Para uma tipologia de associações de palavras do português

Sandra Antunes
Centro de Linguística da Universidade de Lisboa

Abstract

Based on a lexicon of Portuguese multiword expressions, this presentation focuses on an ongoing work that aims at the creation of a typology that describes these expressions taking into account their semantic, syntactic and pragmatic properties. We also plan to annotate each MWE-entry in the mentioned lexicon according to the information obtained from that typology. Our objective is to create a valuable resource, which will allow for the automatic identification MWE in running text and for a deeper understanding of these expressions in their context.

Keywords: Multiword expressions, typology, lexicon, annotation
Palavras-chave: Associação de palavras, tipologia, léxico, anotação

1. Introdução

Após um longo período em que a sintaxe foi considerada a componente mais proeminente da gramática, o estudo do léxico começa a ganhar vigor a partir dos anos 50 do século XX, altura em que começaram a aparecer trabalhos que chamavam a atenção para o facto de as palavras adquirirem o seu significado através das relações que estabelecem com as palavras com as quais coocorrem (Firth, 1955; Coseriu, 1967), sendo possível identificar padrões associativos regulares e complexos que dão informações cruciais sobre os potenciais significados e usos dos itens lexicais (Sinclair, 1991).

Contrariando o que foi durante largos anos a perspetiva dominante, o léxico deixa de ser encarado como um componente estático e independente da gramática. O seu estudo estende-se para além dos aspetos tradicionalmente observados (como os processos de formação de palavras e as relações de sentido, como a sinonímia, antonímia, etc.) e começa--se a prestar particular atenção ao uso recorrente de combinações específicas de palavras por parte dos falantes. De facto, o recurso a dados reais da língua (levado a cabo por investigadores que defendem que a introspeção não é suficiente para analisar e descrever objetivamente uma língua) mostra que, embora os falantes tenham à sua disposição uma multiplicidade de escolhas em termos de associação de itens lexicais para a formação dos seus enunciados, é frequente a utilização de determinadas sequências sintagmáticas que aparecem

***Textos Selecionados, XXVIII Encontro Nacional da Associação Portuguesa de Linguística,* Coimbra, APL, 2013, pp. 79-97, ISBN: 978-989-97440-2-8**

como pré-construídas (princípio idiomático (Sinclair, 1991:110)). A identificação e análise destas expressões (levadas a cabo por autores como Firth (1955), Mel'čuk (1984), Benson *et al*. (1986), Hausmann (1989), Sinclair (1991), Cowie (1994), Fernando (1996), Corpas Pastor (1996), Moon (1998) ou Sag *et al*. (2002), entre outros) mostram que, efetivamente, o léxico também é composto por grupos de palavras mais ou menos previsíveis, não necessariamente fixos, que desempenham um papel importante na estrutura de qualquer língua, contribuindo grandemente para o seu processamento e desenvolvimento[1]. Na verdade, tornou-se cada vez mais evidente que o léxico de uma língua é em grande parte composto por estes grupos pré-construídos: Jackendoff (1997) estima que o número de associações de palavras presente no léxico de um falante é igual ao número de palavras simples e Sag *et al*. (2002) comentam o facto de na WordNet 1.7 (Fellbaum, 1998) 41% das entradas corresponderem a este tipo de expressões.

Partindo do conceito lato de associação de palavras (sequência de palavras que apresenta um elevado grau de coesão sintática e semântica entre os elementos do grupo, abrangendo diferentes tipos de relações sintagmáticas) e tendo em consideração o papel importante que estas expressões desempenham na língua, este artigo apresenta o estudo que está a ser desenvolvido com o objetivo de identificar as principais associações de palavras do português europeu (selecionadas a partir de um léxico previamente extraído de um *corpus* constituído por 50 milhões de palavras) e estabelecer uma tipologia baseada em trabalhos como os dos autores já referidos, bem como uma proposta de anotação do léxico supra mencionado com a informação obtida através da tipologia. Do ponto de vista da linguística computacional, uma vez que estas expressões levantam sérios obstáculos na criação de ferramentas para o processamento da língua natural, crê-se que a criação de um léxico de associações de palavras anotado com informação detalhada sobre o seu comportamento sintático e semântico constituirá um recurso valioso no que respeita à sua deteção e anotação automáticas (Hendrickx *at al*., 2010a).

Deste modo, apresenta-se, na secção 2, uma breve descrição da constituição do *corpus*, da metodologia seguida para a extração automática das associações pertinentes e da

[1] Com o crescente interesse pelo estudo das associações de palavras, surgiram inúmeras propostas de análise deste fenómeno linguístico, resultando numa grande proliferação terminológica que, normalmente, varia de autor para autor. No caso do português, também não há unanimidade no que respeita à denominação e definição deste fenómeno, existindo na literatura vários termos, dos quais se destacam palavras compostas (Cunha & Cintra, 1987), compostos sintagmáticos (Correia & Lemos, 2005), expressões lexicalizadas (Villalva, 2000), fraseologismos/frasemas (Vilela, 2002), expressões fixas (Ranchhod, 2003), unidades lexicais multipalavra (Ranchhod *et al*., 2003; Abalada *et al.*, 2010), unidades pluriverbais (Rio-Torto, no prelo), combinatórias (Antunes *et al.*, 2008), unidades multilexicais (Mendes *et al*., 2012).

organização do léxico. Na secção 3, são discutidos o conceito de associação de palavras adotado e a categorização que está a ser elaborada e, na secção 4, expõe-se uma proposta para a anotação do léxico. A secção 5 encerra este artigo com uma breve conclusão, apontando-se as utilizações possíveis dos resultados.

