# Extraction of Multi-word Expressions from Small Parallel Corpora

Yulia Tsvetkov

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE MASTER DEGREE

University of Haifa
Faculty of Social Sciences
Department of Computer Science

August, 2010

# Extraction of Multi-word Expressions from Small Parallel Corpora

By: Yulia Tsvetkov

Supervised By: Dr. Shuly Wintner

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE MASTER DEGREE

University of Haifa

Faculty of Social Sciences

Department of Computer Science

August, 2010

Approved by: _____

Date:_____

(supervisor)

Approved by: _____

Date:_____

(Chairman of M.A. Committee)

# Contents

# Extraction of Multi-word Expressions from Small Parallel Corpora

Yulia Tsvetkov

## Abstract

Multi-word Expressions (MWEs) are lexical items that consist of multiple orthographic words (e.g., *ad hoc, by and large, New York, kick the bucket*). In this thesis we focus on MWEs with a non-compositional meaning, expressed by their non-literal translation to another language. We present a general methodology for extracting multi-word expressions (of various types), along with their translations, from small parallel corpora.

We first show a technique for fully automatic construction of constantly growing parallel corpora. We propose a simple and effective dictionary-based algorithm to extract parallel document pairs from a large collection of articles retrieved from the Internet, potentially containing manually translated texts. We implemented and tested this algorithm on Hebrew-English parallel texts, and collected a small parallel corpus.

We then automatically align the parallel corpus and focus on *mis*alignments; these typically indicate expressions in the source language that are translated to the target in a non-compositional way. We developed a simple algorithm that proposes MWE candidates (along with their translations) based on such misalignments. We use a large monolingual corpus to rank and filter these candidates. Evaluation of the quality of the extraction algorithm reveals significant

improvements over naïve alignment-based methods. External evaluation shows an improvement in the performance of a machine translation system that uses the extracted dictionary.

# 1 Introduction

Multi-word Expressions (MWEs) are lexical items that consist of multiple orthographic words (e.g., *ad hoc, by and large, New York, kick the bucket*). Sag et al. (2002) define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)", i.e., there is a mismatch between the interpretation of the expression as a whole and the standard meanings of the individual words that make it up.

MWEs are a heterogeneous class of constructions with diverse sets of characteristics, distinguished by their idiosyncratic behavior. Morphologically, some MWEs allow some of their constituents to freely inflect while restricting (or preventing) the inflection of other constituents. In some cases MWEs may allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation. Syntactically, some MWEs behave like words while other are phrases; some occur in one rigid pattern (and a fixed order), while others permit various syntactic transformations. Semantically, the compositionality of MWEs is gradual, ranging from fully compositional to fully idiomatic (Bannard et al., 2003).

Al-Haj (2010) presents a systematic linguistic characterization of MWEs in Hebrew, and provides in a full picture of the diverse properties that Hebrew MWEs exhibit. The substantial variability of MWEs over a wide range of parameters, is demonstrated by the following Hebrew[1] examples (Al-Haj, 2010):

- MWEs can appear as fixed or flexible lexical combinations. As an example of a fixed lexical combination consider (1): the constituents and the order in which they occur in a text are fixed and the expression is continuous. The expression (2), in contrast, contains an open slot that can be filled by a noun phrase, and the order of components can be changed. We therefore view this MWE as an unfixed lexical combination.

[1]To facilitate readability we use a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are *abgdhwzxTiklmns'pcqršt*.

(1) *ap* *‘l* *pi* *kn*
even on mouth thus
‘nevertheless’

(2) *akl* *at* — *bli* *mlx*
ate ACC — without salt
‘easily defeat’ (lit. ‘eat somebody without salt’)

- MWEs can have a variety of part-of-speech (POS) categories, including Noun-Noun compounds (3), Verb-Prepositions (4), Noun-Adjectives (5), (6), Adjective-Nouns (7), Participle-Nouns (8) and Conjunctions (9):

(3) *bit* *spr*
house book
‘school’ (lit. ‘a book house’)

(4) *‘bd* *‘l*
worked on
‘play a trick on’

(5) *‘in* *hr‘*
eye the evil
‘evil eye’ (lit. ‘the evil eye’)

(6) *hxlwnwt* *hgbwhim*
the windows the high
‘upper echelon’ (lit. ‘the high windows’)

(7) *kl* *d‘t*
light mind
‘frivolous’ (lit. ‘ light minded’)

(8) *iwšb* *raš*
sitting head
‘chairman’ (lit. ‘(person) sitting (at) head’)

(9) *ala* *am* *kn*
but if yes
‘unless’

- Semantically, MWEs cover a wide spectrum, from highly idiomatic (10), (11) to completely transparent (12):

(10) *kptwr* *wprx*
button and flower
'fantastic' (lit. 'a button and a flower')

(11) *kmTxwwi* *kšt*
? bow
'a stone's throw' (no literal meaning)

(12) *bdwar* *xwzr*
in mail returning
'by return mail' (lit. 'by returning mail')

They are also extremely prevalent: Jackendoff (1997, page 156) estimates that the number of MWEs in a speakers' lexicon is of the same order of magnitude as the number of single words. Sag et al. (2002) note that this is almost certainly an underestimate, observing that 41% of the entries in WordNet 1.7 (Fellbaum, 1998), for example, are multi-words. In an empirical study, Erman and Warren (2000) found that over 55% of the tokens in the texts they studied were instances of *prefabs* (defined informally as word sequences that are preferred by native speakers due to conventionalization.)

Because of their prevalence and irregularity, MWEs must be stored in lexicons of natural language processing applications. Handling MWEs correctly is beneficial for a variety of applications, including information retrieval (Doucet and Ahonen-Myka, 2004), building ontologies (Venkatsubramanyan and Perez-Carballo, 2004), text alignment (Venkatapathy and Joshi, 2006), and machine translation (MT) (Baldwin and Tanaka, 2004; Uchiyama et al., 2005).

Identifying MWEs and extracting them from corpora is therefore both important and difficult. In Hebrew (which is the subject of our research), this is even more challenging due to two reasons: the rich and complex morphology of the language; and the dearth of existing language resources, in particular parallel corpora, semantic dictionaries and syntactic parsers.

We propose a novel algorithm for identifying MWEs in bilingual corpora, using automatic word alignment as our main source of information. In contrast

to existing approaches, we do not limit the search to one-to-many alignments, and propose an error-mining strategy to detect misalignments in the parallel corpus. We also consult a large monolingual corpus to rank and filter out the expressions. The result is fully automatic extraction of MWEs of various types, lengths and syntactic patterns, along with their translations. We demonstrate the utility of the methodology on Hebrew-English MWEs by incorporating the extracted dictionary into an existing machine translation system.

The main contributions of this thesis are thus a novel algorithm for collecting parallel corpora, and a new alignment-based algorithm for MWE extraction that focuses on misalignments, augmented by validating statistics computed from a monolingual corpus. After discussing related work, we detail in Section 3 a technique for parallel corpus collection, and in Section 4 our methodology for MWE extraction. We provide a thorough evaluation of the obtained results in Section 5. We then extract translations of the identified MWEs and evaluate the contribution of the extracted dictionary. We conclude with suggestions for future research.

