Collocations in Portuguese: A corpus-based approach to lexical patterns

Amália Mendes, Centro de Linguística da Universidade de Lisboa
Sandra Antunes, Centro de Linguística da Universidade de Lisboa

Collocations and, more generally, multiword expressions, have been extensively studied for the English language and a large set of resources are available in terms of linguistic description and tools for language learning. On the contrary, combinatorial resources for Portuguese are scarce, although specific types of collocations, such as light verb constructions, nominal compounds and proverbs, have been the topic of many studies. This chapter reviews different theoretical perspectives on multiword expressions and collocations in Portuguese and presents in more detail the results of the COMBINA-PT project, a corpus-based approach to the study of collocations.

1. Introduction

Lexicalized phrases, as sequences of two or more words that show some degree of fixedness and in some cases an idiomatic meaning, have been the subject of attention of phraseology as well as corpus-based studies (Mel'čuk 1998, Sinclair 1991). Their high frequency in texts has been singled out and a large set of lexical and lexicographic resources are available for the English language. The term multiword expression (MWE) is frequently used to encompass different types of lexical sequences that present some degree of lexicalization that ranges from fully lexicalized idioms to collocations, i.e., co-occurrences between two or more words that tend to be more frequent than expected based on the frequency of each element in a corpus. This may include a diversified set of MWE, such as collocations, nominal compounds, idioms, formulae, proverbs and light verb constructions (see, for instance, typologies presented in Sag et al. (2002) and Cowie (1994, 2001)).

Contrasting with English, combinatorial resources for Portuguese are scarce, although specific types of MWE have been the topic of linguistic studies, and corpus-based perspectives have recently been developed for the European and Brazilian varieties of Portuguese.

We will first review in section 2 the work undertaken on MWE in Portuguese in different grammatical areas and under different theoretical perspectives, ranging from morphology to corpus-based approaches on collocations and lexical bundles. In section 3, we present the results of the COMBINA-PT project, a corpus-based

approach to the study of collocations. In section 4, we briefly revise some experiments in automatic MWE detection and applications in the computational processing of the Portuguese language, as well MWE in the context of Portuguese as a second or foreign language.

2. Approaches to multiword expressions and collocations in Portuguese

The diversity of MWE in Portuguese is addressed in Bacelar do Nascimento (2013), that specifically focuses on the concept of lexicalization as a process that can be assessed and measured through several formal properties (such as lexical distribution, syntactic manipulation, word insertion and inflection) and the full or partial idiomatic nature of the sequence. These formal and semantic properties are complemented by statistical information which is crucial to assess cases that are semantically compositional, that are not fully lexicalized but do occur with unexpected frequency. In fact, lexicalization, being a gradual process, is best captured in terms of a continuum that ranges from free word combinations to preferred co-occurrences, and to totally fixed word sequences.

This results in the absence of clear-cut frontiers between different degrees of lexicalization that directly reflect in the difficulty in establishing non overlapping subtypes of MWE. The distinction between a free sequence of words and a collocation is certainly a challenge, as is the distinction between a collocation and a compound word. On the one hand, the role of idiomaticity is a point where approaches frequently differ: while some will only consider idioms, others will retain compositional sequences as MWE based on formal or statistical criteria – and, indeed, it is clear that many nominal compounds have no idiomatic nature. On the other hand, the formal criteria presented to assess an expression need to be understood as pointing to different degrees of lexicalization, instead of a dual category lexicalized vs. free occurrence. Furthermore, even fully lexicalized and idiomatic MWE that resist lexical and syntactic manipulation tests show surprising results when querying corpus data. The fact that the expression is stored as a lexicalized phrase in our mental lexicon is no guarantee of fixedness, mainly due to the speaker's capacity to "play" with new concepts and to irony effects. This was duly noted in Barlow (2000) when analysing the different syntactic realizations in corpora of the English set phrase *it ain't over until the fat lady sing*s, that illustrates the concepts of syntactic and conceptual blends (where *blending* is a general cognitive process involving the merger of formal and conceptual structures to produce new structures, cf. Turner and Fauconnier (1995)). In Portuguese, Bacelar do Nascimento et al. (2006) analyse, in the COMBINA corpus (a subset of the Reference Corpus

of Contemporary Portuguese, cf. Mendes et al. 2006), how the phrase *no poupar é que está o ganho* 'profit is in saving, lit: in the saving is the profit' is actually used. Besides 3 unchanged occurrences of the phrase, there were other 9 cases where the verb *poupar* was replaced by other verb forms: the synonym *economizar* 'to economize' but also quite semantically distinct verbs such as *atacar* 'to attack', *esperar* 'to wait' and *cooperar* 'to cooperate'. This does not undermine the fixed nature of the expression since, when confronted to the non-canonical version, native speakers immediately acknowledge that it is a version of a frozen expression. It does, however, provide support to consider that these set expressions are stored as such in the mental lexicon while their internal structure is still analyzable.

While Bacelar do Nascimento (2013) reviews the concept of MWE and its diversity, other studies have long focused on specific types of expressions that have undergone some degree of lexicalization. One such case is nominal compounds, which have received much attention in morphological studies of word formation through the process of composition. The attention here is focused on the internal morphological structure of the fixed phrase that reflects in terms of orthography, inflection in number and word stress (Villalva 2003), aspects that directly relate to the lexicalization process, but also to the compositional or non-compositional nature of the compound. Word formation through composition is at the interface between morphology and syntax, and its status in the studies of MWE is not consensual. While studies in phraseology plan to exclude compound words (i.e., the area of morphology) from the the set of linguistic phenomena to address (Moon 1998), corpus-driven perspectives will, on the contrary, include nominal compounds in the study of MWE while attempting to distinguish between collocations and compounds (Firth 1957; Benson et al. 1986a, 1986b; Sag et al. 2002). The difficulty in distinguishing clearly between compounds and free combinations is visible in the work of Baptista (1995) on nominal compounds. Lexicalization is viewed as a gradual process and the application of several criteria reveals different degrees of fixedness that range from totally free to totally fixed sequences. The consequence is that the list of compound nouns provided in Baptista (1995) reflects different levels of frozen expressions, some clearly associated to compounds, while many others are felt to be yet on the path to full lexicalization and usually referred to as collocations. The distinction between, on the one hand, free combinations and collocations and, on the other hand, collocations and compounds is indeed better grasped in terms of a continuum that is revealed by the number of criteria that apply.

Studies undertaken under the Lexicon-Grammar framework established by Gross (1975) have given much attention to frozen structures and sentences, as well as constructions with copulative and support verbs, under the general principle that units of meaning are in fact the sentences and not the words. Lexicalization and

idiomaticity are concepts taken into consideration in these studies and many of the structures that are described fall into subcategories of MWE, such as light verbs constructions that are considered a subtype of syntactically-flexible expressions in Sag et al. (2002). Although studies that focus on the support verb constructions do not target explicitly the study of fixed expressions, the fact that they provide tables with an extensive listing of the lexical elements that co-occur with these verbs as well as a detailed description of their formal and semantic properties gives important information in what concerns MWE. Ranchhod (1983, 1990) addresses structures with the Portuguese verbs *ser* and *estar* 'to be', while Chacoto (2005) describes the support verb *fazer* 'to make'. Ranchhod (1990: 147-148) discusses different levels of compound nouns that have a specific meaning and cooccur with *estar* 'to be', from completely frozen sequences such as *estar na crista da onda* 'to be at your top', *estar nos braços de Morfeu* 'to be asleep', to cases where one of the elements may form a paradigm, as in *estar na flor da idade / juventude / vida* 'to be at his/her prime, lit: to be in the flower of age / youth / life'.

