77057	mercado
196	34	76255	vantagens
197	57	76059	volume
198	24	74991	resíduo
199	41	74979	recuperação
200	173	74421	processo
201	17	74358	domiciliar

202	23	74013	polímeros
203	9	73880	lubrificante
204	9	73880	monômeros
205	34	72076	mistura
206	23	71677	insumos
207	8	71430	akatu
208	8	71430	cartonadas
209	8	71430	eteno
210	8	71430	ktons
211	8	71430	reciclabilidade
212	8	71430	solados
213	94	70151	cerca
214	9	69974	kraft
215	9	69974	reciclou
216	45	69251	naturais
217	11	66652	tapetes
218	25	66075	contaminação
219	18	65823	tintas
220	14	65303	eletrodomésticos
221	8	65174	bata
222	8	65174	moldada
223	8	65174	reciclando
224	28	64991	urbano
225	28	64686	químicos
226	22	64560	kg
227	9	64403	grampos
228	11	64244	ambientalmente
229	11	64244	prensadas
230	36	63979	primas
231	27	63464	combustível
232	17	63240	zinco
233	14	62993	bateria
234	15	62926	pneu
235	7	62501	cartonada
236	7	62501	hidrapulper
237	7	62501	rebites
238	7	62501	recauchutagem
239	19	62074	primário
240	40	61876	cor
241	34	61661	reduzir
242	8	61468	reprocessamento
243	22	61118	orgânica
244	12	61097	adubo
245	12	61097	alimentícios
246	14	60875	ong
247	12	60349	peneiras
248	9	60307	vinila
249	22	60107	empregos
250	16	59966	democratização
251	15	59840	transformados
252	46	59401	índice
253	14	59394	artesanal

254	18	59304	alumina
255	10	59249	prensas
256	17	59229	htm
257	17	59229	pesados
258	24	58666	consciente
259	8	58608	pead
260	8	58608	xaxim
261	9	58590	economizando
262	10	58096	reutilizar
263	13	57948	infinitas
264	15	57072	lavagem
265	10	57010	virgens
266	19	56757	rígido
267	7	56496	abinee
268	7	56496	barrilha
269	7	56496	fenícios
270	7	56496	inservíveis
271	7	56496	papiro
272	15	56332	conhecendo
273	37	56306	matérias
274	32	56179	produzido
275	11	55942	lâminas
276	25	55717	ácido
277	9	55614	bandejas
278	10	55011	refrigerante
279	10	55011	utilidades
280	13	54913	descartados
281	11	54479	casacas
282	34	54432	br
283	31	54385	elétrica
284	8	54222	acondicionamento
285	8	54222	fundidos
286	6	53572	aparistas
287	6	53572	ciclosft
288	6	53572	couchés
289	6	53572	ecoficientes
290	6	53572	morim
291	6	53572	parafinados
292	6	53572	recicloteca
293	6	53572	reviverde
294	6	53572	torneira
295	7	53013	abal
296	7	53013	terefalato
297	8	52447	contaminam
298	14	52334	recolhimento
299	41	52026	civil
300	44	51682	associação
301	15	51661	pedaços
302	10	51559	retalhos
303	11	50630	consumidas
304	12	50435	peneira
305	7	50353	alcalina

306	34	50341	consumidor
307	20	49843	postos
308	16	49742	sólido
309	17	49500	separados
310	8	49434	calorífico
311	10	49335	tiras
312	9	48951	descartadas
313	16	48757	cozinha
314	10	48649	lençóis
315	10	48649	leveza
316	10	48649	níquel
317	15	48503	encaminhado
318	16	48278	recicla
319	83	48248	empresas
320	7	48173	descontaminação
321	11	47903	lâmpada
322	20	47855	aumentando
323	13	47635	garrafa
324	37	47327	aplicações
325	22	47288	estimular
326	9	47210	polipropileno
327	38	47095	solo
328	25	46943	compostos
329	8	46931	comercializam
330	21	46814	faça
331	24	46770	calor
332	10	46731	envelopes
333	50	46584	geração
334	14	46564	consumir
335	11	46428	misturado
336	7	46316	fluorescentes
337	11	45960	microorganismos
338	15	45937	gestor
339	8	45821	durável
340	8	45821	separe
341	35	45332	resistência
342	11	45055	perigosos
343	15	44980	vapor
344	8	44787	jogue
345	8	44787	moldado
346	8	44787	vassouras
347	18	44756	retirada
348	7	44695	fertilizante
349	7	44695	freáticos
350	5	44644	cabides
351	5	44644	craqueamento
352	5	44644	ecoeficiência
353	5	44644	esterco
354	5	44644	latinha
355	5	44644	metalizados
356	5	44644	paracambi
357	5	44644	plastificados