# 2 Related work

## 2.1 Collection of parallel corpora

Most of the existing tools that harvest a parallel corpus from a collection of texts that may contain translated documents are designed as the following pipeline:

1. Detection of Web sites that are likely to have translated materials

2. Extraction of parallel texts from these sites.

STRAND (Resnik, 1998, 1999) is an architecture for structural translation recognition. To detect bilingual Web sites, a search engine query is used to find "parents" and "siblings": Web sites containing links to translated versions of the same site. At the next stage poor candidates are filtered out by comparing the

structure (HTML tags) of two pages and the lengths of the translated texts. In a later version of STRAND (Resnik and Smith, 2003), content based matching of the texts is added. Text similarity is computed as

$$\frac{\#\textit{word-to-word translations}}{\#\textit{word-to-word translations} + \ \#\textit{untranslated words}}$$

To compute the number of translations, Resnik and Smith use a symmetric word-to-word translational model (Melamed, 2000), with additional complexity improvements. This technique was tested on English-French document pairs and reported as competitive to the structure-based approach of STRAND.

In BITS (Bilingual Internet Text Search) (Ma and Liberman, 1999), candidate Web sites are defined by their domain names, e.g., .de sites are considered as candidates in German. Ma and Liberman (1999) assume additionally that 10% of these sites include translations to English, and hence use the entire domain as a set of candidates. To detect parallel documents, the system defines the content similarity for every two texts as follows:

$$sim(A, B) = \frac{\#\textit{translation token pairs}}{\#\textit{tokens in text } A}$$

Translation token pairs within a fixed window in a parallel text are detected using a translation lexicon. Additional filters are applied for document length, similarity of anchors, etc. BITS was used to collect a 63MB corpus of English-German texts.

PTMINER (Chen and Nie, 2000) follows Resnik's technique to identify candidate sites by submitting particular requests to search engines. Then, parallel pairs are detected by filename and text length comparison, language identification and sentence alignment. English-French and English-Chinese corpora were produced with this technique.

To the best of our knowledge, none of the existing techniques was applied to Hebrew. All the architectures discussed above are designed to perform an

unsupervised retrieval of a static snapshot of parallel candidate sites. We believe that this method is likely to miss the most valuable translation sources. In the next section we explain this claim along with an alternative approach: to manually detect candidate sites, and then automatically monitor them over time. Moreover, we describe a novel content-based algorithm for parallel text matching and its application to the Hebrew-English language pair.

## 2.2 Automatic extraction of MWEs

Early approaches to identifying MWEs concentrated on their collocational behavior (Church and Hanks, 1989). Pecina (2008) compares 55 different association measures in ranking German Adj-N and PP-Verb collocation candidates. This work shows that combining different collocation measures using standard statistical classification methods improves over using a single collocation measure. Other results (Chang et al., 2002; Villavicencio et al., 2007) suggest that some collocation measures (especially PMI and Log-likelihood) are in fact superior to others for identifying MWEs. Soon, however, it became clear that mere co-occurrence measurements are not enough to identify MWEs, and their linguistic properties should be exploited as well (Piao et al., 2005). Hybrid methods that combine word statistics with linguistic information exploit morphological, syntactic and semantic idiosyncratic properties to extract idiomatic MWEs.

To enhance the quality of MWE processing, existing linguistico-statistical approaches make use of part-of-speech taggers for handling certain categories of words; lemmatizers are used for recognizing all the inflected forms of a lexical item. Cook et al. (2007), for example, use prior knowledge about the overall syntactic behavior of an idiomatic expression to determine whether an instance of the expression is used literally or idiomatically. They assume that in most cases, idiomatic usages of an expression tend to occur in a small number of canonical forms for that idiom; in contrast, the literal usages of an expression are

less syntactically restricted, and are expressed in a greater variety of patterns, involving inflected forms of the constituents.

Al-Haj and Wintner (2010) focus on morphological idiosyncrasies of Hebrew MWEs, and leverage such properties to automatically identify a specific construction, noun-noun compounds, in a given text. However, Al-Haj and Wintner (2010) do not account for the semantics of the MWEs, which is the focus of our current research.

Semantic properties of MWEs can be used to distinguish between compositional and non-compositional (idiomatic) expressions. Katz and Giesbrecht (2006) and Baldwin et al. (2003) use Latent Semantic Analysis for this purpose. They show that compositional MWEs appear in contexts more similar to their constituents than non-compositional MWEs. For example, the co-occurrence measured by LSA between the expression *'kick the bucket'* and the word *die* is much higher than co-occurrence of this expression and its component words. The disadvantage of this methodology is that to distinguish between idiomatic and non-idiomatic usage of the MWE it relies on the MWE's known idiomatic meaning, and this information is usually absent. In addition, this approach won't work when only idiomatic or only literal usage of the MWE is overwhelmingly frequent.

Van de Cruys and Villada Moirón (2007) use unsupervised learning methods to identify non-compositional MWEs by measuring to what extent their constituents can be substituted by semantically related terms. Such techniques typically require lexical semantic resources that are unavailable for Hebrew.

An alternative approach to using semantics capitalizes on the observation that an expression whose meaning is non-compositional tends to be translated into a foreign language in a way that does not result from a combination of the literal translations of its component words. Alignment-based techniques explore to what extent word alignment in parallel corpora can be used to distinguish between idiomatic expressions and more transparent ones. A significant added

value of such works is that MWEs can thus be both identified in the source language and associated with their translations in the target language. MWE candidates and their translations are extracted as a by-product of automatic word alignment of parallel texts (Och and Ney, 2003).

Villada Moirón and Tiedemann (2006) focus on Dutch expressions and their English, Spanish and German translations in the Europarl corpus (Koehn, 2005). MWE candidates are ranked by the variability of their constituents' translations. To extract the candidates, they use syntactic properties (based on full parsing of the Dutch text) and statistical association measures. Translational entropy (Melamed, 1997) is used as the main criterion for distinguishing between idiomatic expressions and non-idiomatic ones. This approach requires syntactic resources that are unavailable for Hebrew.

Unlike Villada Moirón and Tiedemann (2006), who use aligned parallel texts to *rank* MWE candidates, Caseli et al. (2009) actually use them to extract the candidates. After the texts are word-aligned, Caseli et al. (2009) extract sequences of length 2 or more in the source language that are aligned with sequences of length 1 or more in the target. Candidates are then filtered out of this set if they comply with pre-defined part-of-speech patterns, or if they are not sufficiently frequent in the parallel corpus. Even with the most aggressive filtering, precision is below 40% and recall is extremely low (F-score is below 10 for all experiments). Our setup is similar, but we extract MWE candidates from the aligned corpus in a very different way; and we use statistics collected from a *monolingual* corpus to filter and rank the results.

Zarrieß and Kuhn (2009) also use aligned parallel corpora but only focus on one-to-many word alignments. To restrict the set of candidates, they focus on specific syntactic patterns as determined by parsing both sides of the corpus (again, using resources unavailable to us). The results show high precision but very low recall.

Ren et al. (2009) extract MWEs from the source side of a parallel corpus,

ranking candidates on the basis of a collocation measure (log-likelihood). They then word-align the parallel corpus and naïvely extract the translations of candidate MWEs based on the results of the aligner. To filter out the list of translations, they use a classifier informed by "translation features" and "language features" (roughly corresponding to translation models and language models used in MT). The extracted translation pairs are fed into a baseline Chinese-English MT system and improve BLEU results by up to 0.61 points. While our MWE extraction algorithm is very different, and our translation extraction method is more naïve, we, too, use MT as an external evaluation method for the quality of the extracted translations.

