A Flexible Framework for Collocation Retrieval and Translation from Parallel and Comparable Corpora

Oscar Mendoza Rivera, Ruslan Mitkov and Gloria Corpas Pastor
Research Group in Computational Linguistics, University of Wolverhampton
{o.mendozarivera, r.mitkov}@wlv.ac.uk, gcorpas@uma.es

Abstract

This paper outlines a methodology and a system for collocation retrieval and translation from parallel and comparable corpora. The methodology was developed with translators and language learners in mind. It is based on a phraseology framework, applies statistical techniques, and employs source tools and online resources. The collocation retrieval and translation has proved successful for English and Spanish and can be easily adapted to other languages. The evaluation results are promising and future goals are proposed. Furthermore, conclusions are drawn on the nature of comparable corpora and how they can be better exploited to suit particular needs of target users.

1 Introduction

Multiword expressions (MWEs) are lexical units made up of several words in which at least one of them is restricted by linguistic conventions. One example is the expression *fast food*, in which the word *fast* is arbitrary, as it cannot be replaced with synonyms, such as *quick*, *speedy* or *rapid*. It is thought that a significant part of a language's vocabulary is made up of these expressions: as noted by Biber *et al.* (1999), MWEs account for between 30% and 45% of spoken English and 21% of academic prose, while Jackendoff (1997) goes as far as to claim that their estimated number in a lexicon is of the same order of magnitude as its number of single words. Furthermore, these numbers are probably underestimated: they appear in all text genres, but specialised domain vocabulary, such as terminology, "overwhelmingly consists of MWEs" (Sag *et al.*, 2002, p. 2).

Collocations represent the highest proportion of MWEs (Lea and Runcie, 2002; Seretan, 2011). As such, collocation retrieval has sparked interest in the NLP community (Smadja, 1993; Sag *et al.*, 2002; Lü and Zhou, 2004; Sharoff *et al.*, 2009; Gelbukh and Kolesnikova, 2013). Several methods have been adopted to measure the association strength of collocations, which has achieved favourable results with increases in accuracy (Seretan, 2011). However, a much more limited number of studies have dealt with post-processing of collocations from the perspective of their practical use. Collocation translation, for instance, while a natural follow-up to collocation extraction in this trail of research, still poses a problem for computational systems (Seretan, 2011). Furthermore, while several collocation resources have been put together, such as the multilingual collocation dictionary *MultiCoDiCT* (Cardey *et. al*, 2006), approaches to collocation retrieval and translation lack, in general, the solid theoretical basis of phraseology (Corpas Pastor, 2013).

To address this problem, the present paper describes the development and implementation of a computational tool to allow language learners and translators to retrieve collocations in a source language (SL) and their translations in a target language (TL) from bilingual parallel and comparable corpora. The project focuses on English and Spanish, but the methodology is designed to be flexible enough to be applied to other pairs of languages.

The remainder of this paper is organised as follows: Section 2 discusses the phraseology basis of our project and presents two collocation typologies (one in English and one in Spanish) as well as a comparative grammar. Section 3 provides a brief review of existing techniques for the extraction and translation of collocations. Section 4 presents a new methodology and outlines the implementation of a computational tool based on it. Finally, Section 5 details the results from the experiments set up to evaluate our system, and discusses opportunities for future work.

Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, September 3, 2013.
Monti, J., Mitkov, R., Corpas Pastor, G., Seretan, V,. eds.

2 Phraseology

Collocations are compositional and statistically idiomatic MWEs (Baldwin and Kim, 2010). Like idioms, collocations belong to the set phrases of a language. Unlike them, while the meaning of an idiom is mostly incomprehensible if not previously heard (*to **pay** through the nose*, *cold **turkey***), collocations are compositional: their meanings can be deduced from the meaning of their component words (*to **pay** attention*, *roast **turkey***). However, these are arbitrary. For example, the expression *I **did** my homework* is correct in English, but the expression *I **made** my homework* is not. The choice of using the verb *to do* and not the verb *to make* in this particular example can be thought of as an arbitrary convention. In addition, some collocates exhibit delexical and metaphorical meanings (*to **make** an attempt, to **toy** with an idea*). Similarly, collocations are cohesive lexical clusters. This means that the presence of one or several component words of a collocation in a phrase often suggests the existence of the remaining component words of that collocation. This property attributes particular statistical distributions to collocations (Smadja, 1993). For example, in a sample text containing the words *bid* and *farewell*, the probability of the two of them appearing together is higher than the probability of the two of them appearing individually.

