FROMM, Guilherme . FERRAMENTAS DE ANÁLISE LEXICAL COMPUTADORIZADAS: UMA APLICAÇÃO PRÁTICA . Revista Factus, Taboão da Serra, v. 1, n. 3, p. 153-164, 2004.

FERRAMENTAS DE ANÁLISE LEXICAL COMPUTADORIZADAS:

UMA APLICAÇÃO PRÁTICA

Guilherme Fromm[1]

RESUMO: muitos estudos trabalham com a descrição e a comparação de ferramentas de análise lexical. Nosso estudo pretende, por outro lado, apresentar um exemplo concreto de como trabalhar com esses programas: começamos com um *corpus* de especialidade e dele tentamos, através das ferramentas, retirar os elementos que possam preencher uma ficha terminológica e, posteriormente, montar o verbete de um vocabulário técnico.

Palavras-chave: Ferramentas de Análise Lexical, Lingüística do Corpus, Terminologia, Terminografia.

ABSTRCT: many studies deal with the description and the comparison of lexical analysis tools. Our study intends, otherwise, to work with a real example and how the programs fit according to our necessities: we start with a technical corpus and from it we try, with the tools, to extract elements to fulfill our terminological chart and later, build our technical vocabulary entry.

Keywords: Lexical Analysis Tools, Corpus Linguistic, Terminology, Terminography

Apresentação

Quando pretendemos analisar uma grande quantidade de textos, já há algum tempo, dispomos de programas de tratamento lexical. Esses programas facilitam a vida do pesquisador: caso não existissem, levaria anos para tabular e trabalhar com os dados obtidos em seu *corpus*.

Já existem estudos comparativos entre diversos programas (como em LUCCA e NUNES, 2002), mostrando seus pontos fortes e fracos, suas interfaces e ferramentas

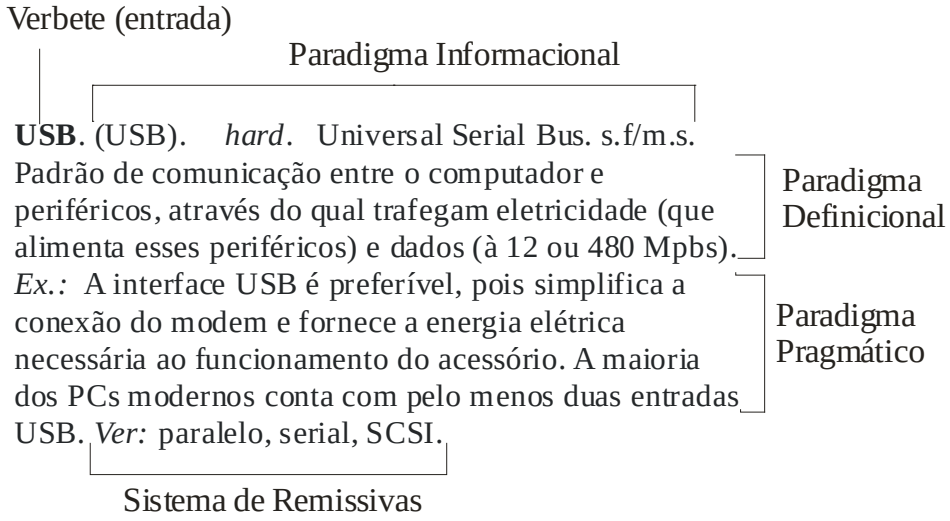
[1] Especialista em Tradução, mestre em Lingüística e doutorando em Língua e Literaturas Inglesa e Norte-Americana. Professor da UNIBAN

.

disponíveis. Deixando a curiosidade de lado, especialmente quanto à questão *qual programa é melhor?*, procuramos, neste trabalho, comparar alguns programas e responder uma pergunta por nós formulada: *qual desses programas é o melhor para determinadas necessidades?*

A necessidade proposta

O objetivo básico proposto, para a aplicação desses programas, é fazer um levantamento de palavras, conceituações e exemplos para a criação de um vocabulário (dentro da diferenciação entre dicionários, vocabulários e glossários, proposta por BARBOSA, 2001) temático monolíngüe. Os programas serão usados para fazer a escolha dentro do *corpus* proposto, levantando sua nomenclatura (baseada, inicialmente, no critério de freqüência), seus termos, suas definições e exemplos. Tudo isso servirá como fonte para o preenchimento de fichas terminológicas (como nos exemplos em FROMM, 2002). Como exemplo, podemos citar um verbete finalizado[2] que uma ficha terminológica, construída a partir de um corpus monolíngüe, pode fornecer:



