# Alignment-based extraction of multiword expressions

**Helena de Medeiros Caseli** · **Carlos Ramisch** · **Maria das Graças Volpe Nunes** · **Aline Villavicencio**

**Abstract** Due to idiosyncrasies in their syntax, semantics or frequency, Multiword Expressions (MWEs) have received special attention from the NLP community, as the methods and techniques developed for the treatment of simplex words are not necessarily suitable for them. This is certainly the case for the automatic acquisition of MWEs from corpora. A lot of effort has been directed to the task of automatically identifying them, with considerable success. In this paper we propose an approach for the identification of MWEs in a multilingual context, as a by-product of a word alignment process, that not only deals with the identification of possible MWE candidates, but also associates some multiword expressions with semantics. The results obtained indicate the feasibility and low costs in terms of tools and resources demanded by this approach, which could, for example, facilitate and speed up lexicographic work.

**Keywords** Automatic identification · Word alignment · Machine translation · Terminology · Multiword expressions · Lexical acquisition · Statistical methods

## 1 Introduction

A multiword expression (MWE) can be defined as any word combination for which the syntactic or semantic properties of the whole expression cannot be obtained from its parts (Sag et al 2002).

Multiword expressions play an important role in Natural Language Processing (NLP) applications, which should not only identify the MWEs but also be able to deal with them when they are found (Fazly and Stevenson 2007). Failing to identify MWEs may cause serious problems for many NLP tasks, especially those envolving some kind of semantic processing. Therefore, there is an enormous need for robust (semi-)automated ways of acquiring lexical information for MWEs (Villavicencio et al 2007).

MWEs are language dependent and culturally motivated, which means that the translation of MWE occurrences is an important challenge for machine translation methods. Different approaches have been proposed for identifying

H. M. Caseli
NILC – Department of Computer Science, Federal University of São Carlos (Brazil)
E-mail: helenacaseli@dc.ufscar.br

C. Ramisch
Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
E-mail: ceramisch@inf.ufrgs.br

M. G. V. Nunes
NILC – ICMC, University of São Paulo (Brazil)
E-mail: gracan@icmc.usp.br

A. Villavicencio
Institute of Informatics, Federal University of Rio Grande do Sul (Brazil), and Department of Computer Science, University of Bath (UK)
E-mail: avillavicencio@inf.ufrgs.br

MWEs in one language (Pearce 2002; Baldwin and Villavicencio 2002; Evert and Krenn 2005; Zhang et al 2006; Villada Moirón and Tiedemann 2006; Villavicencio et al 2007; Van de Cruys and Villada Moirón 2007). However few have investigated this problem in the multilingual context of machine translation, more specifically within the task of automatic word alignment (mainly the models of Brown et al (1993)), which plays a vital role in corpus-based (example-based or statistical) MT approaches.

The automatic word alignment of two parallel texts — a text written in one (source) language and its translation to another (target) language — tries to identify for each word in a source sentence equivalences in the parallel target sentence. Therefore, if a word sequence $S$ ($S = s_1 \dots s_n$ with $n \geq 2$) in one text is aligned to a word sequence $T$ ($T = t_1 \dots t_m$ with $m \geq 1$) in its counterpart, that is $S \leftrightarrow T$, then we can assume that: (a) $S$ and $T$ share some semantic features, and (b) $S$ may constitute a MWE. Then we state that the sequence $S$ will be a MWE candidate if it is aligned with a sequence $T$ composed of one or more words (a $n : m$ alignment with $n \geq 2$ and $m \geq 1$). For example, the sequence of two Portuguese words *academic world* is a MWE candidate because these two words were joined to be aligned with the sequence of two words *ambiente acadêmico* (a 2 : 2 alignment) and also with the single Portuguese word *academia* (a 2 : 1 alignment).

Thus, notice that the alignment-based MWE extraction method does not rely on the conceptual asymmetries between languages since it does not expect that a source sequence of words be aligned with a single target word. The method indeed looks for the sequences of source words that are frequently joined together during the alignment despite the number of target words involved. These MWE candidates may then be automatically validated, and the noisy non-MWE cases among them removed. As a consequence, MWE extraction can benefit from automatic word alignment of parallel texts without prior MWE information.

In this paper, we investigate experimentally whether MWEs and their translations can be identified as a by-product of the automatic word alignment of parallel texts, with reasonable precision rates. We focus on English MWEs and their Portuguese translations. Another important result obtained by this paper is that the word alignment is able to attach semantic information to word and multiword units, by means of their target language counterparts. This approach can help to considerably reduce and accelerate lexicographic work by generating lists of MWEs with their translations, for the construction of bilingual resources, and/or with some semantic information for monolingual resources.

The remainder of this paper is structured as follows. Section 2 briefly discusses MWEs and some previous works on methods for automatically extracting them. Section 3 describes the method proposed to extract MWEs and their translations as a by-product of an automatic word alignment process. Section 4 presents the evaluation methodology and analyses the results and section 5 finishes this paper with some conclusions and proposals for future work.

## 2 Related Work

The term Multiword Expression has been used to describe a large number of distinct but related phenomena, such as phrasal verbs (e.g. *come along*), nominal compounds (e.g. *frying pan*), institutionalised phrases (e.g. *bread and butter*), and many others. They are very frequent in everyday language and, in English, Jackendoff (1997) estimates the number of MWEs in a speaker's lexicon to be comparable to the number of single words. This is reflected in several existing grammars and lexical resources, where almost half of the entries are Multiword Expressions.

However, due to their heterogeneous characteristics, MWEs present a tough challenge for both linguistic and computational work (Sag et al 2002). Some MWEs are fixed, and do not present internal variation, such as *ad hoc*, while others allow different degrees of internal variability and modification, such as *touch a nerve* (*touch/find a nerve*) and *spill beans* (*spill several/musical/mountains of beans*). In terms of semantics, some MWEs are more opaque in their meaning (e.g. *to kick the bucket* as *to die*), while others have more transparent meanings that can be inferred from the words in the MWE (e.g. *eat up*, where the particle *up* adds a completive sense to *eat*). Therefore, providing appropriate methods for the automatic identification and treatment of these phenomena is a real challenge for NLP systems.