## 3 Acquisition of Parallel Corpora

Parallel corpora are crucial resources for NLP applications that require some sort of semantic interpretation: machine translation, automatic lexical acquisition, word sense disambiguation, etc. Collecting corpora, representing and maintaining them are non-trivial tasks. But the main challenge is to find a good source of manually translated parallel texts. An example of such a source is translated literature, but in most cases it cannot be used due to copyright restrictions or fees. Religious texts are not a subject of intellectual property, but their language is often outdated and the domain is too specific. Other examples of possible sources of parallel corpora are translated texts produced by government agencies, software and military manuals, but the language of these documents tends to be technical and domain-specific, and the size of such corpora is limited. Parliamentary proceedings, such as Europarl (Koehn, 2005) or the Canadian Hansards, are large and valuable parallel corpora, although their content is limited to legislative discourse. Unfortunately, such corpora are unavailable for Hebrew and many others medium-density languages (Varga et al., 2005).

Therefore, there is a natural need to search for translated materials on the Web, "a huge fabric of linguistic data often interwoven with parallel threads" (Resnik and Smith, 2003). We describe a novel content-based algorithm to extract parallel articles from a large collection of documents retrieved from the Internet, which potentially contain manually translated texts. We compiled the first Hebrew-English parallel corpus, containing articles on news, politics, sports, economics, literature, etc. We perform a daily crawl of Web sites with dynamic contents (newspaper sites), extending our corpus constantly. The average number of parallel sentences added to our corpora every month is 3625. Evaluation results show that we obtain 100% precision and 86.5% recall (threshold values were chosen to favor precision over recall, since the quality of the corpus is crucial for us while its size is just a matter of time).

Although the experiments were held for Hebrew-English, the proposed method is independent of linguistic knowledge and can be generalized to any other language pair for which a bilingual dictionary is available.

### 3.1 Articles content and availability

In order to retrieve quality parallel corpora, texts should be searched on sites that are not biased to a specific subject and not edited by the same person. In addition, to guarantee the continuous growth of the corpus, sites with dynamic content should be used. Newspaper sites satisfy both conditions: they cover a wide variety of domains: politics, culture, science, sports, arts and leisure, etc.; and new articles are published frequently. Identification of such sites can be done manually, since there are few such sites and even one or two are sufficient to build a good resource. Due to the dynamic nature of these sites the size of the corpus is just a matter of time. Previously proposed techniques for automatic detection by querying search engines are unlikely to find such sites: articles usually do not contain links to their translated version, since these versions are targeted to a different readership. Translated articles can be located

on different domains and maintained by different teams, and their URL does not necessarily contain the title of the article or any other identification of its identity. Therefore, neither HTML structure nor filename are useful features for article comparison, and detection of document pairs can only be done by semantic analysis of the texts.

As a source for building our corpus we use a daily on-line newspaper in Hebrew and its version in English. Not all articles are translated, and some are only translated partially.

### 3.2 Parallel Corpora Builder

Our system, Parallel Corpora Builder (PCB), was developed to collect a parallel corpus from websites with dynamic content which potentially contain translated texts. The system architecture is illustrated in Figure 1. In the following subsections we describe the system in detail.

### 3.3 Web crawling

A Cron job is used to run a crawler several times a day and to harvest all fresh articles. Web crawling of the sites is a purely technical problem. We use a simple script to clean downloaded web pages from HTML tags and extract only text and metadata (date, domain, source URL, etc.)

The following features facilitate the task of collecting newspaper articles:

- To locate links to recently published articles, we use RSS feeds that are usually available on newswire sites.

- On-line newspaper articles commonly contain a link to the print version. We download these pages instead of the original articles, since they usually contain less user interface components such as Javascript, Flash, etc., and therefore require smaller effort to extract the raw text.

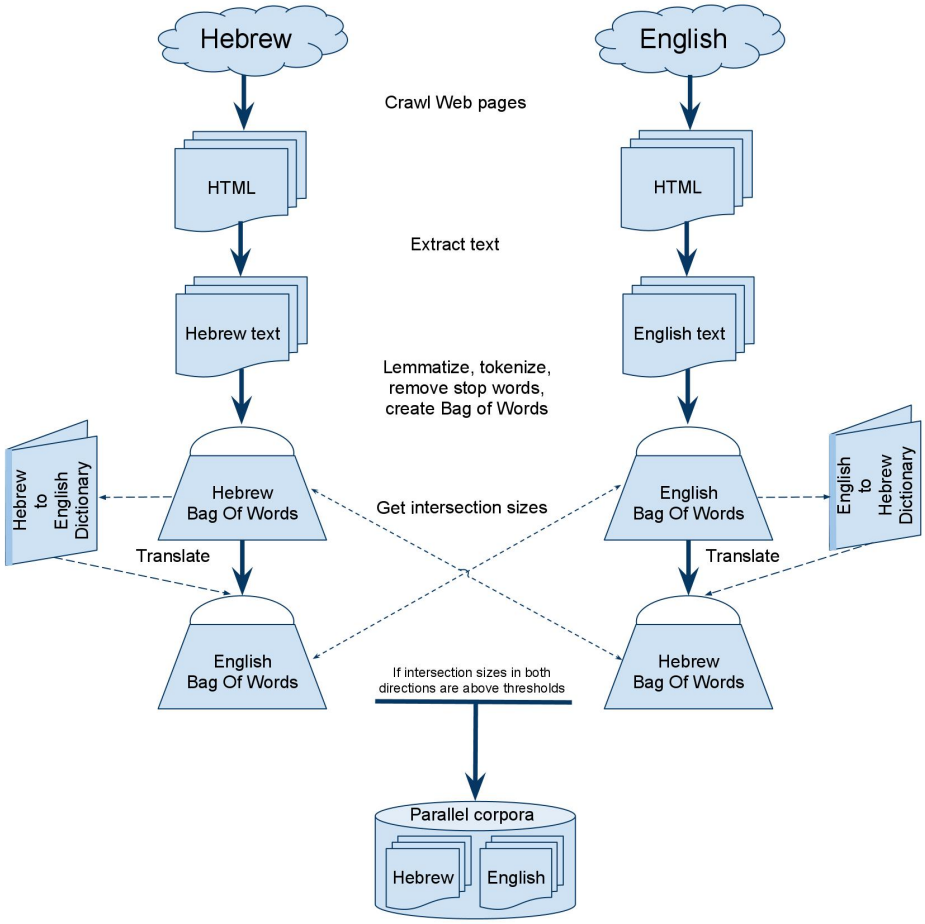


Figure 1: Parallel Corpora Builder (PCB) architecture.

### 3.4 Identification of parallel articles.

We run a content-based comparison of all Hebrew-English document pairs that were collected during the previous month to extract translated documents. Two documents $E, H$ are defined as mutual translations if $E$ contains enough translated terms from $H$ and vice versa. We now detail this process.

