COMBINA-PT:
uma base de dados de combinatórias lexicais do português

Sandra Antunes, Maria Fernanda Bacelar do Nascimento,
Amália Mendes, Luísa Pereira & Tiago Sá
Centro de Linguística da Universidade de Lisboa

1. Introdução

Apresentamos neste artigo os resultados de um projecto[1] que teve como objectivo o levantamento de combinatórias lexicais do português europeu a partir de um *corpus* escrito de 50 milhões de palavras, compilado a partir do *Corpus* de Referência do Português Contemporâneo (CRPC) do Centro de Linguística da Universidade de Lisboa (CLUL). Estas combinatórias consistem em padrões associativos de palavras apresentando diversos graus de fixidez e composicionalidade, desde expressões idiomáticas e fixas a co-ocorrentes privilegiados. O simples recurso à intuição e à introspecção é insuficiente para identificar as associações de palavras de uma língua, sendo determinante a utilização de um *corpus* de grandes dimensões.

O conceito de combinatória foi primeiro definido por Firth (1955) como consistindo na caracterização de uma palavra de acordo com as palavras que tipicamente com ela co-ocorrem. A definição do conceito tem variado consoante os autores e centra-se essencialmente em três propriedades fundamentais associadas à combinatória: a sua fixidez sintáctica, a sua idiomaticidade (embora este não seja um critério obrigatório para muitas abordagens teóricas) e o seu grau de associação lexical, isto é, a maior ou menor tendência para a co-ocorrência dos elementos que compõem o grupo. Para este último critério impõe-se o recurso a dados de frequências (do grupo, mas também de cada elemento do grupo) e, portanto, é essencial a disponibilidade de um *corpus* equilibrado de grandes dimensões (sendo difícil, por exemplo, obter dados pertinentes com um *corpus* de 1 milhão de palavras).

De acordo com a forma como as propriedades acima apresentadas são perspectivadas, as abordagens apresentam uma definição mais restrita ou mais lata de combinatória. Para alguns autores, as combinatórias constituem um tipo de associação de palavras com características restritas (Hausmann, 1979; Mel'cuk, 1984), enquanto para outros abrangem vários tipos de relações sintagmáticas (Sinclair, 1991).

[1] Projecto *Combinatórias Lexicais do Português* (COMBINA-PT), desenvolvido no Centro de Linguística da Universidade de Lisboa, com financiamento da Fundação para a Ciência e a Tecnologia (POCTI/LIN/48465/2002)

***Textos Seleccionados. XXIII Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, APL, 2008, pp. 33-45**

Iremos apresentar de forma mais pormenorizada, na secção 3, a perspectiva seguida no âmbito deste trabalho, mas interessa-nos desde já realçar um dado importante que está na base desta análise e que foi amplamente confirmado com os resultados obtidos, o qual consiste no facto de as línguas naturais seguirem padrões associativos regulares e complexos (Sinclair, 1991). Embora estando teoricamente disponível para o falante uma multiplicidade de escolhas livres em termos de associação de itens lexicais para a formação de frases e enunciados, a observação do uso da língua aponta para a utilização recorrente por parte dos falantes de determinadas sequências sintagmáticas que aparecem como pré-construídas. A isto chama Sinclair (1991) o **princípio idiomático**, sendo a combinatória definida por este autor como "the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening. Collocations can be dramatic and interesting because unexpected, or they can be important in the lexical structure because of being frequently repeated" (Sinclair, 1991: 170).

Os resultados obtidos com este trabalho pretendem contribuir para a compreensão deste princípio idiomático e da forma como se manifesta na língua portuguesa. Parece-nos um tema de especial interesse por se situar no interface entre o léxico e a sintaxe, estando as combinatórias num ponto de tensão entre uma análise enquanto constituintes sintácticos analisáveis, e portanto com estrutura interna reconhecida, e uma análise enquanto unidade não analisável do léxico. Iremos observar na secção 3 alguns dados que nos levarão a retomar esta questão.

