Reconhecimento automático de expressões idiomáticas em corpus – algumas experiências

Graça Fernandes[1], **Jorge Baptista**[1,2]

[1]Faculdade de Ciências Humanas e Sociais – Universidade do Algarve (UALG)
P-8005-139 Faro, Portugal

[2] L²F – Spoken Language Laboratory, INESC-ID Lisboa, Portugal

jbaptis@ualg.pt, gracitafernandes@gmail.com

Abstract. *This paper reports an experience consisting in the application of an electronic dictionary of multi-word, idiomatic, intransitive, verbal expressions of European Portuguese, to large-sized corpus, using finite-state methods and lexicon matrices. It aims at quantifying the advantages and limitations of this method.*

Resumo. *Este artigo apresenta uma experiência de aplicação a um corpus de grandes dimensões de um dicionário electrónico de frases fixas do Português Europeu, com o objectivo de avaliar as vantagens e as limitações deste tipo de metodologia.*

1. Introdução, métodos e resultados.

Neste artigo, descrevemos uma experiência de aplicação a um *corpus* de texto de grandes dimensões (o CETEMPúblico) de um fragmento do dicionário electrónico de frases fixas; trata-se de construções verbais fixas/idiomáticas de tipo intransitivo (sem complemento directo), tais como *dar à sola* (‘fugir’). Neste dicionário, a informação linguística é codificada em matrizes léxico-sintácticas. Estas são intersectadas automaticamente com transdutores de estados finitos para a sua aplicação a textos no reconhecimento deste tipo de expressões. A avaliação que fazemos desses resultados visa apenas a identificação das vantagens e/ou limitações aos métodos de representação formal e de aplicação das gramáticas aos textos que aqui adoptámos.

Baseamo-nos nos dados linguísticos recolhidos e formalizados por Fernandes (2007)[1], que descreve as classes de construções intransitivas das frases fixas do Português Europeu. Na Tabela 1 apresenta-se o conjunto actual destas construções:

Tabela 1.
Classificação das frases fixas com complementos preposicionais do Português Europeu

Classe	Estrutura	Exemplo	Efectivos
CP1	N_0 *V Prep* C_1	*O Mário deu à sola*	625
CPN	N_0 *V Prep* (*C de N*)$_1$	*O Mário não chega aos calcanhares da Ana*	90
CPP	N_0 *V Prep* C_1 *Prep* C_2	*O Mário deu com os burrinhos na água*	200
		Total	915

O método de formalização da informação linguística aqui adoptado para aplicação de dicionários electrónicos a *corpus* baseia-se na intersecção de tabelas

[1] O seu trabalho insere-se num programa mais vasto de elaboração de um léxico-gramática das expressões fixas do Português (Baptista *et al.* 2004, 2005). Esta descrição apoia-se nos trabalhos de M. Gross (1982, 1989) e adopta a perspectiva teórica do Léxico-Gramática, desenvlvido por M. Gross (1996) e baseado na gramática transformacional de operadores de Zellig. S. Harris (1991).

lexicais com transdutores de estados finitos [2]. Dado o número de frases fixas recolhidas, limitamo-nos aqui a fazer algumas observações sobre os resultados obtidos a partir de uma selecção de 100 expressões de uso mais corrente da classe CP1. Para o processamento lexical do texto, utilizámos os recursos linguísticos disponíveis para o Português Europeu distribuídos com o software UNITEX . Constituímos um *corpus* de trabalho a partir da versão disponível *on-line* do CETEMPúblico, utilizando as concordâncias que resultaram da pesquisa de expressões racionais, constituídas, genericamente, pelo lema do verbo e o complemento preposicional fixo, permitindo uma janela de 5 palavras entre estes elementos. Assim, por exemplo, para a expressão <*dar*> *à sola* utilizámos a expressão racional: `[lema="dar"][]{0,5}"à""sola"`.

A partir deste *corpus* de trabalho, constituído por 6.039 extractos, verificou-se (semi-automaticamente) a ocorrência de 5.183 instâncias das expressões-alvo, tendo-se obtido, na aplicação do dicionário electrónico ao corpus, uma **precisão** média de **0,94**. Na secção que se segue, discutimos estes resultados.