Previous works on MWE identification have often used statistical measures alone (Pearce 2002; Evert and Krenn 2005; Zhang et al 2006; Villavicencio et al 2007) or combined with some kinds of linguistic information such as syntactic and semantic properties (Baldwin and Villavicencio 2002; Van de Cruys and Villada Moirón 2007) or automatic word alignment (Villada Moirón and Tiedemann 2006). For instance, Evert and Krenn (2005) compare the use of some statistical measures for MWE identification, and find that the efficacy of a given statistical measure depends on factors like the type of MWEs being targeted for identification, the domain and size of the corpora used, and the amount of low-frequency data excluded by adopting a threshold.

Looking from a different perspective Villavicencio et al (2007) investigate the use of statistical measures (mutual information, permutation entropy and $\chi^2$) for automatically identifying MWEs in general, discussing the influence of the corpus size and nature over the methods. Their results suggest that these different measures have a high level of agreement about MWEs, whether in carefully constructed corpora or in more heterogeneous web-based ones. Moreover, the application of these methods shows that grammar coverage can be significantly increased if MWEs are properly identified and treated.

Among the methods that use additional information along with statistics to extract MWE, the one proposed by Villada Moirón and Tiedemann (2006) seems to be the most similar to our approach. The main difference between them is the way in which word alignment is used in the MWE extraction process. In this paper, the word alignment is the basis of MWE extraction process while Villada Moirón and Tiedemann's method uses the alignment just for ranking the MWE candidates which were extracted on the basis of association measures (log-likelihood and salience) and head dependence heuristic (in parsed data).

Our approach follows to some extent that of Zhang et al (2006), that used error mining methods for the detection of missing lexical entries for MWEs and related constructions, as this paper focuses on the extraction of generic MWEs as a by-product of an automatic word alignment. Another related work is the automatic detection of non-compositional compounds (NCC) by Melamed (1997) in which NCCs are identified by analyzing statistical translation models trained in a huge corpus by a time-demanding process. Both approaches look for sequecences of words that are translated as a unit but while our method takes as MWE candidates any two or more consecutive source words, regardless of whether they are translated as one or more target words, Melameds method does not detect phrases that are translated word-for-word.

To the best of our knowledge this is the first work that investigates to what extent automatic word alignment can be used to extract MWEs, that is, the first alignment-based MWE extraction method. In this way, cost-effective tools for the automatic alignment of texts can generate a list of MWE candidates with their appropriate translations, for bilingual lexicons, or without the translations, for monolingual lexicons.

## 3 Experimental Methodology

In order to verify how the alignment process can contribute to MWE extraction, we propose the following steps. First, a parallel corpus has to be pre-processed to be used as the input for our MWE extraction method. The parallel corpus is sentence-aligned, then PoS (Part-of-Speech) tagged, and finally word-aligned by automatic tools as explained in Section 3.1. Then, for each language, a list of MWE candidates is created by extracting those sequences of two or more words that have the same alignment, that is, that are linked to the same unit in the other language. Each list is filtered to remove unlikely candidates according to some empirical criteria. The extraction method is described in detail in Section 3.2.

As a consequence, from a parallel corpus it is possible to obtain as products of the same process: (1) a list of MWEs for each language as well as (2) the corresponding translation(s) of each MWE. It is important to notice that the translation of a MWE in one language is not necessarily a MWE in the other language. Indeed, a MWE can be translated as a single word, e.g. *eat up* in English as *comer* (*eat*) in Portuguese. Moreover, different occurrences of a MWE can be aligned to distinct translations. For instance, the expression *academic world* in English may be translated into Portuguese as *academia* or as *ambiente acadêmico* depending on the context.

### 3.1 Preprocessing of the Corpus

The corpus used in this experiment is composed of articles of a Brazilian scientific magazine *Pesquisa FAPESP* (journalistic genre and academic-scientific domain)[1], written in Portuguese (pt) and English (en). Table 1 contains the number of texts, sentences and tokens for each language in this test corpus.

The pt–en corpus was sentence-aligned by a version of the Translation Corpus Aligner (TCA) (Hofland 1996) called TCAalign. It relies on several alignment criteria to automatically find the correspondence between source and

[1] *Pesquisa FAPESP* is available at http://revistapesquisa.fapesp.br.

**Table 1** Number of texts, sentences and tokens, for each language, in the test corpus

| Language | Texts | Sentences | Tokens |
|---|---|---|---|
| pt | 646 | 17,397 | 494,391 |
| en | 646 | 17,397 | 532,121 |
| Total | 1,292 | 34,794 | 1,026,512 |

**Table 2** Number of surface forms covered by the original and the extended morphological dictionaries

| Language | Original | Extended |
|---|---|---|
| pt | 128,772 | 1,136,536 |
| en | 48,759 | 61,601 |

target sentences, such as a bilingual anchor word list, words with an initial capital (candidates for proper nouns), special characters (such as question and exclamation marks), cognate words and sentence lengths (Caseli et al 2004). `TCAalign` achieved 97% precision and 98% recall in the sentence alignment of the test corpus. It is important to note that after the automatic alignment, all alignments different from 1 : 1 (only 6% of the total amount) were manually verified before being used in the next preprocessing steps.

The aligned parallel sentences were then PoS-tagged in each language using the corresponding morphological analysers and PoS taggers from `Apertium` (Armentano-Oller et al 2006). The morphological analysis provides one or more lexical forms or analyses (information on lemma, lexical category and morphological inflection) for each surface form (instance of a word in the text) using a monolingual morphological dictionary. The PoS tagger chooses the best possible analysis based on a first-order hidden Markov model (HMM). To improve the coverage of morphological analysis, the original ditionaries were enlarged with entries from Unitex[2] dictionaries as explained by Caseli et al (2006). The number of surface forms covered by the original and the extended versions of morphological dictionaries are shown in Table 2.