2.1 Typologies of collocations

Hausmann (1985) argued the components of a collocation are hierarchically ordered: while the *base* can be interpreted outside of the context of a collocation and can therefore be considered as semantically autonomous, the *collocatives* depend on the base in order to get their full meaning. He also presented a typology of collocations in English based on their syntax (see Table 1). Similarly, Corpas Pastor (1995, 1996) studied, classified, and contrasted collocations for Spanish and English and has proposed her own typology of collocations in these two languages (see Tables 1 and 2). These tables show the base of collocations in bold and use abbreviations borrowed from the tagset of TreeTagger (Schmid, 1994): *VB* stands for verb, *NN* for noun, *RB* for adverb, *JJ* for adjective, and *IN* for preposition. Furthermore, these typologies have been helpful in the development of this project's underlying methodology to extract collocations (see Sections 4.1 and 4.2).

Type	Examples
1. VB + **NN** *(direct object)*	*to express concern,* *to bid farewell*
2. NN or JJ + **NN**	*traumatic experience,* *copycat crime*
3. NN + *of* + **NN**	*pinch of salt,* *pride of lions*
4. RB + **JJ**	*deadly serious, fast asleep*
5. **VB** + RB	*to speak vaguely,* *to sob bitterly*
6. VB + IN + **NN**	*to take into consideration,* *to jump to a conclusion*
7. VB + **NN** *(subject)*	*to break out <war>,* *to crow <a cock>*

Table 1: Typology of collocations in English

Type	Examples
1. VB + **NN** *(direct object)*	*conciliar el sueño,* *entablar conversación*
2. **NN** + JJ or NN	*lluvia torrencial,* *visita relámpago*
3. NN + *de* + **NN**	*grano de arroz,* *enjambre de abejas*
4. RB + **JJ**	*profundamente dormido,* *estrechamente relacionado*
5. **VB** + RB	*trabajar duro, jugar sucio*
6. VB + IN + **NN**	*tomar en consideración,* *poner a prueba*
7. VB + **NN** *(subject)*	*ladrar <un perro>,* *estallar <una guerra>*

Table 2: Typology of collocations in Spanish

2.2 Transfer rules

Bradford and Hill (2000) studied the comparison between the grammar of English and Spanish. Based on their work, we have developed a set of transfer rules (see Table 3) between these two languages which help us translate collocations (see Section 4.4).

English	Spanish
VB + **NN**	VB + **NN**
NN or JJ + **NN**	**NN** + JJ or NN
NN + *of* + **NN**	NN + *de* + **NN**
RB + **JJ**	RB + **JJ**
VB + RB	**VB** + RB
VB + IN + **NN**	VB + IN + **NN**

Table 3: English-Spanish syntax comparison

It is worth noting that these transfer rules are designed to aid us in our own approach to the task of syntactic processing, but they are not all-inclusive. In fact, as is often the case, there are exceptions to the rules. For example, collocations in English such as *copycat crime* (*delito inspirado en uno precedente* or *que trata de imitarlo,* in Spanish) and *to commit suicide* (*suicidarse* in Spanish) cannot be translated using the proposed approach.

3 Related Work

This section presents a brief review of existing techniques for the extraction and translation of collocations. It starts by outlining collocation extraction and then moves to translation.

3.1 Collocation retrieval

Early work on collocation extraction focused on statistical processing. Choueka *et al.* (1983) developed an approach to retrieve sequences of words occurring together over a threshold in their corpora. Similarly, Church and Hanks (1989) proposed a correlation method based on the notion of mutual information. Smadja (1993), however, highlighted the importance of combining statistical and linguistic methods. In recent years, advances have been made (Ramisch *et al.*, 2010; Seretan, 2011), many of them advocating rule-based and hybrid approaches (Hoang, Kim and Kam, 2009), and based on language-specific syntactic structures (Santana *et al.*, 2011) or machine learning of lexical functions (Gelbukh and Kolesnikova, 2013).

