Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Breve estudo sobre requisitos de ferramentas de software para construção de dicionários

J. L. De Lucca

Maria das Graças Volpe Nunes

NILC-TR-02-21

Novembro 2002

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

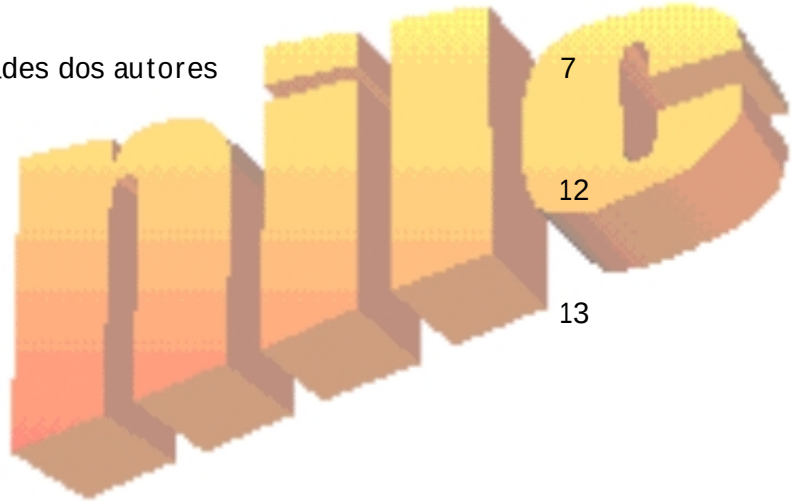
Resumo

O presente relatório técnico tem como objetivo apresentar o resultado de uma pesquisa aleatória, online, em fóruns internacionais de lingüística sobre que ferramentas devem conter um software para fazer dicionários, bem como trazer à luz os resultados de pesquisas voltadas ao usuário de dicionários e o que estes esperam encontrar em uma obra de referência, particularmente os dicionários monolíngües.

Os resultados advindos das pesquisas servirão para nortear o projeto de construção de um software que auxilie o lexicógrafo na compilação de dicionários e vocabulários.

As sugestões recebidas são muito amplas e, se postas em prática, distam muito das atuais ferramentas disponíveis no mercado. A incorporação de todas as ferramentas sugeridas, além das necessidades apontadas por usuários de dicionários monolíngües, proporcionaria um ambiente muito favorável ao lexicógrafo, independentemente de sua língua materna e objetivo da obra, se dicionário monolíngüe ou bilíngüe, vocabulário monolíngüe, bilíngüe ou multilíngüe, com ou sem definições.

Índice



Breve estudo sobre requisitos de ferramentas de software para construção de dicionários

1. Introdução

Quando pensamos em dicionários vêm-nos à mente o quão difícil deve ser a tarefa de elaborá-los. Esta tarefa tão difícil começa com uma escolha simples: o que colocar e o que não colocar em um dicionário. Como afirma HAAS (1962) "A good dictionary is not one only which has all the information who needs; it is also one in which you can find the information you are looking for – preferably in the very first place you look".

Os dicionários, diferentemente da literatura, nascem como resultado do aproveitamento de outros dicionários, com algumas poucas modificações, como supressão, correção ou ampliação da macro e microestruturas. Como bem disse SILVEIRA (1992:49-50) "Como nós sabemos muito bem, os maiores autores de dicionários no mundo são duas senhoras: a tesoura e a goma arábica. Todos os dicionários começam assim, quer dizer, recortando de outros dicionários e acrescentando coisinhas". No Brasil, atualmente, há "autores" de minidicionários especializados nesta função. Em cada época há sempre os dicionários-espelho e, se analisarmos as definições dos minis e pequenos dicionários da língua portuguesa, verificaremos a existência de uma semelhança extraordinária entre eles.

Não estamos pensando neste tipo de autor ao projetarmos um software para auxiliar o lexicógrafo na confecção de dicionários e vocabulários. Pensamos em pesquisadores dispostos a levantar um corpus e, a partir deste, criar seu próprio dicionário.