After PoS-tagging, the pairs of parallel sentences were word-aligned by `GIZA++` (Och and Ney 2000b). `GIZA++` is a statistical word aligner that uses the IBM models (Brown et al 1993) and the Hidden-Markov alignment model (Vogel et al 1996; Och and Ney 2000a) to find the best correspondences between source and target tokens. `GIZA++` (version 2.0) was executed with standard parameters — with iterations of IBM-1, IBM-3, IBM-4 and HMM — and trained with the whole set of 17,397 pairs of pt–en parallel sentences.

The parallel sentences were aligned by `GIZA++` in source–target and target–source directions and, then, the alignments in both directions were merged using the union algorithm proposed by Och and Ney (2003). The alignment error rate of `GIZA++` in our test corpus after the union was 8.61% (Caseli et al 2006).

Figure 1 shows an extract of a pt–en sentence pair in which each surface form (the word as it appears in the text, e.g. the underlined en word *blood*) is followed by the output of the tagger (its lemma and PoS tags, e.g. *blood*<n><sg>) and the alignment produced by the word aligner (the position of the corresponding token on the other side, e.g. 23).

Multiword unit alignments are formed by joining positions of the correspondent tokens, separeted by a "_" character, as in the underlined alignment of Figure 1 between the pt word *pressão* (the 23rd source token) and two en words: *blood* and *pressure* (the 27th and 28th target tokens). This 1 : 2 alignment connects a source single pt word *pressão* to the target en multiword unit *blood pressure*. If there are MWE entries in the morphological dictionaries, they can be recognized by the PoS taggers, as the single 34th en token *heart_attacks*. Nevertheless, the MWE coverage of the morphological dictionaries is usually very limited so that the automatic extraction method proposed in this paper is crucial for and effective in extracting relevant MWEs that are not included in these dictionaries.

## 3.2 Extraction Method

The extraction of both MWEs and their translations from the word-aligned corpus was carried out in two steps. In the first step, MWE candidates are selected through the identification of special alignments as explained in Section 3.2.1. In the second one, empirical rules are applied to discard unlikely candidates as described in Section 3.2.2. The remaining units in the candidate list are considered to be MWEs.

[2] http://www-igm.univ-mlv.fr/~unitex/

| | |
|---|---|
| pt | <s snum=6>Em/Em<pr>:1 média/média<n><f><sg>:2 ,/,<cm>:3 … 20/20<num>:20 %:21 dos/de<pr>+o<det><def><m><pl>:22 homens/homem<n><m><pl>:24 apresentam/apresentar<vblex> <pri><p3><pl>:25 pressão/pressão<n><f><sg>:27_28 alta/alto <adj><f><sg>:26 ,/,<cm>:29 um/um<det><ind><m><sg>:30 fator/fator<n><m><sg>:32 de/de<pr>:33 risco/risco<n><m> <sg>:31 de/de<pr>:33 enfarte/enfarte<n><m><sg>:34 e/e <cnjcoo>:35 derrames/derramar<vblex><prs><p2><sg>:36 cerebrais/cerebral<adj><mf><pl>:36 ./.<sent>:37 </s> |
| en | <s snum=6>On/On<pr>:1 average/average<n><sg>:2 ,/,<cm>:3 … 20/20<num>:18 %:19 of/of<pr>:20 the/the<det><def><sp>:0 men/man<n><pl>:21 showed/show<vblex><past>:22 high/high <adj><sint>:24 blood/blood<n><sg>:23 pressure/pressure<n> <unc><sg>:23 ,/,<cm>:25 a/a<det><ind><sg>:26 risk/risk<n> <sg>:29 factor/factor<n><sg>:27 for/for<pr>:28_30 heart_attacks /heart_attack<n><pl>:31 and/and<cnjcoo>:32 strokes/stroke<n> <pl>:33_34 ./.<sent>:35 </s> |

**Fig. 1** An extract of a pt–en PoS-tagged and word-aligned sentence pair.

### *3.2.1 Extraction of MWE candidates and their translations*

For each language, a list of sequences of two or more words (the MWE candidates) was created from the output of the aligner or the tagger, along with the target words aligned to them (the possible translations).

The candidates produced by the aligner are those in which two or more words have the same alignment, that is, they are linked to the same target unit. For example, in Figure 1, the pt sequence *derrames cerebrais* (at positions 33 and 34 in the pt sentence) is aligned to the 36th en word *strokes*. In the other direction, the en sequence *blood pressure* (at positions 27 and 28 in the en sentence) is aligned to the 23rd pt word *pressão*.

The candidates produced by the PoS tagger are those in which the words are joined by a "_" character as defined in the morphological dictionary. In Figure 1, the en sequence *heart_attacks* (at position 34 in the en sentence) is an example of a sequence generated by the tagger. This en sequence is aligned to the 31st pt word *enfarte*. We make this distinction between the MWE candidates produced by the aligner or by the tagger in order to evaluate precisely the gain obtained by using the automatic word alignment in MWE extraction. The presence of manually defined MWEs (those contained in the dictionaries used by the tagger) would certainly add some noise to the evaluation process.

At the end of the first step, a list of MWE candidates is output along with their possible translations, that is, the target units with which the candidates are aligned. In fact, our method produces, at once, two lists of MWE candidates: (1) the list of MWE candidates in pt along with their translations in en and (2) the list of MWE candidates in en along with their translations in pt.

### *3.2.2 Filtering the candidates*

The list of MWE candidates created in the first step was then filtered to remove those candidates that: (a) match some sequences of PoS tags or words (patterns) empirically defined, or (b) whose frequency is below a certain threshold.