3.2 Parallel corpora

Classic approaches to translation using parallel corpora exploited the concepts of alignment and correspondence at sentence level (Brown *et al.*, 1991; Gale and Church, 1993). Two methods were developed: length-based and translation-based (Varga *et al.*, 2005). Collocation translation using parallel corpora has also been approached using transfer systems that rely on generative grammars, because of the notion that the base of a collocation determines its collocatives (Wehrli *et al.*, 2009) and the assumption that source and target MWEs share their syntactic relation (Lü and Zhou, 2004).

3.3 Comparable Corpora

Parallel resources are generally scarce and in many cases not available at all. The wider availability of comparable texts offers new opportunities to both researchers and translators. While these do not allow for bridging between languages (Sharoff *et al.*, 2009), research suggests (Rapp, 1995) that a word is closely associated with words in its context and that the association between a base and its collocatives is preserved in any language. Fung and Yuen (1998), for instance, argued that the first clue to the similarity between a word and its translation is the number of common words in their contexts. Similarly, Sharoff *et al.* (2009) proposed a methodology that relies on similarity classes.

4 System

The system[1] employs the following three language-independent tools: TreeTagger to POS-tag corpora, the MWEToolkit (Ramisch *et al.*, 2010) to extract collocations according to specific POS-patterns, and Hunalign (Varga *et al.*, 2005) to align corpora at sentence level. Furthermore, it connects online to *WordReference* and uses it as a multilingual translation dictionary and thesaurus. Figure 1 illustrates the architecture of the system; its main modules will be described in greater detail in the following paragraphs.

4.1 Candidate selection module

This module processes the SL corpus in order to format it to comply with the input requirements of the modules that follow in the system pipeline. It represents the linguistic component of the hybrid approach to collocation retrieval. It makes use of both TreeTagger and the MWEToolkit to perform *linguistic pre-processing* in the form of lemmatisation and POS-tagging on the input data, as well as *POS-pattern definition*.

Linguistic processing aims at transforming the input data from a stream of alphanumeric characters to sequences of words, which can be grouped in *n-grams*. It is important to work with lemmas instead of inflected words in order to identify collocations; otherwise, for example, collocations such as *committing murder* and *committed murder* would be treated separately, even though they are obviously the same (whose lemma is *commit murder*). The system relies on TreeTagger to annotate text sentences with both lemma and POS-tagging information. Its output is then transformed into the XML format (see Figure 2) by running a Python script, part of MWEToolkit.

[1] Consisting of a series of Python scripts which handle text and XML representations, and implemented using the wxPython development environment for Mac OSX.