2. Do *corpus* ao léxico

2.1. Constituição do *corpus* e extração automática das associações

Para a realização deste trabalho foi utilizado o léxico de associações de palavras do português europeu constituído no âmbito do projeto COMBINA-PT[2], desenvolvido no Centro de Linguística da Universidade de Lisboa. Para tal, foi extraído do *Corpus* de Referência do Português Contemporâneo[3] um subcorpus equilibrado[4] de registo escrito com cerca de 50 milhões de palavras. A grande dimensão do *corpus* é justificada pelo facto de a extração de associações de palavras requerer uma grande quantidade de dados reais, uma vez que só assim é possível observar, de um modo mais objetivo, os diferentes padrões associativos das expressões, a maior ou menor variação a nível léxico-sintático e a frequência de ocorrência e consequente estabilidade na língua (fator importante para os casos de associações que só mais recentemente começaram aparecer). O *corpus* é composto por vários tipos de discurso, como se pode observar na Tabela 1.

[2] http://www.clul.ul.pt/en/research-teams/187-combina-pt-word-combinations-in-portuguese-language

[3] O CRPC é um *corpus* monitor com cerca de 311 milhões de palavras, constituído por amostragens de diversos tipos de texto de discurso escrito e oral, que dizem respeito às variedades regionais e nacionais do português (https://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc).

[4] Entende-se por *corpus* equilibrado aquele em que os diferentes tipos de discurso representam, de modo proporcional, o uso corrente de uma língua. A inclusão de diferentes registos é essencial para descrever todos os padrões de ocorrência dos itens lexicais, uma vez que as palavras podem apresentar coocorrentes diferentes de acordo com o tipo de discurso em que se inserem. A maior proporção de texto jornalístico (60%) é justificada pelo facto de ser aquele que abrange uma maior diversidade de temas e se dirige a um público mais diversificado, apresentando, por isso, uma linguagem mais próxima da utilizada no quotidiano.

CONSTITUIÇÃO DO *CORPUS*	
JORNAL	**30.000.000**
LIVRO	**10.917.889**
REVISTA	**7.500.000**
VARIA	**1.851.828**
FOLHETO	**104.889**
ACÓRDÃOS DO SUPREMO TRIBUNAL DE JUSTIÇA	**313.962**
DIÁRIOS DA ASSEMBLEIA DA REPÚBLICA	**277.586**
TOTAL	**50.966.154**

Tabela 1. Constituição do *corpus*

Para a extração das associações de palavras foi aplicada, em UNIX, uma ferramenta informática que, a partir de métodos estatísticos de Informação Mútua (IM)[5], permitiu extrair grupos compostos por 2 a 5 palavras, dando para cada grupo informação vária, como: (i) o número de elementos do grupo; (ii) a frequência; (iii) a distância a que ocorrem os elementos do grupo – os grupos formados por 2 palavras podem ser contíguos ou estarem separados por um máximo de 3 palavras (*conjuntura internacional*; *conjuntura económica internacional*); (iv) o valor de IM e consequente ordenação dos resultados; (v) as concordâncias de cada grupo (linhas de contexto no *corpus*).

2.2. Seleção manual das associações significativas

Dado o elevado número de associações extraídas automaticamente do *corpus* (cerca de 1.7 milhões), foi selecionado manualmente apenas um conjunto desses resultados com base nos valores de IM. Tendo em conta trabalhos como os de Evert & Krenn (2001) e Pereira & Mendes (2002), em que se observou que os valores de IM entre 7 e 11 correspondiam à faixa em que ocorriam associações de palavras mais interessantes, procedeu-se à seleção de grupos com valores entre 8 e 10 (Mendes *et al.*, 2006). A validação dos grupos foi feita prestando-se

[5] A medida de associação lexical de Informação Mútua permite estabelecer uma relação entre a frequência do grupo no *corpus* e a frequência de cada elemento do grupo, também no *corpus*, e medir o grau de associação lexical entre os elementos do grupo, i.e., a maior ou menor tendência para a sua coocorrência (Church & Hanks, 1990). Este grau de associação (Índice Combinatório) possibilita uma apresentação hierarquizada das combinatórias segundo a significância que detêm no *corpus*.

particular atenção a determinados critérios, sobre os quais assentam, normalmente, as definições de associação de palavras:

(i) fixidez do grupo a nível léxico-sintático, que pode ser observada através da possibilidade de substituir elementos, inserir modificadores ou alterar a ordem sintagmática ou os traços de género e número;

(ii) especificação semântica, que implica a perda (total ou parcial) do significado literal dos elementos que compõem o grupo;

(iii) ocorrência frequente, que aponta para uma associação preferencial entre os seus elementos, o que pode revelar uma fase inicial do processo de fixidez.