Seleção prévia de programas

Antes de começarmos a analisar quais programas se encaixariam na solução da problemática apresentada pelo estudo, decidimos alguns parâmetros para facilitar nossa pesquisa. Esses programas devem:

[2] Explicação dos paradigmas em FROMM (p.144, 2004).

a. Ler documentos nos formatos .txt, .doc e html: a maioria das bases coletadas de *corpora* são fornecidas nesses formatos;
b. ter uma interface gráfica no sistema operacional Windows, para facilitar as consultas, e que trabalhe no sistema de computador mais encontrado no Brasil: o PC;
c. trabalhar com um leiaute que facilite o intercâmbio de informações entre as ferramentas, além de ser visualmente claro;
d. não apresentar um limite na quantidade de textos analisados, já que as tendências nas áreas de Lingüística de *Corpus* e Terminologia apontam para o uso de *corpora* cada vez maiores;
e. apresentar uma ferramenta para contagem de freqüência das palavras e também para a visualização das mesmas, já que a nossa macroestrutura é baseada nos termos mais comuns encontrados em cada *corpus*;
f. demonstrar algumas estatísticas básicas quando da contagem de palavras, para que possamos ter parâmetros básicos para uma pré-análise;
g. contar com uma ferramenta de análise do *corpus*, já que o mesmo pode não ser homogêneo, e, para um resultado mais preciso, é preciso que ele seja balanceado;
h. disponibilizar uma ferramenta de lematização: muitos termos mostram, dentro do *corpus*, variações quanto às desinências (verbos) e derivações (substantivos e adjetivos), além de grau, número, etc. Essas variações devem ser agrupadas em uma só entrada no vocabulário;
i. apresentar uma ferramenta de concordância, onde possamos visualizar várias linhas com palavras escolhidas e daí tirar exemplificações e definições para os nossos verbetes;
j. ter uma ferramenta de seleção dos termos correspondentes à área: uma simples lista de freqüência não basta para selecionarmos termos de uma área; o programa deve, sim, apresentar uma ordem de freqüência (ainda mais porque é difícil elaborar uma obra com <u>todos</u> os termos existentes de uma área), mas uma freqüência de termos pertinentes àquela área, que se destaquem quanto ao léxico geral da língua.

O corpus utilizado

Trabalharemos aqui com um *corpus* de tamanho reduzido[3], de uma única área, monolíngüe. Esse *corpus* se constitui de textos levantados pelos alunos do Curso de Especialização em Tradução Inglês/Português da FFLCH/USP, coletados e gentilmente cedidos pelo Projeto Comet/USP.

Ele é constituído de dezoito textos sobre o uso de impressoras, retirados de sites ou manuais de instrução fornecidos pelos fabricantes. Todos estão em português. O maior texto possui 2.036 palavras e o menor, 111 palavras[4]. O formato de todos é .txt.

Os programas disponíveis

Existem vários programas disponíveis na Internet. Alguns são só macros para serem trabalhadas no Word ou Excel (Microsoft), como o KWIC[5], outros só funcionam no sistema operacional DOS, como o TACT ou DICTGEN. Muitos têm um leiaute pobre e confuso, como o Range and Word. Tendo em vista a lista de seleção prévia das nossas necessidades, esses programas foram desconsiderados. Trabalharemos somente com programas que apresentem várias ferramentas, dentro dos sistema operacional Windows e que tenham uma visualização relativamente clara: STABLEX, WordSmith Tools, Monoconc e Concordance. Um estudo detalhado de todos os programas acima citados e outros (não incluindo o STABLEX), em uma análise contrastiva, pode ser encontrando em LUCCA e NUNES (2002).

Tendo em vista a nossa necessidade, apresentaremos a seguir uma análise de cada programa estudado e as respectivas contribuições que eles podem fornecer para a elaboração do nosso vocabulário.

STABLEX

O programa, desenvolvido por André Camlong, apresenta várias ferramentas para análise lexical: algumas dentro do próprio programa (listagem de palavras, leitura dos textos, etc.) e outras, as principais, dentro de uma macro a ser executada no programa EXCEL. O objetivo final do STABLEX não é preparar um dicionário, mas sim, através de um cálculo estatístico-paramétrico, analisar o discurso por detrás de um texto. Muitas ferramentas, porém, nos são úteis.

