EXTRACTION OF RESTRICTED LEXICAL COMBINATIONS BY DETECTING NON-COMPOSITIONALITY OF MULTIWORD EXPRESSIONS

EXTRAÇÃO DE COMBINAÇÕES LEXICAIS RESTRITAS PELA DETEÇÃO DA NÃO COMPOSIONALIDADE DE EXPRESSÕES PLURIVERBAIS*

Joana Veloso
UNIVERSIDADE DO MINHO, PORTUGAL
joanasilvaveloso@gmail.com

Alberto Simões
UNIVERSIDADE DO MINHO, PORTUGAL
ambs@ilch.uminho.pt

Álvaro Iriarte S.
UNIVERSIDADE DO MINHO, PORTUGAL
alvaro@ilch.uminho.pt

In this article an evaluation of a method for extracting restricted lexical combinations from parallel corpora by detecting non-compositionality of multiword expressions in translation will be presented. This method presupposes that by finding sequences of words (word bigrams are used) whose translation does not follow a simple word-to-word conversion of the component words, a collocation is probably present.

Keywords: Collocations, Translation, Parallel Corpora

Neste artigo apresentamos uma avaliação sobre um método para extrair combinações lexicais restritas a partir de corpora paralelos, pela deteção da não composicionalidade de expressões pluriverbais na tradução. Este método baseia-se na presunção de que, encontrando sequências de palavras (são usados bigramas de palavras) cuja tradução não siga a tradução palavra por palavra dos seus componentes, é provável estar-se perante uma colocação.

Palavras-chave: Colocações, Tradução, Corpora Paralelos

* We are grateful to the anonymous reviewers for their valuable comments..

0. Introduction

Collocations can be defined in very different ways according to different authors, and the same sequence of words can be considered or not a collocation by different researchers, even when using a similar definition.

A great number of linguists define "collocation" in terms of frequency. That is the criterion used by many authors within the computational linguistics community for corpus-based automatic collocation extraction. However, other linguists (especially lexicologists and lexicographers) believe that the fact that two lexemes frequently co-occur in context is not a good enough reason to be considered a collocation. Mel'čuk *et al.* (1995: 51) state:

> «Kjellmer 1994 illustre un essai d'extration automatique de collocations d'un corpus informatisé sans intervention décisionnelle d'un lexicologue; ce dictionnaire est rempli d'expressions comme *Mr. Smith, was a member, the abilities, a bad thing,* etc. qui n'ont rien à voir avec les collocations. »

Coseriu denies that a lexeme combination is a "lexical solidarity" only because words are often combined. For Coseriu, a lexical solidarity is explained by lexical restrictions based on the linguistic content of a lexeme that drive it to combine with other lexemes.

> «… la probabilidad estadística de las combinaciones no tiene prácticamente nada que ver con las solidaridades y no es prueba de su existencia: *cavallo bianco* es, probablemente, más frecuente que *cavallo sauro*; pero, en el primer caso, la probabilidad de la combinación depende de la realidad extralingüística; en el segundo, en cambio, está dada lingüísticamente, por el contenido de *sauro.*» (Coseriu, 1977: 160).

There are software tools that allow the automatic extraction of frequent lexical combinations, however, these tools do not distinguish between collocations, idioms, free phrases, compound names, or multiword technical terms. It should be noted that "restricted lexical combinations" (especially collocations) with "frequent combinations" of two or more lexemes must not be confused.

It is true that a collocation is also a frequent combination of two or more lexemes but frequent lexical combinations are not always collocations. A lexical combination such as *ler um livro* is a frequent combination, but it is a completely free phrase, a combination of words formed according to the syntactic rules (the verb *ler* can be combined with everything capable of being read). However, the adjectives in lexical combinations such as *atividade febril, mudança radical, vontade louca,* ódio *mortal, amor cego* are fixed, although they have the same meaning (intense, a lot). These lexical combinations are frequent because the choice of adjective is not free, that is, the fact that these two lexemes frequently appear combined is not the cause but the consequence of them being a collocation (Alonso Ramos, 1993).

However, from a lexicographical point of view, these frequent lexical combinations, although they are not collocations in the narrow sense that we use, should have special treatment in dictionaries and lexical databases. In this regard, Cowie´s distinction between *restricted collocations* and *open collocations* is useful (Cowie *et al.*,1984).