The treatment of fixed expressions in the Lexicon-Grammar is discussed in Ranchhod (2003) in what concerns compound nouns, adjectives and adverbs, and also fixed sentences, described as phrase structures that show strong lexical and syntactic restrictions between the verb and at least one of its arguments (or "compound verbs" in Gross (1984), apud Ranchhod (1995: 247)). These fixed sentences, which present however a syntactic structure that is identical to sentences with free word combination, are organized in 14 classes, according to their internal structure. The same approach was applied to Brazilian Portuguese in Vale (2002) and 10 classes were established: the results are close to those of European Portuguese but there are nevertheless some specific syntactic and lexical aspects that deserve contrastive studies. A list of Portuguese proverbs and their lexical and semantic properties was provided in Chacoto (1994), also in the framework of Lexicon-Grammar. Under a different theoretical perspective, the annotation of complex predicates over the *CINTIL* corpus[1] included expressions with support verbs, sometimes referred to as light verbs (Mendes & Pereira 2010).

The lexicographic treatment of different types of (semi-)fixed expressions in Portuguese is addressed in Iriarte Sanromán (2000), following the model established by Igor Mel'čuk in the *Dictionnaire Explicatif et Combinatoire du Français*

1 *CINTIL-Corpus Internacional do Português* is a linguistically interpreted corpus of European Portuguese. It is composed of one million annotated tokens that were manually verified. The annotation comprises information on part-of-speech, open classes lemma and inflection, multiword expressions pertaining to the class of adverbs and to the closed POS classes, and multiword proper names, for named entity recognition <http://cintil.ul.pt>.

Contemporain (DEC) (Mel'čuk et al. 1984-1999). Lexicographic units above the single word are categorized as collocations, phrasemes, semi-phrasemes, quasi-phrasemes and pragmatemes, although the author points out that it is difficult to draw clear frontiers between them. Information on collocations are codified under the lemma of the node and it is argued that bilingual dictionaries should provide listings of equivalent collocations in both languages, either lexically identical or not (examples are given from Portuguese and Spanish).

While most of the approaches in morphology and in the Lexicon-Grammar framework are not intrinsically corpus-based, Corpus Linguistics brings another perspective on the data that reflects directly on MWE. The study of corpus data revealed a tendency for words to co-occur, even when the meaning of the sequence was still compositional, without idiomatic interpretation. While other studies still diverged in terms of including non idiomatic sequences into the category of "fixed expressions", corpus-based approaches drew the attention to the fact that language was composed of "prefabricated chunks" that were still semantically and syntactically analysable (Sinclair 1991). The availability of large sets of data for Portuguese enabled similar approaches and conclusions.

For the Brazilian variety of Portuguese, Sardinha (2014) sets out to verify the extent of the use of collocations (taken as statistically significant co-occurrents) in newspaper texts by using both a specialized newspaper reportage corpus (11,467 tokens) from the *Corpus Brasileiro de Variação de Registro* (*CBVR*; Brazilian Register Variation Corpus) and the general register-diversified *Corpus Brasileiro* (Brazilian Corpus) of 1.1 billion tokens. For each text, the set of potential collocates is compared with the significant collocates found in the reference corpus and the results point to 90 per cent presence of collocations and to minimal variation in the presence of collocations from text to text. Of course, the method used to select a significant collocation will affect the percentage of collocations encountered in texts (indeed numbers that are reported may vary considerably). In this case, the top 2,000 collocates are selected and sorted by logDice. Some of the results of this study bring about the challenge of distinguishing between co-occurrence and collocation: while *federal* is naturally felt as a collocate of *deputado* (*deputado federal* 'federal deputy'), the relation between *federal* and *anunciou* 'announced' is less evident. The CBVR corpus has also been the source of data for the study of lexical bundles, i.e. frequently occurring sequences of words in a corpus of texts from a single register. The concept of lexical bundle is strongly related to their discourse functions across registers and is categorized in three major classes proposed by Biber (2006) for the English language – stance expressions, referential expressions and special functions (politeness and inquiries). The results presented in Sardinha et al. (2014) show, on the one hand, that the 3 categories are successfully applied

to Portuguese and are thus language-independent, and, on the other hand, that lexical bundles vary considerably across register, a probable consequence of lexical routinization.

For European Portuguese, a lexicon of MWE was created based on a balanced corpus of 50 million tokens, a subcorpus of the larger Reference Corpus of Contemporary Portuguese of 311 million tokens. The MWE lexicon includes 14.153 canonical forms and their morphological and syntactic variations (Mendes et al. 2006). We present in section 3 a detailed discussion of several aspects involved in this work: The MWE's extraction process, the methodology adopted for the selection of the expressions and the categorization of the results.

3. COMBINA-PT – Word Combinations in the Portuguese Language

The main goal of this project[2] was to establish a lexicon of Portuguese significant lexical MWE that would cover different types of word combinations. These expressions were extracted from a balanced written corpus, using automatic extraction tools, followed by manual validation.

3.1. The corpus and the extraction tool

For the COMBINA-PT project we used a 50 million word written corpus that was extracted from the Reference Corpus of Contemporary Portuguese[3] (Généreux et al. 2012). Given that a particular word may co-occur with different lexical units according to the type of discourse in which it occurs (Firth 1957; Stubbs 2004), different types of discourse were considered at the time of the compilation of the corpus, so that we could uncover as many different patterns of cooccurrence of a lexical unit as possible. The corpus design is presented in Table 1.

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

3 The *Reference Corpus of Contemporary Portuguese* is a monitor corpus (Sinclair 1991) of 311 million words, constituted by sampling from several types of written and spoken text, comprising all the national and regional varieties of Portuguese <https://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc>.

Table 1. Corpus design

CORPUS DESIGN			
NEWSPAPERS			**29.344.736**
BOOKS	Fiction Technical Didactic	6.237.551 3.827.551 852.787	 **10.917.889**
MAGAZINES	Informative Technical	5.709.061 1.790.939	 **7.500.000**
MISCELLANEOUS			**1.851.828**
LEAFLETS			**104.889**
SUPREME COURT VERDICTS			**313.962**
PARLIAMENT SESSIONS			**277.586**
TOTAL			**50.310.890**

In order to identify the MWE, a PERL program was specifically developed that automatically extracts from the corpus groups of 2 to 5 tokens using Mutual Information (MI) as a statistical measure for sorting the results (Church & Hanks 1990; Pereira 1994)[4]. Several cut-off measures were applied during the automatic process:

(i) exclusion of 2-grams with initial or ending grammatical words, using a stop-list (since our goal consisted on the analysis of lexical associations only, we wanted to exclude grammatical associations (Benson et al. 1986a, 1986b));

(ii) exclusion of n-grams with internal punctuation;

(iii) exclusion of n-grams under a minimum criterion of frequency: 4 occurrences for groups of 3 to 5 tokens, and 10 occurrences for 2-grams;

(iv) 2-grams can be contiguous or be separated by a maximum of 3 tokens, while groups of more than 2 tokens are obligatorily contiguous.