---

[3] Para SARDINHA (1999), esse *corpus* seria considerado pequeno.
[4] Contagem feita pelo Wordsmith Tools, ferramenta Wordlist.
[5] Apesar de não utilizarmos esse programa aqui, ele foi um dos primeiros a trabalhar com a concordância de palavras em contextos diferentes (Key Words In Context) e ainda funciona como modelo para outros.

Como quase todos os programas, o STABLEX apresenta uma listagem de palavras e suas freqüências. Diferente de outros, porém, ele apresenta a distribuição da freqüência texto a texto. Esse simples recurso já pode nos dar algumas idéias da constituição dos textos.

Uma segunda listagem de freqüência mostra a quantidade de palavras que apresentam a mesma distribuição na totalidade dos textos. Podemos verificar, em uma das telas do programa, por exemplo, que as palavras *não*, *opcional* e *rede* apresentam a mesma freqüência de uso no conjunto do *corpus* (35 vezes). Nessa segunda listagem, essas palavras aparecem agrupadas sob a mesma linha (25), que indica haver 3 palavras com freqüência de 35 aparições no *corpus*, perfazendo um total de 105 palavras.

Uma terceira tabela mostra a relação entre as palavras agrupadas por freqüência e o seu peso lexical no *corpus* de estudo e em cada texto. Quando o valor for acima de dois[6], significa que essas palavras têm um uso privilegiado (vocabulário preferencial) por parte do escritor; quando oscilam entre +2 e -2, indicam um uso normal (vocabulário básico); quando for maior que -2, as palavras em destaque têm uso negligenciado (vocabulário diferencial). O teste χ^2 de Fisher indica o grau de normalidade na distribuição lexical, dentro de cada texto, em relação ao conjunto: quanto mais próximo de 0, menos desvios o texto apresenta. A macro apresenta, ainda, vários tipos de tabelas configuráveis, como um gráfico de comparação entre os graus de desvios por textos do *corpus*.

O gráfico da macro indica, por exemplo, que o texto 17 extrapola todos os outros em quantidade de vocabulário preferencial, enquanto os textos 15, 16 e 18 destacam-se através do vocabulário diferencial. O teste do χ^2, porém, nos indica que todos os textos estão dentro do grau de normalidade, tratando-se aqui de um *corpus* equilibrado.

Vantagens

- O teste do $\chi 2$ nos indica como equalizar o *corpus*, tornando-o mais homogêneo. Um *corpus* extremamente heterogêneo demanda um tempo maior de pesquisa e acaba mostrando muitos hápax de áreas não afins, havendo a necessidade, por parte do pesquisador, de "peneirar" os dados. Dentre os programas estudados, é o único que apresenta essa possibilidade (ainda que, acreditamos, ela não tenha sido pensada como tal);

[6] O valor, bastante discutido na base teórica do programa (ZAPAROLLI, 2002), representa um grau de desvio aceitável em relação ao padrão, que é O.

- a metodologia de análise estatística é exaustivamente trabalhada pelo autor;
- quando já feita a homogeneização do *corpus*, o vocabulário preferencial nos apresenta uma pista sobre os termos específicos daquela área.

Desvantagens

- A construção da lista de palavras é muito demorada: se o *corpus* de estudo for extenso e constituído de muitos textos, o processamento se torna muito lento, mesmo em computadores mais potentes;
- o programa aceita uma quantidade pequena de textos (100), impedindo que haja representatividade em termos de tamanho;
- é difícil relacionar as palavras e as análises nelas feitas: as tabelas não são claras por não apresentarem as palavras ou conjuntos de palavras a que se referem, havendo a necessidade constante de trocarmos de tabela em busca de determinada palavra;
- retirar o contexto ao redor de cada palavra para a construção da ficha terminológica exige o trabalho de voltar ao programa e pedir para que ele ache, de acordo com a palavra requerida, o texto de onde ela foi tirada;
- a lematização é feita através de um processo manual de copiar e colar entre as tabelas, o que exige muito tempo por parte do pesquisador.

