Typologies of MultiWord Expressions Revisited: A Corpus-driven Approach[1]

Maria Fernanda Bacelar do Nascimento, Amália Mendes, Sandra Antunes
Centro de Linguística da Universidade de Lisboa
{fbacelar.nascimento, amalia.mendes, sandra.antunes}@clul.ul.pt

Abstract Multiword Expressions (MWEs) have been and are still a challenge for linguistic analysis, lexicography and natural language processing. Several typologies of MWEs have been proposed taking into account several parameters like, for example, its degree of cohesion, its internal variation and its compositional nature. However, the definition of a MWE is still controversial and typologies based on discrete categorization seem to fail to describe a phenomenon with such variation. In this paper, we plan to revise some typologies of MWEs using a corpus-driven approach and to analyse corpus findings and their relation to MWEs categorization.

1. Introduction

In the 50's, Firth (1955) firstly introduced the concept of collocation, defining it as the characterization of a word according to the words that typically co-occur with it. The increasing interest in the study of the lexicon (particularly the description and classification of lexical categories according to their different and possible meanings) allowed the development of several studies that showed that the lexicon does not consist mainly of simple lexical items but appears to be populated with numerous chunks, more or less predictable, though not fixed.

> "On the one hand, *bank* co-occurs with words and expressions such as *money*, *notes*, *loan*, *account*, *investment*, *clerk*, *official*, *manager*, *robbery*, *vaults*, (...). On the other hand, we find *bank* co-occurring with *river*, *swim*, *boat*, *east* (...)" (Hanks, 1987: 127, *apud* Church & Hanks, 1989: 76).

It became notorious that natural languages follow complex regular associative patterns and that the identification of such patterns would give important information on the meanings of the word and its actual uses (Sinclair, 1991). Once they start to be frequently repeated, these word associations tend to correspond to a conventional way of saying things, turning out to be an important aspect in the lexical structure of the language.

[1] This study was undertaken under the Project Word Combinations in Portuguese Language (COMBINA-PT), developed at the Centre of Linguistics of the University of Lisbon under a research grant of the Portuguese Ministry of Science (POCTI/LIN/48465/2002).

> “Several nouns are frequently qualified by the adjective *hard*. We talk of *hard luck*, *hard facts* and *hard evidence*. We can also talk about *strong evidence* but are unlikely to use *strong facts* or *strong luck*; *tough luck* but not *tough facts* or *tough evidence*; *sad facts* but not *sad luck* or *sad evidence*. Of course, it is always possible to depart from the normal patterns of English, so it is not claimed that *sad evidence* can not occur – just that it’s not worth following as a pattern.
> Note that in the above examples of *hard*, there are two rather different meanings. In *hard luck*, *hard* means *unfortunate*, but in *hard facts* and *hard evidence* it means *unlikely to be proved wrong*. Despite this, the patterns of collocation show that the near-synonym *strong* goes only with *evidence*. So, the patterns of collocation are not governed by meaning.” (Sinclair (1987), Introduction to the COBUILD Dictionary, apud Krishnamurthy, 1997: 44-45)

These lexical associations may present different degrees of cohesion, ranging from totally frozen groups, semi-frozen groups or just sets of favoured co-occurring forms. We will use the term Multiword Expressions (MWEs) to refer to this range of different word associations. A number of typologies of these MWEs have been proposed taking into account several parameters, like, for example, their degree of cohesion, internal variation or compositional meaning. However, the exact definition of a collocation or of a MWE is still controversial. While some authors clearly distinguish the phenomenon of collocations from other types of word associations and syntagmatic relations (Hausmann (1979) and Mel’cuk (1984)), others have a broader perspective (Sinclair, 1991). In section 2 we will present these different definitions of MWEs; in section 3 some typologies based on discrete categorization will be reviewed; section 4 will address the corpus-driven methodology; section 5 will discuss the corpus data and how it follows or challenges MWEs typologies.

2. Reviewing some definitions of MWEs

One of the criteria used by some authors to define a MWE relies on its meaning. In this way Hausmann (1979) and Mel’cuk (1984) define collocations as a conventional combination of words, whose meaning can not be predicted by the meaning of the words that compose it. In fact, for Hausmann (1979), a collocation is constituted by a base (*Basis*), that is semantically autonomous, and by a collocator (*Kollocator*) that needs the base in order to get its full meaning. For the author, collocations consist of affine combinations of striking habitualness and have limited combinatorial capacity.

The author distinguishes 8 types of collocations according to the word class of its elements: (1) N + Adj (*célibataire endurci* 'confirmed bachelor'); (2) N$_{\text{(subject)}}$ + V (*la colère s'apaise* 'the anger wears off'); (3) V + N$_{\text{(object)}}$ (*tenir un journal* 'to keep a diary'); (4) V + Adv (*exiger énergiquement* 'to insist firmly'); (5) Adv + Adj (*gravement malade* 'critically ill'); (6) N + (prep) + N (*marché du travail* 'labour market'); (7) V + prep + N (*rougir de honte* 'to blush'); (8) Adj + N (*(dans un) proche avenir* 'in the near future').