4 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.

Table 2 presents the results of the extraction of the MWE *fio de prumo* 'plumb line'.

Table 2. Example of the MW unit *fio de prumo* 'plumb line'

# 6 **fio de prumo** 1 eg(3) og(6) ic(9.844055) fg(6) fe(1877 2290575 71) N(50310890)			
123962906	indicada, para cada ponto, pelo	fio de prumo;	- o @bsentido@b-
123962913	erces e alinhar as paredes com o	fio de prumo.	A casa gandaresa é
123962920	s bastavam para saber utilizar o	fio de prumo	e travar bem os ado
123962927	á o músico António Pinho Vargas,	fio de prumo	(móveis, design, ex
123962934	prumada do edifício, tendo-se o	fio de prumo	prendido num grampo
123962941	nosso comandante!: recto como um	fio de prumo.	Rico homem!... ALB

A wide range of information is available for each group, as illustrated in Table 2:

(i) distance (first number after the MWE in bold);

(ii) number of elements of the group (eg);

(iii) frequency of the group at a specific distance (og);

(iv) MI value (ic);

(v) total frequency of the group in all occurring distances (fg);

(vi) frequency of each element of the group (fe);

(vii) total number of words in the corpus (N);

(viii) concordance lines of the MWE, in KWIC format, together with a reference code corresponding to the position of the expression in the corpus.

A lexical database was designed in SQL, with an Access interface, so as to enable the graphic representation of MWE and to offer a platform for user-friendly manual validation and lemmatization. The candidate list is loaded into the database together with all the associated fields mentioned above. The record of the MWE 'plumb line' (cf. Table 2) in the database is presented in figure 1.

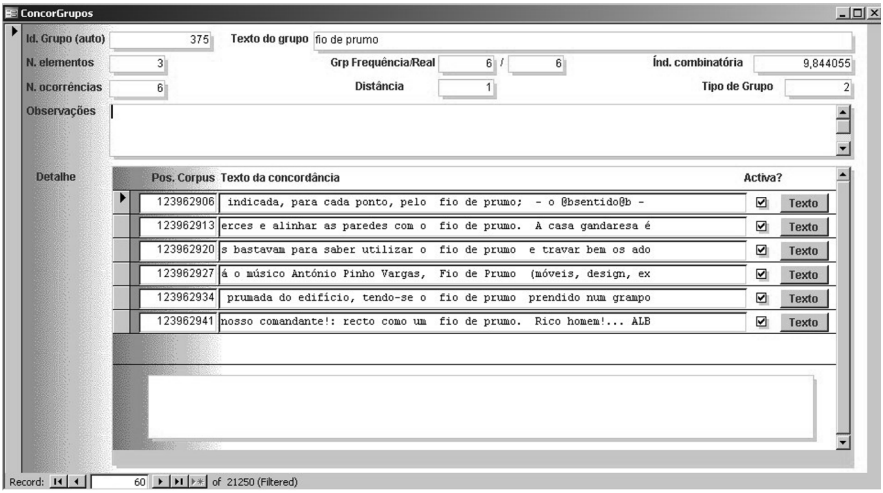

Figure 1. Record for the MWE fio de prumo 'plumb line' in the database

3.2. The selection of MWE: Manual validation and lemmatization

The large candidate list extracted from the corpus (1.7M groups) made it necessary to select only a part of these expressions in order to carry out an efficient manual validation. As previous experiments conducted in the automatic extraction and evaluation of MWE (Evert & Krenn 2001; Bacelar do Nascimento 2000; Pereira & Mendes 2002) showed that there was a higher concentration of good candidates around medium MI values, we first selected a list of nodes that occurred in n-grams with MI values between 8 and 10. We then manually inspected each n-gram that included one of these nodes, by applying several criteria upon which usually relies the definition of a MWE:

(i) lexical and syntactic fixedness, which can be observed through the possibility of replacing elements, inserting modifiers, changing the syntagmatic structure or gender/number features;

(ii) semantic cohesion, which can be observed through the total or partial loss of compositional meaning;

(iii) frequency of occurrence, which means that the expressions may be semantically compositional but occur with high frequency, revealing sets of favoured co-occurring forms, which could point to a path of lexicalization;

(iv) grammatical constituency (nominal, adjectival, adverbial or verbal phrases, or sentences).

Since discrete categorization is difficult to achieve (as has been proven by the abundance of typologies in the literature (Benson et al. 1986a, 1986b; Hausmann 1989; Cowie 1994; Mel'čuk 1998) and a fine-grained categorization of the large set of candidate data would be a time-consuming task to be performed at the initial stage of the work, we applied a first broad categorization into four types of expressions that takes into account the syntactic category and functional value:

(i) nominal, adjectival, adverbial or prepositional phrases that represent a particular function in the sentence or in the nominal phrase (e.g., nominal function: *ar puro* 'fresh air'; adjectival function: *cheio de nove horas* 'fussy'; adverbial function: *sempre a abrir* 'speedily');

(ii) verbal phrases or sentences (e.g., *contrair uma doença* 'contracting the disease'; *cautela e caldos de galinha nunca fizeram mal a ninguém* 'you can never be too careful');

(iii) named entities (e.g., *União Europeia* 'European Union');

(iv) groups that require further attention either because their status as a MWE is unclear, or because the group has more than 5 tokens (maximum of tokens extracted from the corpus), like proverbs, and will be later recovered (e.g.; *não há amor como o primeiro* 'there's no love like the first love').

The expressions that were selected were then indexed under a **group lemma**, i.e., a form that covers all inflected forms of the expression in the corpus (when the expression does not occur in any other inflected variant in the corpus, the group lemma is simply the form that occurred). Note that, as has been pointed out by several authors (Halliday 1966; Mitchel 1971; Sinclair 1991; Fernando 1996; Moon 1998), corpus analysis clearly shows that MWE have different types of internal variation. Following Moon (1998: 122), we assumed that some expressions "have fixed or canonical forms and that variations are to some extent derivative or deviant". The group lemma was then associated to a **main lemma** (single word), which is the lemma under analysis. To illustrate these options, consider the expressions *fruto silvestre* ('wild fruit') and *frutos silvestres* ('wild fruits') that are both associated to the group lemma *FRUTO SILVESTRE* ('wild fruit'). In turn, this group lemma is associated to the main lemma *FRUTO* ('fruit'), since the selection of this expression was undertaken during the analysis of that lemma. Figure 2 illustrates some MWE that were identified for the lemma *fogo* 'fire'.