It will, therefore, be appropriate, in lexicographical practice, to recognize the existence of more or less free/restricted lexical combinations and to register usual or frequent free combinations in dictionaries (Ettinger, 1982; Corpas, 1995).

As we know, stability and semantic specialization are the main features of this type of non-free phrases. Clearly syntactic constraints will be higher for idioms than for collocations. For example, the idiom *perder a cabeça* has more syntactic constraints than the collocation *prestar atenção*, which allows some changes (see Aguilar-Amat, 1993: 67-68). However, from a lexicographical point of view, it is impossible to establish general syntactic rules for all idioms and collocations.

The criteria that allow us to determine if a lexical combination was lexicalized cannot be morphological or syntactic in nature. This has more to do with the consensus and the memory of the linguistic community that uses it:

> «Le critère ultime de définition d'une unité lexicale est bien ici, par excellence, le consensus de la communauté linguistique […], non pas comme en syntaxe ou en morphologie par la reconnaissance d'une bonne formation mais sur la base de la mémorisation.» (Paillard, 1997: 66).

Ultimately, when it comes to distinguishing between free phrases and collocations, in addition to the more or less intuitive perceptions of speakers, the use of a foreign language serves to illustrate that choosing a collocative for a base of collocation is not free (Calderón, 1994: 80; Tomaszczyk, 1983: 45). Therefore, we are of the opinion that the use of a foreign language can be useful to extract collocations from parallel corpora.

For the purpose of this work, our definition of collocation is: *if any of the member words or the complete sequence inherit a different meaning of its/their usual sense when used in conjunction with another word, then this sequence of words is considered to be a collocation.*

Our assumption is that, given a multiword expression t_A and its translation t_B, t_A is a collocation if t_B includes words which are not direct translations of any of t_A words.

To test this assumption, we chose a mid-sized corpus, the European Central Bank corpus from the Per-Fide project (Araújo, Almeida, Simões & Dias, 2010). For the languages, we chose the Spanish/Portuguese language pair. The main reason for this choice is the proximity of the languages, and the bilingual translation dictionaries we had available. We expect to make further tests on this hypothesis with other language pairs in the future, namely including Germanic languages.

As for the text to be analyzed, we selected Spanish word pairs in which one of the words is an adjective and the other is a noun (in any order). For this purpose, the FreeLing (Simões & Carvalho, 2012; Padró & Stanilovsky, 2012) morphological analyzer was used. For each word pair, each possible word translation was looked up in the translation segment. This was possible with the help of a Spanish/Portuguese translation dictionary Apertium (Corbí-Bellot *et al.,* 2005).

If any of the possible translations for both words occur together (in any order) in the Portuguese part, that word pair is discarded. On the other hand, if only one of the words has a translation in the Portuguese part, the Spanish segment is stored, together with a snippet of the Portuguese translation. The next section explains this algorithm in more detail.

The word pairs obtained, together with the respective translation, were manually evaluated for whether they are, or not, restricted combinations. The evaluation section will discuss the details on the manual assessment of the obtained results, explaining the main problems found as well as the future enhancements to the proposed algorithm. Finally, a set of conclusions is drawn from the obtained results.

1. Related Work

The task of identifying restricted lexical combinations, as we will state in this article, is not new. It is a relevant procedure for different tasks on Natural Language Processing (NLP) like, for example, Machine Translation (MT), where idiomatic expressions cannot be translated literally. Even collocations need to be translated with caution. For example, *computer graphics* cannot be translated into Portuguese as *gráficos de computador*, not because it is a wrong translation, but because the term that was coined in the Portuguese community was *computação gráfica* (*graphics' computation).

The easiest way to detect sequences of words likely to be considered as a collocation or, at least, as a compound term, is to use the Mutual Information (MI) or Pointwise Mutual Information (PMI) metrics. As this is just an association measure, authors use these indicators together with other techniques, like the usage of patterns (Guinovart & Simões, 2009). However, by themselves, these two measures are not enough for the extraction of collocations. A prior study (Pavel 2005) presents a vast amount of measures that can be used to detect collocations. Nevertheless, most of them perform badly by themselves, and as presented below, new approaches have been used.