É possível compilar um bom dicionário partindo, não de outros dicionários, mas de um corpus representativo da língua. Basta usar como referência a lista de palavras mais freqüentes em nosso corpus, aliadas a outras não tão freqüentes, mas escolhidas pelo nosso bom senso.

Um software para fazer dicionários deve ter embutidos muitos dos princípios que norteiam a compilação de um dicionário tradicional, em formato papel, mas não os vícios da colagem.

O principal objetivo de um dicionário ou um software para criação de dicionários é atender as expectativas do lexicógrafo e, ao mesmo tempo, do usuário a que se destina.

Para **BIBER** (1998:59-60), são mais proveitosos os programs de uso particular para fazer concordanceamento, pois "Such programs can be designed to achieve higher accuracy in the search, the output can be tailored exactly for individual preferences". Em outras palavras, muitas vezes um programa "caseiro" resolve muito mais do que um software preparado para fazer uma enormidade de coisas... de que você não precisa!

Quando escrevemos programas que atendem nossas necessidades e pecualiridades, é possível incluir operações inexistentes em softwares prontos. Cada trabalho tem características em comum, mas muitos traços são particulares, não atendendo a todos os usuários.

Como afirma **BIBER** (1998:255-256) "Another advantage of writing your own programs is that you can do many analyses more quickly and accurately" e mais: "When you write your own programs, there is no limit to the size of the corpus that can be analized. The program will analize each text in the corpus sequentially, for as long as there are texts to analyse" .

2. As necessidades dos usuários

Algumas pesquisas já foram feitas para descobrir as necessidades dos usuários de dicionários. Estas pesquisas variaram tanto em número de entrevistados, como tipo de dicionário.

Os computadores podem simplificar o trabalho dos autores de dicionários mediante a criação de algumas ferramentas básicas de auxílio ao autor.

Antes de discutirmos o que deveria ter um software para fazer dicionários, do ponto de vista do autor, é, preciso conhecer o que os usuários de dicionários procuram em dicionários monolíngües da língua

Em De Lucca (2001b) é apresentada uma pesquisa feita na cidade de São Paulo, em 1999, junto a seis escolas, duas particulares e quatro públicas, totalizando 763 questionários respondidos, dos quais foi extraída uma amostra de 316, correspondentes ao primeiro ano do nível médio. Eram trinta e seis questões, obrigando ao usuário a escolher entre mais de 160 opções. Dentre as inúmeras questões, uma refere-se particularmente às necessidades dos alunos ao consultar um dicionário, mais especificamente um minidicionário: “O que os alunos procuram quando consultam um minidicionário de língua portuguesa”. Dos 763 questionários respondidos, com grande vantagem para as demais opções, definições obteve 61,3%. Em segundo lugar aparece ortografia com 47,4%. Vale salientar que esta questão era composta por 29 items. Sinônimos vêm logo depois com 39,2%, enquanto antônimos aparece com 16,8%. Em quarto lugar aparecem as gírias, com 30,7%. Sentido figurado aparece em quinto lugar com 23,7%, o campo de utilização da palavra (subject label), 21,2%. Alguns items, contudo, se analisados apenas pelos números, não parecem ser muitos procurados pelos alunos, como é o caso de divisão silábica (6,2%), aumentativos (6,3%), dominutivos (5,4%) e superlativos absolutois sintéticos (6,8%).

Estes resultados, confrontados com os de **BÉJOINT** (1981), com outro público-alvo e para outra finalidade, não-nativos da língua (francês), mostraram certa similaridade.