WordSmith Tools[7]

O programa de Mike Scott apresenta-se como um canivete suíço de análise lexical. É constituído de várias ferramentas, mas três são as principais:

WordList

Faz uma listagem das palavras e apresenta, em uma mesma janela (com cinco abas), diferentes tipos de análise:

- freqüência: listagem de palavras em ordem de freqüência no conjunto do *corpus*;
- listagem alfabética das palavras e suas freqüências;
- estatísticas: apresenta várias estatísticas, como a relação entre tokens e types[8], simples e através de cálculo estatístico;

[7] Versão 4..

- nomes dos arquivos;
- notas extras.

Keywords (Palavras-Chave)

A ferramenta Keywords elabora uma listagem de palavras consideradas chave dentro de um *corpus*. Assim como o STABLEX, essa listagem apresenta as palavras de uso privilegiado (em preto) e aquelas de uso comum (vermelho).

Para a elaboração da mesma, é necessário um outro *corpus*, de exclusão. Esse *corpus* de exclusão deve ser representativo em relação ao léxico geral da língua (leia-se: ele deve ser, de um modo geral, dez vezes maior que o *corpus* analisado) ou em relação ao léxico especializado daquela área.

Concord (Concordâncias)

O programa elabora, a partir da ferramenta Keyword ou de uma busca por uma palavra qualquer (digitada), uma lista de todas as linhas onde ela aparece em todo o *corpus*. Como muitas outras, essa ferramenta imita o leiaute do programa KWIC. Ela é ideal para mostrar regências e convencionalidades (TAGNIN, 1989) que a palavra escolhida (e centralizada no meio da tela) pode apresentar: colocações (combinabilidade dos elementos), binômios, expressões convencionais, expressões idiomáticas, etc. Além disso, no nosso caso, pode fornecer pistas para a montagem da definição da palavra dentro da ficha terminológica, já que muitas vezes essas linhas apresentam aquilo que AUBERT (1996) chama de contextos explicativos e definitórios.

Vantagens

- Rápido, trabalha a construção da lista de palavras a uma taxa de 3 milhões de palavras por minuto;
- visualização clara;
- não há limite de tamanho do texto ou quantidade de textos;
- a construção de sentidos é facilitada pelas concordâncias fornecidas para cada palavra a partir de uma lista de palavras-chave;
- o processo de lematização é simples:
- é o programa que apresenta o maior número de ferramentas.

---

[8] "Na língua inglesa os estatísticos do léxico costumam opor o *token* (ocorrência no texto) ao *type* (lexema referido pela ocorrência formal)." (BIDERMAN, 2001, p.167)

Desvantagens

- O balanceamento do *corpus* tem de ser feito previamente;
- o autor pouco trabalha com a metodologia estatística abordada pelo programa;
- não há como pegar mais de uma linha de texto nas concordâncias, o que pode dificultar a identificação de contextos explicativos;
- para trabalhar com a ferramenta Keywords, é necessário um *corpus* de exclusão, ou seja, significa gastar mais tempo preparando outro *corpus* (ou, no mínimo, tentando encontrar um *corpus* maior).

Concordance

Conforme o nome já explicita, esse programa é basicamente um concordanceador. Ele apresenta também uma listagem de palavras, que pode ser selecionada pela ordem alfabética ou de freqüência. Para mostrar as concordâncias de qualquer palavra, basta selecioná-la na aba esquerda, onde está a listagem, e ela será mostrada na aba direita.

Para visualizar o texto de onde foi tirada uma linha de concordância, basta dar um duplo clique sobre a palavra centralizada. Uma nova janela se abrirá, com o texto correspondente. A ferramenta de lematização é bastante simples, mas deve ser alimentada com todas as palavras base e suas derivadas para que o programa faça a junção das mesmas.

Vantagens

- Rapidez;
- visualização bastante clara;
- ao clicar na concordância, obtemos uma janela de texto, da qual podemos tirar os exemplos;
- boa ferramenta de lematização.

Desvantagens

- Poucas ferramentas;
- metodologia de análise estatística não discutida pelo autor.

Monoconc

O programa, de um modo geral, assemelha-se em uso e quantidade de ferramentas ao Concordance. Assim como todos os outros, apresenta uma listagem das palavras do *corpus* em ordem a ser escolhida pelo consulente: alfabética ou por freqüência.