Probably the bigger challenge is to detect idiomatic expressions, mostly when they can also have a literal meaning (like *break the ice* that can be considered literally or not). Li & Sporleder (2009) present a set of different properties that can be extracted from texts in order to detect if these expressions are being used literally. Properties are very diverse, from the usage of prepositions before or after the expression, to graphs of cohesion between the different sentence components. These properties are then used in a Support Vector Machine algorithm. These same authors (Li & Sporleder, 2010) also worked on the use of Gaussian Mixture Models for this same task. Their evaluation points to 92% precision for the detection of literal expressions, but only 42% to detect non-literal (idiomatic) expressions.

Muzny & Zettlemoyer (2013) also use classification techniques to distinguish between idiomatic and non-idiomatic expressions. For that, they trained a binary perceptron based on two types of features: lexical features, like the usage of capital letters, and graph features, using relations information obtained from WordNet and Wiktionary. The perceptron was training on Wiktionary labeled data, and used the non-labeled data for test purposes. The results go up to 65% of precision, and recall over 52%.

The most relevant study found using multilingual information for the detection of collocation extraction (Seretan & Wehrli, 2006), does not use translation information, but only a parser able to process text in different languages. The extraction method, itself, does not take any real advantage of parallel corpora.

Our approach understands the translation as a function that can, somehow, associate "*semantic"* to each word. Therefore, if the translation ("semantic") is not compositional, we have a candidate collocation.

2. The Hypothesis

The hypothesis we are testing is: if a sequence with two words, an adjective and a noun, is translated by two other words, and only one of them is a translation of the original words found in a translation dictionary, then we have a candidate collocation.

This can be better explained using mathematical syntax. Let us define the *T* function, that translates Spanish words into Portuguese, and the concatenation operator (a dot), which joins two words.

The translation of two words w_a and w_b is considered to be compositional if

(1) $\mathrm{T}(w_\mathrm{a} . w_\mathrm{b}) = \mathrm{T}(w_\mathrm{a}) . \mathrm{T}(w_\mathrm{b})$ or even, $\mathrm{T}(w_\mathrm{a} . w_\mathrm{b}) = \mathrm{T}(w_\mathrm{b}) . \mathrm{T}(w_\mathrm{a})$

Therefore, we are looking for a pair of words (w_a, w_b) in which one of them is an adjective and the other a noun, and whose translation does not follow the equation presented above (1). That is, we want to find w_a and w_b where

(2) $\mathrm{T}(w_a . w_b) = \mathrm{T}(w_a) . w_c \wedge \mathrm{T}(w_a . w_b) = w_c . \mathrm{T}(w_b)$ with $\mathrm{T}(w_a \neq w_c \wedge \mathrm{T}(w_b) \neq w_c$.

The extraction algorithm used is very simple, and its main purpose is to test the hypothesis that the collocation extraction based on non-translation composition is possible. The algorithm starts by iterating over each translation unit in the parallel corpus. A translation unit is composed by a segment S_{SP} for the Spanish language, and a segment S_{PT} for the Portuguese language. Then, each possible bigram from the segment S_{SP} is analyzed using the FreeLing morphological analyzer, looking for a sequence in which one of the words is a noun and the other an adjective. Note that, although FreeLing has modules to do part-of-speech tagging they were not used. Nevertheless, we are aware of the problems this approach arises, and we will discuss them later.

When such a pair of words is found, their possible translation sets are computed. Note that each word can have more than one translation, and, therefore, we need to construct a set of translations. This translation was done using the Apertium translation dictionary. Then, these translation sets are searched in the target language segment S_{PT}. If any of the words from both translation sets occur together, the word pair is discarded.

On the other hand, if one of the words has a translation in the target segment, but the other does not, the Spanish word pair is saved for manual assessment. Together with the word pair, a segment of Portuguese words in the vicinity of the found translation is seized and also stored. This list was then assessed manually.

3. Assessments and Evaluation

The assessment was performed manually using online resources as reference, such as IATE (InterActive Terminology for Europe) (Johnson & Macphail, 2000), and both paper and online Spanish-Portuguese and Portuguese-Spanish dictionaries. A Linguistics MSc student classified each word pair manually into one of the following classes:

- **Error:** used for all entries whose Spanish and Portuguese segments are not related with each other. This happens mainly because the application was not able to find the sequence of words that include the translation of the selected pair of words, or because the original corpus had alignment errors;

- **Free combination:** the pair of words is correctly translated, but it is not a restricted lexical combination (accordingly with the criterion we defined earlier). This happens mostly when a possible word translation is not included in the used translation dictionary;

- **Restricted combination:** the pair of words is correctly translated, and it corresponds to a collocation or a quasi-phraseme.