A base de dados relacional Combina que aqui apresentamos contém as combinatórias de um conjunto de lemas seleccionados e, para cada combinatória, dá informação sobre a sua frequência, a distância entre os elementos da combinatória (quando se aplica), informação estatística que mede a força da associação lexical entre os elementos do grupo através do índice de Informação Mútua (*Mutual Information* (Church & Hanks, 1989)) e contextos da combinatória extraídos do *corpus*, formando um conjunto de concordâncias da expressão.

Os resultados foram manualmente validados e organizados (através de uma lematização não exaustiva), estando disponíveis para consulta na página do CLUL (www.clul.ul.pt).

Apresentam-se no ponto 2 deste artigo a metodologia seguida para a extracção automática dos grupos a partir do corpus (2.1), as ferramentas utilizadas (2.2) e a informação sobre o tratamento posterior dos dados na base relacional Combina (2.3). Em 3 é discutido o conceito de combinatória e em 4 expõe-se o trabalho de organização das combinatórias (que se aproxima do conceito de lematização). Na última secção serão apontados desenvolvimentos futuros e utilizações possíveis dos resultados.

2. Metodologia: *corpus* e ferramentas de extracção e selecção

2.1. *Corpus*

O *corpus* utilizado neste trabalho, e que designamos como *corpus* COMBINA, foi desenhado e compilado a partir do *Corpus* de Referência do Português Contemporâneo

(CRPC)[2], um *corpus* monitor com cerca de 330 milhões de palavras, desenvolvido no CLUL. O *corpus* COMBINA é um *corpus* escrito com 50 milhões de palavras que abrange os registos jornalístico e literário, incluindo, ainda, textos especializados de várias áreas (ver Quadro 1).

CONSTITUIÇÃO DO *CORPUS*			
JORNAL			**30.000.000**
LIVRO	literário	6.237.551	
	técnico	3.827.551	
	didáctico	852.787	**10.818.719**
REVISTA	informativa	5.709.061	
	técnica	1.790.939	**7.500.000**
VARIA			**1.851.828**
FOLHETO			**104.889**
ACÓRDÃOS DO SUPREMO TRIBUNAL			**313.962**
DIÁRIO DA ASSEMBLEIA DA REPÚBLICA			**277.586**
TOTAL			**50.866.984**

Quadro 1: Constituição do *corpus*

2.2. Ferramenta de extracção de combinatórias

Sobre o *corpus* COMBINA foi aplicada, em UNIX, uma ferramenta informática (Concor.cb) que extraiu todos os grupos compostos por 2 a 5 palavras, dando para cada grupo a seguinte informação:

- concordâncias de cada grupo (linhas de contexto no *corpus*);
- número de elementos do grupo;
- frequência do grupo;
- frequência dos elementos do grupo no *corpus*;
- distância a que ocorrem os elementos do grupo no caso de serem compostos por 2 palavras, podendo haver um máximo de 3 palavras entre os elementos do grupo:

(1) a. *conjuntura internacional* (dist. 1)
b. *conjuntura económica internacional* (dist. 2)

Formatadas: Marcas e numeração

Formatadas: Marcas e numeração

[2] O CRPC é um *corpus* escrito e falado compilado no CLUL desde 1998. Este *corpus* está em permanente actualização, inclui diversos tipos de textos e recobre todas as variedades regionais e nacionais do português. Para mais informações, ver a página do projecto em http://www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_crpc.php. Vários *subcorpora* do CRPC podem ainda ser consultados na página do CLUL.

- medida de associação lexical – Informação Mútua (IM), calculada a partir da relação entre a frequência do grupo e a frequência de cada elemento do grupo no *corpus*;
- ordenação dos resultados de acordo com esta medida.

Para reduzir a lista de grupos candidatos a combinatórias extraídos automaticamente, a ferramenta permite estipular uma frequência mínima de ocorrências definida pelo utilizador (para o projecto, foi estabelecida a frequência mínima de 3 para os grupos de 3 a 5 palavras e de 10 para os pares) e ainda, opcionalmente, eliminar grupos que incluam elementos de pontuação. Permite ainda eliminar grupos de duas palavras em que uma delas é uma palavra gramatical, de forma a limitar os resultados a combinatórias lexicais.

