DICIONÁRIOS DE LÍNGUA CORRENTE: ALGUMAS CONSIDERAÇÕES

Regina Reis
INESC

Este artigo pretende chamar a atenção para alguns problemas surgidos durante a fase de criação de um dicionário de máquina a inserir num analisador sintáctico de língua portuguesa, nomeadamente a omissão de vocábulos nos dicionários de língua corrente e a falta de actualização de novos significados em formas já existentes.

O corpus

Para o teste do dicionário recorremos ao nosso corpus (cf. Santos (1992)), digitado manualmente, num total de 14873 ocorrências correspondentes a 4884 formas. Os textos que o constituem são heterogéneos já que não se pretendia fazer qualquer estudo temático de vocabulário, mas simplesmente fazer um estudo sintáctico da língua. Ainda que a sua extensão possa parecer pequena para serem tiradas algumas conclusões, pensamos que o podemos fazer, já que nos parece conter as três propriedades que um corpus deve apresentar (cf Galisson (1976)):

+ ser exaustivo, isto é, independentemente da sua extensão, deve conter todos os enunciados considerados característicos do tipo de discurso a estudar, no nosso caso, a língua corrente;

+ ser representativo, isto é, a partir das características, poderem ser feitas generalizações

+ ser homogéneo, isto é, que contenha o mesmo nível de língua, quando se trata de um levantamento temático; o que não foi o nosso caso, vários são os níveis de lingua utilizados, desde o popular ao literário, ao formal, ao de vulgarização, características que devem estar presentes num dicionário de língua corrente.

Por estas razões os documentos que o constituem foram retirados da imprensa diária, de obras literárias, revistas de vulgarização de ciências, letras e artes, semanários, roteiros, cartas publicitárias, notas internas do INESC, panfletos informativos e alguns enunciados de linguagem oral, e algumas frases fabricadas, etc.

Dada a diversidade das fontes, vários são os domínios de experiência abrangidos, entendendo-se por domínio de experiência o conjunto de taxemas lexicais ligados a um dos numerosos campos de experiência comuns a todos os menbros de uma comunidade, por exemplo, a política, a religião, a economia etc., é pois

diversificada também a temática do corpus.

Para o estudo descrito neste artigo recorremos a um dicionário de língua, de edição actualizada, dentro dos disponíveis no mercado.

Tipologia dos dicionários de língua

Segundo a tipologia dos dicionários apresentada por Bernard Quemada (Quemada (1990)), os dicionários de língua dividem-se em dicionários gerais e dicionários de especialidade. Os primeiros são extensivos e cobrem uma determinada sincronia, numa ou em várias épocas; os segundos são restritivos (selectivos) e tratam de domínios de experiência (vocabulários de especialidade, técnicos e científicos) e caracterizam-se por uma informação sobre o conteúdo (a definição) da vedeta, sem informação sobre a pronúncia, contextos, fonética, etimologia, e morfologia.

O dicionário que utilizámos pertence ao primeiro grupo, dicionário geral, dadas as características do corpus já citadas.

Metodologia

À medida que fomos preenchendo o dicionário de máquina, fomos tomando nota de todos os vocábulos não encontrados no dicionário que utilizámos, para procedermos a uma segunda consulta noutro dicionário. Desta vez escolhemos um dicionário publicado em 1991, em cinco volumes, pelo Círculo de Leitores. O primeiro é o dicionário da Porto Editora, 6ª edição, sem data, mas sabemos tratar-se da última.

Os números

Das 4884 formas distintas (correspondentes a 564 lemas) não foram encontradas respostas satisfatórias para 90 formas (correspondentes a 78 lemas) no dicionário, quer por omissão (ex: *subdomínio*), quer por desactualização de significado (ex: *computador*), o que perfaz cerca de 1.8% em termos de ocorrências e 13.8% em termos de lemas.

De notar que não entraram para esta contagem as 159 ocorrências de nomes próprios e os algarismos.

As formas contabilizadas foram, evidentemente, lematizadas. Por lematização entende-se o modo de agrupamento "standard" das diversas variantes de um mesmo signo, com a finalidade de simplificar apresentação e desse modo facilitar a consulta dos extractos lexicais em geral. Deste modo, os verbos foram contados no infinitivo, os substantivos no masculino singular, bem como os adjectivos.