358	5	44644	pneumáticos
359	5	44644	pueras
360	5	44644	reciclaço
361	6	44621	adesivas
362	6	44621	automotivas
363	6	44621	cáustica
364	6	44621	eps
365	6	44621	reaproveitados
366	6	44621	recarregáveis
367	17	44063	energética
368	30	43973	aproximadamente
369	10	43409	isentos
370	11	43362	metálicas
371	7	43255	baldes
372	30	43002	especiais
373	10	42910	cloreto
374	10	42910	prensa
375	6	42184	condicionar
376	6	42184	policloreto
377	23	42138	limitações
378	44	42092	utilizado
379	11	41803	voluntária
380	12	41699	triagem
381	9	41639	sobras
382	16	41611	supermercados
383	13	41068	voltando
384	9	41057	durabilidade
385	10	41040	espelhos
386	13	40557	usina
387	26	40215	obtenção
388	6	40205	extrusora
389	6	40205	lingotes
390	6	40205	ton
391	12	40200	encaminhados
392	11	40015	residências
393	46	39987	diversos
394	12	39912	fabricados
395	17	39855	urbanos
396	7	39698	lítio
397	14	39690	cobre
398	8	39054	têxtil
399	25	38793	instrumentos
400	17	38790	pó
401	13	38626	separadas
402	6	38530	alcalinas
403	6	38530	vazada
404	9	38423	procure
405	13	38397	eco
406	8	38390	cloro
407	8	38390	pellets
408	114	38301	podem
409	19	38223	gerado

410	11	38069	fabricado
411	9	37944	evite
412	7	37771	tambores
413	93	37709	uso
414	14	37660	transformando
415	28	37464	densidade
416	19	37369	oxigênio
417	18	37338	degradação
418	7	36904	etileno
419	7	36904	vazamento
420	19	36776	gerenciamento
421	14	36525	separar
422	12	36508	obtem
423	5	36314	catação
424	5	36314	eletrólise
425	5	36314	fraldas
426	5	36314	reaproveitadas
427	205	36214	pode
428	7	36092	acrílico
429	7	36092	polimerização
430	6	35790	canos
431	4	35715	auditado
432	4	35715	compostado
433	4	35715	eletrolítica
434	4	35715	gwh
435	4	35715	metallic
436	4	35715	monolúcido
437	4	35715	nPK
438	4	35715	parkes
439	4	35715	refusão
440	4	35715	resíduos
441	4	35715	reutilize
442	4	35715	sequem
443	4	35715	taint
444	8	35454	rígidas
445	11	35212	consome
446	11	34690	cordas
447	11	34690	drenagem
448	11	34690	pisos
449	15	34642	óxido
450	6	34636	contêineres
451	6	34636	embalar
452	7	34604	ácida
453	9	34547	calçados
454	20	34352	capacitação
455	50	34271	atividade
456	13	34261	iso
457	9	34169	hp
458	5	34128	reaproveitáveis
459	25	34005	preservação
460	8	33939	extrusão
461	7	33920	domésticas