2.3. Base de dados relacional Combina

Foi em seguida desenvolvida uma base de dados relacional com plataforma SQL e interface em formato Access, que importa os resultados da extracção de grupos e permite representar as combinatórias, bem como proceder à sua selecção e organização (lematização parcial) e analisar as suas propriedades cruzando a informação dos vários campos (ver Figura 1).

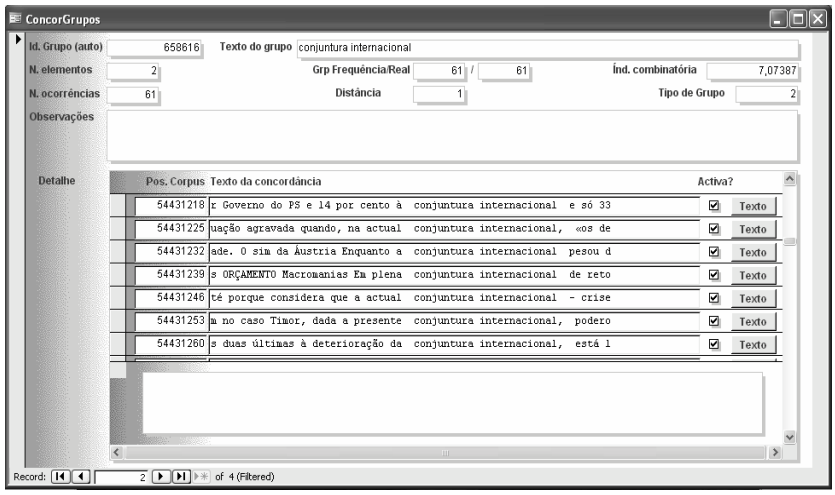

Figura 1: Base de dados COMBINA

Durante a fase de selecção dos grupos, a base de dados dá acesso a vários níveis de informação que são importantes para a decisão de considerar ou não determinado grupo como combinatória: para além dos dados estatísticos (valor de IM, designado "Índice de

Combinatória" na base de dados) e dos dados de frequência, é essencial a observação das ocorrências dos grupos no *corpus*. Casos de contextos que não contêm a combinatória em análise podem ser desactivados como exemplos dessa combinatória, sendo recontada a frequência do grupo, pelo que passa a constar da base a frequência inicial (que deu origem ao cálculo estatístico) e a frequência real após análise dos contextos.

3. Selecção de combinatórias: critérios e discussão

De acordo com a perspectiva *corpus-driven* subjacente a este projecto, pretendeu-se partir dos dados do *corpus* para a identificação dos tipos de combinatórias do português, em vez do percurso contrário que seria procurar no conjunto dos resultados obtidos determinados tipos de combinatórias, já previamente identificados.

Optámos, portanto, por trabalhar com um *corpus* não anotado ao nível das classes de palavras e por utilizar este primeiro trabalho sobre combinatórias do português como uma fase exploratória do próprio conceito em análise, das suas limitações e dos critérios pertinentes para a sua definição, a fim de se estabelecer uma tipologia das combinatórias a partir dos resultados obtidos. Veremos nas conclusões de que forma este projecto contribuiu para este objectivo e quais os desenvolvimentos futuros que pretendemos alcançar, com base neste trabalho preparatório.

Perante o elevadíssimo número de grupos automaticamente extraídos do *corpus* Combina (cerca de 1.7 milhões), optou-se por analisar e validar manualmente um conjunto desses resultados, com base nos valores de IM. Em estudos anteriores sobre este tema (Bacelar do Nascimento, 2000; Pereira & Mendes, 2002) e em estudos realizados durante a fase inicial deste projecto, foi evidenciado não só que os valores de IM mais elevados não são os mais significativos (facto já conhecido na literatura sobre este tema (Evert & Krenn, 2001)[3] como também o facto de os valores de IM entre 7 e 11 corresponderem a uma faixa com maior concentração de combinatórias.