A pesquisa de Béjoint referia-se exclusivamente a dicionários monolíngües de inglês. Cada questionário tinha 21 perguntas que foram respondidas por 122 informantes. Trabalhou com estudantes do segundo e terceiro ano de inglês da Universidade de Lyon. Uma das questões era: "Que tipos de informação você mais procura em seu dicionário?". 87% dos alunos responderam que a informação mais procurada era o sentido da palavra; as informações sintáticas vieram em segundo lugar com 53%; logo após, sinônimos com 52%; ortografia e pronúncia, 25%; variações da linguagem, 19% e finalmente, etimologia com 5%.

Vemos, portanto, que algumas informações tornam-se indispensáveis em qualquer dicionário da língua: definições, sinônimos (desde que não substituam as definições por paráfrase), ortografia, pronúncia (também para os casos de dicionários bilíngües ou para estrangeiros), informações sintáticas, subject labels e vários níveis de linguagem (regionalismos, gírias e, porque não, palavras obscenas).

3. As necessidades dos autores

Segundo nossa pesquisa realizada em um fórum de língüistica na Internet, sem a formalidade e, às vêzes, até constrangimento de um questionário, formulamos a seguinte pergunta:

We would like to hearing from those who have experiences with the compiling dictionaries and vocabularies the following: WHAT you would like, would need, and

would hope of a Dictionary Creation Software. What type of tools would be essential for making dictionaries, vocabularies and other any type of reference work.

Houve mais de duas dúzias de informantes, mas apenas quinze deram respostas conclusivas. A validade da pesquisa está no nível do público, todos envolvidos com lexicografia, especialmente com lexicografia computacional.

Estas foram, em síntese, as sugestões recebidas

Entradas *on the fly*

Pronunciation

Corretor ortográfico compatível com Microsoft Word

Um grande Corpus, em uma ou mais línguas

Listas de freqüência lematizadas

Variantes geográficas de ortografia e pronúncia

Gênero das palavras

Colocações

Estatística

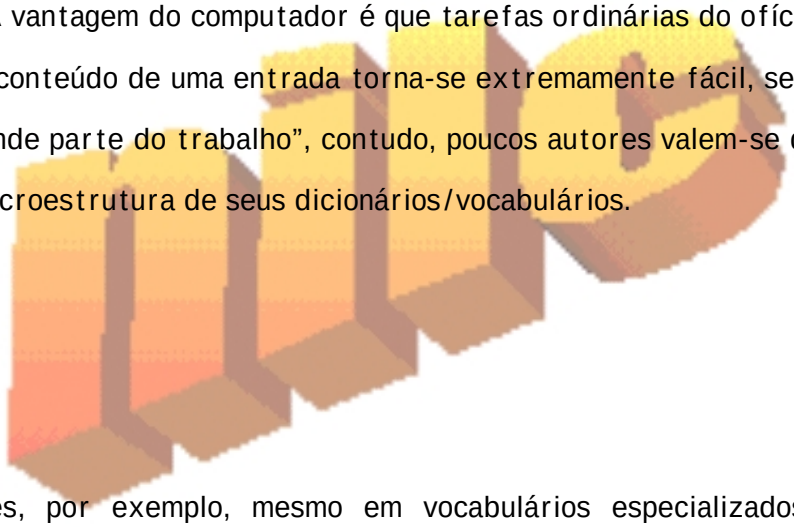
Contextos definitórios

Parece-nos que, tanto lexicógrafos, especialistas e usuários de dicionários mantêm algumas preocupações em comum, como: definições, sinônimos e ortografia. Algumas sugestões vêm ao encontro do que pensamos ser adequado e necessário colocar em um software dessa natureza. Um corpus ou vários corpora, sem dúvida, é indispensável, inclusive com referências cruzadas, explorando o corpus muito além de wordlists (listas de palavras), concordancers (concordanceadores) e combinations (combinações). A experiência também tem mostrado que um corpus com frases curtas adequa-se melhor à exemplificação, especialmente do ponto de vista editorial. Se for possível, e é, separar

os contextos ordinários dos definitórios e, além disso, serem curtos, será de extrema utilidade na confeção de colocações.