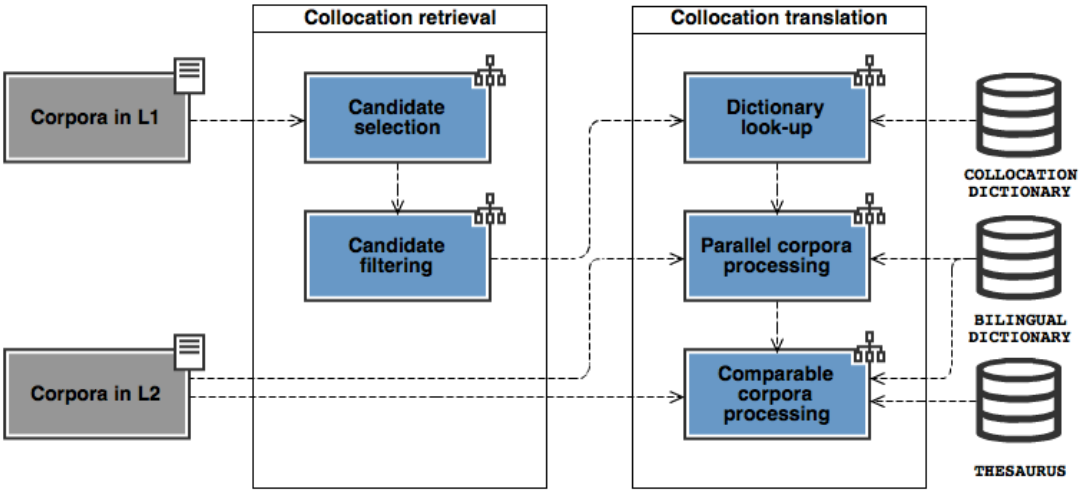

Figure 1: Architectural scheme of the system

POS-pattern definition aims at applying syntactic constraints on collocation candidates. This stage is language-dependent: as long as a language can be POS-tagged and a typology of the most commonly occurring collocations exists for it, POS-patterns can be defined. This task is simplified because the MWEToolkit supports the definition of syntactic patterns of collocations to extract. These can include repetitions, negation, and optional elements, much like regular expressions (see Figure 3, a definition of the English POS-pattern *NN or JJ + NN*). When retrieving collocations, each sentence in the corpus is matched against this set of patterns, and all n-grams which do not comply with any of them are ignored. Patterns that correspond exactly to the typologies of collocations in English and Spanish presented above have been defined (see Section 2.2).

```
<s s_id="0">
   <w surface="Harry" pos="NP" lemma="Harry"/>
   <w surface="unwrapped" pos="VBD" lemma="unwrap"/>
   <w surface="his" pos="PP$" lemma="his"/>
   <w surface="chocolate" pos="NN" lemma="chocolate"/>
   <w surface="frog" pos="NN" lemma="frog"/>
   <w surface="." pos="SENT" lemma="."/>
</s>
```

Figure 2: Sample XML output of TreeTagger

```
<pat>
   <pat repeat="+">
      <either>
         <pat> <w pos="JJ"/> </pat>
         <pat> <w pos="NN"/> </pat>
      </either>
   </pat>
   <w pos="NN"/>
</pat>
```

Figure 3: Example of POS-pattern definition

4.2 Candidate filtering module

This module computes collocation candidates and assigns a weight to each of these according to its probability of representing a collocation. It corresponds to the statistical component of our hybrid approach to collocation retrieval and relies on the MWEToolkit to perform n-gram selection and statistical processing. The toolkit receives two XML files as input: a representation of all sentences in the corpus with all words described by linguistic properties (see Figure 2), and a set of user-defined POS-patterns (see Figure 3). It performs *n-gram selection* by matching each sentence in the corpus against all defined POS-patterns, producing a set of collocation candidates. Once candidates have been extracted, it performs *statistical processing* by computing the frequencies of each candidate's word components from the SL corpus. This information is used to calculate a log-likelihood score for each candidate. Candidates are then ranked according to their scores. Figure 4 presents a sample collocation candidate in English. As can be observed, the toolkit not only extracts the lemma form of a collocation (*lemon drop*), but also the different surface forms it appears in (*lemon drops*).

```
<cand candid="1305">
  <ngram>
    <w lemma="lemon" pos="NN"> <freq value="9" /> </w>
    <w lemma="drop" pos="NN"> <freq value="18" /> </w>
    <freq value="6" />
  </ngram>
  <occurs>
    <ngram>
      <w surface="lemon" lemma="lemon" pos="NN" />
      <w surface="drop" lemma="drop" pos="NN" />
      <freq value="4" />
    </ngram>
    <ngram>
      <w surface="lemon" lemma="lemon" pos="NN" />
      <w surface="drops" lemma="drop" pos="NN" />
      <freq value="2" />
    </ngram>
   </occurs>
</cand>
```

Figure 4: Sample collocation candidate

4.3 Dictionary look-up module

This module connects to the online translation dictionary *WordReference* to attempt a direct translation in TL of a collocation in SL. *WordReference* translation entries include two tables: one for one-word direct translations (*principal translations*), and another for translations of MWEs (*compound forms*). Furthermore, the dictionary lists its translation entries in order from the most common to the least common. A Python script was written to handle the connection to the *WordReference* API. Our task is to look at the *compound forms* table and attempt to find a match for our collocation. If such a match is found, its translation from the HTML is extracted, and presented to the user. If no match is found, then translation will be based on the bilingual corpora presented by the user as input, triggering the *parallel corpora* or the *comparable corpora module* accordingly.

4.4 Parallel corpora module

This module first employs Hunalign to align the input corpora. Next, after *syntactic processing* and *semantic processing*, transformational rules are applied in order to identify the TL translations of all collocations extracted from SL. A sample output of Hunalign is presented in Figure 5: the first column refers to a sentence number in the SL corpus, the second column refers to a sentence number in the TL corpus, and the third column represents a confidence value, or the estimated certainty of the SL-TL pairing.

```
101      96        0.336927
102      97        0.583117
103      98        0.228
104      99        0.229412
105      100       0.226056
```

Figure 5: Sample Hunalign output

Semantic processing consists in identifying the base of a collocation in SL and finding its translation in TL. The POS-tags of the components of the collocation (see Figure 4) will help completely determine its base. This is because the POS-pattern of the collocation should adhere to one of the set of POS-patterns defined previously (see Figure 3). Next, the components representing the base for collocations will be identified following their linguistic model (see Section 2.1). Finally, *WordReference* is employed to retrieve the first three translation entries that match the POS-tag of our base from its *principal translations* table.