Estes resultados da análise levaram-nos a seleccionar um conjunto de grupos com valores de IM entre 8 e 10 sobre os quais procedemos à primeira fase de selecção de combinatórias. A expansão do trabalho de selecção deu-se a partir desses grupos, sendo automaticamente criada na base uma lista de todas as palavras que neles ocorrem. Após lematização, essa lista constituiu, numa segunda fase, o conjunto de lemas a tratar ao nível das combinatórias em que ocorrem na língua portuguesa.

Como vimos, as definições de combinatórias centram-se essencialmente: i) na existência de uma maior fixidez do grupo (que se manifesta na dificuldade em substituir elementos lexicais do conjunto, em alterar a ordem sintagmática ou os traços de género e número e em admitir a não contiguidade dos elementos); ii) no facto de as combinatórias apresentarem especificidades semânticas que as tornam não composicionais; iii)

[3] Neste trabalho, Evert e Krenn apresentam um estudo comparado de algumas das principais medidas de associação lexical utilizadas, mostrando que o IM, embora dê resultados abaixo dos restantes no princípio da lista ordenada de candidatos a combinatórias, apresenta no entanto resultados semelhantes aos restantes índices nas secções seguintes da lista.

no facto de serem grupos extremamente frequentes, o que aponta para uma associação preferencial entre os seus elementos.

É com base nestes critérios que procedemos à análise dos grupos candidatos a combinatórias e iremos apresentar aqui alguns dados que se prendem com estes mesmos critérios. O nosso ponto de partida foi considerar que o conceito de combinatória exclui associações totalmente livres (embora nem sempre seja fácil delimitar esta fronteira), ficando por considerar que tipos de associações lexicais poderiam entrar neste conceito.

No que diz respeito à fixidez sintáctica, a literatura sobre o tema fornece pouca informação sobre a variação que as combinatórias podem admitir e qual o grau de liberdade sintáctica a partir do qual determinado grupo candidato deve ser considerado como uma associação livre. O trabalho sobre o *corpus* e sobre um conjunto tão grande de dados mostra, na verdade, níveis elevados de variação lexical e sintáctica nos grupos estudados. Além disso, as variantes flexionais das palavras do português contribuem em muito para a existência de variação nos grupos estudados, como se apresenta em (2):

(2) a. estou atento a
b. estamos atentos ao
c. estivemos atentos àquela

No caso de (2), para além das variações ao nível do verbo *estar* e do adjectivo *atento*, também encontramos variação que resulta da contracção da preposição *a*, que faz parte da combinatória, com artigos e pronomes, que lhe são exteriores. O estudo destas combinatórias implica abstrair da variação encontrada para determinar uma forma que reúne a maioria destes casos (*estar atento a*), mas sem nunca perder a informação relativa às variantes flexionais que de facto ocorreram no *corpus* e a partir das quais foi elaborado o índice de IM. Embora, por um lado, este índice ganhasse em ser calculado sobre a forma abstracta que abrange as várias realizações da combinatória, também é verdade que os dados mostram que, na generalidade dos casos, a combinatória só é realizada nalgumas ou apenas numa das variantes flexionais possíveis, sendo imprescindível ter em conta essa informação. Tentámos, no tratamento dos resultados, organizar os dados quando estes apresentavam variação, embora mantendo sempre informação sobre os grupos que de facto ocorreram, juntamente com a sua frequência e IM (ver secção 4).

a) Distância

Embora várias definições de combinatórias as considerem sequências contíguas de palavras, a definição proposta por Sinclair e citada no início deste trabalho deixa claro que as combinatórias podem ser compostas por elementos próximos no texto, embora não contíguos, estipulando Sinclair que poderão os elementos da combinatória estar separados por 4 palavras. Esta medida é aparentemente arbitrária, mas o trabalho sobre os dados do *corpus* mostrou que, na verdade, são poucos os grupos significativos com elementos separados por 4 palavras (distância 5), e mesmo por 3 palavras. No entanto, nestes casos, podem considerar-se combinatórias se os elementos do grupo, para além

de ocorrerem distanciados, também ocorrerem em contiguidade. Assim, no âmbito deste trabalho, optou-se por seleccionar apenas grupos com elementos não contíguos quando estes também ocorressem contiguamente (na distância 1) e apresentassem propriedades de combinatórias. Assim, por exemplo, foram seleccionados dois grupos *respirar fundo*, um na distância 2 (*respirar* bem *fundo*) e outro na distância 1 (*respirar fundo*). Aliás, quando os elementos de um grupo ocorrem a várias das distâncias possíveis, é no entanto claro que, tratando-se de uma combinatória, existe um pico de frequência numa dessas distâncias, o que aponta para uma realização preferencial da combinatória numa das variantes.