When in doubt about a combination being considered restricted or free, we took the decision to consider it as a free combination. This means that our evaluation is less favorable to our hypothesis.

From here on, we will discuss each class, providing real examples for each of them.

3.1. Errors

The errors found are from very different kinds, such us from alignment problems, some minor bugs in the algorithm implementation, or the lack of translations from the translation dictionary:

- The use of a morphological analyzer instead of a part-of-speech tagger leads to some examples with verbs misclassified as nouns and/or adjectives. Nevertheless, considering that our hypothesis is the existence of a sequence with a noun and an adjective, the examples classified as a result of this problem are irrelevant for proving it. Table 1[1] shows some of these situations.

Table 1: Examples of extractions where a verbal form was mistakenly interpreted as a noun

anexo figura	presente *anexo figura* um modelo
conjunto presente	Se o *conjunto apresentar*
informe figura	este *relatório consta* de o
certificado falla	verificação de o *certificado falhar*

- The algorithm, when searching the set of words in the context of the found translation broke the segment, losing the interesting part of the translation. This turned the assessment impossible. This was a problem inherited from the bad segmentations performed by other tools like the segmenter, tokenizer and the sentence-aligner. For example, the alignment for "*actividades pesqueras"* computed by the algorithm was "*definitiva de as actividades de* *". Given the missing word (marked by the asterisk) this segment could not be classified correctly, and therefore, fell in the error class. Just like with the case above, we do not have any detail on the validity (or not) of the hypothesis. Table 2 shows further examples of this segmentation problem. All these cases can be safely ignored for the hypothesis test. Between parentheses we show

[1] In these tables, the left column is the Spanish extracted pair, and to the right is the segment extracted from the Portuguese side. In italics we give emphasis to the translation of the terms from the first column.

the missing words. This seemed to be a problem on the corpus segmentation and alignment process.

Table 2: Examples of truncated segments

tabaco crudo	Sector de o tabaco em (*cru*)
derechos humanos	promoção de os direitos de (*o homem*)
productos pesqueros	mercado de os produtos de (*pesca*)
Seguridad Alimentaria	Europeia para a Segurança de (*a alimentação*)
Seguridad Alimentaria	Europeia para a Segurança de (*a alimentar*)

- The algorithm implementation is not prepared to find all occurrences of the word translations. This means that, if two similar pairs occur (like "*cantidad superior"* and "*calidad superior"*) the algorithm will use the first translation pair twice (aligning "*qualidade superior"* with "*cantidad superior"* and not the correct "*calidad superior"*). This is, indeed, a bug introduced by our implementation, but when it was detected it was too late to perform a complete new extraction and restart the manual evaluation. Therefore, they were ignored for our hypothesis test. Table 3 shows some of these examples. In italics, on the right, the aligned segment.

Table 3: Examples of misaligned segments

Comunidad Económica	que instituí a *Comunidade Europeia*
cantidad superior	em uma *quantidade inferior*
tiempo completo	de trabalho a *tempo parcial*
tercera columna	referidas em a *coluna 2*
navegación marítima	afectos a a *navegação aérea*

- The translation, sometimes, uses a pronoun to refer to a noun used on a previous sentence, while the original sentence repeats the noun. See Table 4 for some examples.

Table 4: Examples of misalignments resulted from the use of pronouns

Estado membro	legislação de *esse Estado*
ciertos productos	regime de *esses produtos* .
valores límite	ou de *esses valores* ,
último caso	. Em *esse caso* ,
segundo Estado	legislação de *esse Estado ;*

- Some translation units were not really translated. In some cases the Portuguese version included the text in Spanish, and in some other, in English, as shown in Table 5. Some others, as shown in Table 6, include typos that, not being in the dictionary, activated our hypothesis by mistake.

Table 5: Examples of translation units where at least one of the sides is untranslated

medio ambiente	contaminación del medio ambiente
Autoridades nacionales	Lista de las autoridades nacionales
medio ambiente	en el medio ambiente acuático
Vivo Test	In Vivo Test for Chromosomal

Table 6: Examples of translations with minor typos, mistakenly extracted as collocations

presente artículo	de o *presnete artigo*
legítimo titular	seu *legítimo titual* a ocupar
zona geográfica	específicas em uma *zona geográfirca*
presente Reglamento	O *presidente regulamento* é
proyectos transnacionales	acompanhamento de os *projectos trannacionais*

3.2 Free Combinations

The main interference with the algorithm, which could make it extract free combinations, is the lack of translations from the used translation dictionary. When a word is not found in the dictionary (either as the word not existing in the source language – Table 7; or the target language do not include the used translation – Table 8), the algorithm considers the translation to be incorrect, and therefore, it can be used for our hypothesis. A similar problem occurred with words not correctly lemmatized, and therefore, not found in the translation dictionary.