Mel'cuk (1984) introduces the *Lexical Functions* (LFs) that describe the combinatory properties of lexical units (LUs) in a systematic way. In the process of text production, the speaker has to select lexical units to build his sentences. In this perspective, two types of LUs have to be distinguished: (i) LUs that are selected according to their meaning (**semantically-driven lexical choices**); (ii) LUs that are selected contingent on other LUs (**lexically-driven lexical choices**). This second type of choice is carried out along with two major linguistic relations: a **paradigmatic relation**, that subsume all substitution relations that may hold between lexical units in specific contexts (like the lexemes *young* and *tall*, that are paradigmatically related in the pairs of phrases *young student* and *tall student*), and a **syntagmatic relation**, that holds between lexical units that can co-occur in the same phrase or clause (like *boy* and *ran*, that are syntagmatically related in the phrase *the boy ran*).

"*Lexical Functions* (LFs) are a set of formal tools designed to describe, in a fully systematic and compact way, all types of genuine lexical relations that obtain between LUs of any language" (Mel'cuk, 1996: 38). Formally, LFs correspond to mathematical functions: f(x) = y (where x is the argument/keyword; y is the value).

Examples of LFs:

1. Adjectival LFs: **f** is intense/very; intensification → **Magn**
   a. **Magn**(*malade* 'ill') = *très* 'very', *gravement* 'critically'
   b. **Magn**(*dormer* 'to sleep') = *profondément* 'deeply', *comme une souche* 'like a log'
2. Verbal LFs
   a. **Oper$_1$**(*remarque* 'remark') = *faire* 'to make' [ART~]
   The keyword of **Oper$_1$** is its direct object (*faire un remarque* 'to make a remark')
   b. **Func$_1$**(*aider* 'help') = *vient* 'comes' [*de* 'from' N]

The keyword of **Func$_1$** is its grammatical subject (*l'aide vient de qn* 'aide comes from someone')

c. **Labor$_{12}$**(*note* 'note') = *prendre* 'to take' [N *en* 'in' ~]

The keyword of **Labor$_{12}$** is its indirect object (*prendre qc en note* 'to take note of something')

Also based in the meaning criterion, Cruse (1986) defines a collocation from a different point of view. For the author, "the term collocation will be used to refer to sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent" (Cruse, 1986: 37). The author exemplifies collocations with expressions such as *fine weather*, *torrential rain*, *light drizzle* and *high winds*.

However, the author also points out that collocations also have a semantic cohesion that "is the more marked if the meaning carried by one (or more) of its constituent elements is highly restricted contextually, and different from its meaning in more neutral contexts" (op. cit.: 37). That is the case of *heavy* in expressions like *heavy drinker/smoker/drug-user*. This sense of *heavy* requires narrowly defined contextual conditions and for this sense to be selected, the notion of 'consumption' seems to be a prerequisite. The author claims that we are still in the realms of transparent sequences, because each constituent produces a recurrent semantic contrast:

1. heavy/light (He's a ___ smoker) = heavy/light (They were ___ drinkers)
2. drinker/smoker (He's a heavy ___) = drinker/smoker (They were light ___s)

Another criterion used to define a collocation relies on its fixedness. That is the criterion used by Benson et alii (1986) that claim that there are many fixed, identifiable, non-idiomatic phrases and constructions that may be called recurrent combinations, fixed combinations or collocations. For the authors, collocations fall into two major groups: **grammatical collocations** and **lexical collocations**. A grammatical collocation is a phrase consisting of a lexical word (noun, adjective or verb) and a grammatical word (preposition, article or conjunction), like the expressions *account for*, *adapt to*, *agonize over*, *aim at*. The authors distinguish this type of collocations from what they call **free combinations**, that "consist of elements that are joined in accordance with the general rules of English syntax and freely allow substitution" (Benson et alii, 1986: ix), such as *after lunch*, *at three o'clock*, *in the library*, *on the boat*, that may have a limitless number of possible combinations. Lexical collocations, in contrast to grammatical collocations, are exclusively composed by lexical words, such as *warmest regards* (Adj + N) or *commit murder* (V + N). These are expressions with a high degree of cohesion, since, in the first case, we can not have **hot regards* or **hearty regards*, and, in the second case, the verb *commit* is limited in use to a small number of nouns meaning 'crime' or 'wrongdoing'. The authors also distinguish this type of collocations from **free lexical combinations**, in which the elements are not bound specifically to each other and may occur with other lexical items freely (the expression *condemn murder* is considered a free combination since the verb *condemn* may occur with an unlimited number of nouns, such as *abortion*, *abduction*, *abuse of power*, etc.).