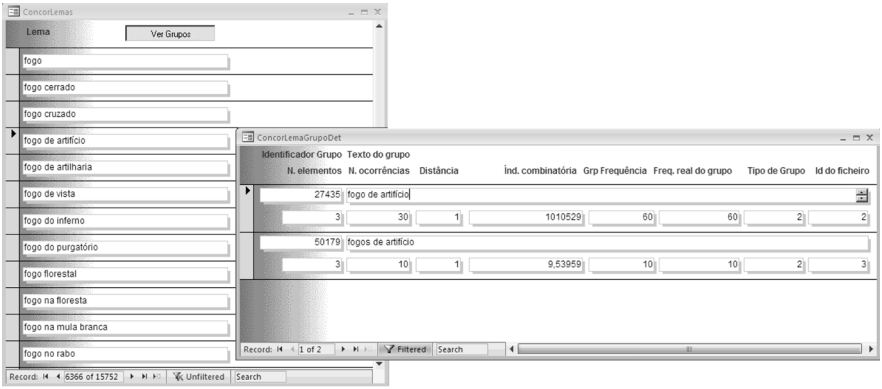

Figure 2. Example of MWE for the lemma *fogo* 'fire': *fogo cerrado* 'barrage fire', *fogo cruzado* 'crossfire', *fogo de artifício* 'firework', *fogo de artilharia* 'artillery fire', etc.

The COMBINA-PT lexicon comprises 1,180 main lemmas, 14,153 group lemmas and 48,154 word combinations[5].

3.3. Analysis of the data: Towards a typology

Although many theoretical models consider that MWE are typically fixed expressions, or expressions that present a small degree of variation, with (semi-)idiomatic meaning, corpus data clearly shows that MWE can vary in many different ways: from lexical to syntactical and structural variation. Furthermore, many expressions considered fixed (like proverbs) exhibit variation, and many compositional expressions are strongly lexicalized.

While manually inspecting the data, we identified several types of variation in both compositional and idiomatic expressions (Bacelar do Nascimento et al. 2006; Hendrickx et al. 2010) and we provide an account of such cases in the following sections, together with the option followed during the attribution of a group lemma.

3.3.1. *Lexical cohesion*

The analysis of the data revealed that lexical variation is one of the most productive

5 The lexicon is available at Meta-Share repository <http://www.meta-net.eu/meta-share> and on the CLUL's website <http://www.clul.ul.pt/sectores/linguistica_de_corpus/manual_combinatorias_online.php>.

types of MWE variation, and it includes lexical alternation, lexical insertion and permutation of elements and truncation.

There is a wide range of **lexical alternation**, where a word can be replaced by another word within a homogenous (synonyms, hypernyms) or heterogeneous lexical set. This alternation can occur with any type of grammatical category, although verb variation seems to be the commonest type. We encode this type of variation in the lexicon by listing the lexical items separated by a slash mark:

(i) alternation of verbs: *ter/adoptar/tomar uma posição* 'to have/to adopt/to take a position';

(ii) alternation of nouns: *arma/faca/lâmina/pau/espada de dois gumes* 'two-edged weapon/knife/blade/stick/sword';

(iii) alternation of adverbs: *ir direitinha/seguramente para o inferno* 'to go directly/surely to hell';

(iv) alternation of adjectives: *escuro/triste/negro como a noite* 'dark/sad/black as night';

(v) alternation of prepositions: *ter em/entre mãos* to have in/between hands' (meaning 'to take into hands')

Also, MWE are not always contiguous and they frequently allow **lexical insertion**. The inserted elements often have an intensification and quantification function, and the extra data corresponds, in most cases, to adverbs and adjectives: *dizer cobras e lagartos* 'bad-mouth, lit: to say snakes and lizards', *dizer sempre cobras e lagartos* 'always bad-mouth', lit: to say always snakes and lizards'; *um murro no estômago* 'a kick in the teeth', *um forte/autêntico murro no estômago* 'a strong/real kick in the teeth'. These expressions were lemmatized under a single form in the lexicon, which corresponded to the contiguous group. Some MWE also allow the insertion of articles and pronouns, and also the contraction of prepositions with these elements: *aprovação da moção* 'adoption of_the$_{[fem\ sing]}$ motion', *aprovação daquela moção* 'adoption of_that_one$_{[fem\ sing]}$ motion'. These expressions were lemmatized with no referential determiner, i.e., without any contraction.

When the expressions allow for the **permutation** of their elements (*estar de mãos e pés atados* / *estar de pés e mãos atados* 'to be tied hand and foot / to be tied foot and hand'), the assignment of the group lemma took into account the most frequent expression.

When MWE are formed by an extensive number of tokens, only part of the expression may be explicitly realized in the corpus. The **truncation** phenomenon usually occurs with proverbs: *deitar cedo e cedo erguer dá saúde e faz crescer*

'early to bed and early to rise makes a man healthy, wealthy and wise'; *deitar cedo e cedo erguer*… 'early to bed and early to rise …'. The group lemma of these occurrences is considered to be the canonical form (without truncation).

3.3.2. *Syntactic cohesion*

Since Portuguese (as well as other Romance languages) is a highly inflected language, the most common form of syntactic variation is **inflection of verbs and nouns**. Portuguese verbs vary on tense, mood, person and number (*pôr a nu* 'to lay bare', *põem a nu* 'they lay bare', *pôs a nu* 'he laid bare', *puseram a nu* 'they laid bare'), while nouns vary on number and gender (*pintado*$_{[masc\ sing]}$ *de fresco*, *pintada*$_{[fem\ sing]}$ *de fresco*, *pintados*$_{[masc\ pl]}$ *de fresco*, *pintadas*$_{[fem\ pl]}$ *de fresco* 'freshly painted'). Following canonical options in lexicography, the group lemma is the infinitive form, with verbal expressions, and the masculine and/or singular form, with nominal expressions. However, when nominal expressions occurred in only one form (regardless of which gender and number), the group lemma is the exact form that occurred (*aguaceiros fracos* 'light rains').

MWE that comprise possessive constructions may also **alternate between prepositional phrases (expressing the possession) or lexicalized possessives or pronouns**: *está nas mãos do governo* 'it's in the hands of the government', *está nas suas mãos* 'it's in his hands', *está nas mãos dele* 'it's in his hands, lit: it's in the hands of him'. These groups, which have slots that could be filled by several elements, were lemmatized with indefinite lexical elements, in capital letters (*estar nas mãos de ALGUÉM* 'to be in the hands of SOMEONE).

As one might expect, syntactic variation occurs especially with verbal MWE, since most admit syntactic alternations: (i) **nominalizations:** *conhecer do recurso* 'to hear an appeal', *conhecimento do recurso* 'hearing of the appeal'; (ii) **alternation between adjectival and prepositional modifiers:** *silêncio mortal* 'deadly silence', *silêncio de morte* 'silence of death'; (iii) **relativization:** *correr riscos* 'to take chances', *os riscos que correm* 'the chances that they take'; (iv) **passivization:** *passar ALGO a pente fino* 'to examine SOMETHING with a fine-tooth comb', *ALGO foi passado a* pente fino 'SOMETHING was examined with a fine-tooth comb'. The syntactic variation may involve changes in POS categories, as in nominalizations and alternations, and in such cases the expressions are codified under different group lemmas, while variation involving word order will keep the canonical group lemma.