Considerou-se, deste modo, que a seleção dos grupos teria de integrar não só critérios lexicais, sintáticos e semânticos, mas também pragmáticos e, por vezes, fonológicos (estes últimos podem desempenhar um papel importante, uma vez que certas expressões podem apresentar aliterações ou restrições relativamente à ordem dos constituintes sem que para isso haja motivação sintática ou semântica aparente), tentando excluir-se associações totalmente livres, embora nem sempre seja fácil delimitar essa fronteira. Tal dificuldade é agravada pelo facto de que, em ambos os casos, as expressões são constituídas por sequências de categorias gramaticais formalmente idênticas (N Adj, N Prep N, Adj N, etc).[6]

Após a seleção das associações mais significativas, estas foram organizadas em lemas principais (ex. *fogo*) e em lemas de grupo, que dizem respeito à forma canónica das associações (ex. *arma de fogo*) e que reúnem as variantes flexionais que ocorreram no *corpus* (ex. *arma de fogo; armas de fogo*). O léxico final é composto por 1180 lemas principais e 14.153 lemas de grupo, que, por sua vez, incluem todos os tipos de variação, num total de 48.154 associações[7].

3. Análise dos dados

Ao contrário do que acontece em línguas como o inglês, o francês, o russo ou o espanhol, para as quais existe uma grande proliferação de estudos sobre associações de palavras, não há, para o português, muitas propostas de definição e categorização deste

[6] A literatura também fornece pouca informação sobre o grau de liberdade léxico-sintática a partir do qual determinado grupo deve ser considerado uma associação livre. Para a distinção entre estes tipos de expressões, teve-se em atenção os critérios sintáticos descritos em Baptista (1994).

[7] O léxico, bem como toda a informação sobre os critérios seguidos para a seleção das associações e a sua organização e lematização, pode ser consultado em:
https://www.clul.ul.pt/sectores/linguistica_de_corpus/manual_combinatorias_online.php

fenómeno linguístico. A maior parte dos trabalhos que existem centram-se, mais particularmente, no estudo de expressões idiomáticas e palavras compostas em particular (Macário Lopes, 1992; Chacoto, 1994; Baptista, 1994; Vilela, 2002; Ranchhod, 2003) e a análise das expressões é feita, quase sempre, de acordo com as propriedades morfossintáticas dos seus constituintes.[8]

Na verdade, pode considerar-se a existência de uma certa dificuldade em identificar e definir este tipo de expressões que parece estar essencialmente relacionada com o facto de as associações de palavras consistirem num fenómeno linguístico que se situa na fronteira entre a gramática e o léxico. Estas expressões são fruto da composição de duas ou mais unidades lexicais autónomas realizada noutras componentes da gramática (sintática, morfológica, discursiva), que se fixa na língua (começa a ser frequentemente utilizada e, consequentemente, institucionaliza-se, i.e., é aceite por toda ou por uma parte significativa da comunidade linguística) e passa a funcionar como uma estrutura sintática e semanticamente autónoma. Este processo de lexicalização é gradual e, em último caso, pode ser acompanhado por uma especialização semântica, ou seja, a expressão deixa de ter significado composicional.

Tendo-se observado que existem associações de palavras com diferentes graus de coesão sintático-semântica, e tendo ainda em conta critérios de frequência e índice estatístico, começou-se a desenvolver uma tipologia que vai desde expressões totalmente composicionais, mas que apontam para coocorrências preferenciais, até às expressões idiomáticas e, supostamente, com maiores restrições a nível flexional, lexical e sintático.

Deste modo, do ponto de vista semântico, as expressões começaram por ser distribuídas em três categorias:

(i) significado composicional – expressões cujo significado pode ainda ser calculado através do significado de todos os seus elementos (ex. *vestido de noiva*);

(ii) significado parcialmente idiomático – expressões que já sofreram algum grau de especialização semântica, mas em que ainda se consegue atribuir significado composicional a alguns dos seus elementos (ex. *sorriso amarelo*);

(iii) significado idiomático – expressões semanticamente opacas, com significado global, em que não é possível atribuir significado composicional a nenhum dos seus elementos (ex. *a ferro e fogo*).

[8] Note-se, no entanto, a existência de alguns estudos que focam a identificação e anotação de predicados complexos (construções formadas com verbos-suporte), como Hendrickx *et al*. (2010b) e Duran *et al*. (2011).

Note-se, contudo, que, como seria de esperar, no que respeita à intuição semântica, verificou-se alguma dificuldade na avaliação da composicionalidade/idiomaticidade de algumas expressões. Enquanto nalguns casos não há dúvidas em considerar que o significado de algumas combinações é perfeitamente composicional ou idiomático, noutros pode existir uma maior flutuação nessa avaliação, que resulta principalmente de dois fatores: (i) a caraterística polissémica da língua (das diferentes aceções que uma palavra pode ter, é necessário considerar quais são as que denotam significados composicionais e quais são as que já representam significados figurados. Se se considerar como significado composicional apenas o significado prototípico das palavras, i.e., aquele que normalmente ocorre em primeiro lugar nas entradas dos dicionários, estar-se-á perante uma definição muito restrita de composicionalidade que levará a que grande parte das expressões seja considerada idiomática); (ii) a consciência da motivação subjacente à criação das expressões, que pode conduzir a uma avaliação no sentido da composicionalidade. Verifica-se, assim, que as fronteiras entre os diferentes tipos de combinações são muito ténues, não permitindo definições e limitações precisas.