Finally, Sinclair (1991) considers that in order to explain the way in which meaning arises from language text we have two principles of interpretation: **the open-choice principle** (where the speaker has a very large number of complex choices and the only restraint is grammaticalness) and **the idiom principle** (where the speaker has available a large number of semi-preconstructed phrases that constitute single choices, reflecting a natural tendency to economy of effort). In fact, it has been observed that the speaker actually uses his memory and routine, and that his discourse corresponds to single choices presented in the idiomatic principle. For the author, collocations illustrate the idiom principle. Words appear to be chosen in pairs or groups and these may not be necessarily adjacent. According to the author, a collocation is "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).

3. Reviewing some typologies of MWEs

As can be seen above, the different definitions of a collocation presented by different authors show that this is not a consensual topic and this controversy is also reflected in the different proposals of typologies.

Hausmann (1989) proposes the typology presented in figure 1.

$$\textbf{Word-combinations}\left\{\begin{array}{l}\text{fixed}\left\{\begin{array}{l}\text{idioms (ex : \textit{laver la tête à qn})}\\ \qquad\qquad\text{' to reprimand someone'}\end{array}\right\}\\ \text{non} - \text{fixed}\left\{\text{combinations}\left\{\begin{array}{l}\text{counter - affine} \rightarrow \text{counter - creation (ex : \textit{la route se rabougrit})}\\ \qquad\qquad\qquad\qquad\text{' the road languished'}\\ \text{affine} \rightarrow \text{collocation (ex : \textit{faire une promenade})}\\ \qquad\qquad\qquad\text{' to take a walk'}\\ \text{free} \rightarrow \text{co - creation (ex : \textit{acheter une maison})}\\ \qquad\qquad\qquad\text{' to buy a house'}\end{array}\right\}\right\}\end{array}\right\}$$

Figure 1: Hausmann’s classification of word-combinations

This typology relies essentially in the distinction between fixed and non-fixed expressions. Whether the first one only comprises the idioms, the second registers three types of non-fixed expressions ranging from counter-creations (or “poetic metaphors” (Lakoff, 1993)), collocations (cf. Hausmann’s definition in section 2.) and co-creations (semantically motivated combinations).

A different approach is introduced by Mel’cuk (1996) who distinguishes between free combinations (relations that hold between lexemes in a phrase with a purely compositional semantics) and non-free combinations (relations that hold between lexemes in a phrase whose semantics has to be partially or entirely derived from the phrase as a whole). In what non-free combinations are concerned, the author also distinguishes those which definitely do not have a compositional meaning from what he calls ‘pragmatemes’, i.e., pragmatically constrained combinations where the phrases in question are semantically freely composable but unexchangeable in specific contexts by any other synonymous expression (ex: *best before*).

Returning now to non-free combinations with non-compositional meaning, these are called ‘semantic phrasemes’ and are subclassified by the author into ‘full phrasemes’, or idioms (whose semantics is completely opaque and its meaning can not be obtained from the meaning of the constituent lexemes (ex: [*to*] *cool one’s head*; [*to*] *speal the beans*)), ‘quasi-phrasemes’ (whose semantics is partially obtainable from the meanings of its constituent lexemes, but contains, however, an additional meaning that can not be derived from those meanings (ex: *start a family*)) and ‘semi-phrasemes’, or collocations (cf. Mel’cuk’s definition in section 2.).

A general overview of Mel'cuk's typology is presented in figure 2.

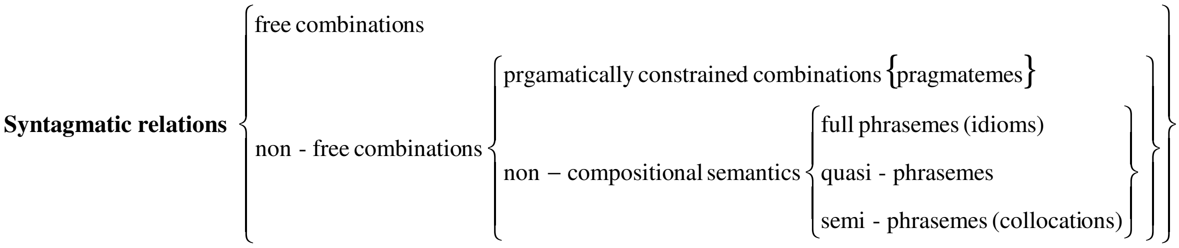


Figure 2: Mel'cuk's typology of syntagmatic relations

Viegas et alii (1998) argue for a continuum perspective, ranging from free-combining words (totally compositional meaning) to semantic collocations, idiosyncrasies and idioms (non-compositional meaning):

— **free-combining words** (*a wonderful man*);

— **semantic collocations** (*a fast car*; *a long book* (cf. Pustejovsky (1995) account of such expressions by the use of a coercion operator));

— **idiosyncrasies**

- **restricted semantic co-occurrence** (the meaning of the collocation is semi-compositional. "There is an entry in the lexicon for the base (...), whereas we cannot directly refer to the sense of the semantic collocate in the lexicon, as it is not part of its senses. We assign the co-occurrence a new semi-compositional sense, where the sense of the base is composed with a new sense for the collocate. (...) For instance (...), a *heavy smoker* is someone who smokes a lot, and not a 'fat' person. (...) we do not have in our lexicon for heavy a sense for 'a lot' (...)" (Viegas et alii, 1998:1329-1330);
- **restricted lexical co-occurrence** (the meaning of the collocate is compositional but has a lexical idiosyncrasy behavior. "(...) there are entries in the lexicon for the base and the collocate, with the same senses as in the co-occurrence. (...) What we are capturing here is a lexical idiosyncrasy or in other words, we specify that we should prefer this particular combination of words" (op. cit.: 1330). It is the case of expressions such as *rancid butter* or *sour milk*);

— **idioms** (*to kick the bucket*).