3.3.3. *Structural cohesion*

Usually, pure idioms are lexically and syntactically fixed. However, the creative use of language leads to idioms' manipulation, where only part of the expression matches the canonical form (cf. the discussion of the expression *no poupar é que está o ganho* in section 2). But despite being different expressions, the speakers recognize them as a version of a particular frozen expression, and still perceived them as a single unit, with the same meaning as the canonical form: **canonical form**: *Deus escreve direito por linhas tortas* 'God writes straight with crooked lines'; **manipulated idiom**: *um jornalista escrevia direito por linhas tortas* 'a journalist wrote straight with crooked lines'. These expressions were lemmatized according to the exact form in which they occurred.

We also observed expressions with free slots, where part of the lexicon may vary without any apparent limits, while the other part remains fixed, as illustrated in Table 3 with the expression *NP[definite, singular, feminine] é a mãe de todas as N[plural, feminine]*.

Table 3. Example of an expression with free slots

a educação 'education'		*civilizações* 'civilizations'
a arte 'art'	é a mãe de todas as	*ciências* 'sciences'
a tecnologia 'technology'	'is the mother of all'	*vitórias* 'victories'
a liberdade 'liberty'		*virtudes* 'virtues'

3.3.4. *Semantic cohesion*

MWE also present different degrees of semantic cohesion that range from (i) totally compositional meaning, where all the words keep their literal meaning: *casamento de conveniência* 'marriage of convenience'; to (ii) partially idiomatic meaning, where some words keep their literal meaning while others have an idiomatic meaning that results from this particular combination and can not be replaced for any synonym: *saúde de ferro* 'excellent health, lit: iron health'; and to (iii) totally idiomatic meaning, where the meaning of the expression is not equivalent to the sum of the individual meanings of the words: *pés de galinha* 'crow's feet'. Also, since lexicalization is a result of a gradual process, expressions resulting from metaphoric processes may present different degrees of cohesion. The compositional and idiomatic meanings may still coexist and be synchronically observable, until they progressively depart from each other and, ultimately, the literal meaning may cease

to occur. It is the case of expressions such as *porto de abrigo*, which has the literal meaning 'harbor', and the idiomatic meaning 'safe haven'.

However, we are aware that determining whether a MWE has a compositional or idiomatic meaning is not a straightforward task, since one may find notorious difficulties regarding the evaluation of the meaning of certain expressions. This difficulty seems to be linked to two major factors: (i) the polysemous nature of the words (it is necessary to establish a boundary between compositional and figurative meanings, and if we adopt a restrictive definition, where the literal meaning of a word is the its first prototypical meaning, it will trigger us to consider a large number of MWE as idiomatic); (ii) the awareness of the semantic motivation that had led to the idiomatic meanings, which depends on cultural and social factors, and that will lead us to classify a lot of expressions as compositional.

Another difficulty that arises from semantic analysis, particularly regarding compositional expressions that may undergo full lexical and syntactic variability, is related to the perennial question of defining dividing lines between free combinations and MWE. For those cases, frequency and statistical data played a central role. In their classification of MWE, Sag et al. (2002) pay particular attention to 'institutionalized phrases', which they define as "semantically and syntactically compositional, but statistically idiosyncratic" (idem: 7). This means that, despite having compositional meaning and allowing lexical-syntactic variation, there are some expressions that occur with markedly high frequency than any other alternative lexicalization of the same concept, becoming a conventionalized way of saying things. This is particularly observable in binomials, which may include elements that share a specific domain (*fome e miséria* 'hunger and poverty'; *fumo e fogo* 'smoke and fire') or express an antonym relation (*públicas e privadas* 'public and private'; *guerra e paz* 'war and peace'), and that seem to be truly lexicalized. Also, there are some expressions with adverbial intensification where some adjectives usually occur with a specific adverb (*absolutamente indispensável* 'absolutely indispensable'; *completamente falso* 'completely false'). Frequency data also allow us to notice that there are some expressions where one element is preferred to any other synonym, which may reveal that they may be in their way to a possible fixedness (for example, *desvendar o mistério* 'to unravel the mystery' occurred with a much high frequency than the similar expressions *descobrir o mistério* 'to discover the mystery' or *resolver o mistério* 'to solve the mystery').

3.3.5. *Deriving a typology from corpus data*

As we mentioned in section 2, many proposals of MWE typologies use scales to best accommodate the kind of variation that we illustrate above: MWE are classi-

fied based on a continuum ranging from compositional expressions, with a higher degree of variation (free combinations), at one end, to totally idiomatic expressions, with a lower degree of variation (pure idioms), at the other end, with various intermediate categories (collocations, transitional collocations, compounds, etc.) that demonstrate the overlap between all the categories. As pointed out by Bolinger (1977: 168): "There is no clear boundary between an idiom and a collocation or between a collocation and a freely generated phrase – only a continuum with greater density at one end and greater diffusion at the other".

Given all the factors mentioned above, and as a further step in the analysis and description of these expressions, it is our aim to establish a typology for Portuguese MWE that encloses as much linguistic information as possible, such as: (i) syntactic (flexibility/fixedness); (ii) semantic (transparency/opacity); (iii) lexical (POS of the constituents and the expression itself); (iv) grammatical (grammatical function); (v) discursive (discourse function, such as organizing, informing, evaluating, etc.); (vi) pragmatic (frequency of occurrence). A preliminary proposal, presented in Antunes & Mendes (2013), tries to take into account different types of MWE, and is grounded on a semantic criterion that motivates a first level of categorization. A fine-grained categorization is driven by morphosyntactic and syntactic criteria (based on a continuum that ranges from non-fixed to fixed expressions) and discourse function (Cowie 1994, 2001; Mel'čuk 1998). In what concerns the morphosyntactic and syntactic variation, and considering, for instance, that even semi-lexicalized and idiomatic verbal expressions typically admit variation in verb inflection, a distinction based on the expression's POS was also established.

(i) Expressions with compositional meaning

a) favoured co-occurring forms / collocations / institutionalized phrases – nominal, adjectival or verbal expressions with full lexical and syntactic variation that occurred with higher frequency in the corpus than any other lexicalization of the same concept, revealing a tendency to co-occur in certain contexts: *razão especial* 'special reason' (87 occurrences) vs. *motivo particular* 'particular reason' (2 occurrences); *lufada de ar fresco* 'breath of fresh air' (35 occurrences) vs. *baforada de ar fresco* 'puff of fresh air' (3 occurrences); *condenar ao fracasso* 'to doom to failure' (26 occurrences) vs. *votar ao fracasso* 'to vote to failure' (6 occurrences);

b) restricted collocations – nominal, adjectival or verbal expressions that present a certain degree of lexical fixedness, since one of the constituents cannot be replaced by synonyms or semantically related words, while the other constituent may be part of a lexical set that fills a distributional paradigm: *concurso de fotografia*/*beleza*/*montras*/*etc.* 'photo/beauty/win-

dow dressing contest' vs. *#competição de fotografia/beleza/montras/etc.* 'photo/beauty/window dressing competition'; *pedir ajuda/informações* 'to ask for help/information' vs. *#perguntar ajuda/informações* 'to question for help/information';

c) compound nouns – expressions that represent a single concept and that do not allow lexical or syntactic variation, except for number/gender inflection or a very small variation in the distributional paradigm: *cama/camas de casal* 'double bed/beds', *cama/camas de solteiro* 'single bed/beds';

