

# Lexico-Semantic Multiword Expression Extraction

Tim Van de Cruys & Begoña Villada Moirón

University of Groningen

## Abstract

This paper describes a fully unsupervised and automated method for the large-scale extraction of multiword expressions (MWEs) from large corpora. The method takes into account the non-compositionality of MWEs; the intuition is that a noun within a MWE cannot easily be replaced by a semantically similar noun. To implement this intuition, a noun clustering is automatically extracted (using distributional similarity measures), which gives us clusters of semantically related nouns. Next, a number of statistical measures – based on selectional preferences – is developed that formalize the intuition of non-compositionality. The ratio of individual noun preference over cluster preference shows how likely a particular expression is to be a MWE (i.e. whether or not an individual noun accounts for all the preference of a certain cluster). Our approach has been tested on Dutch, and has been both manually and automatically evaluated.

## 1 Introduction

MWEs are expressions whose linguistic behaviour is not predictable from the linguistic behaviour of their component words. Baldwin (2006) characterizes the idiosyncratic behavior of MWEs as “a lack of compositionality manifest at different levels of analysis, namely, lexical, morphological, syntactic, semantic, pragmatic and statistical”. One property that seems to affect MWEs the most is semantic non-compositionality. MWEs are typically non-compositional. As a consequence, it is not possible to replace the content words of a MWE by semantically related words. Take for example the expressions in (1) and (2):

- (1) a. break the vase
- b. break the cup
- c. break the dish
- (2) a. break the ice
- b. \*break the snow
- c. \*break the hail

Expression (1) is fully compositional. Therefore, *vase* can easily be replaced with semantically related nouns such as *cup* and *dish*. Expression (2), on the contrary, is non-compositional; it is impossible to replace *ice* with semantically related words, such as *snow* and *hail*. Note that we assume a dual classification of expressions into compositional and non-compositional instances; we ignore the possibility that expressions fall in a continuum between compositionality and non-compositionality with many fuzzy cases in between. By ‘fuzzy cases’ we refer to expressions that are neither fully compositional nor fully non-compositional; such expressions may involve metaphoricity or figurative language.

Due to their non-compositionality, current proposals argue that MWEs need to

be described in the lexicon (Sag, Baldwin, Bond, Copestake and Flickinger 2002). In most languages, electronic lexical resources (such as dictionaries, thesauri, ontologies) suffer from a limited coverage of MWEs. To facilitate the update and expansion of language resources, the NLP community would clearly benefit from automated methods that extract MWEs from large text collections. This is the main motivation to pursue an automated and fully unsupervised MWE extraction method.

## 2 Previous work

Recent proposals that attempt to capture semantic compositionality (or lack thereof) employ various strategies. Approaches evaluated so far make use of dictionaries with semantic annotation (Piao, Rayson, Mudraya, Wilson and Garside 2006), wordNet (Pearce 2001), automatically generated thesauri (Lin 1999, Fazly and Stevenson 2006, McCarthy, Keller and Carroll 2003), vector-based methods that measure semantic distance (Baldwin, Bannard, Tanaka and Widdows 2003, Katz and Giesbrecht 2006), translations extracted from parallel corpora (Villada Moirón and Tiedemann 2006) or hybrid methods that use machine learning techniques informed by features coded using some of the above methods (Venkatapathy and Joshi 2005).

Pearce (2001) describes a method to extract collocations from corpora by measuring semantic compositionality. The underlying assumption is that a fully compositional expression allows synonym replacement of its component words, whereas a collocation does not. Pearce measures to what degree a collocation candidate allows synonym replacement. The measurement is used to rank candidates relative to their compositionality.

Building on Lin (1998), McCarthy et al. (2003) measure the semantic similarity between expressions (verb particles) as a whole and their component words (verb). They exploit contextual features and frequency information in order to assess meaning overlap. They established that human compositionality judgements correlate well with those measures that take into account the semantics of the particle. Contrary to these measures, multiword extraction statistics (log-likelihood, mutual information) poorly correlate with human judgements.