Finally, a more complex typology, created from a natural language processing point of view in order to avoid overgeneration, idiomaticity and parsing problems, is presented by Sag et alii (2002), and it covers the following types of expressions:

1. **Lexicalized Phrases** – word combinations that present at least partially idiosyncratic syntax or semantic or contain words which do not occur in isolation. They can be subclassified into:
   - **a) Fixed Expressions –** immutable expressions that are fully semantically and syntactically lexicalized, like *in short*, and that do not undergo neither morphosyntactic variation (**in shorter*) nor internal modification (**in very short*).
   - **b) Semi-fixed Expressions** – expressions that present constraints on word order and composition, but undergo some degree of lexical variation. These expressions can be subclassified into:
     - **(i) Non-decomposable Idioms** – expressions that have a non-compositional meaning (*kick the bucket*) and that are not subject to syntactic variability (**kick the great bucket*). The only type of variation observable is inflection (*kicked the bucket*).
     - **(ii) Compound Nominals** – Syntactically unalterable units that can inflect for number, like *car park* or *part of speech*.
     - **(iii) Proper Names** – Expressions syntactically idiosyncratic where one of the elements may be optionally ellidable (*the Oakland Raiders → the Raiders*).
   - **c) Syntactically-flexible Expressions –** Expressions that exhibit a much wider range of syntactic variability than fixed or semi-fixed expressions. These expressions can be subclassified into:
     - **(i) Verb-particle Constructions –** Constructions that consist of a verb and one or more particles, such as *look up* or *fall of*. In some cases these verb-particle constructions may take a NP argument between or following the verb and particle(s) (*call Kim up*; *call up Kim*). However, other cases are compatible with only one realizations (*fall of a truck*; **fall a truck of*). Adverbs can often be inserted between the verb and the particle (*fight bravely on*).

(ii) Decomposable idioms – Expressions that do not have a compositional meaning but tend to be syntactically flexible to some degree (*sweep under the rug*).

(iii) Light Verbs – Constructions highly idiosyncratic, like *make a mistake* or *give a demo*, where is difficult to predict which light verb combines with a given noun (**do a mistake*; **give a demo*). These constructions are subject to full syntactical variability, like passivization (*a demo was given*), extraction (*how many demos did Kim give*) and internal modification (*give a revealing demo*).

2. **Institutionalized Phrases –** Expressions that are syntactically and semantically compositional but statistically idiosyncratic, like *traffic light*, *fresh air* or *kindle excitement*. Given the strict compositionality, it would be expected the same concept to be expressible in other ways (like *traffic director* or *intersection regulator*). The idiosyncrasy of these expressions are statistical rather than linguistic in that they are observed with much more higher frequency than any other lexicalization of the same concept. As institutionalized phrases are fully compositional, they undergo full syntactic variability.

In order to provide a contribution for the study and classification of MWEs in Portuguese language, the project Word Combinations in Portuguese Language (COMBINA-PT), developed at the Centre of Linguistics of the University of Lisbon (CLUL), aims at the creation of a large lexical database of European Portuguese MWEs automatically extracted through the analysis of a large corpus of naturally occurring data, statistical interpreted with lexical associations measures and validated by hand. The availability of large amounts of textual data and corpus-driven analysis enables adequate descriptions of the concrete use of language, which would remain impossible if researchers only rely on introspection and native speaker intuition.

4. Corpus-driven methodology

For MWEs extraction, a corpus of 50M tokens was compiled, using a 330M tokens monitor corpus of Portuguese language developed at CLUL, the *Reference Corpus of Contemporary Portuguese* (CRPC)[2]. The COMBINA-PT corpus of 50M tokens is a balanced written corpus covering newspapers, books, magazines and journals and other documents (see Table 1 below).

CORPUS CONSTITUTION			
NEWSPAPERS			**30.000.000**
BOOKS	Fiction	6.237.551	
	Technical	3.827.551	
	Didactic	852.787	**10.818.719**
MAGAZINES AND JOURNALS	Informative	5.709.061	
	Technical	1.790.939	**7.500.000**
MISCELLANEOUS			**1.851.828**
LEAFLETS			**104.889**
SUPREME COURT VERDICTS			**313.962**
PARLIAMENT SESSIONS[3]			**277.586**
TOTAL			**50.866.984**

Table 1: Constitution of the corpus.