2. Discussão

De um modo geral, pode dizer-se que as principais dificuldades de reconhecimento das frases fixas no corpus resultam, como já se esperava, da inserção de elementos livres entre os elementos constitutivos da expressão fixa, dado que estas inserções não foram (intencionalmente) previstas no grafo de referência. Algumas destas inserções são apenas elementos formais, como por exemplo as aspas, que, em textos escritos, frequentemente assinalam o carácter idiomático destas expressões: *[...] Candidatos a bombeiros* ***«brincaram» com o fogo*** [...] Ext 176357 (soc, 95b). Contudo, outras inserções são elementos lexicais, observando-se frequentemente a inserção de advérbios de natureza variada entre o verbo e o complemento preposicional fixo: [...] ***batemos*** *envergonhadamente* ***em retirada*** *para os copos do bar* [...] Ext 1321159 (nd, 94a); [...] *os madridistas* ***escaparam*** *ontem* ***de boa*** [...] Ext 309547 (des, 96b). Uma vez que estas inserções não foram (intencionalmente) descritas no grafo de referência, as frases fixas não foram identificadas. Seria fácil adaptar o grafo de forma a admitir essa inserção, mas tal não era nosso objectivo aqui.

Noutros casos, o não reconhecimento da expressão fixa deve-se a fenómenos de coordenação: [...] *os militantes laranjas trataram de* ***dar ao talher*** *e* ***ao dente*** [...] Ext 1178536 (eco, 92b); *Muito* ***darão à língua*** *e* ***ao dente****, eructando aos microfones a salvação estes mestres do Aviz.* Ext 1271901 (nd, 94a). A possibilidade de coordenar os complementos fixos de duas expressões idiomáticas distintas, construídas com o mesmo verbo, leva-nos a considerar que, só no âmbito de uma análise sintáctica das frases em que estas expressões fixas ocorrem (e não por uma mera identificação de cadeias de palavras) será possível identificá-las com rigor e adequação.

Também nas construções com pronome indefinido *–se*, com um valor genérico (‘alguém’, ‘as pessoas em geral’), e com função de sujeito: *No Ocidente chamamos-lhe brindes, mas na Geórgia* ***não*** se ***brinca em serviço****.* Ext 38967 (nd, 93a) , o sistema não reconhece o pronome, que acaba por constituir uma inserção entre os elementos componentes da expressão fixa. Uma vez mais, só a análise sintáctica correcta de *–se*

[2] Seguimos a proposta inicialmente apresentada por Senellart (1998) e desenvolvida por Silberztein (2000) a partir da versão 3.1 do software INTEX, e mais tarde implementada no software UNITEX (versão 1.2.; Paumier 2004). Por falta de espaço, remetemos o leitor para o manual deste último.

bem como da construção em que este se encontra, permitiria a correcta identificação da expressão fixa.

Um caso geral de não reconhecimento ou de reconhecimento incompleto das expressões fixas acontece na situação em que o verbo está acompanhado de um verbo auxiliar. Os auxiliares formam com o verbo principal um complexo verbal, razão por que deveriam ser reconhecidos como parte integrante da expressão fixa: Diana [...] detesta «junk food» [...], *apesar de na noite anterior* ***se ter metido nos copos***, [...] Ext 635725 (clt, 94a). Note-se que a expressão *<meter>-se nos copos* é reconhecida nos casos em que o verbo não apresenta auxiliar. Obviamente, na medida em que não dispomos de um módulo de análise dos verbos auxiliares, o reconhecimento das expressões em que o verbo se encontra nesta situação apenas foi feito parcialmente. Já no caso das expressões intrinsecamente negativas (Fernandes e Baptista, *no prelo*), considerámos que o advérbio de negação (*Neg*) faz parte da expressão e deveria, por isso, ser reconhecido. Ora, como o verbo auxiliar pode ocorrer entre *Neg* e o verbo principal, situação que não está descrita no grafo de referência, tal impossibilita o reconhecimento da expressão fixa: *É que advogado, mesmo de mota,* ***não pode brincar em serviço*** [...] Ext 484626 (eco, 95a):

Em alguns casos, o grafo reconhece correctamente a sequência-alvo mas esta faz parte de uma expressão mais longa. Na medida em que esta sequência maior não consta da matriz, poderia considerar-se que os resultados do processo de identificação da expressão fixa apenas foram alcançados parcialmente. É o caso da expressão N_0 *cortar nas despesas*, em que se verifica que o elemento *C*: *despesas* pode surgir acompanhado de diversos modificadores, formando até nomes compostos com alguns destes. De facto, agumas destas combinações, tais como *despesas de investimento, despesas públicas, despesas sociais, despesas correntes, despesas salariais, despesas de funcionamento* e *despesas orçamentais*, são muito provavelmente termos compostos (da economia em geral ou de contabilidade pública). Outras combinações (*despesas de supermercado, despesas com os meios aéreos*) relevam da sintaxe particular do nome predicativo *despesas*. Ambas as situações apontam no sentido de, nesta expressão, apenas a cabeça do grupo nominal (*despesas*) ser fixa com o verbo (*cortar*), podendo o elemento *C* fazer parte de (ser o elemento nuclear de) nomes compostos ou apresentar complementos próprios das construções predicativas que determina. Naturalmente, tais situações implicam a análise sintáctica do contexto alargado em que a expressão se insere, o que está para além dos objectivos a que aqui nos propusemos.