Table 7: Basic examples where the dictionary lacked a entry for one of the words

Segundo resultado	resultado : *segundo resultado* :
primer trimestre	em o *primeiro trimestre* de
presente Directiva	requisitos de a *presente directiva*
primeros párrafos	os dois *primeiros parágrafos* podem
presente apêndice	o *presente apêndice* , os

Table 8: Basic examples where the dictionary lacked one of the synonyms

Texto pertinente	EEE) (*Texto relevante*
siguiente texto	a *seguinte redacção* :
Medidas vigentes	*Medidas existentes*
última fabricación	a última *data de fabrico*
Conocimientos sucintos	; *Conhecimentos sumários* de as

Finally, there is yet another problem, related with the *textual deixis*, where there is a reference to a different position in the text that, different languages refer to in different ways. Examples are *cuadro anterior*/*quadro acima*, *fórmula anterior*/*fórmula acima* and *criterios anteriores*/*critérios acima.* As *anterior* and *acima* are not direct translations, the algorithm extracted them as restricted combinations (although they are free combinations).

3.3 Restricted Combinations and Collocations

Other than the correct restricted combinations, there are two special kinds that should be mentioned:

- There are situations where the pair of words in Spanish has a single word translation in Portuguese, either because in Portuguese one of the words is usually omitted, or because there is a word with the complete meaning of the two Spanish words. This situation was named *reduction* and happens a few times. Examples are shown in Table 9. These were considered restricted combinations. The best examples from Table 9 are the first and the last. In Portuguese, and although there is the concept of *meio ambiente*, it is usually used only as *ambiente.* And in the case of *cigarros pequeños*, Portuguese as a word for that: *cigarrilhas*. These situations were validated manually in IATE (InterActive Terminology for Europe).

Table 9: Examples of reductions: situations where two words are correctly translated by only one word

medio ambiente	protecção de o *ambiente* ,
auditoría medioambientales	de ecogestão e *auditoria* (
titular opositor	parte de o *titular* .
trabajo anual	uma unidade de *trabalho* ,
cigarros pequeños	*cigarrilhas* e cigarros ,

- There is another situation with nouns (mostly geographic) that were mistakenly extracted, as shown in Table 10. These were extracted because of the way the nouns are translated. This table shows three columns. The first two are Spanish and Portuguese, and the third, a direct Portuguese translation of the Spanish term.

Table 10: Examples of nouns whose translation was a problem for the algorithm

Sudeste Asiático	Ásia de o Sudeste	*Sudoeste Asiático*
República Federal	originários de a República Federativa	*República Federal*
continental español	Espanha continental	*continental espanhol*

- Given that we decided to analyze bigrams, there are situations where the bigram is part of a bigger restricted combination. This is usually easy to detect given the specific area of the used corpora, and given that the Portuguese segment includes more words than the two existing in Spanish. Table 11 shows examples.

Table 11: Cases of restricted combinations with more than two words

gestión medioambiental	o *sistema de gestão ambiental*
producción homogénea	A *unidade de produção homogénea*
política agrícola	de a *política agrícola comum*
ejecución forzosa	de *medidas de execução forçada*
residuos radiactivos	*Gestão de os resíduos radioactivos*

When the algorithm returned interesting results, returning restricted lexical combinations, we were unable to distinguish between collocations and other types of restricted combinations, like quasi-phrasemes and idioms (this last type was not found, probably given the type of the used corpus).