As estatísticas extraídas do corpus necessitam ir além da simples média aritmética. Particularmente interessa-nos ver médias aritméticas ao lado de desvio-padrão e comparação de wordlists (listas de palavras) através do Coeficiente de Pearson.

HAENSCH (1982:435-436) ponderou que as fontes de um lexicógrafo devem ser "um corpus representativo da língua ou área de especialidade, a alfabetização ou ordenação das entradas. A vantagem do computador é que tarefas ordinárias do ofício como trocar ou modificar o conteúdo de uma entrada torna-se extremamente fácil, sem necessidade de refazer grande parte do trabalho", contudo, poucos autores valem-se de corpus para confeção da macroestrutura de seus dicionários/vocabulários.



Muitos editores, por exemplo, mesmo em vocabulários especializados preferem a inclusão de informações gramaticais, como o gênero, por exemplo. Ficam muito felizes quando colocamos subject labels para diferenciar as áreas ou subáreas. Também aceitam remissivas e falsos cognatos. Definições, quando necessário, são bem vindas. A Fig 1. é um fragmento do dicionário De Lucca (2001), de vocabulário especializado, e que contém todas estas informações.

monopoly profit 402

e precio *m* monopolístico
i prezzo *m* di monopolio
p preço *m* de monopólio
d Monopolpreis *m*

4984 monopoly profit
(*microecon.*)
f profit *m* de monopole
e beneficio *m* monopolístico
i rendita *f* monopolistica
p lucro *m* monopolista
d Monopolgewinn *m*

4985 monopoly tax
(*microecon.*)
f taxe *f* de monopole; droit *m* de régie
e derecho *m* de monopolio; tasa *f* de monopolio
i tassa *f* di monopolio
p taxa *f* de monopólio
d Monopolsteuer *f*

* **monopoly trading situation** → **1632**

4986 monopsonistic position
(*microecon.*)
f position *f* monopsonistique
e posición *f* monopsonística
i posizione *f* monopsonistica
p posição *f* monopsonística
d monopsonistische Position *f*

4987 monopsony
(*microecon.*)
[Monopoly of buyers, a market situation in which there is only a single buyer.]
f monopsone *m*
e monopsonio *m*
i monopsonio *m*
p monopsônio *m*
d Monopson *n*; Angebotsmonopol *n*

4988 Monte Carlo method
(*math. econ.*)
f méthode *m* de Monte-Carlo
e método *m* Monte Carlo
i metodo *m* di Monte-Carlo
p método *m* Monte Carlo
d Montecarlo Methode *f*

4989 monthly average
(*math. econ.*)
f moyenne *f* mensuelle
e promedio *m* mensual
i media *f* mensile
p média *f* mensal
d Monatsdurchschnitt *m*

4990 moratory *adj*
(*macroecon.*)
[Temporary suspension of a payment, esp. public debt.]
f moratoire
e moratorio
i moratorio
p moratório
d moratorisch

* **mortality ratio** → **1651**

4991 mortgage *v*
(*fin. econ.*)
f hypothéquer
e hipotecar
i ipotecare
p hipotecar
d mit einer Hypothek belasten

4992 mortgage
(*law econ.*)
[A legal transfer of ownership, but not possession of property, from a debtor to a creditor.]
f hypothèque *f*
e hipoteca *f*
i ipoteca *f*
p hipoteca *f*
d Hypothek *f*

4993 mortgage bank
(*fin. econ.*)
f banque *f* de crédit hypothécaire; caisse *f* de crédit hypothécaire
e banco *m* hipotecario
i banca *f* di credito ipotecario
p banco *m* hipotecário
d Hypothekenbank *f*

* **mortgage bond** → **4995**

4994 mortgaged development
(*econ. d. and growth*)
f développement *m* hypothéqué
e desarrollo *m* hipotecado
i sviluppo *m* ipotecato
p desenvolvimento *m* hipotecado
d hypothekarische Entwicklung *f*