We use morphological analysis tools for Hebrew (Itai and Wintner, 2008) and for English (Minnen et al., 2001) to reduce inflected forms of words to a common base form. Then, after tokenization, lemmatization and stop word re-

moval, each article is represented by its bag of words (BOW). We then generate a BOW that represents the translation of this article to the parallel language by translating (using a dictionary) each word in the article. A translated BOW is usually much larger than the one in the original language, since all possible translations of each word are added. We use the same dictionary in both directions. Given a Hebrew-English pair of texts, we have

- $H$ - the BOW of the Hebrew text
- $H2E$ - the BOW of translations of $H$ to English
- $E$ - the BOW of the English text
- $E2H$ - the BOW of translations of $E$ to Hebrew

the two texts are identified as mutual translations and added to the parallel corpus if they satisfy the following formula:

$$(\frac{|H \cap E2H|}{|H|} > T_{Heb}) \ and \ (\frac{|E \cap H2E|}{|E|} > T_{Eng})$$

where $T_{Heb}$ and $T_{Eng}$ are threshold values for Hebrew and English documents, respectively, determined empirically based on data collected in the first month.

Our experiments show that if text similarity is computed only in one direction, many false positives are added, and tuning the threshold value does not resolve this problem: for tighter thresholds, translated texts are filtered out along with the false positives. Bidirectional similarity check shows a dramatic improvement in translation detection resulting in perfect precision. In addition, the bidirectional approach is useful to filter out partially translated texts.

Moreover, to achieve perfect precision, we also remove texts that have more than one parallel document (this is a very rare case). The only case of such a scenario is when these articles are very closely related in subject.

Since we compare all possible pairs of documents, complexity may become a serious obstacle for large amounts of data. To solve this problem we rely on

the fact that translated articles are published on the site in relatively close time intervals. We split the downloaded data to groups, stamped by the time they appeared on the Web site. Then, we run the pair detection algorithm monthly: every month we collect on average about 1500 articles in Hebrew and 600 in English, and comparison of all pairs is feasible.

### 3.5 Evaluation

The evaluation was performed on Hebrew and English articles collected during 3 months. As we mention above, we deliberately favor precision over recall, and our system was designed to filter out all suspicious documents. To compute the recall, we ran our system with lower thresholds and manually checked the results to identify undetected translations. Table 1 details the evaluation results.

| Month | English articles | Hebrew articles | Parallel articles | Detected parallel articles | Precision | Recall |
|---|---|---|---|---|---|---|
| 07 | 624 | 1530 | 168 | 145 | 100% | 86.3% |
| 08 | 548 | 1486 | 172 | 149 | 100% | 86.6% |
| 09 | 600 | 1341 | 165 | 143 | 100% | 86.7% |
| average | 573 | 1452 | 168 | 145 | 100% | 86.5% |

Table 1: PCB evaluation

The main advantage of our algorithm is its simplicity: without sophisticated heuristics or probabilistic models, we use the naive BOW comparison and achieve excellent results.

## 4 Extracting MWEs from parallel corpora

### 4.1 Methodology

We propose an alternative approach to existing alignment-based techniques for MWE extraction. Using a small bilingual corpus, we extract MWE candidates from noisy word alignments in a novel way. We then use statistics from a large monolingual corpus to rank and filter the list of candidates. Finally, we extract

the translation of candidate MWEs from the parallel corpus and use them in an MT system.

## 4.2 Motivation

Parallel texts are an obvious resource from which to extract MWEs. By definition, idiomatic expressions have a non-compositional meaning, and hence may be translated to a single word (or to an expression with a different meaning) in a foreign language. The underlying assumption of alignment-based approaches to MWE extraction is that MWEs are aligned across languages in a way that differs from compositional expressions; we share this assumption. However, existing approaches focus on the results of word alignment in their quest for MWEs, and in particular consider 1:$n$ and $n$:$m$ alignments as potential areas in which to look for them. This is problematic for two reasons: first, word alignment algorithms have difficulties aligning MWEs, and hence 1:$n$ and $n$:$m$ alignments are often noisy; while these environments provide cues for identifying MWEs, they also include much noise. Second, our experimental scenario is such that our parallel corpus is particularly small, and we cannot fully rely on the quality of word alignments, but we have a bilingual dictionary that compensates for this limitation. In contrast to existing approaches, then, we focus on *misalignments*: we trust the quality of 1:1 alignments, which we verify with the dictionary; and we search for MWEs exactly in the areas that word alignment *failed* to properly align, not relying on the alignment in these cases.

Moreover, in contrast to existing alignment-based approaches, we also make use of a large monolingual corpus from which statistics on the distribution of word sequences in Hebrew are drawn. This has several benefits: of course, monolingual corpora are easier to obtain than parallel ones, and hence tend to be larger and provide more accurate statistics. Furthermore, this provides validation of the MWE candidates that are extracted from the parallel corpus: rare expressions that are erroneously produced by the alignment-based technique can

thus be eliminated on account of their low frequency in the monolingual corpus.

Specifically, we use a variant of pointwise mutual information (PMI) as our association measure. While PMI has been proposed as a good measure for identifying MWEs, it is also known not to discriminate accurately between MWEs and other frequent collocations. This is because it promotes collocations whose constituents rarely occur in isolation (e.g., typos and grammar errors), and expressions consisting of some word that is very frequently followed by another (e.g., *say that*). However, such cases do not have idiomatic meanings, and hence at least one of their constituents is likely to have a 1:1 alignment in the parallel corpus; we only use PMI *after* such alignments have been removed.

An added value of our methodology is the automatic production of an MWE translation dictionary. Since we start with a parallel corpus, we can go back to that corpus after MWEs have been identified, and extract their translations from the parallel sentences in which they occur.

Finally, alignment-based approaches can be symmetric, and ours indeed is. While our main motivation is to extract MWEs in Hebrew, a by-product of our system is the extraction of *English* MWEs, along with their translations to Hebrew. This, again, contributes to the task of enriching our existing bilingual dictionary.

### 4.3 Resources

Our methodology is in principle language-independent and appropriate for medium-density languages (Varga et al., 2005). We assume the following resources: a small bilingual, sentence-aligned parallel corpus; large monolingual corpora in both languages; morphological processors (analyzers and disambiguation modules) for the two languages; and a bilingual dictionary. Our experimental setup is Hebrew-English. We use the small parallel corpus described in Section 3 (Tsvetkov and Wintner, 2010) which consists of 19,626 sentences, mostly from newspapers. Some data on the parallel corpus are listed

in Table 2 (the size of our corpus is very similar to that of Caseli et al. (2009)).

| | English | Hebrew |
|---|---|---|
| Number of tokens | 271,787 | 280,508 |
| Number of types | 14,142 | 12,555 |
| Number of unique bi-grams | 132,458 | 149,668 |

Table 2: Statistics of the parallel corpus

We also use data extracted from two monolingual corpora. For Hebrew, we use the morphologically-analyzed MILA corpus (Itai and Wintner, 2008) with part-of-speech tags produced by Bar-Haim et al. (2005). For English we use Google's Web 1T corpus (Brants and Franz, 2006). Data on the Hebrew corpus are provided in Table 3.

| | |
|---|---|
| Number of tokens | 46,239,285 |
| Number of types | 188,572 |
| Number of unique bi-grams | 5,698,581 |

Table 3: Statistics of the Hebrew corpus

Finally, we use a bilingual dictionary consisting of 78,313 translation pairs. Some of the entries were collected manually, while others are produced automatically (Itai and Wintner, 2008; Kirschenbaum and Wintner, 2010).