Note-se ainda que este critério semântico utilizado na divisão das associações permite encontrar o mesmo tipo de associação (ex. nomes compostos) em diferentes categorias, tentando, deste modo, observar-se os diferentes graus de lexicalização que detém na língua.

Dentro de cada categoria, as expressões são ainda caraterizadas de acordo com a categoria gramatical, ou o valor funcional, e a fixidez.

Do ponto de vista da fixidez, as expressões foram analisadas do seguinte modo:

(i) expressões fixas – a expressão não apresenta variação;

(ii) expressões semifixas – a expressão apresenta apenas flexão verbal ou nominal, caraterística de línguas altamente flexionais, como o português;

(iii) expressões com variação – a expressão pode apresentar variação:

- morfossintática: flexão; modificação da estrutura sintática (nominalização, passivização, relativização, pronominalização, possessivização, ocorrência em estruturas predicativas, alternância entre a ocorrência de modificadores adjetivais e preposicionais, alternância entre a ocorrência em estruturas sem negação e com negação);
- lexical: permutação de elementos; inserção de modificadores (artigos, pronomes, quantificadores, advérbios, adjetivos); redução e extensão; substituição lexical (por sinonímia,

por antonímia, dentro de um campo lexical homogéneo, dentro de um campo lexical heterogéneo, entre preposições, entre elementos simples e elementos com sufixos avaliativos, realização livre de argumentos/complementos).

Na verdade, a análise de um conjunto tão grande de dados mostrou níveis elevados de variação lexical e sintática nos grupos estudados, como se pode observar nos exemplos presentes nas categorias que seguidamente se apresentam.

3.1. Tipologia

A tipologia encontra-se, assim, dividida em três categorias principais, que, por sua vez, são analisadas de acordo com critérios sintáticos, lexicais e pragmáticos.

Expressões com significado composicional

- **coocorrentes privilegiados** (expressões com elevada frequência de ocorrência no *corpus* e que, apesar de manterem o significado literal de todos os seus elementos e terem um fraco grau de lexicalização, formam uma unidade de uso, mostrando que certas palavras apresentam uma tendência para ocorrer com determinadas outras, em certos contextos)

greve de fome; *pão de centeio*; *folha caduca*; *café solúvel* (nominal)

morrer de fome; *desvendar o mistério*; *declarar guerra*; *firmar um acordo* (verbal)

No que respeita à fixidez, esta parece ser a classe que apresenta maior variação. Além da flexão, é possível encontrar:[9]

(i) inserção de modificadores (*greve massiva de fome*; *pão negro de centeio*; *morreríamos todos de fome*; *desvendar finalmente o mistério*);

(ii) substituição lexical (*pão de centeio/milho/trigo*; *greve de fome/zelo*; *morrer de/à fome*; *desvendar/descobrir/decifrar o mistério*; *declarar guerra/greve/falência*; *firmar um acordo/protocolo/contrato*);

(iii) modificação da estrutura sintática através da possibilidade de ocorrência em estruturas predicativas (*o pão é de centeio*; *o café é solúvel*), nominais (*declaração de guerra*), passivas (*o acordo foi firmado*; *o mistério foi desvendado*), relativas (*o mistério que desvendaram*) ou com pronomes (*desvendaram-no*).

[9] Nas categorias em que ocorre maior variação, a lista apresentada não é exaustiva, estando normalmente limitada aos exemplos apresentados.

- **expressões institucionalizadas** (expressões que se distinguem das anteriores por serem estatisticamente idiossincráticas, i.e., ocorrerem com muito mais frequência do que qualquer outra possível lexicalização do mesmo conceito)

  lufada de ar fresco; *motivo de força maior* (nominal)

  condenar ao fracasso; *abrir um precedente* (verbal)

  o que tem de ser tem muita força; *de boas intenções está o inferno cheio* (frase)

Além da flexão, verifica-se:

(i) substituição lexical, uma vez que há lexicalizações alternativas (*lufada/baforada/rajada/corrente de ar fresco*; *motivo/razão/caso de força maior*; *condenar/votar ao fracasso*; *abrir/criar um precedente*);

(ii) inserção de elementos (*abrir mais um precedente*; *de boas intenções, como dizem os Lísicos, está o inferno cheio*);

(iii) permutação de elementos (*condenar ao fracasso este projeto* / *condenar este projeto ao fracasso*);

(iv) modificação da estrutura sintática através de pronominalização (*condenou-o ao fracasso*), passivização (*o precedente foi aberto*) ou nominalização (*abertura de um precedente*).