Existe, igualmente, não contiguidade dos elementos do grupo em casos em que o complemento de combinatórias como *pôr em causa* ocorre no interior do grupo, como (pôr alguma coisa em causa; pô-lo em causa). Uma vez que a identificação dos grupos é feita de forma automática, tal variação na forma que a combinatória assume faz com que esta ocorra em vários grupos com baixa frequência, o que pode levar a que a combinatória não apareça nos resultados, caso se estabeleça um limite mínimo de frequência baixo. Neste tipo de casos, em que não existe contiguidade e ocorre no interior da combinatória um elemento que dela não faz parte, seleccionámos os vários grupos para ser possível abranger as variantes de uma determinada combinatória, organizando-as posteriormente (ver secção 4).

b) Variação sintáctica

Algumas combinatórias mais fixas não admitem variação ao nível da ordem sintáctica dos seus elementos, como em casos paradigmáticos do tipo *esticar o pernil* ou *perder os sentidos*, que não aceitam passivização (3) ou relativização (4).

(3) a. * o pernil foi esticado
    b. * os sentidos foram perdidos
(4) a. * o pernil que foi esticado
    b. * os sentidos que foram perdidos

Estes são casos em que a não analisabilidade da combinatória é mais evidente e em que a fixidez sintáctica está acompanhada de idiomaticidade, isto é, a combinatória funciona como uma unidade tanto a nível sintáctico como semântico.

No entanto, a grande maioria dos grupos analisados ocorre em construções sintácticas como a passiva ou a relativa. Retomando a expressão *pôr em causa*, esta admite passivização (5), tal como *correr riscos* (6) que ocorre ainda com o nome *risco* no singular precedido de artigo definido e seguido de complemento preposicionado (7) e, no plural, numa construção relativa (8), precedido de artigo definido.

(5) ser posto em causa
(6) foram corridos riscos desnecessariamente
(7) correr o risco de
(8) os riscos que correm

Embora admitam maior variação sintáctica e mostrem, portanto, que são analisáveis internamente, estas são, no entanto, associações lexicais preferenciais, como mostra a dificuldade em substituir os elementos do grupo.

(9) # colocar em causa
(10) # sofrer riscos

c) Padrão léxico-semântico

Embora fosse objectivo deste trabalho restringir o levantamento às combinatórias lexicais do português, alguns dos grupos que constam da base de dados apontam para padrões sintáctico-semânticos que constituem estruturas gramaticais preferenciais, aspecto extremamente pertinente no âmbito do estudo das combinatórias (Sinclair & Renouf, 1991). Nestes casos, a variação lexical (que não permite identificar uma combinatória lexical regular de forma automática) aponta, no entanto, para um padrão léxico-semântico que pode ser lexicalizado através de vários tipos de estruturas. É o caso, por exemplo, do grupo *revelar pormenores* que nos contextos do *corpus* ocorre sempre precedido de um elemento ou de uma expressão com valor de negação (exemplos (11)), mostrando que a combinatória *revelar pormenores* ocorre no padrão: [*NEG*] *revelar pormenores*.