Table 12: Examples where the algorithm worked

Disposiciones legales	disposições legislativas
mercado interior	mercado interno
cadena alimentaria	Cadeia Alimentar
Derecho interno	direito nacional
derechos humanos	direitos fundamentais
contingentes arancelarios	contingentes pautais
fronteras interiors	fronteiras internas
entidad contratante	entidade adjudicante
días hábiles	dias úteis
años naturales	anos civis
persona física	pessoa singular
peso neto	peso líquido
años naturales	anos civis
precio neto	preço líquido
personas jurídicas	pessoas colectivas
amarillo oscuro	amarelo torrado
sentencia firme	sentença transitada
petróleo crudo	ou resíduos de petróleo bruto
atún rojo	atum rabilho
correo normal	correio ordinário
historiales médicos	processos médicos

3.4 Analysis

When applied to the European Central Bank corpora, our approach extracted more than 40.000 pairs of candidate collocations. They were evaluated by exhaustion (instead of evaluating a sample of random entries, the evaluator tagged each one of the extracted candidates). This means the evaluation is not affected by sample bias. This, together with the fact that the evaluator gave preference to free combinations over restricted combinations, means that this evaluation is the baseline of the algorithm.

Table 13 presents the number of cases found and classified according to each of the previously mentioned classes. If we ignore the cases of errors, nouns and reductions, we can note that restricted combinations are one quarter of the total number of found combinations.

Table 13: Summary of cases found for each category.

Category	Number of Occurrences	Percentage
Free Combinations	19 428	42,15 %
Restricted Combination	6 447	13,99 %
Errors	19 281	41,83 %
Reductions	914	0,04 %
Nouns	19	0,0004 %

4. Conclusions

The first reaction to the results was of discontent, as a lot of free combinations were found. As soon as the examples were analyzed was realized: firstly, the translation resource lacks coverage, and secondly, the algorithm used misses the correct lemmatization for some words. These two reasons can be fixed (or made better) using other approaches or tools for the lemmatization, and using other dictionaries or even probabilistic translation dictionaries (Simões & Almeida, 2003; Simões, Almeida & Ramos Carvalho, 2013) to enrich the translation coverage.

Nevertheless, most of the situations found are easy to correct, and, therefore, further experiments should be performed before considering the method inadequate. In fact, will be interesting to see how this approach performs in a less noisy corpus, with better dictionaries, and with other languages.

This data, being manually classified, can be used to train machine learning algorithms. For statistical machine translation, it is possible to denote/specify which segments should be reused directly without any change (when they are idiomatic), and which segments can be generalized, allowing some of the words to be replaced, and reusing the translation structure. For the extraction of further collocations from other corpora, this data can be used to train a supervised machine learning algorithm, or just be used as a golden standard for this kind of system.

Analyzing the results obtained, the initial starting hypothesis should be reformulated. Using this approach, restricted combinations, and not just a specific type of restricted combinations such as the case of collocations, are detectable. Of course, authors like Mel'čuk (1995) define formal types for each one of these restricted lexical combinations. The problem is the non-existence of a clear distinction between them. Some lexical combinations will be classified differently according to the way the linguist decomposes semantically the expression.

There is another problem with our hypothesis, when a restricted combination coincides in the two languages being analyzed, because they can be mistakenly considered free lexical combinations. Examples of this problem are *de segunda mano/em segunda mão, ódio mortal/ódio mortal*, *amor ciego/amor cego*. We expect that this may not be a problem when using parallel corpora including more distinct languages (Portuguese/English, Spanish/English, etc.).

References

Aguilar-Amat, A. (1993): *Las colocaciones de nombre y adjetivo. Un paso hacia una teoria léxico-semántica de la traducción.* Tese de doutoramento [microfichas]. Barcelona: Universitat Autònoma de Barcelona, Servei de Publicacions.

Alonso Ramos, M. (1993). *Las Funciones Léxicas en el modelo lexicográfico de I. Mel'čuk.* Tese de doutoramento. Madrid: UNED.

Araújo, S., Almeida, J.J., Simões, A. & Dias, I. (2010). Apresentação do projecto Per-Fide: Paralelizando o português com seis outras línguas. *Linguamática, 2*(2), 71-74.

Calderón Campos, M. (1994). *Sobre la elaboración de diccionarios monolingües de producción. Las definiciones, los ejemplos y las colocaciones léxicas*. Granada: Universidad de Granada.

Corbí-Bellot, A. M., Forcada, M.L., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Ramírez, G., Sánchez-Martínez, F. *et al.* (2005). An open-source shallow-transfer machine translation engine for the romance languages of Spain. *Proceedings of the European Association for Machine Translation, 10th Annual Conference* (pp. 79-86). Budapest: European Association for Machine Translation.