A partir de agora usaremos o termo "palavras não encontradas" para descrever os lemas das formas que não constavam no dicionário no nosso estudo, quer omissas ou sem o sentido desejado.

Classificação gramatical

Quanto à sua classificação gramatical, dividem-se em:

adjectivos	29
verbos	2
siglas, abreviaturas	10

De entre os substantivos, considerámos:

lexias simples	19
lexias compostas	11
lexias complexas	15

Entendemos por lexias simples os vocábulos plenos; por lexias compostas as formas tradicionalmente chamadas palavras compostas, constituidas por vocábulos separados por hífen; e por lexias complexas, formas constituídas por dois ou mais vocábulos plenos mas que correspondem a um único conceito (por ex., *sistema monetário europeu*).

A contabilização das lexias complexas foi apenas feita para as que se encontravam no corpus em letras maiúsculas.

Registámos ainda 12 estrangeirismos, 61 formas de numerais cardinais e ordinais (incluindo o artigo indefinido singular), e um hapax, "enculturação". Por hapax entende-se uma palavra que ocorre uma única vez num corpus e não tem continuidade na história de uma língua.

De entre os adjectivos :

lexias simples	17
lexias compostas	12

Além disso,

calão	2
estrangeirismos	2

Os verbos pertencem ao calão: *xinar* e *xibar*.

Razões aventadas para a sua omissão no dicionário

Lexias simples

Analisando as formas omissas, verificámos :

Termos técnicos

Em dois ou três casos são formas de domínios específicos da ciência e da técnica que ainda não passaram pelo processo de vulgarização (por processo de vulgarização entende-se a passagem de um termo de determinado campo especializado para o domínio público), como *ameiada*, *fractais* e *balética*.

Neologismos de forma

Os neologismos de forma correspondem a novas unidades lexicais criadas para corresponder a conceitos novos, como por exemplo *telecópia*, *videotexto*, *dopado*, *imagística*, *interdisciplinaridade*.

Neologismos semânticos

Neste caso, ao vocábulo já existente é acrescentado um novo sema que lhe dá uma significação nova; os exemplos que encontrámos foram:

acetatos, usados como "transparentes para uma apresentação pública", que apenas consta como forma pertencente ao domínio da química;

angular, termo técnico de fotografia, que apenas está registado como pertencendo ao domínio da geometria;

nave, termo da aeronáutica, que aparece apenas como termo da arquitectura, da náutica e da geografia.

Lexias compostas

Algumas não aparecem, enquanto que outras também já de uso corrente estão patentes. Por ex. *fim-de-semana* aparece, mas *dia-a-dia* e *anti-inflamatório* encontram-se omissos.

Lexias complexas

Ainda que não tenhamos feito um estudo detalhado deste caso, podemos afirmar que estão omissas algumas como *corpo de baile*, *lei de bases*, *região autónoma*, *ciências básicas*, *tribunal constitucional*, *assembleia da república*.

Alguns problemas relativos a adjectivos

Alguns adjectivos não aparecem enquanto que o substantivo a que se refere está atestado. Ex. *artesanal* (*artesanato* consta), *conjuntural* (mas *conjuntura*).

Por outro lado, algumas lexias compostas não aparecem enquanto que os dois elementos que as compõem estão atestados: *científico-tecnológico*, *anti-inflamatório*, etc. De notar que *anti-infeccioso*, por exemplo, se encontra no dicionário.

Estrangeirismos

Verificámos com agrado que o apêndice ao dicionário contendo vocábulos estrangeiros, já lexicalizados na língua portuguesa, foi actualizado, conforme é dito no preâmbulo. No entanto, formas que há muito entraram no léxico corrente não foram encontradas, como por exemplo: *jogging*, *marketing*, *software*, *hardware*, *suite*, etc..

Siglas

A par do apêndice acima referido foi introduzido um novo para as siglas, no entanto algumas, há tanto tempo institucionalizadas, não constam: é o caso de CTT e LNEC.

Consulta a outro dicionário

Dado não termos encontrado todos os vocábulos no primeiro dicionário, consultámos o segundo tal como foi referido no início deste artigo.