- **nomes compostos** (expressões nominais que representam uma ideia simples)

  noite de núpcias; cama de casal; tumor maligno; barco rabelo[10]*; cidade invicta*[11]; *idade do ferro*

Do ponto de vista sintático, estas expressões são fixas. Contudo, dada a extensão de algumas expressões, é frequente encontrá-las truncadas (*deitar cedo e cedo erguer*). Adicionalmente, ao contrário do que se poderia prever, pode também ocorrer substituição lexical (*no poupar / anunciar / atacar / descontar / esperar / provar / comparar / economizar é que está o ganho*). Esta variação ilustra a criatividade lexical dos falantes que, ao reconhecerem a expressão canónica, são capazes de analisar os seus componentes e substituir um elemento específico do grupo. As expressões que denotam entidades (*idade do ferro*) são

[10] Estas expressões são normalmente denominadas solidariedades lexicais: o significado de um dos elementos da expressão já inclui o significado do outro elemento que com ele coocorre (Coseriu, 1967; Sanromán, 2000).
[11] Nomes perifrásticos (Sanromán, 2000).

fixas. As restantes, além da flexão, podem apresentar um pequeno paradigma distribucional (*cama de casal/solteiro; tumor maligno/benigno*), pelo que podem ocorrer em estruturas predicativas (*a cama é de casal; o tumor é benigno*). Apesar da existência desta variação, estas expressões são bastante mais coesas do ponto de vista sintático e semântico do que as anteriores, optando--se, assim, pela sua inclusão nesta categoria.

- **expressões com verbo-suporte** (estruturas V N, em que os constituintes mantêm a sua grelha argumental, mas em que o verbo é semanticamente esvaziado, ou perde parte do seu conteúdo semântico, passando o nome a ser o predicado principal. Normalmente, a estrutura pode ser parafraseada pelo verbo pleno do qual o nome deriva, no caso dos nomes deverbais, ou por um verbo associado ao nome)

dar um passeio; *fazer uma demonstração*; *pôr uma questão; ter em consideração*

Estas expressões apresentam grande variação a nível lexical e sintático, podendo observar-se, além da flexão:

(i) substituição lexical (*dar/fazer um passeio*; *pôr/colocar uma questão*).

(ii) inserção de modificadores (*dar um longo passeio*; *ter em devida consideração*);

(iii) modificação da estrutura sintática (*a demonstração foi feita*; *a demonstração que fizeram*; *fizeram-na*).

- **provérbios**

deitar cedo e cedo erguer dá saúde e faz crescer; quem te avisa teu amigo é; no poupar é que está o ganho (frases)

Estes dados põem em causa a nossa conceção dos provérbios e aforismos como unidades fixas da língua e levanta a questão sobre se existem de facto expressões totalmente invariáveis.

Expressões com significado parcialmente idiomático

- **expressões composicionais com significado adicional associado que não é derivável a partir dos significados dos constituintes** (Mel'čuk, 1998)

deitar as mãos à cabeça (+ desespero); *abrir a boca* (+ falar/bocejar) (verbal)

Nesta categoria estão apenas presentes as expressões verbais, que, além da flexão, podem apresentar inserção de modificadores (*deitou logo as mãos à cabeça; abriu finalmente a boca*) e substituição lexical (*deitar/levar/lançar as mãos à cabeça*).

- **nomes compostos**

a) expressões com significado adicional (cf. categoria anterior)

câmara de gás (+ morte); *campo de concentração* (+ trabalhos forçados/morte); *cinturão negro* (+ grau de habilidade em artes marciais)

b) expressões em que o significado idiomático resulta de uma combinação particular (o mesmo significado não volta a ocorrer quando o elemento em causa se combina com outros constituintes)

sorriso amarelo (amarelo = forçado); *arma branca* (branco = constituído de lâmina)*; drogas pesadas* (pesado = que causa dependência física e psicológica);

c) expressões em que o significado figurado resulta do estatuto polissémico dos seus constituintes (esta categoria distingue-se da anterior uma vez que o mesmo significado pode ocorrer quando o elemento em causa se combina com outros constituintes)

saúde de ferro; *vontade de ferro* (que lembra este metal pela sua resistência)

fio de sol; *fio de luz*; *fio de fumo* (linha contínua)

tédio mortal; *silêncio mortal* (intenso; profundo)

d) entidades (períodos históricos, personalidades, instituições)

dama de ferro; *guarda de ferro*

Apesar de estas expressões serem bastante coesas do ponto de vista léxico-sintático, observou-se que algumas estruturas podem apresentar inserção de modificadores (*sorriso muito amarelo*; *drogas mais pesadas*) e variação entre elementos simples e elementos com sufixos avaliativos (*sorrisinho amarelo*; *fiozinho de sol; saudinha de ferro*) e entre estruturas com modificadores adjetivais e preposicionados (*silêncio mortal* / *silêncio de morte*).

As expressões que denotam entidades são fixas e terão tido origem como as expressões apresentadas em c).

Expressões com significado idiomático

- **expressões transpostas para um campo semântico que lhes é alheio**

balde de água fria; pescadinha de rabo na boca; *faca de dois gumes* (nominal)

levar o barco a bom porto; *deitar água na fervura*; *estar de mãos e pés atados*; *ter olhos na cara* (verbal)

a ferro e fogo; a torto e a direito; a sangue frio (adverbial)

de cara lavada; de pedra e cal; *de se lhe tirar o chapéu* (valor adjetival)

As expressões com valor adverbial e adjetival são fixas. Nas restantes categorias, além da flexão, algumas expressões podem apresentar:

(i) substituição lexical (*faca/arma/espada/pau de dois gumes*; *levar/conduzir o barco a bom porto*; *deitar/pôr/colocar/lançar água na fervura*);

(ii) inserção de modificadores (*deitar alguma água na fervura*).