Similarly, *syntactic processing* consists in finding the translations of the collocatives in the TL corpus. It requires the output of both the *candidate filtering* module, which is an XML file containing a set of SL collocations (see Figure 4) and that of Hunalign presented above (see Figure 5). It also requires, as input, the TL corpus, which is a translation of the SL corpus. We implemented an algorithm that first reads the SL corpus and finds all sentences where a collocation appears, and then performs these tasks for each of the retrieved SL sentences:

- Read the output of Hunalign and match the SL sentence to its TL counterpart, where the translation of the collocation should appear.
- Expand this TL sentence to a window of five sentences to be extracted and analysed, to make up for any Hunalign precision error.
- For each of the translations in TL of the collocation's base, obtained during semantic processing, go through our window of sentences, one sentence at a time, and look for the presence of the translation within it. If a match is found, it means the translation of the collocatives in TL should also be present within the sentence.
- POS-tag the matching TL sentence using TreeTagger.
- Apply a transfer rule (see Table 3) to obtain the translation of the collocatives in TL.

4.5 Comparable corpora module

This module computes similarity classes in order to find the TL translations of all extracted SL collocations via *query expansion*, *query translation*, and *context generalisation*.

Query expansion produces a generalisation of the SL collocation's context by computing two different similarity classes, one centred on the base of the collocation, and another on its context (two open-class words that appear before it, and two after it). Computing similarity classes requires the use of a thesaurus. For English, we use WordNet, and obtain the first five synsets of the same POS-tag of any given open-class word. As for Spanish, *WordReference* is made use of. Our first similarity class, the one centred on the base of the collocation, will thus consist of up to six words, the original base itself and up to five synonyms. Correspondingly, our second similarity class will consist of up to 24 words: the four context words we retrieved, and up to five synonyms for each of them.

Next in the pipeline process is *query translation*, which computes a translation class, an expansion of the target language translations of the words that make up our original similarity class. Here again, we rely on *WordReference* as our de facto bilingual dictionary and thesaurus. For each of our two similarity classes, we iterate through all of their words, look up each via the *WordReference* API and retrieve up to five translation entries that match their POS-tags, and then further expand these by retrieving up to five thesaurus entries for each. This means that our first translation class, the one centred on the base of the collocation, will contain up to 30 translations for each of the (up to) six words of its similarity class, which totals up to 180 words. Similarly, our second translation class, centred on context words, will contain up to 720 words.

Finally, *context generalisation* aims at finding TL translations of a SL collocation by comparing context similarities. We first determine the POS-pattern of our SL collocation, and then see if any of the words in the translation class of its base corresponds with the base of any of the TL collocations of the same POS-pattern. If a match is found, we compute a similarity class for the context of the matched TL collocation and we see if it has any elements in common with the context of the SL collocation. If it does, we present it to the user as a potential translation of the collocation from the original text.

5 Evaluation

The choice of our experimental corpora was made completely on the basis of the profiles of the target users of our system: language learners and translators. Reading in a target language is an integral component of any language-learning process. We chose *Harry Potter and the Philosopher's Stone* and its translation into Spanish, *Harry Potter y la Piedra Filosofal*, to exemplify this. Similarly, professional translators usually specialise in a certain domain of translation, and therefore must translate technical terminology on a regular basis. Thus, we chose the *Ecoturismo corpus*[2], a collection of multilingual parallel and comparable corpora on tourism and tourism law, as it represents a real-life example of the technical documents a translator works with.

5.1 Experimental setup

Two bilingual annotators, fluent in English and Spanish, reviewed the output of our system after processing both experimental corpora. They assigned a score to the translations the system offered for each collocation according to a five-point scale (with 5 representing an excellent translation). Precision and recall are estimated from these scores for each case study.