A different approach proposed by Villada Moirón and Tiedemann (2006) measures translational entropy as a sign of meaning predictability, and therefore non-compositionality. The entropy observed among word alignments of a potential MWE varies: highly predictable alignments show less entropy and probably correspond to compositional expressions. Data sparseness and polysemy pose problems because the translational entropy cannot be accurately calculated.

Fazly and Stevenson (2006) use lexical and syntactic fixedness as partial indicators of non-compositionality. Their method uses Lin's (1998) automatically generated thesaurus to compute a metric of lexical fixedness. Lexical fixedness measures the deviation between the pointwise mutual information of a verb-object phrase and the average pointwise mutual information of the expressions resulting from substituting the noun by its synonyms in the original phrase. This measure is similar to Lin's (1999) proposal for finding non-compositional phrases. Separ-

ately, a syntactic flexibility score measures the probability of seeing a candidate in a set of pre-selected syntactic patterns. The assumption is that non-compositional expressions score high in idiomaticity, that is, a score resulting from the combination of lexical fixedness and syntactic flexibility. The authors report an 80% accuracy in distinguishing literal from idiomatic expressions in a test set of 200 expressions. The performance of both metrics is stable across all frequency ranges.

In this study, we are interested in establishing whether a fully unsupervised method can capture the (partial or) non-compositionality of MWEs. The method should not depend on the existence of large (open domain) parallel corpora or sense tagged corpora. Also, the method should not require numerous adjustments when applied to new subclasses of MWEs, for instance, when coding empirical attributes of the candidates. Similar to Lin (1999), McCarthy et al. (2003) and Fazly and Stevenson (2006), our method makes use of automatically generated thesauri; the technique used to compile the thesauri differs from previous work. Aiming at finding a method of general applicability, the measures to capture non-compositionality differ from those employed in earlier work.

### 3 Methodology

In the description and evaluation of our algorithm, we focus on the extraction of verbal MWEs that contain prepositional complements, although the method could easily be generalized to other kinds of MWEs.

In our semantics-based approach, we want to formalize the intuition of non-compositionality, so that MWE extraction can be done in a fully automated way. A number of statistical measures are developed that try to capture the MWE's non-compositional bond between a verb-preposition combination and its noun by comparing the particular noun of a MWE candidate to other semantically related nouns.

#### 3.1 Data extraction

The MWE candidates (verb + prepositional phrase) are automatically extracted from the *Twente Nieuws Corpus* (Ordelman 2002), a large corpus of Dutch newspaper texts (500 million words), which has been automatically parsed by the Dutch dependency parser Alpino (van Noord 2006). Next, a matrix is created of the 5,000 most frequent verb-preposition combinations by the 10,000 most frequent nouns, containing the frequency of each MWE candidate.<sup>1</sup> To this matrix, a number of statistical measures are applied to determine the non-compositionality of the candidate MWEs. These statistical measures are explained in §3.3.

#### 3.2 Clustering

In order to compare a noun to its semantically related nouns, a noun clustering is created. These clusters are automatically extracted using standard distributional

---

<sup>1</sup>The lowest frequency verb-preposition combination (with regard to the 10,000 nouns) appears 3 times

similarity techniques (Weeds 2003, van der Plas and Bouma 2005). First, dependency triples are extracted from the *Twente Nieuws Corpus*. Next, feature vectors are created for each noun, containing the frequency of the dependency relations in which the noun occurs.<sup>2</sup> This way, a frequency matrix of 10K nouns by 100K dependency relations is constructed. The cell frequencies are replaced by point-wise mutual information scores (Church, Gale, Hanks and Hindle 1991), so that more informative features get a higher weight. The noun vectors are then clustered into 1,000 clusters using a simple K-means clustering algorithm (MacQueen 1967) with cosine similarity. During development, several other clustering algorithms and parameters have been tested, but the settings described above gave us the best EuroWordNet similarity score (using Wu and Palmer (1994)).