Corpas Pastor, G. (1995): *Un estudio paralelo de los sistemas fraseológicos del inglés y del español.* Tese de doutoramento, Universidad Complutense de Madrid, 1994 [microfichas]. Málaga: Universidad de Málaga: Servicio de publicaciones e intercambio científico.

Coseriu, E. (1977). *Principios de semántica estructural.* Madrid: Gredos.

Cowie, A.P., Mackin, R., & McCaig, I.R. (1984). Oxford *Dictionary of Current Idiomatic English, vol. I-II. General Introduction.* Oxford, OUP.

Ettinger, S. (1982). Formación de palabras y fraseología en la lexicografía. In Haensch, G., Wolf, L., Ettinger, S. & Werner, R. (Eds.), *La lexicografía. De la lingüística teórica a la lexicografía práctica.* (pp. 233-258). Madrid: Gredos.

Guinovart, X. G. & Simões, A. (2009). Parallel corpus-based bilingual terminology extraction. In L'Homme M-C & Szulman, S. (Eds.), *8th International Conference on Terminology and Artificial Intelligence*, Toulouse, France, November, 18-20.

Johnson, I., & Macphail, A. (2000). IATE–Inter-Agency Terminology Exchange: Development of a Single Central Terminology Database for the Institutions and Agencies of the European Union. In Choi, K-S (Ed.), *Proceedings of the Workshop on Terminology resources and computation, LREC 2000 Conference.* Athènes, Grèce.

Li, L. & Sporleder, C. (2009). Classifier Combination for Contextual Idiom Detection Without Labelled Data. In Koehn, P. & Mihalcea, R. (Eds.), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 315-323). Singapore: Association for Computational Linguistics.

Li, L. & Sporleder, C. (2010). Using Gaussian Mixture models to Detect Figurative Language in Context. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 297-300). Los Angeles, California: Association for Computational Linguistics.

Mel'čuk, I. (1995). Phrasemes in Language and Phraseology in Linguistics. In Everaert, M., van Der Linden, E.J., Scheak, A. & Schzender (Eds.), *Idioms: Structural and Psychological Perspectives*. (pp. 167-232). Hillsdale-New Jersey Hove-U.K.: Lawrence Erlbaum Associates.

Mel'čuk, I., Clas, A. & Polguère, A. (1995). *Introduction à la Lexicologie Explicative et Combinatoire.* Louvain-la-Neuve: Duculot.

Muzny, G. & Zettlemoyer, L. S. (2013). Automatic Idiom Identification in Wiktionary. *In the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (pp. 1417-1421), Seattle, Washington, USA.

Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality**.** In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.* Istanbul, Turkey.

Paillard, M. (1997). Co-texte, collocations, lexique. In Guimier (Ed.), Co-texte et calcul du sens. *Actes de la table ronde tenue à Caen les 2 et 3 février 1996* (pp 63-71). Caen: Presses Universitaires de Caen.

Pavel, P. (2005). *An Extensive Empirical Study of Collocation Extraction Methods*. In Proceedings of the ACL Student Research Workshop (pp. 13-18). Stroudsburg, PA, USA: Association for Computational Linguistics.

Seretan, V. & Wehrli, E. (2006). *Accurate Collocation Extraction Using a Multilingual Parser*. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (pp. 953-960). Stroudsburg, PA, USA: Association for Computational Linguistics.

Simões, A. & Almeida, J.J. (2003). NATools - a statistical word aligner workbench. *Procesamiento del Lenguaje Natural, 31,* 217-224.

Simões, A., Almeida, J.J. & Ramos Carvalho, N. (2013). Defining a probabilistic translation dictionaries algebra. In Correia, L., Reis, L.P., Cascalho, J., Gomes, L., Guerra, H., & Cardoso, P. (Eds.), *XVI Portuguese Conference on Artificial Intelligence – EPIA*. (pp. 444-455), Angra do Heroismo, Portugal.

Simões, A. & Carvalho, N. (2012). Desenvolvimento de aplicações em Perl com FreeLing 3. *Linguamática, 4*(2), 87-92.

Tomaszczyk, J. (1983). On Bilingual Dictionaries. The Case for Bilingual Dictionaries for Foreign Language Learners. In Hartmann (Ed.), *Lexicography: Principles and Practice* (pp. 41-51). London: Academic Press.

[Recebido em 13 de fevereiro de 2016 e aceite para publicação em 4 de abril de 2016]