### 4.4 Preprocessing the corpora

Automatic word alignment algorithms are noisy, and given a small parallel corpus such as ours, data sparsity is a serious problem. To minimize the parameter space for the alignment algorithm, we attempt to reduce language specific differences by pre-processing the parallel corpus. The importance of this phase should not be underestimated, especially for alignment of two radically different languages such as English and Hebrew (Dejean et al., 2003).

Hebrew, like other Semitic languages, has a rich, complex and highly productive morphology. Information pertaining to gender, number, definiteness,

person, and tense is reflected morphologically on base forms of words. In addition, prepositions, conjunctions, articles, possessives, etc., may be concatenated to word forms as prefixes or suffixes. This results in a very large number of possible forms per lexeme. Consequently, a single English word (e.g., the noun *advice*) can be aligned to hundreds or even thousands of Hebrew forms (e.g., *lycth* "to-her-advice"). As *advice* occurs only 8 times in our small parallel corpus, it would be almost impossible to collect statistics even on simple 1:1 alignments without appropriate tokenization and lemmatization.

We therefore tokenize the parallel corpus and then remove punctuation. We analyze the Hebrew corpus morphologically and select the most appropriate analysis in context. Adopting this selection, the surface form of each word is reduced to its base form, and bound morphemes (prefixes and suffixes) are split to generate stand-alone "words". We also tokenize and lemmatize the English side of the corpus, using the Natural Language Toolkit package (Bird et al., 2009).

Then, we try to remove some language-specific differences automatically. We remove frequent function words: in English, the articles *a*, *an* and *the*, the infinitival *to* and the copulas *am*, *is* and *are*; in Hebrew, the accusative marker *at*. These forms do not have direct counterparts in the other language.

For consistency, we pre-process the monolingual corpora in the same way. We then compute the frequencies of all word bi-grams occurring in each of the monolingual corpora.

### 4.5 Identifying MWE candidates

The motivation for our MWE identification algorithm is the assumption that there may be three sources to misalignments (anything that is not a 1:1 word alignment) in parallel texts: either MWEs (which trigger 1:$n$ or $n$:$m$ alignments); or language-specific differences (e.g., one language lexically realizes notions that are realized morphologically, syntactically or in some other way

in the other language); or noise (e.g., poor translations, low-quality sentence alignment, and inherent limitations of word alignment algorithms).

This motivation induces the following algorithm. Given a parallel, sentence-aligned corpus, it is first pre-processed as described above, to reduce the effect of language-specific differences. We then use Giza++ (Och and Ney, 2003) to word-align the text, employing *union* to merge the alignments in both directions. We look up all 1:1 alignments in the dictionary. If the pair exists in our bilingual dictionary, we remove it from the sentence and replace it with a special symbol, '*'. Such word pairs are not parts of MWEs. If the pair is not in the dictionary, but its alignment score as produced by Giza++ is very high (above 0.5) and it is sufficiently frequent (more than 5 occurrences), we add the pair to the dictionary but also retain it in the sentence. Such pairs are still candidates for being (parts of) MWEs.

Figure 2-a depicts a Hebrew sentence with its word-by-word gloss, and its English translation in the parallel corpus. Here, *bn adm* (*son-of man*) "person" is a MWE that cannot be translated literally. After pre-processing (Section 4.4), the English is represented as "and i tell her keep away from person" (note that *to* and *the* are deleted). The Hebrew, which is aggressively segmented, is represented as in Figure 2-b. Note how this reduces the level of (morphological and orthographic) difference between the two languages. Consequently, Giza++ finds the alignment depicted in Figure 2-c. Once 1:1 alignments are replaced by '*', the alignment of Figure 2-d is obtained.

If our resources were perfect, i.e., if word alignment made no errors, the dictionary had perfect coverage and our corpora induced perfect statistics, then all remaining text (other than the special symbol) in the parallel text would be part of MWEs. In other words, all sequences of remaining source-language words, separated by '*', are MWE candidates. As our resources are far from perfect, further processing is required in order to prune these candidates. For this, we use association measures computed from the monolingual corpus.

a. *wamrti* *lh* *lhzhr* *mbn* *adm* *kzh*
and-I-told to-her to-be-careful from-child man like-this
"and I told her to keep away from the person"

b. *w* *ani* *amr* *lh* *lhzhr* *m* *bn* *adm* *k* *zh*
and I tell to-her to-be-careful from child man like this

c. *w* *ani* *amr* *lh* *lhzhr* *m* *bn adm* *k* *zh*
and I told her keep away from person {} {}

d. * * * * *lhzhr* * *bn adm* *k* *zh*
* * * * keep away * person

Figure 2: Example sentence pair (a); after pre-processing (b); after word alignment (c); and after 1:1 alignments are replaced by '*' (d)

### 4.6 Ranking and filtering MWE candidates

The algorithm described above produces sequences of Hebrew word forms (free and bound morphemes produced by the pre-processing stage) that are not 1:1-aligned, separated by '*'s. Each such sequence is a MWE candidate. In order to rank the candidates we use statistics from a large *monolingual* corpus. We do *not* rely on the alignments produced by Giza++ in this stage.

We extract all word bi-grams from the remaining candidates. Each bi-gram is associated with its PMI-based score, computed from the monolingual corpus. We use $\text{PMI}^k$, a heuristic variant of the PMI measure, proposed and studied by Daille (1994). $k$, the exponent, is a frequency-related factor, used to demote collocations with low-frequency constituents. The value of the parameter $k$ can be chosen freely ($k > 0$) in order to tune the properties of the PMI to the needs of specific applications. We conducted experiments for k = 0.1, 0.2, ... , 3 and found k = 2.7 to give the best results for our application. Interestingly, about 20,000 of the candidate MWEs are removed in this stage because they do not occur at all in the monolingual corpus.

We then experimentally determine a threshold (see Section 5). A word sequence of any length is considered MWE if all the adjacent bi-grams it contains

score above the threshold. Finally, we restore the original forms of the Hebrew words in the candidates, combining together bound morphemes that were split during pre-processing; and we restore the function words. Many of the candidate MWEs produced in the previous stage are eliminated now, since they are not genuinely multi-word in the original form (i.e., they were single words split by tokenization).

Refer back to Figure 2-d. The sequence *bn adm k zh* is a MWE candidate. Two bi-grams in this sequence score above the threshold: *bn adm*, which is indeed a MWE, and *k zh*, which is converted to the original form *kzh* and is hence not considered a candidate. We also consider *adm k*, whose score is low. Note that the same aligned sentence can be used to induce the *English* MWE *keep away*, which is aligned to a single Hebrew word.

### 4.7 Results

As an example of the results obtained with this setup, we list in Table 4 the 15 top-ranking extracted MWEs. For each instance we list an indication of the type of MWE: person name (PN), geographical term (GT), noun-noun compound (NNC) or noun-adjective combination (N-ADJ). Of the top 100 candidates, 99 are clearly MWEs,[2] including *mzg awir* (*temper-of air*) "weather", *kmw kn* (*like thus*) "furthermore", *bit spr* (*house-of book*) "school", *šdh t'wph* (*field-of flying*) "airport", *tšwmt lb* (*input-of heart*) "attention", *ai apšr* (*not possible*) "impossible" and *b'l ph* (*in-on mouth*) "orally". Longer MWEs include *ba lidi biTwi* (*came to-the-hands-of expression*) "was expressed"; *xzr 'l 'cmw* (*returned on itself*) "recurred"; *ixd 'm zat* (*together with it*) "in addition"; and *h'crt hkllit šl haw"m* (*the general assembly of the UN*) "the UN general assembly".