Note that our clustering algorithm is a hard clustering algorithm, which means that a certain noun can only be assigned to one cluster. This may pose a problem for polysemous nouns. On the other hand, this makes the computation of our metrics straightforward, since we do not have to decide among various senses of a word. In future work, we want to investigate the use of soft clustering algorithms, that take into account the various senses of a noun.

### 3.3 Measures

The measures used to find MWEs are inspired by Resnik’s method to find selectional preferences (Resnik 1993, Resnik 1996). Resnik uses a number of measures based on the Kullback-Leibler divergence, to measure the difference between the prior probability of a noun class  $p(c)$  and the probability of the class given a verb  $p(c|v)$ . We adopt the method for particular nouns, and add a measure for determining the ‘unique preference’ of a noun given other nouns in the cluster, which, we claim, yields a measure of non-compositionality. In total, four measures are used, the latter two being the symmetric counterpart of the former two.

#### 3.3.1 Verb preference

The first two measures,  $A_{v \rightarrow n}$  (equation 2) and  $R_{v \rightarrow n}$  (equation 3), formalize the unique preference of the verb<sup>3</sup> for the noun. Equation 1 gives the Kullback-Leibler divergence between the overall probability distribution of the nouns and the probability distribution of the nouns given a verb; it is used as a normalization constant in equation 2. Equation 2 models the actual preference of the verb for the noun.

$$(1) \quad S_v = \sum_n p(n | v) \log \frac{p(n | v)}{p(n)}$$

$$(2) \quad A_{v \rightarrow n} = \frac{p(n | v) \log \frac{p(n|v)}{p(n)}}{S_v}$$

<sup>2</sup>e.g. dependency relations that qualify *apple* might be ‘object of *eat*’ and ‘adjective *red*’. This gives us dependency triples like  $\langle \textit{apple}, \textit{obj}, \textit{eat} \rangle$ .

<sup>3</sup>we will use ‘verb’ to designate a prepositional verb, i.e. a combination of a verb and a preposition.

When  $p(n|v)$  is 0,  $A_{v \rightarrow n}$  is undefined. In this case, we assign a score of 0.

Equation 3 gives the ratio of the verb preference for a particular noun, compared to the other nouns that are present in the cluster.

$$(3) \quad R_{v \rightarrow n} = \frac{A_{v \rightarrow n}}{\sum_{n' \in C} A_{v \rightarrow n'}}$$

When  $R_{v \rightarrow n}$  is more or less equally divided among the different nouns in the cluster, there is no preference of the verb for a particular noun in the cluster, whereas scores close to 1 indicate a ‘unique’ preference of the verb for a particular noun in the cluster. Candidates whose  $R_{v \rightarrow n}$  value approaches 1 are likely to be non-compositional expressions.

### 3.3.2 Noun preference

In the latter two measures,  $A_{n \rightarrow v}$  and  $R_{n \rightarrow v}$ , the direction of preference is changed: they model the unique preference of the noun for the verb. Equation 4 models the Kullback-Leibler divergence between the overall probability distribution of verbs, and the distribution of the verbs given a certain noun. It is used again as a normalization constant in equation 5, which models the preference of the noun for the verb.

$$(4) \quad S_n = \sum_v p(v | n) \log \frac{p(v | n)}{p(v)}$$

$$(5) \quad A_{n \rightarrow v} = \frac{p(v | n) \log \frac{p(v|n)}{p(v)}}{S_n}$$

When  $p(v|n)$  is 0,  $A_{n \rightarrow v}$  is undefined. In this case, we again assign a score of 0.

Equation 6 gives the ratio of noun preference for a particular verb, compared to the other nouns that are present in the cluster.

$$(6) \quad R_{n \rightarrow v} = \frac{A_{n \rightarrow v}}{\sum_{n' \in C} A_{n' \rightarrow v}}$$

Both measures have the same characteristics as the ones that model verb preference. If a noun shows a much higher preference for a verb than the other nouns in the cluster, we expect that the candidate expression tends towards non-compositionality.