(iii) permutação de elementos (*estar de mãos e pés atados* / *estar de pés e mãos atados*);

(iv) alternância entre a ocorrência em estruturas sem negação e com negação (*ter olhos na cara* / *não ter olhos na cara*).

Note-se que, no caso das expressões verbais, apesar da ocorrência de variação, alguns constituintes são fixos (*água na fervura*; *o barco a bom porto*). Ressalvando-se esta caraterística, optou-se, no entanto, por analisar estas expressões na sua totalidade, uma vez que, em caso de substituição por sinónimos ou paráfrases, toda a estrutura é substituída (*deitar água na fervura* = acalmar).

- **nomes compostos**

flor de estufa; braço de ferro; pés de galinha; sangue fresco; prato forte; pera doce

Apesar da ocorrência de flexão, estas expressões apresentam um grau bastante elevado de fixidez, observando-se apenas alguma alternância entre elementos simples e elementos com sufixos avaliativos (*florzinha de estufa; bracinho de ferro*).

É importante salientar o facto de que a transposição de uma expressão para outro campo semântico (feito, normalmente, através de processos semânticos como a metáfora ou a metonímia) é um processo sincrónico e semanticamente motivado, pelo que, num determinado

momento, as estruturas apresentam simultaneamente o significado literal e o idiomático. No entanto, com o passar do tempo, as expressões podem perder o significado literal e, enquanto nalguns casos a motivação é facilmente reconhecida, noutros a interpretação do significado idiomático vai depender da capacidade que os falantes têm em recuperar essa motivação, capacidade essa que será feita em função de fatores culturais, sociais e idioletais. Note-se igualmente que, a nível estatístico, nos casos em que as expressões mantêm ambos os significados, tem-se verificado que o idiomático é aquele que ocorre com mais frequência.

- **expressões que contêm elementos que não ocorrem fora da combinação** (Vilela, 2002)
  sem chus nem bus; por um triz; de cor; por artes de berliques e berloques

  Esta expressões são fixas.

- **provérbios**

de noite todos os gatos são pardos; grão a grão enche a galinha o papo; *água mole em pedra dura tanto dá até que fura*

Tal como acontece com os provérbios composicionais, também neste caso é possível encontrar truncação (*água mole em pedra dura*) e substituição lexical (*de noite/à noite/à hora em que todos os gatos são pardos; grão a grão enche a galinha/muita gente o papo*).

Numa primeira análise da variação léxico-sintática destas expressões, é possível adiantar que as associações de palavras parecem apresentar comportamentos distintos de acordo com a sua estrutura sintática. Assim, enquanto as expressões frásicas (provérbios, aforismos, etc.), por norma, não permitem variação a nível sintático (a única variação que ocorre é lexical, nomeadamente a substituição de elementos, resultante da criatividade dos falantes), as expressões verbais admitem um elevado grau de variação a nível morfossintático, que, no entanto, parece ficar mais reduzido à medida que vão adquirindo um significado mais figurado. Por outro lado, as expressões nominais levantam problemas mais específicos, uma vez que os grupos composicionais têm um comportamento semelhante ao dos idiomáticos e nem sempre é fácil distingui-los. Os modificadores do nome podem denotar diferentes

relações (‘feito de’, ‘parte de’, ‘serve para’, etc.) que, por sua vez, podem definir o tipo de significado (literal ou idiomático) da expressão.

4. Anotação do léxico

O próximo passo da investigação consiste em anotar o léxico de associações de palavras referido na secção 2 com toda a informação constante na tipologia descrita na secção 3. Deste modo, cada uma das entradas das associações no léxico será enriquecida com informação respeitante a: (i) forma canónica da expressão (nos casos de ocorrência de substituição lexical, opta-se pela estrutura mais frequente); (ii) definição, no caso de expressões idiomáticas, através de sinónimos ou paráfrases; (iii) categoria gramatical da expressão e dos seus elementos; (iv) tipo de significado (composicional, parcial ou totalmente idiomático) e possíveis significados adicionais; (v) possível variação léxico--sintática; (vi) propriedade dos constituintes (obrigatórios, opcionais, livres, etc.).

Um dos principais objetivos da anotação deste léxico tem que ver com a criação de um recurso útil para a área da linguística computacional, uma vez que contribuirá para o desenvolvimento de sistemas de reconhecimento automático de associações de palavras em texto corrido. Nesta área, além do problema da idiomaticidade, a ocorrência de variação lexical e/ou sintática também levanta grandes dificuldades à correta deteção destas expressões. Uma vez que essa variação terá de ser devidamente anotada, apresentamos uma breve proposta de anotação, no léxico, de alguns casos de variação.