(11) a. **não** revelar pormenores
b. **sem** revelar pormenores
c. **escusando-se a** revelar pormenores
d. **Ainda é cedo para** revelar pormenores

d) Variação lexical

Mesmo as expressões geralmente mais fixas, como os aforismos, podem apresentar variações inesperadas, quando se observam os contextos do *corpus*. Por exemplo, a expressão *no poupar é que está o ganho* só ocorreu 3 vezes no *corpus*, isto é, ocorreu o número mínimo de vezes estabelecido neste projecto para a selecção de um grupo com mais de 2 palavras. Identificámos, no entanto, várias ocorrências de uma expressão aproximada, em que varia o primeiro elemento verbal (frases 12):

(12) a. no anunciar é que está o ganho
b. no atacar é que está o ganho
c. no descontar é que está o ganho
d. no prejuízo é que está o ganho
e. no esperar é que está o ganho
f. no provar é que está o ganho
g. no comparar é que está o ganho
h. no economizar é que está o ganho

Embora expressões como *no poupar é que está o ganho* sejam unidades fixas do nosso léxico mental, os dados mostram que os falantes são capazes de analisar componencialmente esta unidade, substituindo um elemento específico do grupo. Esta variação não põe em causa a natureza fixa da expressão aforística pois, quando confrontados com estas variantes, os falantes do português reconhecem a expressão canónica que está subjacente às frases apresentadas em (12). No entanto, estes dados põem em causa a nossa concepção dos aforismos e de outros tipos de expressões compostas por vários elementos lexicais como unidades totalmente fixas e não analisáveis e levanta a questão sobre se existem de facto combinatórias totalmente invariáveis. Levanta, ainda, outra questão importante, que é a da organização do léxico dos falantes, já que estamos perante expressões que têm propriedades de unidade lexical e de constituinte sintáctico.

e) Frequência e medida estatística de associação lexical

De facto, como foi referido, existe um conjunto alargado de expressões que apresentam possibilidades extensas de variação a nível flexional e sintáctico, embora manifestem dificuldades na substituição dos elementos que as compõem. Um grande conjunto dos grupos analisados não apresenta, de facto, alto grau de fixidez sintáctica nem idiomaticidade (e pode até admitir alguma variação lexical de um ou mais dos seus elementos). No entanto, são casos em que existe uma co-ocorrência preferencial dos elementos lexicais do grupo, quer por corresponderem a determinada realidade extra-linguística frequente (*pão com manteiga*) quer por corresponderem a uma forma preferencial de expressar determinado conceito (*onda de assaltos*). Nestes casos, a informação de frequência permite destacar uma expressão como especialmente frequente na língua e o índice estatístico IM pode mostrar que os elementos da expressão têm de facto tendência para ocorrerem juntos. Estas co-ocorrências preferenciais (*onda de assalto*, *absolutamente indispensável*) expressam com frequência relações semânticas que apontam para formas de organização do léxico (*insultos e ameaças*, *críticas e acusações*, *ganhos e perdas*, *públicas e privadas*, *trabalhadores e empregadores*).

Como se depreende da discussão sobre os critérios que estiveram subjacentes à definição do conceito de combinatória e sua selecção na base, optou-se no âmbito deste projecto por um conceito lato, que abrange expressões com diferente composição interna, diversos graus de fixidez e significados idiomáticos mas também composicionais. Assim, foram seleccionadas:

- expressões com valor de provérbio e de aforismo, que formam unidades frásicas completas e apresentam fortes restrições flexionais, lexicais e sintácticas (*quem canta, seus males espanta*), embora, como foi referido, estas também possam apresentar grande variação nos dados do *corpus*;
- expressões idiomáticas que apresentam grandes restrições lexicais e sintácticas, embora possam variar ao nível flexional (*esticar o pernil*, *perder os sentidos*);
- expressões parcialmente idiomáticas que admitem variação sintáctica (*correr riscos*);

- expressões composicionais que admitem variação lexical (*onda*/*maré*/*vaga de assaltos*);
- expressões composicionais que apontam para co-ocorrências preferenciais pela sua frequência e índice estatístico (*erros e imprecisões*; *defesa do consumidor*; *objecto de consumo*; *contacto telefónico*).

Os dados disponíveis no âmbito deste projecto estão a ser trabalhados no quadro de uma proposta de tipologia das combinatórias do português[4]. Neste trabalho, que também tem como objectivo apresentar uma proposta de tratamento lexicográfico destas unidades, irá propor-se uma organização destes vários tipos de expressões, de acordo com o seu significado – consoante seja composicional (soma dos significados literais dos elementos do grupo), resulte da conjunção de elementos literais e figurados, seja idiomático ou adquira, por exemplo, um uso pragmático específico – e com a relação que se estabelece entre estes tipos de expressões e as suas restrições formais ao nível flexional, lexical e sintáctico.