d) light verbs – V N expressions that may undergo full syntactic variation (passivization, relativization, lexical insertion), but where the noun is used in the normal sense while the verb meaning appears to be partially bleached. These expressions also present some lexical restrictions since it may be difficult to predict which verb combines with a given noun: *ter influência* 'to have influence', *ter muita influência* 'to have a major influence', *a influência que teve* 'the influence that it had', *#fazer influência* 'to make influence';

e) grammatical combinations / discourse connectives – expressions that represent prepositional, adverbial or conjunctional phrases, or other structures that function as text organizers: *por outro lado* 'on the one hand'; *no que respeita a* 'with regard to';

f) formulae (Cowie 1994) – expressions at the discourse level, with a pragmatic function, such as conversational formulae, which exhibit a high degree of fixedness: *com os melhores cumprimentos* 'best regards'; *até à vista* 'see you soon';

g) proverbs and clichés – expressions or sentences that give advice or express something that is generally true: *o que tem de ser tem muita força* 'whatever will be will be'; *de boas intenções está o inferno cheio* 'the road to hell is paved with good intentions'. Even though these expressions are traditionally considered lexically and syntactically fixed, as has been mentioned before, the creative use of language may lead to several kinds of idiom's manipulation.

(ii) Expressions with partial idiomatic meaning (Mel'čuk's quasi-phrasemes and semi-phrasemes)

a) verbal expressions, which may undergo full syntactic variation, and that have an additional meaning that can not be derived from the meaning of its parts: *deitar as mãos à cabeça* 'to put one's hands on the head' (it can also mean 'despair').

b) compound nouns:

(i) nominal expressions with an additional meaning: *campo de concentração* ‘concentration camp’ (it can also means ‘death’);

(ii) nominal expressions where the meaning of one of the constituents does not occur in any other combination: *sorriso amarelo* ‘yellow smile’ (yellow = wry);

(iii) nominal expressions where the meaning of one of the constituents may occur in different combinations (polysemous words): *café fresco* ‘fresh coffe’, *pão fresco* ‘fresh bread’ (fresh = recent);

(iv) periphrastic nouns (Iriarte Sanromán 2000): *continente negro* ‘black continent’ (Africa);

(v) named entities (public figures, historical periods, etc.): *dama de ferro* ‘iron lady’; *guarda de ferro* ‘iron guard’.

(iii) Expressions with idiomatic meaning

a) metaphors – expressions transposed to another semantic field by metaphoric process (except compound nouns). These expressions can show different degrees of fixedness:

(i) fixed expressions: *a ferro e fogo* ‘by fire and sword’;

(ii) expressions that only allow for inflectional variation: *tirar a barriga de misérias* ‘to eat your fill’, *tirámos a barriga de misérias* ‘we ate our fill’;

(iii) expressions that admit lexical variation, allowing the substitution of one of the constituents by other words (like a synonym, an hypernym or an antonym): *cair*/*vender*/*calhar*/*saber*/*ser que nem ginjas* ‘to fall/to sell/to happen/to taste/to be like sour cherries’ (meaning ‘to go down very well’);

(iv) compound nouns – nominal metaphoric expressions: *sangue fresco* ‘fresh blood’; *prato forte* ‘main dish’ (meaning ‘great advantage’). Besides inflection (*a simpatia foi um dos pratos fortes* ‘the sympathy was one of the great advantages’), these compounds also allow some variation, like lexical insertion (*precisamos de sangue mais fresco na equipa* ‘we need more flesh blood in the team’.

(v) proverbs: *grão a grão enche a galinha o papo* ‘grain by grain the hen fills its belly’ (meaning ‘little strokes fell great oaks’).

It should be noted that this semantic division implies that the same type of MWE may occur in different categories (such as compounds or proverbs, which can be found in the compositional or (partially) idiomatic categories), which could help to highlight the process of lexicalization and semantic reinterpretation of some expressions. Although we are trying to draw this dividing line, we are aware that the semantic evaluation of certain expressions does not allow accurate divisions.

4. Automatic identification of MWE and applications

The previous sections made explicit the difficulty in isolating MWE from free word combinations. It is especially the case with collocations, which present the lower level of lexicalization. Collocations are typically defined as a statistically significant co-occurrence of two (or more) words and the issue lies in defining the "statistical" criteria. The application of lexical association measures doesn't isolate collocations from non collocations, but rather gives us a sorted set of candidates. The question is then to establish where to draw the line. Several studies have evaluated and compared different methods of automatic extraction of MWE. However, as Villavicencio et al. (2007: 1034) point out, "given the heterogeneousness of the different phenomena that are considered to be MWEs, there is no consensus about which method is best suited for which type of MWE, and if there is a single method that can be successfully used for any kind of MWE".

The results of the COMBINA project provided a lexicon of MWE that were manually selected, based on their syntactic and semantic properties but also supported by frequency and Mutual Information, and this is a good starting point for us to evaluate automatic measures against manual inspection. The results showed that there was a considerable set of MWE that were selected although both MI and frequency were low (Antunes & Mendes 2014). The MI values between 5-10 account for around 50% of our gold dataset. Almost 25% of equally valid MWE receive values between 10-15, and almost 19% values between 1 and 5. A threshold between 5-15 MI value accounts for almost 80% of our gold dataset. However, lower values still include a high number of significant MWE, proving that one can actually find significant MWE throughout all the range of values, although in different proportions (notice that Evert & Krenn (2001: 190) pointed out that MI's precision remains almost constant or even increases slightly over the data). An automatic selection process would have to deal with the bottleneck of correctly identifying the remaining 20% of significant MWE. Looking at a specific set of lemmas, we compared MI values with T-test and Log-Likelihood: some collocations were better ranked with these two measures than with MI but results were quite similar in general. It is certainly necessary to combine different statistical measures for

a better automatic identification, in the sense that "the individual performances of these measures may well be improved if they are combined together, offering different insights into the problem" (Ramisch et al. 2008: 53). Nevertheless some cases of MWE seem to resist any statistical measure, although our intuition of native speaker and, to a certain degree, distributional criteria will pinpoint them as collocations. Another approach would then be to test different measures against the different subtypes established in our typology (cf. section 3).

Another comparison of statistical measures focused specifically on fixed expressions with idiomatic meaning compiled under the Lexicon-Grammar framework (Baptista et al. 2012). The expressions had the following syntactic constitution: (i) *N0 V Prep C1* (where *N0* stands for a free subject and *C1* represents a prepositional complement with one or more words, like *ir para o galheiro* 'to ruin'; *chegar a bom porto* 'to succeed'); (ii) *N0 V Prep C2* (where *C2* represents a complex nominal, like *ir para a quinta das tabuletas* 'to die'). The authors evaluated the use of T-test, χ^2 and MI for automatically identifying MWE from a 189M word newspaper corpus. However, the authors noted that approximately only half of the expressions of their list occur in the corpus (which probably results from the specific type of the expressions) and that that fact will hamper their identification based on statistical measures. Regarding the matching cases, the authors conclude that χ^2 presents better results than both T-test (which is not suitable for small data) and MI (which may be efficient regarding collocations, but is not appropriated for fixed expressions).