[2]This was determined by two annotators.

| Hebrew | Gloss | Type |
|---|---|---|
| *xbr hknst* | MP | NNC |
| *tl abib* | Tel Aviv | GT |
| *gwš qTip* | Gush Katif | NNC-GT |
| *awpir pins* | Ophir Pines | PN |
| *hc‘t xwq* | Legislation | NNC |
| *axmd Tibi* | Ahmad Tibi | PN |
| *zhwh glawn* | Zehava Galon | PN |
| *raš hmmšlh* | Prime Minister | NNC |
| *abšlwm wiln* | Avshalom Vilan | PN |
| *br awn* | Bar On | PN |
| *mair šTrit* | Meir Shitrit | PN |
| *limwr libnt* | Limor Livnat | PN |
| *hiw‘c hmšpTi* | Attorney General | N-ADJ |
| *twdh rbh* | thanks a lot | N-ADJ |
| *rcw‘t ‘zh* | Gaza Strip | NNC-GT |

Table 4: Results: extracted MWEs

# 5 Evaluation

MWEs are notoriously hard to define, and no clear-cut criteria exist to distinguish between MWEs and other frequent collocations. In order to evaluate the utility of our methodology, we conducted three different types of evaluations (two types of internal evaluation, and an external evaluation) that we detail in this section.

## 5.1 Internal evaluation

First, we use a small annotated corpus of Hebrew noun-noun constructions (Al-Haj and Wintner, 2010). The corpus consists of 463 high-frequency bigrams of the same syntactic construction; of those, 202 are tagged as MWEs (in this case, noun compounds) and 258 as non-MWEs. This corpus consolidates the annotation of three annotators: only instances on which all three agreed were included. Since it includes both positive and negative instances, this corpus facilitates a robust evaluation of precision and recall. Of the 202 positive

examples, only 121 occur in our parallel corpus; of the 258 negative examples, 91 occur in our corpus. We therefore limit the discussion to those 212 examples whose MWE status we can determine, and ignore other results produced by the algorithm we evaluate.

On this corpus, we compare the performance of our algorithm to four baselines: using only $\text{PMI}^{2.7}$ to rank the bi-grams in the parallel corpus; using $\text{PMI}^{2.7}$ computed from the monolingual corpus to rank the bi-grams in the parallel corpus; and using Giza++ 1:$n$ alignments, ranked by their $\text{PMI}^{2.7}$ (with bi-gram statistics computed once from parallel and once from monolingual corpora). 'MWE' refers to our algorithm. For each of the above methods, we set the threshold at various points, and count the number of true MWEs above the threshold (true positives) and the number of non-MWEs above the threshold (false positives), as well as the number of MWEs and non-MWEs below the threshold (false positives and true negatives, respectively). From these four figures we compute precision, recall and their harmonic mean, $f$-score, which we plot against (the number of results above) the threshold in Figure 3. Clearly, the performance of our algorithm is consistently above the baselines.

Second, we evaluate the algorithm on more datasets. We compiled three small corpora of Hebrew two-word MWEs. The first corpus, **PN**, contains 785 person names (names of Knesset members and journalists), of which 157 occur in the parallel corpus. The second, **Phrases**, consists of 571 entries, beginning with the letter $x$ in the Hebrew Phrase Dictionary of Rosenthal (2009), and a set of 331 idioms we collected from internet resources. Of those, 154 occur in the corpus. The third set, **NN**, consists of the positive examples in the annotated corpus of noun-noun constructions described above.

Since we do not have negative examples for two of these sets, we only evaluate recall, using a threshold reflecting 2750 results. For each of these datasets, we report the number of MWEs in the dataset (which also occur in the parallel corpus, of course) our algorithm detected. We compare in Table 5 the recall of

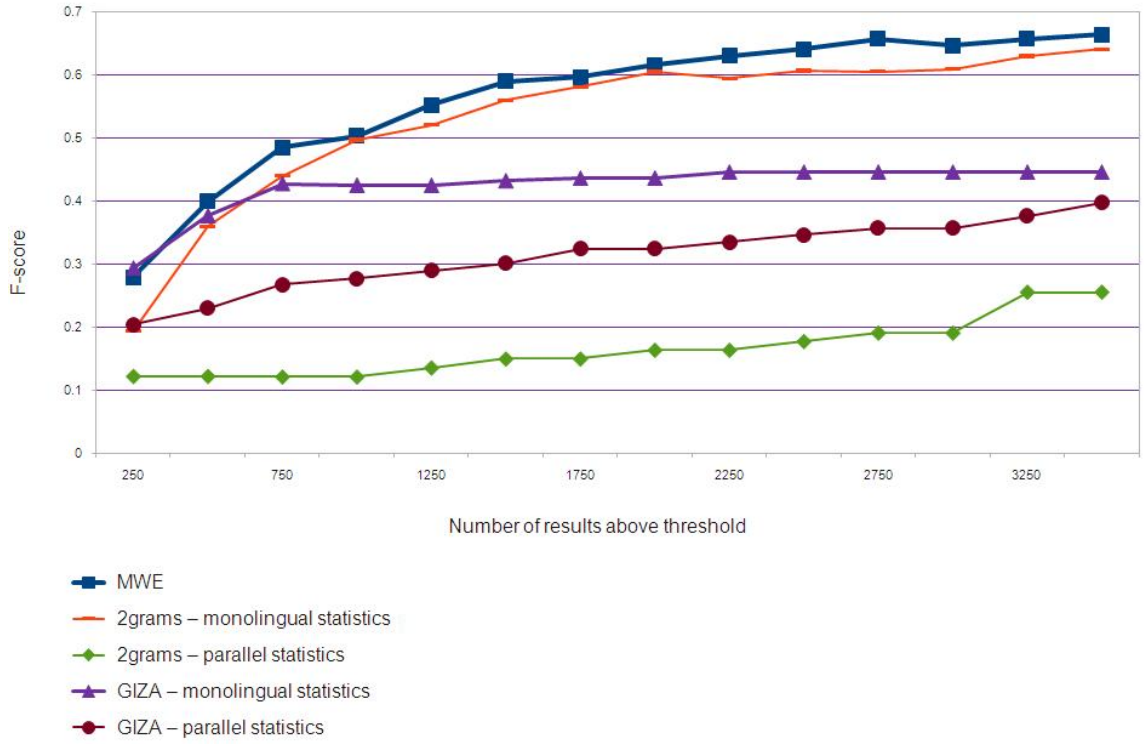


Figure 3: Evaluation results compared with baselines: noun-noun compounds

our method (MWE) to Giza++ alignments, as above, and list also the upper bound (UB), obtained by taking all above-threshold bi-grams in the corpus.

| Method | **PN** | | **Phrases** | | **NN** | |
|---|---|---|---|---|---|---|
| | # | % | # | % | # | % |
| UB | 74 | 100 | 40 | 100 | 89 | 100 |
| MWE | 66 | 89.2 | 35 | 87.5 | 67 | 75.3 |
| Giza | 7 | 9.5 | 33 | 82.5 | 37 | 41.6 |

Table 5: Recall evaluation

## 5.2 External evaluation

An obvious benefit of using parallel corpora for MWE extraction is that the translations of extracted MWEs are available in the corpus. We use a naïve approach to identify these translations. For each MWE in the source-language

sentence, we consider as translation all the words in the target-language sentence (in their original order) that are aligned to the word constituents of the MWE, as long as they form a contiguous string. Since the quality of word alignment, especially in the case of MWEs, is rather low, we remove "translations" that are longer than four words (these are frequently wrong). We then associate each extracted MWE in Hebrew with all its possible English translations.