462	29	33777	responsáveis
463	24	33568	promover
464	11	33445	energético
465	33	33402	ar
466	13	33344	transformado
467	25	33337	http
468	16	33031	aparelhos
469	23	33006	novamente
470	11	32970	dicas
471	31	32841	www
472	9	32740	economizar
473	7	32651	fundes
474	27	32640	cadeia
475	6	32634	autopeças
476	6	32634	simbologia
477	11	32507	descoberto
478	10	32426	consumido
479	12	32391	conscientização
480	5	32371	arames
481	5	32371	regenerada
482	5	32371	transformadores
483	15	32335	evitando
484	20	32333	econômicos
485	34	32277	equipamentos
486	10	32156	tóxicos
487	15	32063	aquecimento
488	7	32061	adesivos
489	12	32005	fios
490	19	31968	gera
491	6	31753	automotivos
492	6	31753	parafusos
493	9	31747	soda
494	8	31722	sensibilização
495	7	31497	lubrificantes
496	19	31391	poluição
497	8	31316	coco
498	10	31120	totalizando
499	9	31119	ps
500	13	30978	armazenamento
501	7	30956	clique
502	7	30956	notebook
503	24	30909	entidades
504	5	30896	filmadoras
505	5	30896	fotocópias
506	5	30896	revalorização
507	4	30734	eletroímã
508	4	30734	laminados
509	4	30734	objectos
510	7	30437	banheiro
511	9	30222	cosméticos
512	433	30187	são
513	6	30174	cortados

514	35	30151	representa
515	21	30149	veículos
516	16	30139	areia
517	34	29974	química
518	11	29954	poliestireno
519	7	29939	indesejáveis
520	7	29939	reutilizado
521	13	29892	coletados
522	48	29890	economia
523	15	29878	plásticas
524	5	29624	aglutinador
525	5	29624	brinquedoteca
526	5	29624	consumimos
527	5	29624	escovas
528	5	29624	reaproveitar
529	5	29624	trituração
530	11	29562	banho
531	11	29562	executiva
532	11	29369	fabricar
533	10	29223	esteira
534	16	29036	vegetal
535	12	28842	resfriamento
536	6	28792	flocos
537	6	28792	reutilizados
538	19	28780	veja
539	49	28624	setor
540	7	28552	energéticos
541	7	28552	radiografias
542	5	28505	borrachas
543	5	28505	eletroeletrônicos
544	5	28505	esquadrias
545	5	28505	jogos
546	5	28505	petroquímicas
547	22	28492	cores
548	8	28435	recuperadas
549	18	28317	gerados
550	4	28123	catador
551	4	28123	compactas
552	4	28123	crepe
553	4	28123	hidreto
554	4	28123	lascas
555	4	28123	vidreira
556	35	27710	eua
557	7	27705	biodegradáveis
558	7	27705	móido
559	6	27564	flandres
560	9	27562	telhas
561	5	27506	biogás
562	5	27506	operacionalizar
563	5	27506	prensada
564	8	27492	recuperado
565	25	27417	necessária

566	7	27302	estanho
567	12	27146	bacia
568	10	27123	elétricos
569	6	26997	automobilística
570	6	26997	php
571	6	26997	poupar
572	19	26976	site
573	3	26786	abeaço
574	3	26786	abir
575	3	26786	abrapex
576	3	26786	aciarias
577	3	26786	arcondicionado
578	3	26786	acv
579	3	26786	ampolas
580	3	26786	apliquim
581	3	26786	aramado
582	3	26786	banjo
583	3	26786	biorremediação
584	3	26786	bracelpa
585	3	26786	cabeçote
586	3	26786	carbonatada
587	3	26786	ciclonatural
588	3	26786	coletaseletiva
589	3	26786	cooperando
590	3	26786	deville
591	3	26786	ecoeficiente
592	3	26786	economize
593	3	26786	ecopontos
594	3	26786	enfardadas
595	3	26786	esterços
596	3	26786	estruque
597	3	26786	garagens
598	3	26786	garrações
599	3	26786	héroult
600	3	26786	iai
601	3	26786	isotônicos
602	3	26786	lingote
603	3	26786	maxiquim
604	3	26786	microcenários
605	3	26786	moveleira
606	3	26786	naoreciclaveis
607	3	26786	onduladas
608	3	26786	papeleiras
609	3	26786	parkesine
610	3	26786	pelbd
611	3	26786	picotadas
612	3	26786	polieteno
613	3	26786	prolata
614	3	26786	radiadores
615	3	26786	recicle
616	3	26786	reflorestadas
617	3	26786	refugos