Apart from the need of an adequate description and typology of MWE in terms of lexicology and lexicography, the subset of fixed or semi-fixed and idiomatic expressions is a challenge for Natural Language Processing tools. Part-of-speech and parsing annotation require syntactic input to deal with those specific word sequences that should be analyzed as a single unit, and semantic information on the meaning of the expression. This leads us to propose a model specifically designed for the annotation of idiomatic MWE expressions in running texts, where idiomatic MWE in texts would be linked to their entry in the COMBINA-PT lexicon, enriched with syntactic and semantic classification and capable of storing information on the possible variation of the expression (at morphological, lexical and syntactic levels) (Hendrickx et al. 2010).

Finally, MWE prove to be crucial in L2 studies, as L2 learners frequently struggle to choose the right combination of words and eventually produce errors related to the lexical-grammatical, semantic or stylistic aspects of MWE (Nesselhauf 2005; Paquot 2013). The new learner corpus of Portuguese COPLE2 (Mendes et al. 2015) provided data for such an evaluation. A Contrastive Interlanguage Analysis of the subcorpus of Chinese, English and Spanish learners of Portuguese L2 pointed to the crucial effects of transfer in the production of MWE, especially in

what concerns collocations (Antunes & Mendes 2015).

5. Final remarks

We presented in this chapter an overview of some of the linguistic studies on MWE in Portuguese, trying to target both European and Brazilian results. We didn't aim to be exhaustive and it would certainly be an impossible goal to achieve due to the constraints of our work. We did however intend to provide a representative picture of the recent studies that targeted the topic of fixed expressions. The diversity of perspectives that are referred along the chapter is a direct consequence of the broad concept at hand. It is certainly inevitable when dealing with a phenomenon that is pervasive to morphology, lexicon and syntax. The contribution of different grammatical areas and theoretical approaches, despite the proliferation of terminology involved, provides new light over a difficult topic that seems to elude attempts of categorization. What every MWE has in common is the fact of undergoing a process of lexicalization, however fixed it may be. This process may apply to smaller or larger units, with different syntactic and discourse functions, subject to polysemy and metaphorical uses, as any unit of language. The multifaceted types of MWE come to no surprise under this point of view and help shed light to the process of lexicalization itself.

The availability of large corpora has complemented traditional typologies of fixed and idiomatic expressions by calling our attention to language use in terms of prefabricated chunks as expressed by the idiom principle (Sinclair 1991), but also by exposing the extensive variation allowed by expressions traditionally referred to by their canonical form.

While different studies have focused on Portuguese fixed expressions, and several teams have reported on experiments in automatically extracting MWE from corpora, this has not yet reflected in large available lexical and lexicographic resources that would benefit applications in the computational processing of Portuguese and applications in learning and teaching Portuguese as a foreign language.

Acknowledgments

This work was partially supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/LIN/UI0214/2013. The authors would like to thank the anonymous reviewers for their comments.

References

Antunes, Sandra & Amália Mendes 2013. MWE in Portuguese: Proposal for a typology for annotation in running text. *Proceedings of the Ninth Workshop on Multiword Expressions*, North American Chapter of the Association for Computational Linguistics, Atlanta, Georgia, USA, 87–92.

Antunes, Sandra & Amália Mendes 2014. An evaluation of the role of statistical measures and frequency for MWE identification. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, May 26-31, 4046–4051. Reykjavik, Iceland.

Antunes, Sandra & Amália Mendes 2015. Portuguese multiword expressions: data from a learner corpus. *Third Learner Corpus Research Conference*, poster. Nijmegen, The Netherlands.

Bacelar do Nascimento, Maria Fernanda 2000. Exemples de combinatoires lexicales établis pour l'écrit et l'oral à Lisbonne. *Corpus, Méthodologie et Applications Linguistiques*, ed. Mireille Bilger, 237–261. Paris: H. Champion et Presses Universitaires de Perpignan.

Bacelar do Nascimento, Maria Fernanda 2013. Processos de Lexicalização. *Gramática do Português*, orgs. Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Mota, Luísa Segura & Amália Mendes, Vol. I, 215–246. Lisboa: Fundação Calouste Gulbenkian.

Bacelar do Nascimento, Maria Fernanda, Amália Mendes & Sandra Antunes 2006. Typologies of multiword expressions revisited: A Corpus-driven Approach. *Spoken Language Corpus and Linguistic Informatics*, eds. Yuji Kawaguchi, Zaima Susumu & Toshihiro Takagaki, Coll. Usage-Based Linguistic Informatics, Vol.V, 227–244. Amsterdam/Philadelphia: Jonh Benjamins Publishing Company.

Baptista, Jorge 1995. *Estabelecimento e Formalização de Classes de Nomes Compostos*. MA Dissertation. Lisboa: Faculdade de Letras da Universidade de Lisboa.

Baptista, Jorge, Oto Araújo Vale & Nuno Mamede 2012. Identificação de expressões fixas em corpora: até onde podem ir os métodos estatísticos? *Caminhos da Linguística de Corpus*, eds. Tania M. G. Shepherd, Tony Berber Sardinha & Marcia Veirano Pinto, 177–190. Campinas, SP: Mercado de Letras.

Barlow, Michael 2000. Usage, blends and grammar. *Usage Based Models of Language*, eds. Michael Barlow & Suzanne Kemmer, 315–345. Stanford: CSLI.

Benson, Morton, Evelyn Benson & Robert Ilson 1986a. *The BBI Combinatory Dictionary of English: A guide to word combination*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Benson, Morton, Evelyn Benson & Robert Ilson 1986b. *Lexicographic Description of English*. Studies on Language Companion Series, Vol, 14. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Biber, Douglas 2006. *University Language: A Corpus-Based Study of Spoken and Written*

Registers. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bolinger, Dwight 1977. *Meaning and Form*. London/New York: Longman.

Chacoto, Lucília 1994. *Estudo e Formalização das Propriedades Léxico-Sintácticas das Expressões Fixas Proverbiais*. MA Dissertation. Lisboa: Faculdade de Letras da Universidade de Lisboa.

Chacoto, Lucília 2005. *O verbo* fazer *em construções nominais predicativas*. Ph.D Dissertation. Faro: Universidade do Algarve.

Church, Kenneth Ward & Patrick Hanks 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1): 22–29.

CINTIL-Corpus Internacional do Português. Faculdade de Ciências da Universidade de Lisboa e Centro de Linguística da Universidade de Lisboa. Version 1. 2006. http://cintil.ul.pt.

Corpus Brasileiro. Centro de Pesquisas, Recursos e Informação de Linguagem. Pontifícia Universidade Católica de São Paulo. 2013. Retrieved from http://corpusbrasileiro.pucsp.br/cb/Acesso.html. Last accessed 1 December 2015.

Corpus Brasileiro de Variação de Registro. Sardinha, Tony Berber 2014. Looking at collocations in Brazilian Portuguese through the Brazilian Corpus. *Working with Portuguese Corpora*, eds. Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira, 9–32. London: Bloomsbury.

Cowie, Anthony P. 1994. Phraseology. *Encyclopedia of Language and Linguistics*, ed. Ronald E. Asher, 3168–3171. Oxford: Pergamon.

Cowie, Anthony P. 2001. Speech formulae in English: Problems of analysis and dictionary treatment. *Making Senses: From Lexeme to Discourse. In Honour of Werner Abraham*, eds. Geart Van der Meer & Alice G. B. ter Meulen, 1–12. Groningen: Center for language and Cognition.