Concordance

O concordanceador do Monoconc difere-se dos outros programas, basicamente, por apresentar, ao clicarmos na linha desejada, o texto ao qual ela pertence na metade superior da janela. Assim como os outros, podemos selecionar quantas palavras queremos destacar ao redor daquela que serve de base: as palavras em vermelho indicam as colocações mais comuns, dentro de um arco de até duas palavras para a esquerda ou direita, que se associam ao termo desejado.

Distribuição da palavra no corpus

Dentre algumas estatísticas que o programa apresenta, a distribuição da palavra dentro do *corpus*, por texto, é uma delas. Existe também a possibilidade de exibição da distribuição da palavra dentro de cada texto, no parágrafo em que ela se encontra e outras possibilidades de combinação de análise.

Vantagens

- Rapidez na análise;
- Facilidade na visualização dos textos, tornando o trabalho de preenchimento de fichas terminológicas mais eficiente.

Desvantagens

- poucas ferramentas: basicamente listagem de palavras e concordâncias;

Considerações Finais

Não existe, geralmente, um programa de análise lexical que atenda a todas as necessidades de um pesquisador. Para tanto, o pesquisador teria que criar o seu próprio programa, o que exige conhecimento aprofundado de diferentes sistemas operacionais e suas vantagens e desvantagens e ainda conhecimento de programação.

Essa questão acaba por repetir-se na nossa análise. Precisaríamos usar, para alcançar o objetivo proposto, três programas:

- O STABLEX para equalizar o *corpus* e verificar o vocabulário preferencial;
- o WORDSMITH para levantar as palavras-chave e visualizar algumas estatísticas;
- o CONCORDANCE ou o MONOCONC para levantar as concordâncias e os exemplos.

Para termos uma maior precisão, poderíamos fazer uma análise contrastiva entre as palavras apresentadas como vocabulário preferencial no STABLEX e como palavras-chave no WordSmith Tools. Embora os critérios de análise estatística dos programas sejam diferentes, eles, teoricamente, deveriam apresentar uma listagem similar. Essa análise contrastiva nos daria maior certeza para o levantamento da terminologia de uma área. O preenchimento de fichas terminológicas, feito a partir das concordâncias verificadas, seria o passo final do trabalho.

A questão da lematização, entretanto, acaba não sendo resolvida por nenhum programa. Aqueles que têm as melhores ferramentas para esse tipo de análise, como o WordSmith e o Concordance, exigem um trabalho manual muito grande por parte do pesquisador, o que pode inviabilizar o trabalho se o *corpus* for muito grande. Ajudaria muito se os autores vendessem, além dos programas, bancos de dados de palavras que podem ser lematizadas.

Bibliografia

AUBERT, F. H. **Introdução à metodologia da pesquisa terminológica bilíngüe**. São Paulo: Humanitas Publicações-FFLCH-USP, 1996.

BARBOSA, M. A. Dicionário, vocabulário, glossário: concepções. In: ALVES, I. M. (Org.). **A constituição da normalização terminológica no Brasil**. 2 ed. São Paulo: FFLCH/CITRAT, 2001.

BARLOW, M. **MonoConc Pro**. V. 2.2. Houston: Athelstan, 2002.

BIDERMANN, M.T.C. **Teoria Lingüística**. 2. ed. São Paulo: Martins Fontes, 2001.

CAMLONG, A. **Stablex**. Paris: e/a, 2003.

FROMM, G. **Proposta para um modelo de glossário de informática para tradutores**. Dissertação de Mestrado. São Paulo: FFLCH/USP, 2002.

_____ . Obras lexicográficas e terminológicas: definições. In: **Revista FacTuS**. Taboão da Serra: FTS, nº 2, 2004.

LUCCA, J.L. de & NUNES, M.G.V. **Breve estudo sobre requisitos de ferramentas de software para construção de dicionários**. São Carlos: NILC/ICMC/USP, 2002.

SARDINHA, T. B. **O que é um corpus representativo?** São Paulo: Inédito, 1999.

SCOTT, M. **WordSmith Tools**. v 4. Oxford: OUP, 2004.

TAGNIN, S. O. **Expressões idiomáticas e convencionais**. São Paulo: Ática, 1989.

WATT, R.J.C. **Concordance**. v 3. 2002.

ZAPPAROLI, Z. M. & CAMLONG, A. **Do Léxico ao Discurso pela Informática**. São Paulo: Edusp/FAPESP, 2002.