The filtering patterns are language dependent and were defined, in previous experiments, after a manual analysis of output candidates that did not correspond to true MWEs. Table 3 shows the set of eigth patterns used to filter the en candidates (1st column) and some examples of false positive MWEs filtered by them (2nd column). Since the sentences were PoS-tagged automatically, even the incorrectly tagged sequences can match these patterns. The filter is not error-free, so sequences that are true MWEs can be erroneously filtered (e.g. *from A to Z*, *from day to day*, *would give anything*, *My God*, *his Majesty*, *I beg your pardon*) being considered false negatives.

Finally, a threshold of 2 occurrences was empirically defined to remove the infrequent candidates (that occur less than twice) in the parallel sentences. Table 4 shows some examples of en MWE candidates along with their frequencies and an indication of the tool that identified it (its source): the aligner (A) or the tagger (T). The possible pt translations for each candidate are also shown along with their frequencies in the test corpus. From this table it is possible to see that the PoS-tagger (T) can fail not only in the PoS tagging process but also in the identification of MWE candidates since de

**Table 3** PoS and word sequences used for filtering en MWE candidates

| Pattern beggining with | Filtered candidates |
|---|---|
| determiner | a detector, a cure, an increase, the american, the atimospheric institute |
| auxiliary verb | does exist, did not, did you, had become, will be, will gain, would allow |
| pronoum | he called, he argues, their children, his life, these are, this spirit |
| adverb | widely studied, publicly stored, not yet, since then, under suspicion |
| conjunction | as smoke cover, or produces, as in workers, and yet, and hence |
| are, is, was, were | are already, are a result, is to, were able, was formed |
| that, what, when, which, who, why | that are, that varies, what was, why do, which lasts, who responds |
| from, to, of | from them, from Bahia, to build, to the, of cell, of our, of this |

**Table 4** Examples of MWE candidates en and their pt translations output by the aligner (A) or the tagger (T)

| MWE candidates | | | Possible translations | |
|---|---|---|---|---|
| | Frequency | Source | | Frequency |
| able_to | 2 | A | consegue_se | 1 |
| | | | consegue | 1 |
| academic_world | 6 | A | academia | 5 |
| | | | ambiente_acadêmico | 1 |
| accompanied_by | 6 | A | acompanhada_de | 2 |
| | | | acompanhados | 2 |
| | | | acompanhado | 1 |
| | | | acompanhado_de | 1 |
| a_hundred | 11 | T | cem | 9 |
| | | | centenário | 1 |
| | | | 100_projetistas | 1 |
| hoped_for | 3 | A | esperados | 2 |
| | | | esperadas | 1 |
| human_being | 7 | T | ser_humano | 6 |
| | | | pessoa | 1 |

**Table 5** Number of en MWE candidates which were extracted in this experiment

| en MWE candidates | Number |
|---|---|
| Identified by the aligner or the tagger (1st step) | 37,267 |
| Filtered by PoS/word patterns (2nd step) | 27,402 |
| Filtered by threshold (2nd step) | 8,609 |
| Final Set | 1,256 |

sequence *a hundred* is a false positive. The PoS patterns used for filtering may not be able to discard these false positives since their tags may not match the filtering patterns. In the example given, *a hundred* was tagged as *num* (numeral) in spite of beginning with a determiner.

Table 5 sumarizes the number of en MWE candidates which were extracted in this experiment. From the first step (see section 3.2.1) we obtained a list of 37,267 sequences of two or more words identified by the aligner or by the tagger. From this list of en candidates, 27,402 were filtered by the patterns of Table 3. From the remaining candidates, 8,609 were excluded because their frequencies were lower than the minimum threshold (2). The 1,256 remaining candidates were evaluated as explained in the next section.

## 4 Evaluation and Results

To evaluate the efficacy of the proposed method, an automatic comparison was performed using two reference dictionaries composed of multiword expressions, followed by an analysis by human experts. In this paper we evaluated the 1,256 `en` MWE candidates extracted as described in section 3.2. The methodology consisted of the following steps:

1. **Resource-based evaluation**
   The 1,256 `en` MWE candidates were first lemmatised by RASP system (Briscoe and Carroll 2002), and then compared to the MWEs defined in the reference dictionaries. For this purpose we used the Cambridge International Dictionary of English (CIDE) (Procter 1995), the Cambridge International Dictionary of Phrasal Verbs (CIDPV) (paperback edition 1997) and a list of phrasal verbs automatically collected from both the British National Corpus (Burnard 2000) and the World Wide Web using the methods described in (Villavicencio 2005). This evaluation was done to verify the existence of the (lemmatized) MWE candidate in at least one of the dictionaries, and if found, the candidate was considered to be a true MWE.
   317 candidates (25.2% of the total amount) were found in at least one reference dictionary.[3] In the absence of MWEs translations in these dictionaries, human experts evaluated all possible translations of the 317 true MWEs, as discussed in the next step. As the coverage of each of these dictionaries may be low, as discussed by Villavicencio (2005) for Verb-Particle Constructions, the second step is necessary to analyse the remaining 939 candidates (74.8%).
2. **Human analysis**
   The MWE candidates that were not found in any reference dictionary were analysed by two non-native human experts who also verified the correctness of the corresponding translations of these candidates. The judges classified each of the 939 candidates as true, if it is a multiword expression, or false, otherwise. For the judges, a sequence of words was considered a MWE mainly if it was: (1) a proper name, (2) a phrasal verb or (3) a sequence of words for which the meaning cannot be obtained by compounding the meanings of its words. Furthermore, they also classified each possible translation as true or false, according to how acceptable they were.
   The judgments of both judges were compared and a disagreement of approximately 11% on multiwords and translations was verified. This disagreement was also measured by the kappa ($\kappa$) measure (Carletta 1996), being $\kappa = 0.768$ for multiwords and $\kappa = 0.761$ for translations, which does not prevent conclusions to be drawn. According to Carletta (1996), among other authors, a value of $\kappa$ between 0.67 and 0.8 indicates a good agreement.