Evert, Stefan & Brigitte Krenn 2001. Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 188–195. Toulouse, France.

Fernando, Chitra 1996. *Idioms and Idiomaticity*. Oxford: Oxford University Press.

Firth, John Rupert 1957. A Synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis.*

Généreux, Michel, Iris Hendrickx & Amália Mendes 2012. A large Portuguese corpus online: Cleaning and preprocessing. *Computational Processing of the Portuguese Language. Proceedings of the 10th International Conference PROPOR2012*, eds. Helena Caseli, Aline Villacencio, António Teixeira & Fernando Perdigão, 113–120. Berlin, Heidelberg: Springer-Verlag.

Gross, Maurice 1975. *Méthodes en syntaxe*. Paris: Hermann.

Gross, Maurice 1984. Une classification des phrases 'figées' du français. *De la Syntaxe à la Pragmatique, Linguisticae Investigationes Supplementa*, eds. Pierre Attal and Claude Muller, Vol. 8, 141–180, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Halliday, Michael A. K. 1966. Lexis as a linguistic level. *In memory of J. R. Firth*, eds. Charles Ernest Bazell, John Cunnison Catford, Michael Alexander Kirkwood Halliday & Robert Henry Robins, 148–162. London: Longmans.

Hausmann, Franz Josef 1989. Le dictionnaire de collocations. *Wörterbücher, dictionaries, dictionnaires. Ein internationales Handbuch zur Lexikographie*, eds. Franz Josef Hausmann, Herbert Ernst Wiegand & Ladislav Zgusta, 1010–1019. Berlin: de Gruyter.

Hendrickx, Iris, Amália Mendes & Sandra Antunes 2010. Proposal for multi-word expression annotation in running text. *Proceedings of the 4th Linguistic Annotation Workshop*, ACL, 152–156. Uppsala, Sweden.

Iriarte Sanromán, Álvaro 2000. *A Unidade Lexicográfica. Palavras, Colocações, Frasemas, Pragmatemas*. Ph.D Dissertation. Braga: Universidade do Minho.

Mel'čuk, Igor 1998. Collocations and lexical functions. *Phraseology. Theory, Analysis, and Applications*, ed. Anthony P. Cowie, 23–53. Oxford: Oxford University Press.

Mel'čuk, Igor et al. 1984–1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I–IV*. Montréal: Les Presses de L'Université de Montréal.

Mendes Amália, Sandra Antunes, Maria Fernanda Bacelar do Nascimento, João Miguel Casteleiro, Luísa Pereira& Tiago Sá 2006. COMBINA-PT: A large corpus-extracted and hand-checked lexical database of Portuguese multiword expressions. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 1900–1905. Genoa, Italy.

Mendes, Amália, Sandra Antunes & Anabela Gonçalves 2015. COPLE2 – Corpus of Portuguese FL/L2. *Third Learner Corpus Research Conference*, poster. Nijmegen, The Netherlands.

Mendes, Amália & *Sílvia Pereira 2010*. Anotação de predicados complexos num corpus de português. *Actas del XXXIX Simpósio de la Sociedad Española de Lingüística*, eds. Pablo Cano López, Soraya Cortiñas Ansoar, Beatriz Dieste Quiroga, Isabel Fernández López & Luz Zas Varela. Santiago de Compostela. Unidixital (CD-Rom).

Mitchell, Terence Frederick 1971. Linguistic 'goings-on': collocations and other lexical matters arising on the syntactic record. *Archivum Linguisticum* 2 (new series): 35–69.

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

Nesselhauf, Nadja 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins Publishing Company.

Paquot, Magali 2013. Lexical bundles and L1 transfer effects. *Language Learning and technologt* 14(2): 30–49.

Pereira, Luísa Alice Santos 1994. *Como se combinam as palavras? Contributo para um Dicionário de Combinatórias do Português*. M.A. Dissertation. Lisboa: Faculdade de Letras da Universidade de Lisboa.

Pereira Luísa & Amália Mendes 2002. An electronic dictionary of collocations for European

Portuguese: Methodology, results and applications. *Proceedings of the 10th International Congress of the European Association for Lexicography*, Vol. II, 841–849. Copenhagen, Denmark.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart & Aline Villavicencio 2008. An evaluation of methods for the extraction of multiword expressions. P*roceedings of the LREC Workshop towards a Shared Task for Multiword Expressions*, 50–53. Marrakech, Morocco.

Ranchhod, Elisabete 1983. On the support verbs *ser* and *estar* in Portuguese. *Lingvisticae Investigationes* 7(2): 317–353. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ranchhod, Elisabete 1990. *Sintaxe dos Predicados Nominais com Estar*. Lisboa: Instituto Nacional de Investigação Científica, Centro de Linguística da Universidade de Lisboa.

Ranchhod, Elisabete 2003. O Lugar das Expressões 'Fixas' na Gramática do Português. *Razões e Emoção. Miscelânea de Estudos oferecida a Maria Helena Mira Mateus*, eds. Ivo Castro & Inês Duarte, 239–254. Lisboa: Imprensa Nacional Casa da Moeda.

Reference Corpus of Contemporary Portuguese. Centro de Linguística da Universidade de Lisboa. Version 2.3. 2012. Retrieved from http://www.clul.ul.pt/en/resources/183-crpc#cqp. Last accessed 1 December 2015.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger 2002. Multiword Expressions: A Pain in the Neck for NLP. *Proceedings of CICLING-2002*, ed. A. Gelbukh, 1–15.

Sardinha, Tony Berber 2014. Looking at Collocations in Brazilian Portuguese Through the Brazilian Corpus. *Working with Portuguese Corpora*, eds. Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira, 9–32. London: Bloomsbury.

Sardinha, Tony Berber, Rosana de Barros Silva & Telma de Lurdes São Bento Ferreira 2014. Lexical Bundles in Brazilian Portuguese. *Working with Portuguese Corpora*, eds. Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira, 33–67. London: Bloomsbury.

Sinclair, John 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, Michael 2004. A quantitative approach to collocations. *Phraseological Units: Basic Concepts and their Applications*, ICSELL 8, eds. David John Allerton, Nadja Nesselhauf & Paul Skandera, 107–119. Switzerland: Schwabe Verlag Basel.

Turner, Mark & Gilles Fauconnier 1995. Conceptual integration and formal expression. *Metaphor and Symbolic Activity* 10(3): 183–203.

Vale, Oto Araújo 2002. *Expressões Cristalizadas do Português do Brasil: Uma Proposta de Tipologia*. Ph.D Dissertation. São Paulo: Universidade Estadual Júlio Mesquita Filho.

Villalva, Alina 2003. Formação de palavras: composição. *Gramática da Língua Portuguesa*, eds. Maria Helena Mira Mateus, Ana Maria Brito, Inês Duarte, Isabel Hub Faria, *Sónia Frota, Gabriela Matos, Fátima Oliveira, Marina Vigário* & Alina Villalva, 5ª edição revista e aumentada, 969–983. Lisboa: Editorial Caminho.

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart & Carlos Ramisch 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1034–1043. Prague.