Não encontrámos 85 formas, correspondentes a 66 lemas.

Verificámos que os neologismos de forma encontram-se em maior número, enquanto que os semânticos estão mais desactualizados. É o caso de *computador*, que não consta enquanto máquina, logo *computacional* e *computorizado* também não; igualmente, *congelador*, como parte de um frigorífico, não consta também; *informática*, idem. Em contrapartida, encontrámos vocábulos inexistentes no outro. Curioso foi verificar o verbo *cleopatrizar*, mas *cleópatra* não consta.

Conclusão

Não foi nossa intenção fazer uma crítica aos dicionários existentes, mas sim chamar a atenção para alguns problemas que surgem quando

recorremos a um dicionário de língua corrente.

Poderemos dar apenas algumas sugestões: por exemplo, no prefácio do dicionário ser dada a explicação da nomenclatura, tal como está no segundo dicionário que consultámos em relação à não existência de "advérbios de modo" (advérbios em -*mente* segundo os autores) na sua macro estrutura.

Da mesma forma, deviam ser explicitados os critérios segundo os quais palavras derivadas (tais como adjectivos deverbais, nominalizações, palavras contendo prefixos produtivos, etc.) aparecem ou não no dicionário. Dever-se-ia também incluir o domínio a que pertence a vedeta sempre que se trate de um vocábulo que, apesar de vulgarizado, remeta para as ciências, as artes, a linguística, etc.

Bibliografia

Galisson, Robert. (1976) *Dicionário de Didáctica das Línguas*, Almedina, Coimbra.

Rey, Alain. (1971) *Le Lexique: images et modèles*, Colin, Paris.

Muller, Charles. (1974) "La lematisation, essai d´analyse mathématique" in *Travaux de Linguistique et de Litérature*, Strasbourg, vol.12, 1.

Muller, Charles. (1977) *Principes et méthodes de statististique lexical*, Hachette, Paris.

Quemada, Bernard. (1990) "Les données lexicographiques et l'órdinateur" in *Cahiers de Lexicologie*, nº 56-57.

Salem, André. (1987), *Pratique des Segments Repétés: Essai de Statistique Textuelle*, Klincksieck, Paris.

Santos, Diana (ed.) (1992) *Processamento de corpora de texto no INESC*. Relatório INESC, 1992.

Vilela, Mário. (1983), *A definição nos dicionários portugueses*, Asa, Porto.

APÊNDICE: Palavras não encontradas

Será indicado com 1 ou 2 em que dicionário. Se não aparecer qualquer indicação, estavm omissas nos dois. As siglas apenas foram verificadas no 1º dicionário.

ABC
abstraccionista
acetato
alemão-federal
amarelo-vivo
ambiental
ameiada (2)
ança (1)
angular
anti-inflamatório
antiautomatismo
apoiante
arquetípico (1)
artesanal (1)
autoconsumo (2)

Av.
bailatória
balética
barthesiano (1)
baudelairiana (1)
bicarbonata (1)
bué
CDS
Œ
ciências-básicas (1)
científico-tecnológico
co-textual
colonialismo
computacional
computador (2)
computorizado
conjuntural
corpus
CTT
decepcionante (1)
desvelamento
dia-a-dia
disposicional
dopado
ecran (2)
empresarial (1)
endurance
entrelaçamento
Ex.
fabulatória (1)
FENPROF
fónico-rítmico
formante
fractal
fundamentalismo (1)
hardware
histórico-sociológico
IBM
ideológico (2)
imagistíca
inalcançável
indefinição
INESC
interdisciplinar
interdisciplinaridade (2)
jogging
Km
LNEC
macro-sistema
mailing
marketing
mestrando
minero-medicinal
monitorização
morfosintáctica (1)
multicritério
multiequipamento
multifacetado (1)
nave
núcleo-síntese
oitocentista
paragramática (1)
pentáculo (1)
plurisignificação
pontilista (2)
pós-cubista
pós-modernidade
prosódico-versificatório
recálculo (1)
rocket (1)
rodeo (1)
scanner
software
Sr.
subdomínio
suite
súlfera (1)
supermagnético (2
telecópia
videotexto (1)
xibar
xinar