A program specifically developed to extract MWEs (CONCOR.CB) was then applied on the corpus in order to automatically extract all groups of 2, 3, 4 or 5 tokens. The following information is provided for each group:

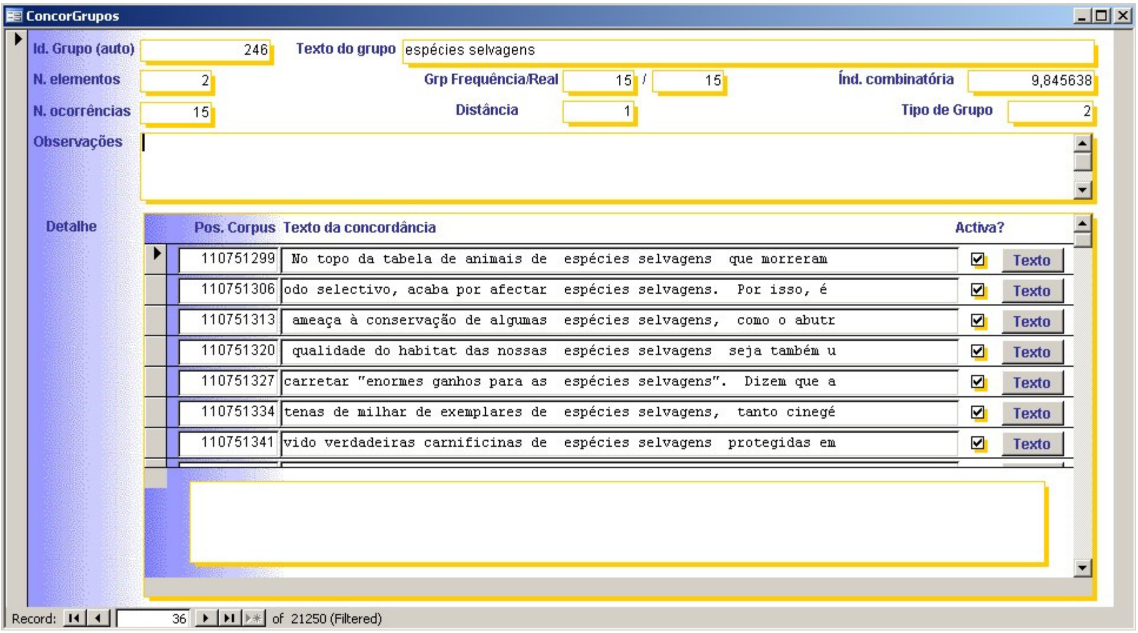
- Number of elements of the group;
- Distance between the group elements: groups of 2 tokens can be contiguous or be separated by a maximum of 3 tokens, while groups of more than 2 tokens are contiguous;
- Frequency of the group at a specific distance;
- Total frequency of the group in all occurring distances;
- Frequency of each element of the group;
- Total number of tokens in the corpus;
- Concordances lines (KWIC format) of the MWE in the corpus, together with an index code pointing to its exact occurring position in the corpus.

[2] CRPC is a written and spoken monitor corpus compiled at CLUL since 1998 and comprises all the national and regional varieties of Portuguese (http://www.clul.ul.pt/english/sectores/projecto_crpc.html).

[3] Parliament sessions are considered written data since the spoken sessions undergo extensive revision when transcribed.

- Lexical association measure: groups automatically extracted are statistically analysed using a selected association measure and are afterwards sorted. The tool allows the user to select which measure to apply, and was first run with Mutual Information (MI). MI calculates the frequency of each group in the corpus and crosses this frequency with the isolated frequency of each word of the group, also in the corpus (Church & Hanks 1989).

The large candidate list extracted from the corpus and the need of effective ways to reduce noise made it necessary to implement several cut-off options. With the first option we eliminated groups with internal punctuation, while with the second we eliminated word pairs with first or final grammatical word using a stop-list (to rule out non-lexical associations). The third option eliminated groups under a selected total minimum frequency: 4 for groups of 3 to 5 tokens, and 10 for 2-token groups. The final candidate list obtained still comprises the considerable number of 1.751.377 MW units. A lexical database was designed in MySQL format so as to enable the representation of MW units and to offer a platform for user-friendly manual validation. An example of a record represented in the database is presented in figure 3. For more information on the extraction and validation process, see Mendes et alii (2006) and Antunes et alii (2006).