4995 mortgage debenture; mortgage bond
(*fin. econ.*)
[A secured bond upon a specific part of a company's assets.]
f obligation *f* hypothécaire; lettre *f* de gage
e abonaré *m* de hipoteca; obligación *f* hipotecaria; cédula *f* hipotecária
i obbligazione *f* ipotecaria; lettera *f* ipotecaria
p debênture *f* hipotecária; título *m* hipotecário; título *m* mobiliário garantido por hipoteca
d hypothekarisch gesicherte Schuldverschreibung *f*; Hypothekenobligation *f*; Pfandbrief *m*

Fig. 1 Elsevier's Economics Dictionary (sample page)

Assim, ao lado do gênero das entradas, poderiam ser incluídos alguns ou todos os items sugeridos, embora este modelo, com tantos traços lingüísticos, raramente tem sido seguido pelos autores.

Os contextos definitórios são extremamente úteis para compreensão das *headwords* (palavras-guia) e raramente são encontrados, portanto justifica-se plenamente a criação de uma rotina para separar estes contextos dos demais.

Um corretor ortográfico, como indicaram alguns, é uma ferramenta indispensável para o lexicógrafo e uma carência total nos atuais softwares de concordâncias.

Quantos aos dicionários monolíngües, há inúmeros modelos a seguir. O Chambers Schoold Dictionary (Fig. 2), por exemplo, é um modelo com inúmeros traços lingüísticos; os vocábulos de origem estrangeiros são acompanhados de sua pronúncia; as definições são geralmente por paráfrases; aqueles vocábulos que geram dúvidas freqüentes possuem um destaque em azul; alguns vocábulos vêm acompanhados de uma síntese histórica. Não há, contudo, collocations.

Fig. 2. Chambers Schoold Dictionary (sample page)

4. Conclusão

Do ponto de vista lingüístico, parece-nos haver algumas coisas indispensáveis de serem consideradas:

Um corpus parece ser o início de todo e qualquer trabalho que vise à criação de ferramentas computacionais para elaboração de dicionários. A partir de um corpus ou corpora será possível estabelecer uma lista das palavras e expressões mais freqüentes da língua. Estima-se que um corpus de 5.000.000 de palavras seja o mínimo aceitável para quem pretende construir uma lista confiável de cerca de 40.000 lemas.

Como conjugar as definições, os sinônimos/antônimos, corretor ortográfico, divisão silábica, pronúncia e outros traços requeridos, tanto por lexicógrafos como por usuários, será uma tarefa extremamente excitante, especialmente por ser indispensável a qualquer base lexicográfica ou terminológica, como vimos pelas pesquisas acima, tornando-se, como sugere o título de um famoso livro de lingüística, "The art and craft of lexicography".

5.Bibliografia

BÉJOINT, H. 1981. “The Foreign Student’s Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills.”Applied Linguistics 2:207-22.

BIBER, D. et all. 1998. Corpus Linguistics – Investigating Language Structure and Use. Cambridge. Cambridge University Press.

DE LUCCA, J. L. 2001a. Elsevier’s Economics Dictionary. Amsterdam. Elsevier Science.

DE LUCCA, J. L. 2001b. Minidicionários da língua portuguesa: análise léxico-estatística, crítica e contrastiva das macro e microestruturas e sugestão de modelo. (Tese de Doutoramento, FFLCH/USP, São Paulo)

HAAS, M. 1962. “What belongs in a bilingual dictionary”. In Householder and Saporta, 1962.

HAENSCH, G. et all. 1982. La Lexicografía – De la Lingüística teórica a la lexicografía práctica. Madrid. Editorial Gredos.

LANDAU, S. I. 1984. Dictionaries: The Art and Craft of Lexicography. New York. Scribner.

PRESSMAN, R.S**.** 1995. Engenharia de Software. São Paulo. Makron Books.

SILVEIRA, Ênio. 1991. Editando o editor. São Paulo. Edusp.