To illustrate these results, Table 6 presents the same examples from Table 4 but now along with their respective evaluations given by the reference dictionaries (D) and by both judges (J1 and J2). We can see in this table the false positive *a hundred* marked as a false candidate (a non-MWE) by D, J1 and J2.

In order to calculate the percentage of true candidates among the 1,256, two approaches can be followed, depending on what criteria one wants to emphasize: precision or recall. To emphasize the precision, one should consider as genuine MWEs only those candidates classified as true by *both* judges, on the other hand, to emphasize the recall, one should consider also those candidates classified as true by *just one of them*. So, in the following tables both values are shown as the lower (the first value) and the upper (the second value) bounds of an interval, respectively.

Following Piao et al (2006), Table 7 presents the set of candidates divided into frequency classes. This table shows the number (#) and the percentage (%) of MWE candidates classified as true by both (the lower bound) or at least one (the upper bound) human judge and also those candidates classified as true in the resource-based evaluation. Considering the 317 candidates classified as true during the resource-based evaluation, and the 302 candidates classified as true by both judges, and the 144 classified as true by at least one of them, the percentage of true candidates ranges from 49.28% (317 + 302 = 619 out of 1,256) to 60.75% (619 + 144 = 763 out of 1,256). The highest precision (71.71%) was obtained for the frequency range between 10 and 99. Examples of high-frequency (freq >= 100) false MWEs are those output by the tagger — *as a* (freq = 337) and *as an* (freq = 100) — and by the aligner — *in a* (freq = 174), *in this* (freq = 205) and *years ago* (freq = 169).

These conclusions corroborate those by Piao et al (2006) in which Chinese MWE are extracted using a statistical tool achiving precisions ranging from 61.16% to 68.82% according to different search window lengths. The highest precision reached by their method was also in the frequency range between 10 and 99 (76.36%).

[3] For example: “artesian wells”, “black hole” and “botanical gardens” are found in CIDE, “clean up”, “consist of” and “depend on” are found in CIDPV.

**Table 6** Examples of MWE candidates and their translations output by the aligner (A) or the tagger (T) after evaluation

| **MWE candidates** | | | | **Possible translations** | | |
|---|---|---|---|---|---|---|
| | **D** | **J1** | **J2** | | **J1** | **J2** |
| able_to | F | T | T | consegue_se | F | F |
| | | | | consegue | T | F |
| academic_world | F | T | T | academia | T | T |
| | | | | ambiente_acadêmico | T | T |
| accompanied_by | F | F | T | acompanhada_de | T | T |
| | | | | acompanhados | F | F |
| | | | | acompanhado | F | F |
| | | | | acompanhado_de | T | T |
| a_hundred | F | F | F | cem | T | T |
| | | | | centenário | F | F |
| | | | | 100_projetistas | F | F |
| hoped_for | T | – | – | esperados | T | T |
| | | | | esperadas | T | T |
| human_being | F | T | T | ser_humano | T | T |
| | | | | pessoa | T | F |

**Table 7** Evaluation of MWE candidates

| Frequency | # Candidates | # True MWEs | % True MWEs |
|---|---|---|---|
| >= 100 | 12 | 7–8 | 58.33%–66.67% |
| 10 → 99 | 152 | 92–109 | 60.53%–71.71% |
| 3 → 9 | 480 | 252–307 | 52.50%–63.96% |
| 2 | 612 | 268–339 | 43.79%–55.39% |
| Total | 1,256 | 619–763 | 49.28%–60.75% |

**Table 8** Evaluation of MWE candidates generated by the aligner or by the tagger

| ***MWE candidates generated by the aligner*** | | | |
|---|---|---|---|
| Frequency | # Candidates | # True MWEs | % True MWEs |
| >= 100 | 4 | 1–2 | 25.00%–50.00% |
| 10 → 99 | 118 | 66–78 | 55.93%–66.10% |
| 3 → 9 | 453 | 236–283 | 52.10%–62.47% |
| 2 | 595 | 257–325 | 43.19%–54.62% |
| Total | 1,170 | 560–688 | 47.86%–58.80% |
| ***MWE candidates generated by the tagger*** | | | |
| Frequency | # Candidates | # True MWEs | % True MWEs |
| >= 100 | 8 | 6–6 | 75.00%–75.00% |
| 10 → 99 | 34 | 26–31 | 76.47%–91.18% |
| 3 → 9 | 27 | 16–24 | 59.26%–88.89% |
| 2 | 17 | 11–14 | 64.71%–82.35% |
| Total | 86 | 59–75 | 68.60%–87.21% |

The next sections describe some experiments carried out to measure the precision of the proposed extraction method according to: (4.1) the output of the tagger or the aligner, (4.2) the types of MWE and (4.3) the possible translations of the true MWEs.

## 4.1 Tagger X Aligner

According to Table 8, the PoS tagger has the highest precision, outputting more true MWEs than the lexical aligner: 68–87% vs. 47–58%. This result was expected since the MWEs output by the PoS tagger were defined manually in the morphological dictionaries. However, the tagger has a much lower recall as the number of true MWEs it identified (59–75) is 9 times lower than the number of true MWEs extracted by the aligner (560–688).