Id. Grupo (auto): 246
Texto do grupo: espécies selvagens
N. elementos: 2
Grp Frequência/Real: 15 / 15
Índ. combinatória: 9,845638
N. ocorrências: 15
Distância: 1
Tipo de Grupo: 2
Observações:

Pos. Corpus	Texto da concordância	Activa?
110751299	No topo da tabela de animais de espécies selvagens que morreram	☑
110751306	odo selectivo, acaba por afectar espécies selvagens. Por isso, é	☑
110751313	ameaça à conservação de algumas espécies selvagens, como o abutr	☑
110751320	qualidade do habitat das nossas espécies selvagens seja também u	☑
110751327	carretar "enormes ganhos para as espécies selvagens". Dizem que a	☑
110751334	tenas de milhar de exemplares de espécies selvagens, tanto cinegé	☑
110751341	vido verdadeiras carnificinas de espécies selvagens protegidas em	☑

Record: 36 of 21250 (Filtered)

Figure 3: Record for the collocation *espécies selvagens* ‘wild species’ in the database

5. Typologies meet corpus data: definitions of MWEs challenged

As we saw in Section 1, definitions of MWE are essentially based on two fundamental criteria: syntactic fixedness and semantic non compositionality, although typologies of MWEs present different views regarding the second criterion since compositional groups are also viewed in some literature as being a type of MWEs. A third criterion, also much discussed, is frequency of occurrence and statistical information.

5.1. Lexical and Syntactic fixedness

The more restrictive definitions consider that a MWE must present a certain degree of syntactic fixedness, but the exact degree of variation that a MWE can undergo without ceasing to be one has not been established and, when working with corpus data, we find high levels of lexical and syntactic variation.

The first and obvious variation in a highly inflected language like Portuguese, as well as in other romance languages, is inflected variation, together with contractions of prepositions and articles/pronouns. For example, in the group *estar atento a* 'to be attentive to', the verb can vary in person, number and time, the adjective varies in gender and number and the prepositional element can be contracted with articles and pronouns, giving a large set of possibilities (e.g., *estou atento à* 'I'm attentive to_the[fem, sg]', *estamos atentos ao* 'we are attentive to_the[masc, sg]', *estivemos atentos àquela* 'we were attentive to_that_one[fem]' - contracted elements are connected in our English translation). To cover all possible realizations of the MW expression lemma *estar atento a* implies recovering and organizing all different word forms that the group comprises.

MWE fixedness is usually related to contiguous realization of the group elements, but when we observe corpus data, it becomes obvious that, especially in the case of MWEs including a verb form, non contiguity is extremely frequent. In most cases, an adverbial element can be inserted, like the group *respire fundo* 'breathe deeply', that also occurs as: *respire bem fundo* 'breathe very deeply'. In these cases, should we consider the existence of one MWE *respire fundo*, with possible variations, or of two independent MWEs? The question becomes even more difficult to answer when facing another group occurring in the corpus and clearly related: *respire profundamente* ‘breathe profoundly’.

With verbal expressions, contiguity is also challenged when verb complements occur inside the MWE: the MWE *pôr em causa* (‘to question’, literally: ‘to put in cause’) will require a direct object that will mostly occur in post-MWE position *pôr em causa* [*algo*] ‘to question [something]’, although it can be lexicalized inside the MWE (*pôr* [algo] *em* causa ‘to [something] question’) and pronominalized as well (as in the corpus occurrence *pô-**lo** em causa* ‘to question it’, literally: ‘to put it in cause’).

A similar process occurs in the case of the MWEs comprising possessive constructions, where the prepositional phrase expressing possession can be lexicalized as a possessive pronoun inside the MWE. For example, the following occurrences: *está nas mãos do governo* ‘(it) is in the hands of the government’, *está nas mãos da Assembleia* ‘(it) is in the hands of the Assembly’, *está nas nossas mãos* ‘(it) is in our hands’, *está nas vossas mãos* ‘(it) is in your hands’ are in fact all realizations of two abstract structures: *estar nas mãos de* [*X*] ‘to be in the hands of [X]’, *estar nas* [*POS*] *mãos* ‘to be in [POS] hands’, where the varying elements, the nominal phrase and the possessive pronoun, are expressed with placeholders. The two structures are obviously related and might be seen as corresponding to a single MWE with syntactic alternation, but it is also true that there is not always correspondence. For example, if the possessive element expresses the first or second person: *está nas minhas / tuas mãos* ‘(it) is in my / your hands’, the structure with prepositional phrase is not available: **está nas mãos de mim / ti* ‘(it) is in the hands of me / you’. As these examples show, hands-on work with a high number of MWEs candidate list raises the difficult question of determining, when faced with high lexical and syntactic variation, which MWEs are in fact realized in the corpus.