618	3	26786	reprocessada
619	3	26786	rerrefinadoras
620	3	26786	retornáveis
621	3	26786	retornável
622	3	26786	saudecoletiva
623	3	26786	spaghetti
624	3	26786	sucateiro
625	3	26786	tachinhas
626	3	26786	triturado
627	3	26786	vidrarias
628	3	26786	vulcanizada
629	11	26710	evita
630	5	26603	nylon
631	5	26603	refinarias
632	11	26546	eletricidade
633	11	26546	proporciona
634	6	26458	recuperam
635	78	26351	através
636	8	26326	elos
637	4	26224	escova
638	4	26224	injeta
639	4	26224	mantas
640	4	26224	otimizando
641	4	26224	varal
642	10	26168	calcário
643	7	26167	difundir
644	7	26167	película
645	7	26167	sapato
646	16	26127	produzidas
647	6	25945	caldeiras
648	6	25945	esponjas
649	17	25931	carbono
650	5	25780	ici
651	5	25780	navegadores
652	8	25779	transformadas
653	17	25674	equivalente
654	33	25419	impacto
655	11	25293	fibra
656	8	25252	libera
657	93	25081	desenvolvimento
658	69	25030	importante
659	5	25024	inertes
660	5	25024	panelas
661	6	24985	geradora
662	13	24880	quilos
663	12	24806	parceiros
664	7	24800	encaminhadas
665	4	24717	carpetes
666	4	24717	contacto
667	4	24717	higiênico
668	4	24717	laminado
669	4	24717	odores

670	4	24717	tabor
671	11	24701	saiba
672	39	24596	utilização
673	6	24536	automotiva
674	6	24536	sucos
675	9	24490	limpo
676	18	24471	placas
677	19	24443	adequada
678	5	24324	inerte
679	7	24170	cilindro
680	11	24129	contribuindo
681	11	24129	transparência
682	6	24104	importadores
683	6	24104	vazias
684	9	24099	perfis
685	12	24050	ganhos
686	58	24039	you
687	19	24025	contribuir
688	21	23974	agentes
689	10	23927	geram
690	7	23867	chineses
691	5	23674	contaminando
692	5	23674	louças
693	5	23674	revestidos
694	13	23545	especificações
695	24	23544	setores
696	4	23465	brisas
697	4	23465	desagregado
698	4	23465	guardanapos
699	4	23465	somarmos
700	10	23299	equivale
701	6	23289	superam
702	12	23208	reduzindo
703	16	23170	camadas
704	10	23145	censo
705	449	23075	ou
706	5	23067	sulfúrico
707	11	22911	plástica
708	7	22725	metálico
709	13	22708	fusão
710	8	22680	selo
711	5	22498	artesanais
712	7	22456	dispostas
713	7	22456	manuseio
714	4	22394	celulóide
715	4	22394	divisórias
716	4	22394	granulado
717	4	22394	hélices
718	4	22394	linho
719	4	22394	prefira
720	4	22394	retire
721	23	22358	venda

722	3	22311	alumino
723	3	22311	atoxidade
724	3	22311	atóxico
725	3	22311	derretidas
726	3	22311	evitamos
727	3	22311	floco
728	3	22311	higiênicos
729	3	22311	polivinila
730	38	22283	ações
731	8	22262	pedras
732	5	21961	espere
733	5	21961	lavado
734	5	21961	rascunhos
735	5	21961	sujo
736	9	21958	sódio
737	10	21833	vendidos
738	35	21749	custo
739	7	21684	limpas
740	23	21541	transformação
741	4	21456	centígrados
742	4	21456	incineradores
743	4	21456	rotor
744	5	21455	coletores
745	7	21437	cia
746	13	21428	decomposição
747	19	21373	produz
748	6	21165	abastecer
749	17	21058	destino
750	7	20960	grânulos
751	16	20890	iniciativas
752	13	20869	cooperativa
753	6	20849	espuma
754	9	20846	ecológica
755	15	20721	org
756	8	20714	esgotos
757	25	20701	idades
758	9	20694	carvão
759	4	20624	balde
760	4	20624	estocado
761	4	20624	fardo
762	4	20624	molhada
763	4	20624	seringas
764	6	20543	sujeira
765	5	20520	odor
766	5	20520	películas
767	15	20493	excesso
768	7	20278	litro
769	6	20245	conserva
770	6	20245	descartado
771	3	20102	acabamentos
772	3	20102	adesiva
773	3	20102	bijuterias