3. Conclusão e perspectivas futuras

Os resultados preliminares já obtidos permitem avançar, ainda que provisoriamente, com algumas conclusões. Em primeiro lugar, verifica-se que para a maioria das expressões idiomáticas que aqui estudámos não se levantam dificuldades de maior ao seu reconhecimento automático por meio de métodos de estados finitos como os que aqui empregámos. A elevada precisão (94%) resulta, sobretudo, do facto de, pela sua elevada fixidez interna, os elementos por que estas expressões são formadas aparecem geralmente em sequência contíguas, não interrompidas por quaisquer inserções. São justamente essas inserções de elementos livres, externos à construção fixa (em que predominam os adverbiais, construções com verbos auxiliares, emprego de pronomes indefinidos, etc.) o principal factor responsável tanto pelo erros de reconhecimento como pelo silêncio (12%). Parece, assim, claro que uma melhoria importante dos resultados poderia ser alcançada se a identificação das expressões idiomáticas tivesse

lugar após uma análise sintáctica prévia do texto, a qual permitisse a identificação dos constituintes inseridos entre os elementos componentes das frases fixas. Naturalmente, beneficiámos do facto de, nestas expressões, não se observarem transformações como pronominalizações, permutas, etc. A extensão do estudo a estas situações deverá permitir aferir melhor a adequação dos métodos de representação/reconhecimento automático aqui utilizados, bem como aferir *quantitativamente* a importância destas expressões em diferentes géneros textuais.

6. Referências.

Baptista, J. (2004) "Frozen vs. Compositional Sequences", in *Journal of Applied Linguistics. Special Issue on Lexicon-Grammar. Papers presented at the Lexicon-Grammar Workshop*, Edited by E. Laporteand Ting-an Cheng, p. 81-94 [in Chinese; English version available].

Baptista, J., Correia, A. and Fernandes, G. (2004) "Frozen Sentences of Portuguese: Formal Descriptions for NLP"., Proceedings of the Workshop on *Multiword Expressions: Integrating Processing, International Conference of the European Chapter of the Association for Computational Linguistics*, p. 72-79, Barcelona, ACL.

Baptista, J., Correia, A. and Fernandes, G. (2005) "Léxico-gramática das frases fixas do portugués europeo. Breve presentación", in *Cadernos de Fraseoloxía Galega* 7, p. 41-53, Santiago de Compostela: Xunta de Galicia.

Fernandes, G. (2007) *Léxico-Gramática das Frases Fixas do Português Europeu. Contruções Intransitivas* (Tese de Mestrado), Faro, Universidade do Algarve/FCHS.

Fernandes, G. and Baptista, J. (in print) "Frozen Sentences with Obligatory Negation: Linguistic Challenges for Natural Language Processing", (to appear in *Cadernos de Fraseoloxía Galega* 8).

Gross, M. (1982) "Une classification des phrases "figées" du français", in *Revue Québécoise de Linguistique* 11-2, p. 151-185, Montréal, UQAM.

Gross, M. (1989) *Les Expressions Figées, Une description des expressions françaises et ses conséquences théoriques.* RT nº 8, PRC-IL, Paris, Université Paris 7/LADL.

Gross, M. (1996) "Lexicon-Grammar" in *Concise Encyclopaedia of Syntactic Theory*, Edited by K. Brown and J. Miller, p. 224-259, Oxford, Pergamon Press.

Harris, Z. S. (1991) *A Theory of Language and Information. A Mathematical Approach*, Oxford, Clarendon Press.

Paumier, S. (2004) *Unitex - manuel d'utilisation*, Paris, Univ. Marne-la-Vallée, http://www-igm.univ-mlv.fr/~unitex/UnitexManual.pdf, May 4, 2005.

Senellart, J. (1998) "Reconnaissance automatique des entrées du lexique-grammaire des phrases figées", in *Le Lexique-Grammaire. Travaux de Linguistique* 37, Edited by B. Lamiroy, p. 109-125, Bruxelles, Duculot.

Silberztein, M. (2000) *Intex* (Manual). Paris: ASSTRIL/LADL.