Variação lexical

- Inserção de modificadores – os elementos inseridos (que, normalmente, têm uma função enfática e não pertencem à forma canónica), não serão considerados constituintes da associação e não serão marcados no léxico (*dizer <u>sempre</u> cobras e lagartos*; *não ter <u>de facto</u> mãos a medir*; *meter <u>bem</u> a mão na consciência*).
- Substituição lexical – a variação é restringida a um grupo limitado de alternativas que serão marcadas como ‘constituintes obrigatórios da associação e membros da lista’ (*comer/vender/levar/comprar/tomar/obter/impingir gato por lebre*).
- Realização livre de constituintes – os elementos com diferentes realizações lexicais serão marcados, por exemplo, como pronomes (ALGUÉM, ALGUM) ou sintagmas específicos (NP,

PP) (*estar nas mãos de ALGUÉM*). Há, contudo, casos em que alguns constituintes podem variar livremente, enquanto outros permanecem fixos (*a educação é a mãe de todas as civilizações* / *a arte é a mãe de todas as ciências* / *a liberdade é a mãe de todas as virtudes*). Estes casos são anotados do mesmo modo que os anteriores (*ALGO é a mãe de todas as NOUN-PL*). No que respeita aos casos de criatividade lexical (observada, por exemplo, nos provérbios), os elementos que não pertençam à forma canónica serão marcados como tal ('diferente da forma canónica').

Variação sintática

- Ocorrência de pronomes e possessivos – os pronomes e possessivos que não façam parte da forma canónica da expressão serão marcados como pertencentes à associação, mas terão uma etiqueta de opcionalidade (*está nas suas mãos*; *está nas mãos dele*).
- Passivização – os verbos auxiliares não serão etiquetados como fazendo parte da associação (*passar ALGO a pente fino* / *ALGO foi passado a pente fino*).

No que respeita a aplicações práticas deste léxico, Hendrickx *et al*. (2010a) defendem a sua utilização para a anotação de expressões idiomáticas no *corpus* CINTIL[12]. As autoras propõem indexar as expressões encontradas no *corpus* à respetiva entrada no léxico. A indexação de cada associação à sua forma canónica (no léxico) permitiria detetar mais facilmente todas as ocorrências de uma determinada expressão e observar a sua variação no *corpus*. Para garantir a deteção de todas as variantes de uma expressão, o processo de anotação combinaria a identificação automática com a validação manual. A revisão manual de todo o *corpus* permitiria, igualmente, dar conta de expressões que ainda não estivessem incluídas no léxico. Nesses casos, cada nova expressão seria manualmente inserida no léxico.

5. Conclusão

Este artigo apresentou uma breve descrição do estudo que está a ser feito para identificar os diferentes tipos de associações de palavras do português europeu (com base num léxico extraído a partir de dados reais da língua) e descrever, tão detalhadamente quanto possível, o

[12] O *corpus* CINTIL (http://cintil.ul.pt) é um *corpus* de 1 milhão de palavras, composto por diferentes tipos de discurso escrito e oral e anotado com informação sobre a classe morfossintática, o lema, a flexão das classes abertas, as locuções pertencentes à classe dos advérbios e às classes fechadas e os nomes próprios multipalavra (para o reconhecimento de entidades nomeadas).

seu comportamento sintático-semântico, não descurando as suas propriedades pragmáticas. Durante a análise, deparámo-nos com dois desafios importantes: a avaliação do significado de algumas expressões (composicional ou idiomático) e a tentativa de dar conta de toda a variação lexical e sintática que as expressões pudessem apresentar. Toda a informação constante na tipologia apresentada será utilizada, posteriormente, para a anotação do léxico de associação de palavras utilizado. No que respeita à análise linguística presente na tipologia, pretende-se que esta encontre aplicações práticas em diversas áreas, como a lexicografia (o lexicógrafo poderá selecionar as expressões pertinentes e descrever o seu comportamento, tendo em conta o verdadeiro uso que os falantes fazem da língua), a psicolinguística (a hipótese de que o cérebro humano se encontra mais bem equipado para a memorização do que para o processamento conduz ao frequente recurso a estas expressões para a fluência do discurso e a simplificação da comunicação) ou a didática do português L2 (o uso da expressão apropriada torna o discurso mais natural). Relativamente ao léxico anotado, pretende-se que represente um valioso recurso na área da linguística computacional e que contribua para o desenvolvimento de ferramentas para a identificação automática destas expressões, aperfeiçoando-se, assim, os sistemas de tradução automática, extração e recuperação de informação, desambiguação de palavras, sumarização e geração automática de textos.

Referências

Abalada, S.; Cabarrão, V. & Cardoso, A. (2010) Proposta de Classificação Semântica de Unidades Lexicais Multipalavra Nominais. In *Textos Seleccionados do XXV Encontro Nacional da APL*. Porto, pp. 81-94.

Antunes, S.; Bacelar do Nascimento, M. F.; Mendes, A.; Pereira, L. & Sá, T. (2008) COMBINA-PT: uma base de dados de combinatórias lexicais do português. In *Textos Selecionados do XXIII Encontro Nacional da APL*. Évora, pp. 33-45.

Baptista, J. (1994) *Estabelecimento e Formalização de Classes de Nomes Compostos*. Dissertação de mestrado, Universidade de Lisboa.

Benson, M.; Benson, E. & Ilson, R. (1986) *The BBI Combinatory Dictionary of English: a guide to word combination*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Chacoto, L. (1994) *Estudo e Formalização das Propriedades Léxico-Sintácticas das Expressões Fixas Proverbiais*. Dissertação de mestrado, Universidade de Lisboa.

Church, K. & Hanks, P. (1990) Word Association Norms, Mutual Information and Lexicography. *Proceedings of the 27th Meeting of ACL.* Vancouver, Canada, pp. 76-83.

Corpas Pastor, G. (1996) *Manual de Fraseologia Española*. Madrid: Gredos.