774	3	20102	celulósicas
775	3	20102	conservantes
776	3	20102	desligue
777	3	20102	forros
778	3	20102	garde
779	3	20102	maleabilidade
780	3	20102	moer
781	3	20102	picotados
782	3	20102	playcenter
783	3	20102	recycling
784	3	20102	silo
785	5	20087	aproveitado
786	5	20087	prateleira
787	13	20064	gerando
788	13	20064	produtiva
789	14	19993	associações
790	4	19875	mangueiras
791	7	19845	colorido
792	7	19845	moinho
793	54	19813	construção
794	44	19752	processos
795	13	19719	químico
796	5	19673	escasso
797	5	19673	vedação
798	7	19635	armazenar
799	7	19635	pigmentos
800	61	19621	milhões
801	9	19539	coletadas
802	7	19428	caminhões
803	7	19428	escritórios
804	8	19336	abnt
805	8	19336	fitas
806	5	19278	processam
807	5	19278	tóxicas
808	7	19225	cerveja
809	4	19195	comestíveis
810	4	19195	escorrer
811	4	19195	fricção
812	4	19195	geladeiras
813	4	19195	transformadoras
814	8	19174	pano
815	10	19166	grãos
816	10	19050	anualmente
817	11	19019	móveis
818	8	19014	polímero
819	5	18900	panela
820	18	18883	índices
821	6	18875	rígidos
822	14	18843	comprar
823	616	18833	por
824	78	18717	país
825	13	18645	mercados

826	33	18644	utilizados
827	7	18637	beneficiamento
828	45	18590	média
829	4	18572	aquecidos
830	4	18572	feltro
831	4	18572	moídos
832	4	18572	nafta
833	4	18572	refino
834	3	18538	babelônios
835	3	18538	bambu
836	3	18538	conduítes
837	3	18538	iogurtes
838	3	18538	jogá
839	3	18538	mineradas
840	3	18538	pak
841	3	18538	pergaminho
842	3	18538	regenerado
843	5	18537	geladeira
844	64	18530	qualidade
845	12	18467	facilidade
846	7	18448	algodão
847	13	18407	gerada
848	14	18404	umidade
849	6	18375	armazenado
850	23	18339	europa
851	16	18236	impactos
852	5	18189	download
853	5	18189	egípcios
854	30	18174	composição
855	56	18168	após
856	249	18111	sua
857	12	18038	articulação
858	70	18012	programa
859	4	17998	facas
860	4	17998	repentino
861	6	17899	devolver
862	6	17899	recipiente
863	7	17899	separa
864	2	17857	abastecedor
865	2	17857	acesse
866	2	17857	achetudoeregiao
867	2	17857	aco
868	2	17857	adquirirmos
869	2	17857	aerador
870	2	17857	aísa
871	2	17857	amaciante
872	2	17857	amassadas
873	2	17857	ambev
874	2	17857	apme
875	2	17857	atiradas
876	2	17857	automobilísticos
877	2	17857	baldios

878	2	17857	batedeira
879	2	17857	chaparia
880	2	17857	cleaners
881	2	17857	clips
882	2	17857	compotas
883	2	17857	containers
884	2	17857	contâiners
885	2	17857	coopamare
886	2	17857	copinho
887	2	17857	corrugado
888	2	17857	crepom
889	2	17857	criolita
890	2	17857	custosas
891	2	17857	desfazemos
892	2	17857	desvulcanizantes
893	2	17857	desvulcanização
894	2	17857	dickson
895	2	17857	docs
896	2	17857	duracell
897	2	17857	ecomarapendi
898	2	17857	eletroímãs
899	2	17857	elquisa
900	2	17857	empilhe
901	2	17857	enfardamento
902	2	17857	engarrafadores
903	2	17857	ensacados
904	2	17857	entintada
905	2	17857	envasada
906	2	17857	envaze
907	2	17857	enxagüe
908	2	17857	enxergá
909	2	17857	esprema
910	2	17857	fabrik
911	2	17857	fumaças
912	2	17857	fumegante
913	2	17857	garrafeiras
914	2	17857	gaseificação
915	2	17857	glassine
916	2	17857	guache
917	2	17857	hyatt
918	2	17857	iar
919	2	17857	ibs
920	2	17857	impositiva
921	2	17857	intercale
922	2	17857	irmp
923	2	17857	isri
924	2	17857	i°
925	2	17857	kastrup
926	2	17857	kilo
927	2	17857	lithium
928	2	17857	lnk
929	2	17857	maioneses