Note that the measures for verb preference and the measures for noun preference are different in nature. It is possible that a certain verb only selects a restricted set of nouns, while the nouns themselves can co-occur with many different verbs. This brings about different probability distributions. In our evaluation, we want to investigate the impact of both preferences.

### 3.3.3 Lexical fixedness measure

For reasons of comparison, we also evaluated the lexical fixedness measure – based on pointwise mutual information – proposed by Fazly and Stevenson (2006).<sup>4</sup> The lexical fixedness is computed following equation 7

$$(7) \quad Fixedness_{lex}(v, n) = \frac{PMI(v, n) - \overline{PMI}}{s}$$

where  $\overline{PMI}$  stands for the mean given the cluster, and  $s$  for the standard deviation. Note that Fazly and Stevenson (2006) use the  $M$  most similar nouns given a certain noun, while we use all nouns in a cluster. This means that our  $M$ -value varies.

### 3.4 Example

In this section, an elaborated example is presented, to show how our method works. Take for example the two MWE candidates in (3):

- (3) a. in de smaak vallen  
in the taste fall  
to be appreciated
- b. in de put vallen  
in the well fall  
to fall down the well

In the first expression, *smaak* cannot be replaced with other semantically similar nouns, such as *geur* ‘smell’ and *zicht* ‘sight’, whereas in the second expression, *put* can easily be replaced with other semantically similar words, such as *kuil* ‘hole’ and *krater* ‘crater’.

The first step in the formalization of this intuition, is the extraction of the clusters in which the words *smaak* and *put* appear from our clustering database. This gives us the clusters in (4).

- (4) a. **smaak:** *aroma* ‘aroma’, *gehoor* ‘hearing’, *geur* ‘smell’, *gezichtsvermogen* ‘sight’, *reuk* ‘smell’, *spraak* ‘speech’, *zicht* ‘sight’
- b. **put:** *afgrond* ‘abyss’, *bouwput* ‘building excavation’, *gaatje* ‘hole’, *gat* ‘hole’, *haat* ‘gap’, *hol* ‘cave’, *kloof* ‘gap’, *krater* ‘crater’, *kuil* ‘hole’, *lacune* ‘lacuna’, *leemte* ‘gap’, *valkuil* ‘pitfall’

Next, the various measures described in §3.3.1 and §3.3.2 are applied. Resulting scores are given in tables 1 and 2.

Table 1 gives the scores for the MWE *in de smaak vallen*, together with some other nouns that are present in the same cluster.  $A_{v \rightarrow n}$  shows that there is a clear preference (.12) of the verb *val in* for the noun *smaak*.  $R_{v \rightarrow n}$  shows that there is

<sup>4</sup>Fazly and Stevenson (2006) combine the lexical fixedness measure with a measure of syntactic flexibility. Here, we only compare our method to the former measure, concentrating on non-compositionality rather than syntactic rigidity.

MWE candidate	$A_{v \rightarrow n}$	$R_{v \rightarrow n}$	$A_{n \rightarrow v}$	$R_{n \rightarrow v}$
val#in smaak	.12	1.00	.04	1.00
val#in geur	.00	.00	.00	.00
val#in zicht	.00	.00	.00	.00

Table 1: Scores for MWE candidate *in de smaak vallen* and other nouns in the same cluster

a unique preference of the verb for the particular noun *smaak*. For the other nouns (*geur*, *zicht*, ...), the verb has no preference whatsoever. Therefore, the ratio of verb preference for *smaak* compared to the other nouns in the cluster is 1.00.

$A_{n \rightarrow v}$  and  $R_{n \rightarrow v}$  show similar behaviour. There is a preference (.04) of the noun *smaak* for the verb *val in*, and this preference is unique (1.00).