Moreover, we have found that 25 out of the 86 MWE candidates output by the PoS tagger (29%) can also be generated by the aligner. 21 of these MWEs candidates were obtained from 1 : $n$ alignments (in which a single pt word is aligned

**Table 9** Percentage of true MWEs according to some PoS patterns

| ***Most accurate PoS patterns*** | | | |
|---|---|---|---|
| Pattern | # Candidates | # True MWEs | % True MWEs |
| A+N | 148 | 97–129 | 65.54%–87.16% |
| N+N | 165 | 113–133 | 68.48%–80.61% |
| V+P | 208 | 185–203 | 88.94%–97.60% |
| Total | 521 | 395–465 | 75.82%–89.25% |
| ***Less accurate PoS patterns*** | | | |
| P+D | 52 | 2–9 | 3.85%–17.31% |
| .+PN | 56 | 4–6 | 7.14%–10.71% |
| Total | 108 | 6–15 | 5.56%–13.89% |

**Table 10** Examples of true and false MWEs according to some PoS patterns

| **Pattern** | **True MWEs** | **False MWEs** |
|---|---|---|
| A+N | artesian_wells, black_hole, botanical_gardens, crude_oil, roman_empire | american_authorities, analogous_substances, actual_fact, good_measure |
| N+N | cotton_plant, data_bank, density_currents, doctorate_degree, end_users | magazine_science, members_staff, salt_solution, may_edition |
| V+P | clean_up, close_to, consisted_of, depend_on | brought_with, take_with, learning_over |
| P+D | in_that, at_that | behind_this, by_him, between_the, during_these |
| .+PN | made_it, makes_it | do_you, for_which, feed_themselves, in_it |

to an en MWE) such as the MWEs *according to*, *amino acid*, *away from* and *up to*. Other 4 candidates were derived from 2 : 2 alignments (after the union of the pt–en and en–pt alignments output by GIZA++): *european union*, *great britain*, *traffic accident* and *united states*.

## 4.2 Types of MWE

Following Piao et al (2006), we applied a post-PoS-filter to the set of MWE candidates to get the frequency distribution of some PoS patterns. Five PoS patterns were considered:

1. adjective + noun (A+N)
2. noun + noun (N+N)
3. verb + preposition/particle (V+P)
4. preposition + determiner (P+D)
5. some categories such as verb and preposition + pronoun (.+PN)

Table 9 shows that the first three patterns represent 41.48% (521) of the total amount of extracted candidates (1,256) and that they can be extracted with 75–89% of precision. On the other hand, the last two patterns (almost 9% of the total amount of extracted candidates) can be filtered during extraction since they are likely to be false MWEs.

Piao et al (2006) have obtained 93.64% and 91.46% precision, respectively, for the first two types of MWEs extracted for Chinese. The other patterns presented in Table 9 were not considered by those authors. The high precision values for the V+P class suggests that our method performs specially well in dealing with verb-particle constructions. This result reflects the nature of the patterns found in this particular language, English, in which V+P constructions are very frequent. As future work, we will also look at other languages (like Portuguese, for example) to investigate specific PoS patterns for them.

Table 10 presents examples of MWE candidates classified as true or false for each pattern of Table 9. In this table the examples of true MWEs for the less accurate patterns (P+D and .+PN) were considered as such by the resource-based evaluation, since these examples were found in at least one reference dictionary.

**Table 11** Evaluation of the translations for true MWEs: all of them or just the most frequent

| ***All possible translations for true MWEs*** | | | |
|---|---|---|---|
| Frequency | # Translations | # True translations | % True translations |
| >= 100 | 142–147 | 37–50 | 26.06%–34.01% |
| 10 → 99 | 582–693 | 229–319 | 39.35%–46.03% |
| 3 → 9 | 555–661 | 305–420 | 54.95%–63.54% |
| 2 | 390–491 | 198–304 | 50.77%–61.91% |
| Total | 1,669–1,992 | 769–1,093 | 46.08%–54.87% |
| ***The most frequent translations for true MWEs*** | | | |
| Frequency | # Translations | # True translations | % True translations |
| >= 100 | 7–8 | 6–7 | 85.71%–87.50% |
| 10 → 99 | 96–115 | 67–88 | 69.79%–76.52% |
| 3 → 9 | 355–426 | 226–297 | 63.66%–69.72% |
| 2 | 390–491 | 198–304 | 50.77%–61.91% |
| Total | 848–1,040 | 497–696 | 58.61%–66.92% |

## 4.3 Translations

The human judges also evaluated all the possible translations for the whole set of 1,256 candidates. Only the possible translations for the candidates classified as true MWEs were considered for this analysis. The evaluation was performed by (1) considering all the possible translations, and (2) considering only the most frequent translations. The results are shown in Table 11.

Since the number of possible translations changes when we only consider the true MWEs classified by both judges (the lower bound) or if we also include those by just one of them (the upper bound), in Table 11, all figures are presented in relation to these bounds. As expected, the approach of selecting only the most frequent translations produced better results (58–66% of true translations) than the approach of considering all possible translations (46–54% of true translations). This confirms the feasibility of the approach to automatically assign a good translation for each MWE candidate.

## 4.4 Comparison with baseline

As we described in section 2, current methods of MWE extraction usually try to rank a list of annotated candidates, so that genuine MWEs are ranked better than false candidates. Before going any further, we underline that, unlike the baseline method, our technique does not start from a pre-processed list, but tries to automatically identify true MWEs, extracting them directly from a corpus along with their translations. Therefore, a direct comparison with other measures that use standard metrics of precision and recall is not straightforward, since it would require the costly and time-consuming manual annotation of a potentially large corpus. Instead, the alternative that we adopt is to perform a dictionary-based evaluation. With respect to recall, we perform a subjective evaluation based on the judgments of the results obtained through both methods by a human annotator.

We use a set of standard statistical association measures as our baseline approach In order to obtain comparable results, we first extract all the n-grams from the English part of the corpus, where we limited the evaluation to $n = 2$ (24065 candidates). We then apply to these bigrams the same POS and threshold filters described in section 3.2.2. For each candidate bigram, the measures described in figure 2 are used to estimate the degree of association between its words [4]. These measures compare the co-occurence frequency of two words with their individual frequencies, since a genuine expression will present higher correlation between $w_1$ and $w_2$ than a random combination of words.