930	2	17857	mercuriais
931	2	17857	metacrílico
932	2	17857	metalizadas
933	2	17857	microisís
934	2	17857	monocamada
935	2	17857	mvc
936	2	17857	nilópolis
937	2	17857	novelis
938	2	17857	oscip
939	2	17857	panetone
940	2	17857	parafinado
941	2	17857	persianas
942	2	17857	pev
943	2	17857	plastificado
944	2	17857	pregadores
945	2	17857	pressione
946	2	17857	problemão
947	2	17857	radiais
948	2	17857	rayovac
949	2	17857	reamur
950	2	17857	redtenbachen
951	2	17857	refletivo
952	2	17857	reingelmann
953	2	17857	reprocessado
954	2	17857	retalho
955	2	17857	sacaria
956	2	17857	scrap
957	2	17857	segregada
958	2	17857	sindirrefino
959	2	17857	socioambientais
960	2	17857	sofás
961	2	17857	solado
962	2	17857	sírios
963	2	17857	taste
964	2	17857	translúcidos
965	2	17857	tritutados
966	2	17857	usgbc
967	2	17857	varta
968	2	17857	vazadouros
969	2	17857	vergalhões
970	2	17857	vernizes
971	2	17857	whinfield
972	2	17857	ímã
973	5	17854	aquecido
974	9	17733	usinas
975	40	17683	programas
976	107	17668	além
977	25	17569	utilizadas
978	5	17531	asfalto
979	5	17531	inventado
980	5	17531	sintética
981	4	17466	graxa

982	12	17458	design
983	12	17458	gases
984	6	17444	minimização
985	6	17444	parafina
986	9	17377	contida
987	7	17377	prensagem
988	12	17377	reduz
989	7	17377	transformada
990	3	17318	bronze
991	3	17318	descontaminar
992	3	17318	molduras
993	3	17318	pneumático
994	3	17318	radiográficas
995	3	17318	recolhida
996	3	17318	separadores
997	3	17318	terciária
998	3	17318	torneiras
999	3	17318	vinil
1000	6	17225	contenham
1001	5	17220	bolhas
1002	5	17220	destilação
1003	14	17157	somos
1004	6	17010	posturas
1005	4	16970	canetas
1006	4	16970	enviam
1007	4	16970	mates
1008	4	16970	miniatura
1009	4	16970	opaco
1010	4	16970	quebradas
1011	4	16970	têxteis
1012	6	16800	conheça
1013	10	16734	confeção
1014	12	16668	mineração
1015	5	16630	contaminado
1016	5	16630	sabão
1017	8	16580	proveniente
1018	13	16559	facilitar
1019	7	16558	limpa
1020	4	16505	carcaças
1021	4	16505	circulou
1022	4	16505	decompor
1023	4	16505	domiciliares
1024	4	16505	louça
1025	4	16505	versatilidade
1026	8	16452	casca
1027	6	16393	impressos
1028	5	16350	confira
1029	5	16350	misturados
1030	5	16350	refrigeração
1031	3	16317	conservas
1032	3	16317	impermeabilidade
1033	3	16317	margarina

1034	3	16317	mexa
1035	3	16317	newsletter
1036	3	16317	plastificantes
1037	3	16317	termofixos
1038	8	16199	bebida
1039	11	16194	gerações
1040	10	16166	futuras
1041	5	16078	infinitamente
1042	4	16069	borra
1043	4	16069	importador
1044	4	16069	mecanicamente
1045	4	16069	propaga
1046	4	16069	separá
1047	7	15946	ecológicos
1048	9	15943	velhos
1049	20	15932	perda
1050	6	15813	moldagem
1051	6	15813	tanque
1052	10	15802	aproveitamento
1053	16	15731	vantagem
1054	20	15719	etapas
1055	8	15709	térmico
1056	17	15688	milhão
1057	16	15679	comercialização
1058	4	15658	aroma
1059	4	15658	húmus
1060	4	15658	rasa
1061	6	15627	industrializados
1062	13	15616	produtivo
1063	14	15577	básico
1064	5	15560	retornou
1065	5	15560	transmissores
1066	7	15509	conexões