5.2 Experimental results

100 English collocations were retrieved from the *Harry Potter* corpus. 12 collocations were successfully translated directly, using *WordReference*, such as *to talk rubbish*, *to speak calmly*, *fast asleep*, and *to lean against the wall*. Out of the remaining 88 collocations, 10 could not be translated at all, and 78 were translated using our approach to processing parallel corpora. Table 4 summarises these results (*A* stands for annotator, *WR* for *WordReference*, and *AVG* for average).

A	WR	1	2	3	4	5	AVG
#1	12	0	0	11	16	51	4.51
#2		0	0	8	17	53	4.58

Table 4: Parallel corpora result scores

As it can be observed, we obtained a high average score of 4.55 for the quality of translations retrieved from parallel corpora. Moreover, only 10 collocations out of the original 100 could not be translated, yielding an equally high score for recall, of 90%.

Similarly, 100 Spanish collocations were retrieved from the *Ecotourism* corpus. 15 of them were translated using *WordReference*; all of these were of the Spanish POS-patterns *NN + JJ or NN + de + NN*, such as *transporte público*, *asistencia técnica*, and *viaje de negocios*. Out of the 85 remaining collocations, 15 could not be translated at all, and the other 70 received translation suggestions found in the comparable corpora. Table 5 summarises the results.

A	WR	1	2	3	4	5	AVG
#1	15	7	14	19	17	13	3.21
#2		8	15	21	15	11	3.09

Table 5: Comparable corpora result scores

Despite the rather low average score of 3.15 for the quality of translations, we managed to provide translation suggestions to 85% of the collo-

[2] Compiled in the framework of the R&D project *Espacio Único de Sistemas de Información Ontológica y Tesaurus sobre el Medio Ambiente: ECOTURISMO* (Spanish Ministry of Education, FFI2008-06080-C03-03).

cations. We can conclude that by imposing flexible constraints on the matching process performed during the task of context generalisation, we obtain average translations for a high number of collocations. These constraints refer to the size of our context window and the number of thesaurus entries we retrieve for each original word during query expansion. Improving our precision score would mean strengthening these constraints, but this would also result in a lower recall. Moreover, in this particular case, recall of the output is more relevant than precision because our suggested translations, even if not always excellent, might offer translators a useful hint for correctly translating collocations.

5.3 Discussion and future work

Against the background of the limitations of the current version of our system, we propose the following future improvements. First, we exploit the nature of collocations as cohesive lexical clusters, but disregard the linguistic property of semantic idiomaticity that differentiates them from other MWEs, such as idioms. Our system cannot, therefore, differentiate between collocations and other MWEs in terms of compositionality. Secondly, we would like to provide better integration between the stages of collocation extraction and collocation translation. Currently, the former relies on TreeTagger and the MWEToolkit, while the latter makes use of Hunalign. This means that all users would also have to have access to these three tools; this poses no significant problem because all of them are open source, and readily available online, but it would be simpler to integrate the tasks performed by these tools into our system in order to increase its ease of use. Finally, we would like to investigate the use of the web as a corpus to find proficient ways of using information offered by search engines.

The expected final users of our system correspond to one of two groups: professional translators and language learners. However, as aforementioned, further fine-tuning of the system might be worthwhile in order to better address the specific needs of these particular user groups. Working with comparable corpora is not highly reliable because of its noisy nature. We opted to impose flexible constraints on the matching process performed during the last stage of comparable corpora processing, context generalisation, in order to increase the recall of our system. As stated before, this would be better suited to translators, who could benefit from the translation suggestions offered by our system to find the most adequate translation of a collocation. Language learners, however, are probably more interested in learning very precise translations for several collocations, rather than translation suggestions for a large number of collocations. A way forward would be to adjust the comparable corpora algorithm so it can impose stronger constraints during the task of context generalisation, to the benefit of language learners.

Future research goals could include (1) providing better integration between the different stages of the project, (2) finding a way to further exploit the use of the web as a corpus to aid in the processes of collocation retrieval and translation, (3) demonstrating the flexibility of our framework by adjusting our system to work with several other languages, and (4) tailoring the constraints imposed by our system to better meet the needs of our final users.

Acknowledgements

This project was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT programme.

References

Baldwin, T., and Kim, S. N. (2010). Multiword Expressions. In: *Handbook of Natural Language Processing*, second edition. Boca Raton, FL.

Biber *et al.* (1999). *Longman Grammar of Spoken and Written English*. Longman, Harlow.