4. Organização das combinatórias em lemas

As combinatórias seleccionadas foram organizadas de forma a identificar, num primeiro nível, um lema de grupo que reúne as variantes flexionais que ocorreram no *corpus* sob uma forma única e, num nível superior, um lema principal que corresponde ao lema a partir do qual a combinatória foi seleccionada (cf. Quadro 2).

LEMA
abastecimento
LEMA DE GRUPO
posto de abastecimento
GRUPO
posto de abastecimento
num “Honda Civic”, assaltaram o **posto de abastecimento** “Galp”, i
riação, com carácter urgente, do **posto de abastecimento**. Há dez d
comercial portuguesa. Num outro **posto de abastecimento** local, os
GRUPO
postos de abastecimento
, afectado significativamente os **postos de abastecimento** localiza
de adição decorrer nos próprios **postos de abastecimento**, mas à r

Quadro 2: Organização da combinatória *posto de abastecimento*

[4] Tese de doutoramento de Sandra Antunes, inscrita na Faculdade de Letras da Universidade de Lisboa e desenvolvida com bolsa de doutoramento da FCT (SFRH/BD/24905/2005).

Como foi referido acima, em muitos casos só uma das possíveis variantes flexionais de uma combinatória está realizada no *corpus*. É importante manter essa informação no tratamento da combinatória, pelo que o lema de grupo corresponde, nestes casos, à forma que ocorreu no *corpus*[5]. Assim, os grupos *abastecimento de armas*, *motivos florais* e *pontos controversos*, por exemplo, que só ocorreram nestas formas específicas, têm lemas de grupo idênticos, não sendo portanto feita qualquer lematização.

Nos casos de combinatórias que admitem variação lexical num dos seus elementos, é frequente que a frequência de cada uma das variantes lexicais não seja suficientemente alta para que seja seleccionada automaticamente como grupo candidato a combinatória. No entanto, a parte invariável da combinatória pode atingir a frequência mínima necessária, permitindo que na fase de validação manual sejam identificadas as variantes do grupo. É o caso das combinatórias *ter em devida consideração* e *tomar em devida consideração*, que não ocorreram individualmente com a frequência mínima estipulada, mas que foram identificadas a partir dos contextos do grupo *em devida consideração*, o qual totalizava a frequência das duas combinatórias e foi, portanto, automaticamente seleccionado. Nestes casos, o lema de grupo atribuído integra as variantes lexicais realizadas no *corpus*, pelo que, no caso exemplificado, o lema atribuído é ***ter/tomar em devida consideração***.

5. Conclusão

Esta base de dados de combinatórias extraídas de um *corpus* escrito de português é um ponto de partida tanto para a discussão dos critérios fundamentais geralmente utilizados para definir o que constitui uma combinatória e a forma como estes critérios se combinam, como para a consequente revisão das tipologias de combinatórias existentes e do tratamento lexicográfico destas unidades. Os resultados poderão igualmente permitir avaliar os procedimentos de extracção automática dos grupos candidatos a combinatórias e as várias medidas de associação lexical existentes para validação das combinatórias significativas.

No futuro, pretende-se continuar este trabalho com base na revisão tipológica que esta primeira fase permitiu, anotando o *corpus* com informação de classe de palavras com o anotador desenvolvido pelo Grupo de Linguística de *Corpus* do CLUL e aprofundar a análise dos dados disponíveis com base em tipos produtivos identificados no âmbito deste projecto.

Para além de permitir uma reflexão sobre o conceito de combinatória, este trabalho poderá constituir, ainda, uma importante fonte de informação para áreas como a lexicografia, o ensino do português, a psicolinguística ou a linguística computacional.

[5] Assim, este processo não corresponde inteiramente à lematização, visto que não é aplicado de forma exaustiva.

Bibliografia

Antunes, S., M. F. Bacelar do Nascimento, J. M. Casteleiro, A. Mendes, L. Pereira, T. Sá (2006) A Lexical Database of Portuguese Multiword Expressions. In VIEIRA, R. et alii (eds) *PROPOR 2006*. Berlin: Springer-Verlag, LNAI 3960, pp. 238-243.