We compare the precision of the alignment-based extraction method with the precision of the association measures using the resources described in section 4, where only the candidate MWEs identified by the method that are listed in the resources are considered to be true positives, following Baldwin and Villavicencio (2002). This provides an automatic basis for comparison that does not require a human annotator. It looks only at precision using a very strict gold-standard,

[4] Evert and Krenn (2005) give a detailed description of standard measures and their application to MWE identification, and more material may also be found on www.collocations.de

$$
\begin{aligned}
PMI &= log_2 \frac{p(w_1 w_2)}{p_\emptyset(w_1 w_2)} \\
MI &= \sum_{u_i \in \{w_i, \neg w_i\}} p(u_1 u_2) log_2 \frac{p(u_1 u_2)}{p_\emptyset(u_1 u_2)} \\
\chi^2 &= \sum_{u_i \in \{w_i, \neg w_i\}} \frac{(p(u_1 u_2) - p_\emptyset(u_1 u_2))^2}{p_\emptyset(u_1 u_2)} \\
t &= \frac{p(w_1 w_2) - p_\emptyset(w_1 w_2)}{\sqrt{p(w_1 w_2)}} \\
Dice &= 2 * \frac{p(w_1 w_2)}{p(w_1) + p(w_2)}
\end{aligned}
$$

**Fig. 2** Independence or null hypothesis is $p_\emptyset(w_1 w_2) = p(w_1)p(w_2)$. We approximate $p(s)$ of a sequence of words $s$ by its relative frequency $c(s)/N$, where $c(s)$ is the function that counts how many times the words in $s$ occur contiguously in a corpus of size $N$.

**Table 12** Examples of extracted MWEs and rank according to statistical measures

| **MWE candidates** | | | | **Baseline rank** | | | | |
|---|---|---|---|---|---|---|---|---|
| | **D** | **J1** | **J2** | PMI | MI | $\chi^s$ | *t* | *Dice* |
| able_to | F | T | T | 9309 | 6832 | 7872 | 51 | 4966 |
| academic_world | F | T | T | 4480 | 20285 | 18668 | 1744 | 2041 |
| accompanied_by | F | F | T | 7742 | 9626 | 9255 | 1317 | 8812 |
| a_hundred | F | F | F | 6342 | 16105 | 18500 | 1591 | 3829 |
| hoped_for | T | – | – | 4235 | 14710 | 15072 | 3986 | 5725 |
| human_being | F | T | T | 3792 | 21210 | 18638 | 2687 | 1976 |

as dictionaries have a limited coverage for MWEs. As a consequence, the results reported are likely to be an under-estimate, with many true MWEs being potentially evaluated as false cases if not listed in the resources. A threshold of 1256 top candidates in the rank is defined to be equal to the number of MWEs extracted by the alignment-based method from the corpus (this measure is also known as *precision at N* or simply *P@N*). We remind that 317 of the MWEs extracted by the aligner were attested in a dictionary, leading to a value of P@1256=25.2%. The association measures achieved a value of P@1256 ranging from 0.2% for *MI* to 8.9% for *Dice* [5]. Since this values are under-estimated, they should not be interpreted as a performance measure (in which case our method would be three times more precise than the best association measure), but they help us to give an idea of the heterogeneity of the compared tasks: while related work shows that association measures perform well in filtering pre-processed MWE candidate lists, we propose a method that performs especially well in extracting MWEs directly from corpora.

Since we do not have manual annotation for the whole corpus, we cannot compare the recall of the baseline with the recall of our method. Instead, we manually compare a small sample of the output of both methods. Therefore, we ranked all the candidates according to each one of the measures and inspected (a) the rank of the example MWEs used in table 4 and (b) the characteristics of the top retrieved candidates. For the former, we can see in table 12 that none of the measures distinguishes the true and false instances in this example, since the pair *a hundred* has both asymmetric translation and high statistical correlation between the terms. If we inspect the top candidates for MI and $\chi^2$, we realize that they contain a function word (e.g. *harmed the*, *advocating the*, *handles the*). On the other hand, PMI and Dice seem to prefer very rare MWEs, like proper names (e.g. *eurico gaspar*, *érico vanucci*) or foreign names (e.g. *epinephelus niveatus*, *cryptomeria japonica*).

Currently, we are unable to evaluatemeasure the recall of our method, so we acknowledge the fact that it could be relatively low. However, the previous analysis shows that the type of information captured by frequency-based methods and by our alignment-based method are of a different nature, suggesting that they should be combined together in order to improve the coverage of the resources build upon the extracted MWEs. Additionally, this preliminar analysis tells us that the baseline approach, when used to identify generic MWEs in corpora of limited size, is very sensitive to low and high

[5] If we consider all the 24065 bigrams, only 936 (3.9%) were found in at least one dictionary.

frequencies and cannot correctly capture the MWEs in the text. Frequency-based methods are based on n-grams and are thus limited to little values of *n*, since for small corpora, higher values of *n* tend to introduce noise in statistical measures, besides being very time-consuming, as performance depends exponentially on *n*. The alignment-based extraction method proposed in this paper is able to identify and extract true MWEs and their translations without suffering from the problems of frequency-based methods and, given that enough parallel text is available, without constrains limiting the size of the extracted MWEs to a certain n-gram window.

## 5 Conclusions and Future Work

This paper presented a method for extracting multiword expressions and their translations as a by-product of automatic lexical alignment. A set of varied experiments obtained promising results. For example, if we limit our extraction method to only those candidates that occur at least 10 and at most 99 times, in our test corpus, we obtain 152 English multiword expressions with an expected precision of 60.53%–71.71%. Furthermore, these 152 MWEs are accompained by 96–115 translations with an expected precision of 69.79%–76.52% — when we consider only the most frequent possible translations.

Finally, if we are interested in only some types of MWEs, we can apply a post-PoS-filter to select those candidates that match some PoS patterns: adjective+noun (148 candidates with an expected precision of 65.54%–87.16%), noun+noun (165 candidates with an expected precision of 68.48%–80.61%) and verb+preposition/particle (208 candidates with an expected precision of 88.94%–97.60%). The application of the post-PoS-filter also revealed some patterns that contribute to depreciate the performance of our method. These patterns are candidates to be excluded in future experiments.