Bradford, W., and Hill, S. (2000). *Bilingual Grammar of English-Spanish Syntax*. University Press of America.

Brown P., Lai J., and Mercer R. (1991). Aligning Sentences in Parallel Corpora. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, Canada, pp. 169-176.

Cardey, S., Chan, R. and Greenfield, P. (2006). The Development of a Multilingual Collocation Dictionary. In: *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, Sydney, pp. 32-39.

Choueka, Y., Klein, T., and Neuwitz, E. (1983). Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus. In: *Journal for Literary and Linguistic Computing*, 4(1): pp. 34-38.

Church, K. W., and Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. In: *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pp. 76-83.

Corpas Pastor, G. (1995). *Un Estudio Paralelo de los Sistemas Fraseológicos del Inglés y del Español.* Málaga: SPICUM.

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

Corpas Pastor, G. (2013). Detección, Descripción y Contraste de las Unidades Fraseológicas mediante Tecnologías Lingüísticas. Manuscript submitted for publication. In *Fraseopragmática,* I. Olza and E. Manero (eds.). Berlin: Frank & Timme.

Fung, P., and Yuen, Y. (1998). An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In: *Proceedings of the 17th International Conference on Computational Linguistics*, pp. 414-420.

Gale W., and Church K. (1993). A Program for Aligning Sentences in Bilingual Corpora. In: *Journal of Computational Linguistics*, 19: pp. 75-102.

Gelbukh A., and Kolesnikova O. (2013). Expressions in NLP: General Survey and a Special Case of Verb-Noun Constructions. In *Emerging Applications of Natural Language Processing: Concepts and New Research*, S. Bandyopadhyay, S. K. Naskar, and A. Ekbal (eds.). Hershey: Information Science Reference. IGI Global.1-21.

Hausmann, F. (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: *Lexikographie und Grammatik*, (Lexicographica, series maior 3), ed. H. Bergenholtz and J. Mugdan. Tübingen: Niemeyer. 175-186.

H.H. Hoang, S.N. Kim, M.Y. Kan, A Re-examination of Lexical Association Measures, In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP*, Singapour: ACL and AFNLP. 31-39.

Jackendoff, R. (2007). *Language, Consciousness, Culture: Essays on Mental Structure.* The MIT Press.

Lea D. and Runcie, M. (2002). *Oxford Collocations Dictionary for Students of English*. Oxford University Press.

Lü, Y. and Zhou, M. (2004). Collocation Translation and Acquisition Using Monolingual Corpora. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04)*, pp. 167-174.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010). MWEToolkit: A Framework for Multiword Expression Identification. In: *Proc. of LREC'10 (7th International Conference on Language Resources and Evaluation).*

Ramisch, C. (2012). A Generic Framework for Multiword Expressions Treatment: from Acquisition to Applications. In: *Proceedings of ACL 2012 Student Research Workshop*, pp. 61-66.

Rapp, R. (1995). Identifying Word Translations in Nonparallel Texts. In: *Proceedings of the 35th Conference of the Association of Computational Linguistics*, pp. 321-322. Boston, Massachusetts.

Sag, I. et al. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (COCLing-2002)*, pp. 1-15.

Santana, O. *et al*. (2011). Extracción Automática de Colocaciones Terminológicas en un Corpus Extenso de Lengua General. In: *Procesamiento del Lenguaje Natural*, (47): pp. 145-152.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Seretan, V. (2011). *Syntax-Based Collocation Extraction (Text, Speech and Language Technology)*, 1st Edition. Springer.

Sharoff, S., Babych, B., & Hartley, A. (2009). "Irrefragable answers" using comparable corpora to retrieve translation equivalents. In: *Language Resources and Evaluation*, 43(1): pp. 15-25.

Sinclair, J., and Jones, S. (1974). English Lexical Collocations: A study in computational linguistics. In: *Cahiers de lexicologie*, 24(2): pp. 15-61.

Smadja, F. (1993). Retrieving collocations from text: Xtract. In: *Computational Linguistics*, 19(1): pp. 143-177.

Varga, *et al.* (2005). Parallel corpora for medium density languages. In*: Proceedings of the RANLP 2005*, pp. 590-596.

Wehrli, E., Nerima, L., and Scherrer, Y. (2009). Deep linguistic multilingual translation and bilingual dictionaries. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 90-94.