The result is a bilingual dictionary containing 2,955 MWE translation pairs, and also 355 translation pairs produced by taking high-quality 1:1 word alignments (Section 4.5). We used the extracted MWE bilingual dictionary to augment the existing (78,313-entry) dictionary of a transfer-based Hebrew-to-English statistical machine translation system (Lavie et al., 2004b). We report in Table 6 the results of evaluating the performance of the MT system with its original dictionary and with the augmented dictionary. The results show a statistically-significant ($p < 0.1$) improvement in terms of both BLUE (Papineni et al., 2002) and Meteor (Lavie et al., 2004a) scores.

| Dictionary | BLEU | Meteor |
|---|---|---|
| Original | 13.69 | 33.38 |
| Augmented | 13.79 | 33.99 |

Table 6: External evaluation

As examples of improved translations, a sentence that was originally translated as "His teachers also hate to the Zionism and besmirch his HRCL and Gurion" (fully capitalized words indicate lexical omissions that are transliterated by the MT system) is translated with the new dictionary as "His teachers also hate to the Zionism and besmirch his Herzl and David Ben-Gurion"; a phrase originally translated as "when so" is now properly translated as "likewise"; and several occurrences of "down spring" and "height of spring" are corrected to "Tel Aviv".

### 5.3 Error analysis

Our MWE extraction algorithm works as follows: translated texts are first sentence aligned. Then, Giza++ is used to extract 1-to-1 word alignments, that are then verified by the dictionary and replaced by '*', if the correct word translation is available. This process filters out candidates that have compositional meaning and, therefore, are not considered MWEs (in our algorithm, a non-compositional meaning of a bi-gram is expressed by its non-literal translation to the parallel language). Sequences of words separated by '*'s are considered MWE candidates. At each step of the application errors may occur that lead to false identification of non-MWEs. We manually annotated the top 1000 bi-gram MWEs extracted by the algorithm and identified 121 false positives. Analysis of these false positives reveals the error sources detailed below. In Table 7 we summarize the statistics of the error sources.

| Error source | **False positives** | |
|---|---|---|
| | # | % |
| Translation quality of the parallel corpus | 46 | 38.02 |
| Sentence alignment errors | 19 | 15.7 |
| Word alignment errors | 21 | 17.36 |
| Noise introduced by preprocessing | 29 | 23.97 |
| Incomplete dictionary | 4 | 3.31 |
| Parameters of the algorithm | 2 | 1.65 |

Table 7: Error sources statistics

**Translation quality of the parallel corpus**

Whereas the sentences are indeed translations, the translations are, to a large extent, non-lexical, in the sense that context is used in order to extract the meaning and deliver it in different wording. As the result, it is sometimes hard or even impossible to align words based on the sentence alone.

**Sentence alignment errors**

1. We use a purely statistical sentence aligner to align sentences based on their length and token co-occurrence information. As a result, some sentences of similar length may incorrectly be marked as mutual translations. Of course, most of the word sequences in such sentences cannot be aligned and hence become MWE candidates.

2. The output of the sentence aligner contains only 1-to-1 sentence translations. As our parallel corpora include non-lexical translations, that sometimes can only be expressed in terms of 1-to-2, or 2-to-1 translated sentences, the sentence aligner may output 1-to-1 alignment, where one of the sentences is only a partial translation of another. The non-translated part of the sentence may contain false MWE candidates.

**Word alignment errors**

Sometimes a word sequence has a translation, but it is not aligned properly. Possible reasons for such errors are:

1. Insufficient statistics of word co-occurrence due to the small size of the parallel corpus

2. Errors caused by bidirectional translation merge (we employ union to merge the translations in both directions (Och and Ney, 2003)). Often the alignment is correct only in one direction, but we lose this information after merging the alignment; this often happens in very long sentences. Another example of the problematic alignment caused by bi-directional merge is cases in which the word aligner proposes N:1 alignment; usually these N words contain the correct sequence or a part of the sequence and the correct analysis of the bi-directional alignments may help filter out the incorrect parts (i.e., the analysis of the intersection of N and M sequences, where M:1 is Hebrew-to-English and N:1 is English-to-Hebrew alignments de-

tected by the word alignment tool).

**Noise introduced by preprocessing**

1. Errors caused by morphological analysis and disambiguation tools may lead to wrong tokenization, or to the extraction of an incorrect base form from the surface form of the word. As the result, the extracted citation form cannot be aligned to its translation, and correctly aligned word-pairs cannot be found in the dictionary. For example, the bi-gram *bniit gdr* is translated as *building fence*. Stemming on the English side produces the erroneous base form *build* for the word *building*. Word alignment correctly aligns the words *bniih* (a noun) and *build* (a verb), but such a pair does not exist in the dictionary, which contains the following pairs: *bnh-build* (verb), and *bniih-building* (noun).

2. An additional source of errors stems from language specific differences in word order between the languages: e.g., *txnt rkbt* is consistently translated as *railway station*; the correct alignment would be *txnh—station*, *rkbt—railway* but due to the different word order in the two languages, and to the fact that both phrases are frequent collocations, Giza++ proposes the alignment *txnh—railway*, *rkbt—station* (these pairs are not in the dictionary and, therefore, the bigram *txnt rkbt* is falsely identified as an MWE). Such problems can be handled with more sophisticated preprocessing that eliminates language specific differences, where not only morphology and function words are taken into account, but also language-specific word order.

**Incomplete dictionary**

If sentence and word alignment results are correct, and the correct word-to-word translation exists, but the translated pair is not in the dictionary,

the word sequence may erroneously be considered an MWE candidate.

**Parameters of the algorithm**

1. Setting the threshold too high causes bi-grams that are subsequences of the longer MWEs to be false positives. For example, the non-MWE, compositional bi-gram *lšlm ms*, which is a subsequence of the MWE *lšlm ms šptiim* (pay lip service), was mistakenly extracted as MWE, since the score of the bi-gram *ms šptiim* is lower than the threshold.

2. During error analysis we revealed the following algorithm drawback: false MWE candidates that occur several times in the parallel corpus are selected to be MWE candidates only in a minority of these occurences. For example, there are twelve occurrences of the bi-gram *nšia hmdinh* (president of the state) in the parallel corpus, but only twice does it appear as a candidate bi-gram, due to two sentences in which the translation of this bi-gram is missing (due to the non-literal or incorrect sentence translation). From this we conclude that the algorithm can also be improved, if the candidates would be selected from bi-grams that have no translation in the parallel language in a majority of their occurrences. We leave this improvement for future work.