Coseriu, E. (1967) Lexikalische Solidaritäten. *Poetica 1*. pp. 293-303.

Correia, M. & Lemos, L. S. P. (2005) *Inovação Lexical em Português*. Edições Colibri e Associação de Professores de Português, Lisboa.

Cowie, A. (1994) Phraseology. In Asher, R. E. (ed.) *The Encyclopedia of Language and Linguistics.* Pergamon, Oxford, pp. 3168-3171.

Cunha, C. & Cintra, L. F. L. (1987) *Nova Gramática do Português Contemporâneo*. Edições Sá da Costa, Lisboa.

Duran, M. S.; Ramish, C.; Aluísio, S. M. & Villavicencio, A. (2011) Identifying and Analyzing Brazilian Portuguese Complex Predicates. *Proceedings of the Workshop on Multiword Expressions*. ACL. Portland, Oregon, USA, pp. 74-82.

Evert, S. & Krenn, B. (2001) Methods for the Qualitative Evaluation of Lexical Association Measures. *Proceedings of the 39th Meeting of ACL.* Toulouse, France, pp. 188-195.

Fellbaum, C. (1998) *An WordNet Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Fernando, C. (1996) *Idioms and Idiomaticity*. Oxford University Press, Oxford.

Firth, R. John (1955) Modes of meaning. *Papers in Linguistics 1934-1951*. London, Oxford University Press, pp. 190-215.

Hausmann, F. J. (1989) Le dictionnaire de collocations. In Hausmann, F. J, H. E. Wiegand & L. Zgusta (eds.) *Wörterbücher, dictionaries, dictionnaires. Ein internationals Handbuch zur Lexikographie*. de Gruyter, Berlin, pp. 1010-1019.

Hendricks, I.; Mendes, A. & Antunes, S. (2010a) Proposal for Multi-word Expression Annotation in Running Text. *Proceedings of the 4th Linguistic Annotation Workshop.* ACL. Uppsala, Sweden, pp. 152-156.

Hendricks, I.; Mendes, A.; Pereira, S.; Gonçalves, A. & Duarte, I. (2010b) Complex Predicates annotation in a corpus of Portuguese. *Proceedings of the 4th Linguistic Annotation Workshop.* ACL. Uppsala, Sweden, pp. 100-108.

Jackendoff, R. (1997) *The Architecture of the Language Faculty*. The MIT Press, Cambridge, MA.

Macário Lopes, A. C. (1992) *Texto Proverbial Português: elementos para uma análise semântica e pragmática*. Dissertação de doutoramento, Universidade de Coimbra.

Mel'čuk, I. (1984) *Dictionnaire explicatif et combinatoire du français contemporain*. Les Presses de L'Université de Montreal, Montréal, Canada.

Mel'čuk, I. (1998) Collocations and Lexical Functions. In Cowie, A. (ed.), *Phraseology. Theory, Analysis, and Applications*. Oxford University Press, Oxford, pp. 23-53.

Mendes, A.; Antunes, S.; Bacelar do Nascimento, M. F.; Casteleiro, J. M. C.; Pereira, L. Pereira & Sá, T. (2006) COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. *Proceedings of the 5fth International Conference of LREC*. Genoa, Italy, pp. 1900-1905.

Mendes, A.; Généreux, M.; Hendrickx, I.; Pereira, L.; Bacelar do Nascimento, M. F. & Antunes, S. (2012) CQPWeb: uma nova plataforma de pesquisa para o CRPC. In *Textos Selecionados do XXVII Encontro Nacional da APL*. Lisboa, Portugal.

Moon, R. (1998) *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford Studies in Lexicography and Lexicology, Clarendon Press, Oxford.

Pereira, L. & Mendes, A. (2002) An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications. *Proceedings of the 10th International Congress of EURALEX*. Copenhagen, Denmark, vol. II, pp. 841-849.

Ranchhod, E. (2003) O Lugar das Expressões 'Fixas' na Gramática do Português. In Castro, I. & I. Duarte (eds.) *Razões e Emoção. Miscelânea de Estudos oferecida a Maria Helena Mira Mateus*. Imprensa Nacional Casa da Moeda, Lisboa, pp. 239-254.

Rio-Torto, G. (no prelo) Unidades Pluriverbais. In M. H. Moura Neves (org.) *As interfaces da gramática*. Araraquara, UNESP Editora.

Sag, I.; Baldwin, T.; Bond, F.; Copestake, A. & Flickinger, D. (2002) Multiword Expressions: A Pain in the Neck for NLP. *Proceedings of the Third International Conference of COLing*. Mexico City, Mexico.

Sanromán, A. I. (2000) *A Unidade Lexicográfica. Palavras, Colocações, Frasemas, Pragmatemas*. Dissertação de doutoramento, Universidade do Minho, Braga.

Sinclair, J. (1991) *Corpus, Concordance and Collocation*. Oxford University Press, Oxford.

Villalva, A. (2000) *Estruturas Morfológicas: unidades e hierarquias nas palavras do português*. In *Textos Universitários de Ciências Sociais e Humanas*. Fundação Calouste Gulbenkian, Lisboa.

Vilela, M. (2002) *Metáforas do Nosso Tempo*. Almedina, Coimbra.