5.2. Syntactic Alternations

Just like in the precedent example of possessive constructions, fixedness is also challenged by the syntactic variation of most MWEs comprising a verb, since most admit syntactic alternations like passive or relative constructions. For example, the same expression *pôr em causa* [algo] can undergo passivization of the direct object: *ser posto em causa* ‘to be questioned’. The verbal expression *correr riscos* ‘to take chances’ can also undergo passivization, with elevation of *riscos* to subject position:

foram corridos riscos desnecessariamente ‘chances were taken unnecessarily’ (in this example, subject will preferably occur in post-verbal position). Relativization of the word *riscos* also occurs in the corpus: *os riscos que correm* ‘the chances that they take’, and takes morpho-syntactic variation even further since the MWE *correr riscos*, with no article, is then obligatorily realized with a definite article *os riscos* ‘the chances’. The MWE *correr riscos* also occurs with *riscos* in singular and preceded by definite article *correr o risco de* ‘to take the chance of’, so that three different lexical realizations are presented in the corpus: *correm riscos*, *os riscos que correm*, *correm o risco de*. While the second one is more directly related to the first, via relativization, and would be considered a variant of the MWE *correr riscos* (despite the insertion of a plural definite article), the third one will be considered a separate MWE *correr o risco de* [X], since *risco* occurs in singular form and is usually followed by a complement (prepositional phrase). While some variation corresponds to syntactic alternations of a MWE, other will point to the existence of another MWE, although clearly related to the first one.

5.3. Semantic patterns

Corpus data also shows cases where lexical variation in one specific position points to a specific semantic pattern that can be lexicalized as very different elements. For example, the verbal expression *revelar pormenores* ‘to reveal details’ always occurs in the corpus preceded by elements expressing a negative value, that can be a single adverb *não* ‘no’ or *sem* ‘without’ or complex sequences like *ainda é cedo para* ‘(it) is still early to’, variants of the structure [*NEG*] *revelar pormenores* ‘[NEG] reveal details’.

não	revelar pormenores	‘**not** to reveal details’
sem	revelar pormenores	‘**without** revealing details’
escusando-se a	revelar pormenores	‘**avoiding to** reveal details’
Ainda é cedo para	revelar pormenores	‘**(it) is still early to** reveal details’

Figure 4: Concordances of the expression [*NEG*] *revelar pormenores* ‘[NEG] reveal details’

These interesting patterns of semantic and syntactic co-occurrence go beyond lexical variation among the same morpho-syntactic category and point to the existence of MWEs that are a complex combination of fixed elements and a semantically constrained structural position.

5.4. Lexical variation

Definitions and typologies of MWEs usually associate fixedness and non compositionality as criteria for identifying MWEs. However, our corpus data show that MWEs considered as frozen, like idioms, can show a surprising level of lexical (and sometimes syntactic) variation. The following idiomatic expression *No poupar é que está o ganho* 'In the saving is the profit/Profit is in saving' forms a sentence that occurs 3 times in the corpus, while several other corpus occurrences show that one position inside this MWE allows large lexical variation:

No	**poupar**	é que está o ganho.	'Profit is in **saving**.'
No	**anunciar**	é que está o ganho.	'Profit is in **announcing**.'
No	**atacar**	é que está o ganho.	'Profit is in **attacking**.'
No	**descontar**	é que está o ganho.	'Profit is in **discounting**.'
No	**prejuízo**	é que está o ganho.	'Profit is in **losing**.'
No	**esperar**	é que está o ganho.	'Profit is in **wainting**.'
No	**provar**	é que está o ganho.	'Profit is in **tasting**.'
No	**cooperar**	é que está o ganho.	'Profit is in **cooperating**.'
No	**comparar**	é que está o ganho.	'Profit is in **comparing**.'
No	**economizar**	é que está o ganho.	'Profit is in **economizing**.'

Figure 5: Lexical variation of the expression *no poupar é que está o ganho* 'profit is in the saving'

Although expressions like *No poupar é que está o ganho* are clearly frozen in our mental lexicon, corpus shows that speakers do substitute some parts of the expression when using it. This does not undermine the idiomatic nature of the expression since, when confronted to the non canonical versions, Portuguese speakers immediately acknowledge that it is a version of a frozen expression. However, it does challenge our conception of idioms as the MWEs showing the highest degree of fixedness, leaving the question of whether there exist a totally frozen type of MWEs, that typologies consider to be at one end of the continuum of fixedness. This lexical variation of even the most idiomatic expressions raises questions regarding automatic identification of MWEs in the corpus: as mentioned in Section 4, a threshold was established as a cut-off measure, eliminating groups under the minimum frequency of four, and thus eliminating the expression *No poupar é que está o ganho*. (This expression can be recovered by the smaller group *é que está o ganho*, occurring twelve times.)

This internal lexical variation shows clearly that this MWE, although perceived as a single unit, do have internal structure and is analysed as such by the speakers.

5.5. Non-compositionality

The task of determining whether a MWE has compositional or non compositional meaning is also not straightforward in many cases. Non compositional meaning would imply that the meaning of the expression is not equivalent to the sum of the words individual meanings. However, in cases like *preencher um vazio* 'to fill emptiness [in a psychological sense]', the MWE can be considered compositional if the meaning of *preencher* 'to fill' and *vazio* 'emptiness' are not assumed as being only physical, which they are not. Establishing the compositional nature of a MWE is thus a task that presumes that one knows what is the meaning or the meanings of each element of the group, not a smaller task.

Some expressions are still compositional but also gain a pragmatic value, like the case of *podes crer* 'you bet' (literally: (you) can believe) that really expresses that someone can believe what was previously expressed by another speaker, but that also expresses a subjective attitude from the speaker, an attitude of strong assertion in informal contexts of dialogue or conversations.