## 6 Conclusions and Future Work

We described a methodology for extracting multi-word expressions from parallel corpora. The algorithm we propose capitalizes on semantic cues provided by ignoring 1:1 word alignments, and viewing all other material in the parallel sentence as potential MWE. It also emphasizes the importance of properly handling the morphology and orthography of the languages involved, reducing wherever possible the differences between them in order to improve the quality

of the alignment. We use statistics computed from a large monolingual corpus to rank and filter the results. We use the algorithm to extract MWEs from a small Hebrew-English corpus, demonstrating the ability of the methodology to accurately extract MWEs of various lengths and syntactic patterns. We also demonstrate that the extracted MWE bilingual dictionary can improve the quality of machine translation.

This work can be extended in various ways. While several works address the choice of association measure for MWE identification and for distinguishing between MWEs and other frequent collocations, it is not clear which measure would perform best in our unique scenario, where candidates are produced by word (mis)alignment. We intend to explore some of the measures discussed by Pecina (2008) in this context. The algorithm used for extracting the translations of candidate MWEs is obviously naïve, and we intend to explore more sophisticated algorithms for improved performance. Also, as our methodology is completely language-symmetric, it can be used to produce MWE candidates in English. In fact, we already have such a list of candidates, whose quality we will evaluate in the future. Finally, as our main motivation is high-precision, high-recall extraction of Hebrew MWEs, we would like to explore the utility of combining different approaches to the same task (Al-Haj and Wintner, 2010) under a unified framework.

## References


Hassan Al-Haj. Hebrew multiword expressions: Linguistic properties, lexical representation, morphological processing, and automatic acquisition. Master's thesis, University of Haifa, February 2010.

Hassan Al-Haj and Shuly Wintner. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, August 2010.

Timothy Baldwin and Takaaki Tanaka. Translation by machine of complex nominals: Getting it right. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 89–96. Association for Computational Linguistics, 2003.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. A statistical approach to the semantics of verb-particles. In Diana McCarthy Francis Bond, Anna Korhonen and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, 2003. URL `http://www.aclweb.org/anthology/W03-1809.pdf`.

Roy Bar-Haim, Khalil Sima'an, and Yoad Winter. Choosing an optimal architecture for segmentation and POS-tagging of Modern Hebrew. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 39–46, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W05/W05-0706`.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA, 2009.

Thorsten Brants and Alex Franz. Web 1T 5-gram version 1.1. LDC Catalog No. LDC2006T13, 2006. URL `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13`.

Helena Caseli, Aline Villavicencio, André Machado, and Maria José Finatto. Statistically-driven alignment-based multiword expression identification for technical domains. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 1–8, Singapore, August 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W09/W09-2901`.

Baobao Chang, Pernilla Danielsson, and Wolfgang Teubert. Extraction of translation unit from Chinese-English parallel corpora. In *Proceedings of the first SIGHAN workshop on Chinese language processing*, pages 1–5, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1118824.1118825.

Jiang Chen and Jian-Yun Nie. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, pages 21–28, Morristown, NJ, USA, 2000. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/974147.974151.

Kenneth. W. Church and Patrick Hanks. Word association norms, mutual information and lexicography (rev). *Computational Linguistics*, 19(1):22–29, 1989.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions (MWE 2007)*, pages 41–48, Prague, Czech Republic, June 2007.

Béatrice Daille. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7, 1994.

Herve Dejean, Eric Gaussier, Cyril Goutte, and Kenji Yamada. Reducing parameter space for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts*, pages 23–26, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1118905.1118910.

Antoine Doucet and Helana Ahonen-Myka. Non-contiguous word sequences for information retrieval. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 88–95, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Britt Erman and Beatrice Warren. The idiom principle and the open choice principle. *Text*, 20(1):29–62, 2000.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press, 1998.

Alon Itai and Shuly Wintner. Language resources for Hebrew. *Language Resources and Evaluation*, 42:75–98, March 2008.

Ray Jackendoff. *The Architecture of the Language Faculty*. MIT Press, Cambridge, USA, 1997.

Graham Katz and Eugenie Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, July 2006. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W06/W06-1203`.

Amit Kirschenbaum and Shuly Wintner. A general method for creating a bilingual transliteration dictionary. In *Proceedings of The seventh international conference on Language Resources and Evaluation (LREC-2010)*, May 2010.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X, Phuket, Thailand*, 2005.

Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. The significance of recall in automatic metrics for mt evaluation. In Robert E. Frederking and Kathryn Taylor, editors, *AMTA*, volume 3265 of *Lecture Notes in Computer Science*, pages 134–143. Springer, 2004a. ISBN 3-540-23300-8.

Alon Lavie, Shuly Wintner, Yaniv Eytani, Erik Peterson, and Katharina Probst. Rapid prototyping of a transfer-based Hebrew-to-English machine translation system. In *Proceedings of TMI-2004: The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, October 2004b.

Xiaoyi Ma and Mark Liberman. BITS: A method for bilingual text search over the web. In *Machine Translation Summit VII, Singapore*, 1999. doi: http://www.ldc.upenn.edu/Papers/MTSVII1999/BITS.ps.

I. Dan Melamed. Measuring semantic entropy. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics*, pages 41–46, 1997.

I. Dan Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26:221–249, 2000.

Guido Minnen, John Carroll, and Darren Pearce. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223, 2001. ISSN 1351-3249. doi: http://dx.doi.org/10.1017/S1351324901002728.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135.

Pavel Pecina. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*, 2008.

Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. Comparing and combining a semantic tagger and a statistical tool for mwe extraction. *Computer Speech and Language*, 19(4):378–397, 2005. ISSN 0885-2308. doi: http://dx.doi.org/10.1016/j.csl.2004.11.002.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 47–54, Singapore, August 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W09/W09-2907`.

Philip Resnik. Parallel strands: A preliminary investigation into mining the web for bilingual text. In *AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 72–82, London, UK, 1998. Springer-Verlag. ISBN 3-540-65259-0.

Philip Resnik. Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 527–534, Morristown, NJ, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3. doi: http://dx.doi.org/10.3115/1034678.1034757.

Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Computational Linguistics*, 29:349–380, 2003.

Ruvik Rosenthal. *Milon HaTserufim (Dictionary of Hebrew Idioms and Phrases)*. Keter, Jerusalem, 2009. In Hebrew.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico, 2002.

Yulia Tsvetkov and Shuly Wintner. Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of The seventh international conference on Language Resources and Evaluation (LREC-2010)*, May 2010.

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. Disambiguating Japanese compound verbs. *Computer Speech & Language*, 19(4):497–512, October 2005.

Tim Van de Cruys and Begoña Villada Moirón. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-1104`.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. In *Proceedings of RANLP'2005*, pages 590–596, 2005.

Sriram Venkatapathy and Aravind Joshi. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, July 2006.

Shailaja Venkatsubramanyan and Jose Perez-Carballo. Multiword expression filtering for building knowledge. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 40–47, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Begoña Villada Moirón and Jörg Tiedemann. Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context*. Association for Computational Linguistics, 2006.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 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 (EMNLP-CoNLL)*, pages 1034–1043, 2007. URL `http://www.aclweb.org/anthology/D/D07/D07-1110`.

Sina Zarrieß and Jonas Kuhn. Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 23–30, Singapore, August 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W09/W09-2904`.