Bacelar do Nascimento, M. F. (2000) Exemples de combinatoires lexicales établis pour l'écrit et l'oral à Lisbonne. In Bilger, M. (ed.) *Corpus, Méthodologie et Applications Linguistiques*. Paris: H. Champion et Presses Universitaires de Perpignan 2000, pp. 237-261.

Braasch, A. & S. Olsen (2000) Toward a Strategy for a Representation of Collocations – Extending the Danish PAROLE-lexicon. *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece, 31 May – 2 June 2000, vol. II, pp. 1009-1016.

Butler, C. S. (1998) Collocational Frameworks in Spanish. *International Journal of Corpus Linguistics* 3(1), pp. 1-32.

Calzolari, N. et alii (2002) Towards Best Practice for Multiword Expressions in Computational Lexicons. *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Canary Islands; Spain, 29 May – 31 May 2002, pp. 1934-1940.

Church, K. W. & P. Hanks (1989) Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1), pp. 22-29.

Clear, J. (1993) From Firth principles: Computational tools for the study of collocation. In Baker, M., G. Francis & E. Tognini-Bonelli (eds.) *Text and technology: In honour of John Sinclair*. Amsterdam: John Benjamins.

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

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

Firth, J. (1957) A Synopsis of Linguistics Theory, 1930-1955. *Studies in Linguistic Analysis.* Oxford Philological Society. (reprinted in Palmer, F. (ed.) (1988) *Selected Papers of J. R. Firth*. Harlow: Longman.

Hausmann, F. J. (1979) Un dictionnaire des collocations est-il possible? In *Travaux de Linguistique et de Littérature XVII* (1).

Hausmann, F. J. (1989) Le dictionnaire des collocations. In Hausmann, F. J. et alii (eds.) *Wörterbücher: ein internationales Hanbuch zur Lexicographie. Dictionaires. Dictionaires*. Berlin/New-York: De Gruyter, pp. 1010-1019.

Kjellmer, G. A. (1994) *Dictionary of English Collocations*. Oxford: Oxford University Press.

Krenn, B. (2000a) CDB – A Database of Lexical Collocations. *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece, 31 May – 2 June 2000, vol. II, pp. 1003-1008.

Krenn, B. (2000b) Collocation Mining: Exploiting Corpora for Collocation Identification and Representation. *Proceedings of KONVENS 2000*, Ilmenau, Deutschland.

Krishnamurthy, R. (1997) Keeping good company: Collocation, Corpus and Dictionaries. In *Cicle de Conferències 95-96*. Institut Universitari de Lingüistica Aplicada, Universitat Pompeu Fabra, Barcelona, pp. 31-56.

Mackin, R. (1978) On collocations: Words shall be known by the company they keep. In *Honour of A. S. Hornby*. Oxford: Oxford University Press, pp. 149-165.

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

Mendes, A., S. Antunes, M. F. Bacelar do Nascimento, J. M. Casteleiro, L. Pereira, T. Sá (2006) COMBINA-PT: a Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. *Proceedings of the V International Conference on Language Resources and Evaluation – LREC2006*. Genoa, May 22--28 2006.

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

Pereira, L. A. Santos (1994) *Como se combinam as palavras? Contributo para um Dicionário de Combinatórias do Português*, Tese de Mestrado, Faculdade de Letras da Universidade de Lisboa.

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger (2002) Multiword Expressions: A Pain in the Neck for NLP. In Gelbukh, A. (ed.) *Proceedings of CICLing-2002*, Mexico City, Mexico.

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

Sinclair, J. & A. Renouf (1991) Collocational Frameworks In English. In Aijmer, K. and B. Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Harlow: Longman, pp. 128-143.

Viegas, E., S. Beale & S. Nirenburg (1998) The Computational Lexical Semantics of Syntagmatic Relations. In *Proceedings of the 17th International Conference on Computational Linguistics*. Montreal, Quebec; Canada, Volume II, pp. 1328-1332.