In terms of evaluation, as with other methods, a full analysis of recall would require that the MWEs to be detected from the corpus were known beforehand, through manual annotation of the corpus. However, depending on the size of the corpora this becomes impracticably costly both in terms of labour and time. The alternative explored in this paper is based on MWE dictionaries and, even if the evaluation of the results is limited by the coverage of the lexical resource, we showed that our method is better than standard association measures in extracting MWEs directly from corpora.

Although the proposed method depends on the availability of parallel corpora, it provides a straightforward way of identifying MWE candidates, that traditional statistical based methods may not detect, as discussed in the previous section. Therefore, if such corpora are available these approaches can be combined together, complementing each other for more comprehensive results. In addition, as parallel corpora are becoming increasingly available for a larger number of languages, this requirement becomes less restrictive, as the applicability of this method for other languages also increases.

Future works include the repetition of this experiment with the same 17,397 pairs of pt–en parallel sentences, but without PoS tagging them. By doing this we aim at excluding completely the influence of incorrect tags for the method. Another proposal for future work is to evaluate the pt MWE candidates not considered in this first experiment. In relation to the MWE extraction algorithm, we would like to experiment with possible ways of combining standard statistical methods with alignmen-based extraction, for instance using association measures to rank the MWEs candidates output by the lexical aligner.

**Acknowledgements** We thank the financial support of the Brazilian agencies FAPESP (02/13207-8) and CNPq (550388/2005-2). We also thank Mônica Saddy Martins for helping in the evaluation process.

## References

Armentano-Oller C, Carrasco RC, Corbí-Bellot AM, Forcada ML, Ginestí-Rosell M, Ortiz-Rojas S, Pérez-Ortiz JA, Ramírez-Sánchez G, Sánchez-Martínez F, Scalco MA (2006) Open-source Portuguese-Spanish machine translation. In: Proceedings of the VII Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR-2006), Itatiaia-RJ, Brazil, pp 50–59

Baldwin T, Villavicencio A (2002) Extracting the unextractable: A case study on verb-particles. In: Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002), Taipei, Taiwan, URL http://lingo.stanford.edu/pubs/tbaldwin/conll2002.pdf

Briscoe T, Carroll J (2002) Robust accurate statistical annotation of general text. In: Proceedings of LREC-2003

Brown P, Della-Pietra V, Della-Pietra S, Mercer R (1993) The mathematics of statistical machine translation: parameter estimation. Computational Linguistics 19(2):263–312

Burnard L (2000) User Reference Guide for the British National Corpus. Tech. rep., Oxford University Computing Services

Carletta J (1996) Assessing agreement on classification tasks: the kappa statistics. Computational Linguistics 22(2):249–254

Caseli HM, Silva AMP, Nunes MGV (2004) Evaluation of Methods for Sentence and Lexical Alignment of Brazilian Portuguese and English Parallel Texts. In: Proceedings of the SBIA 2004 (LNAI), Springer-Verlag, Berlin Heidelberg, 3171, pp 184–193

Caseli HM, Nunes MGV, Forcada ML (2006) Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. Machine Translation 20:227–245

Evert S, Krenn B (2005) Using small random samples for the manual evaluation of statistical association measures. Computer Speech and Language 19(4):450–466

Fazly A, Stevenson S (2007) Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Prague, pp 9–16

Hofland K (1996) A program for aligning English and Norwegian sentences. In: Hockey S, Ide N, Perissinotto G (eds) Research in Humanities Computing, Oxford University Press, Oxford, pp 165–178

Jackendoff R (1997) Twistin' the night away. Language 73:534–59

Melamed ID (1997) Automatic Discovery of Non-Compositional Compounds in Parallel Data. In: eprint arXiv:cmp-lg/9706027, pp 6027–+

Och FJ, Ney H (2000a) A comparison of alignment models for statistical machine translation. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING -2000), Saarbrücken, Germany, pp 1086–1090

Och FJ, Ney H (2000b) Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting of the ACL, Hong Kong, China, pp 440–447

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Computational Linguistics 29(1):19–51

Pearce D (2002) A comparative evaluation of collocation extraction techniques. In: Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain, pp 1–7

Piao SSL, Sun G, Rayson P, Yuan Q (2006) Automatic Extraction of Chinese Multiword Expressions with a Statistical Tool. In: Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006), Trento, Italy, pp 17–24

Procter P (1995) Cambridge International Dictionary of English. Cambridge University Press

Sag IA, Baldwin T, Bond F, Copestake A, Flickinger D (2002) Multiword expressions: A pain in the neck for nlp. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2002), Springer-Verlag, London, UK, (Lecture Notes in Computer Science), vol 2276, pp 1–15

Van de Cruys T, Villada Moirón B (2007) Semantics-based Multiword Expression Extraction. In: Proceedings of the Workshop on A Broader Prespective on Multiword Expressions, Prague, pp 25–32

Villada Moirón B, Tiedemann J (2006) Identifying idiomatic expressions using automatic word-alignment. In: Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006), Trento, Italy, pp 33–40

Villavicencio A (2005) The availability of verb-particle constructions in lexical resources: How much is enough? Journal of Computer Speech and Language Processing 19

Villavicencio A, Kordoni V, Zhang Y, Idiart M, Ramisch C (2007) Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, pp 1034–1043

Vogel S, Ney H, Tillmann C (1996) HMM-based word alignment in statistical translation. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING-1996), Copenhagen, pp 836–841

Zhang Y, Kordoni V, Villavicencio A, Idiart M (2006) Automated multiword expression prediction for grammar engineering. In: Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Association for Computational Linguistics, Sydney, Australia, pp 36–44, URL `http://www.aclweb.org/anthology/W/W06/W06-1206`