Since lexicalization is the result of a gradual process, a specific word sequence can present different degrees of cohesion, synchronically observable. For example, a sequence like *fazer a cama* can be: a free combination with compositional meaning (to built a bed); a fixed combination but still compositional since the meaning of the expression is deduced from the meaning of its elements (to make/arrange the bed); and a strongly lexicalized expression, with non-compositional meaning (to frame someone).

5.6. Frequency and statistical data

We mentioned above that frequency is a much discussed criteria for MWEs identification. When applied to MWEs like *no poupar é que está o ganho*, that we expected to be totally frozen but was not, the impact of frequency for the identification of this particular type of MWE is clearly negative, since low frequency of the group in its original form makes it non recognizable via frequency. However, in the case of MWEs that show a lower degree of lexical and syntactic fixedness as well as a compositional meaning, like the case of preferred co-occurring forms that correspond to a usual way of saying something, then frequency and statistical information is an important criteria to identify those lexical associations and is part of the definition of those units. Those MWEs tend to express semantic relationships: semantic domain sharing, like *insultos e ameaças* 'insults and threats', *críticas e acusações* 'criticisms and accusations', *competências e atribuições* 'competences and atributions'; antonymy, like *ganhos e perdas* 'profits and losses', *fixos e móveis* 'fixed and mobile', *públicas e privadas* 'public and private'; complementarity, like *trabalhadores e empregadores* 'workers and employers'; or adverbial intensification with a specific adverb *absolutamente indispensável* 'absolutely indispensable'.

Looking at MWEs occurring in the corpus also gives us important information on the most frequent types of MWEs. For example, in what concerns verbal expressions, two different kinds are extremely frequent: those involving a verb with its internal complement, like the MWE *correr riscos* 'to take chances', and those involving what is usually called a light verb, like *pôr em causa* 'to question', with the light verb *pôr* 'put'. However, a very infrequent type of verbal MWE is the one involving a verb and its subject, like the examples *correm rumores* 'rumours are flying around' and *os exemplos abundam* 'examples abound'.

5. Conclusion

Large corpus data gives us important information on MWEs since it makes visible lexical and syntactic variation that speakers are not always conscious of and challenge our intuitive native speakers' beliefs on the total fixedness of at least certain types of MWEs. This corpus-driven and usage-based information has two important consequences to the study of MWEs: a revision of the fundamental criteria that define what constitutes a MWE: fixedness, non-compositionality and frequency; the study of their applicability to different subtypes of MWEs.

Besides the issues on fixedness degree and compositional meaning, the study of these MW expressions allows to identify associative patterns that characterizes a word according to: (i) co-occurrence patterns (systematic co-occurrence with particular lexical items in a contiguous or non-contiguous form); (ii) grammatical patterns (systematic co-occurrence with a certain verb class, with specific temporal verb forms or with certain syntactic constructions); (iii) paradigmatic patterns (hyperonymy, homonymy, synonymy or antonymy phenomena); (iv) discursive patterns (strong associations in one language register can be a weak association in another register).

The ultimate goal is to establish a proposal of a corpus-driven typology of MWEs for Portuguese language taking into account the three main criteria discussed above, as well as morphosyntactic properties of the expressions.

References

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 (2006) *PROPOR 2006*, LNAI 3960, Berlin, Springer-Verlag, pp. 238-243.

Bahns, J. (1993) "Lexical collocations: a contrastive view", *ELT Journal*, 47:1, pp. 56-63.

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

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*, vol. 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 and E. Tognini-Bonelli (eds.) *Text and technology: In honour of John Sinclair*, Amsterdam, John Benjamins.

Cruse, A. (1986) *Lexical Semantics*, Cambridge, Cambridge University Press.

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.

Heid, U. (1998) “Towards a corpus-based dictionary of German noun-verb collocations”, *Euralex 98 Proceedings*, Université de Liège, Belgique.

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.

Lakoff, G. (1983) “The Contemporary Theory of Metaphor”, in Ortony, A. (ed.) *Metaphor and Thought*, Cambridge, Cambridge University Press, pp: 202-251.

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*, Les Presses de L'Université de Montréal, Montréal, Canada.

Mel'cuk, I. (1996) "Lexical Functions : A Tool for the Description of Lexical Relations in a Lexicon", in Wanner L. (ed.), *Lexical Functions in Lexicography and Natural Language Processing*. Studies in Language Companion Series (SLCS), Amsterdam/Philadelphia, John Benjamins Publishing Company, pp. 37-102.

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.

Pearce, D. (2002) "A Comparative Evaluation of Collocation Extraction Techniques", *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain, pp. 13-18.

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*, M.A. Thesis, Faculty of Letters, University of Lisbon, ms.

Pustejovsky, J. (1995) *The Generative Lexicon*, Cambridge/Massachussets, The MIT Press, Massachussets Institute of Technology.

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

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

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.