MWE candidate	$A_{v \rightarrow n}$	$R_{v \rightarrow n}$	$A_{n \rightarrow v}$	$R_{n \rightarrow v}$
val#in put	.00	.05	.00	.05
val#in kuil	.01	.11	.02	.37
val#in kloof	.00	.02	.00	.03
val#in gat	.04	.71	.01	.24

Table 2: Scores for MWE candidate *in de put vallen* and other nouns in the same cluster

Table 2 gives the scores for the instance *in de put vallen* – which is not a MWE – together with other nouns from the same cluster. The results are quite different from the ones in table 1.  $A_{v \rightarrow n}$  – the preference of the verb for the noun – is quite low in most cases, the highest score being a score of .04 for *gat*. Furthermore,  $R_{v \rightarrow n}$  does not show a unique preference of *val in* for *put* (a low ratio score of .05). Instead, the preference mass is divided among the various nouns in the cluster, the highest preference of *val in* being assigned to the noun *gat* (.71).<sup>5</sup>

The other two scores show again a similar tendency;  $A_{n \rightarrow v}$  – the preference of the noun for the verb – is low in all cases, and when all nouns in the cluster are considered ( $R_{n \rightarrow v}$ ), there is no ‘unique’ preference of one noun for the verb *val in*. Instead, the preference mass is divided among all nouns in the cluster.

After assessing the values of the four different measures, our method would propose *in de smaak vallen* as a non-compositional expression and therefore, MWE; on the other hand, the method would consider *in de put vallen* as compositional, thus a non-MWE.

<sup>5</sup>Note that this expression is ambiguous: it can be used in a literal sense (*in een gat vallen*, ‘to fall down a hole’) and in a metaphorical sense (*in een zwart gat vallen*, ‘to get depressed after a joyful or busy period’).

## 4 Results and evaluation

In this section, our automatic method is extensively evaluated. In the first part, we present the results of our quantitative evaluation – including both an automatic evaluation (using Dutch lexical resources) and a manual evaluation (carried out by human judges). The second part is a qualitative evaluation, indicating the advantages and the drawbacks of our method.

### 4.1 Quantitative evaluation

#### 4.1.1 Automatic evaluation

The MWEs that are extracted with the fully unsupervised method described above are automatically evaluated by comparing the extracted MWEs to handcrafted lexical databases. Since we have extracted Dutch MWEs, we are using the two Dutch resources available: the Referentie Bestand Nederlands (RBN, (Martin and Maks 2005)) and the Van Dale Lexicographical Information System (VLIS) database. Precision and recall are calculated with regard to the MWEs that are present in our evaluation resources. Among the MWEs in our reference data, we consider only those expressions that are present in our frequency matrix: if the verb is not among the 5,000 most frequent verbs, or the noun is not among the 10,000 most frequent nouns, the frequency information is not present in our input data. Consequently, our algorithm would never be able to find those MWEs.

The first six rows of table 3 show precision, recall and f-measure for various parameter thresholds with regard to the measures  $A_{v \rightarrow n}$ ,  $R_{v \rightarrow n}$ ,  $A_{n \rightarrow v}$  and  $R_{n \rightarrow v}$ , together with the number of candidates found (n). The last line shows the highest values we were able to reach by using the lexical fixedness score.

parameters					precision	recall	f-measure
$A_{v \rightarrow n}$	$R_{v \rightarrow n}$	$A_{n \rightarrow v}$	$R_{n \rightarrow v}$	n	(%)	(%)	(%)
.10	.80	–	–	3175	16.09	13.11	14.45
.10	.90	–	–	2655	17.59	11.98	14.25
.10	.80	–	.80	2225	19.19	10.95	13.95
.10	.90	–	.90	1870	20.70	9.93	13.42
.10	.80	.01	.80	1859	20.33	9.69	13.13
.20	.99	.05	.99	404	38.12	3.95	7.16
$Fixedness_{lex}(v, n)$			3.00	3899	15.14	9.92	11.99

Table 3: Evaluation results compared to RBN & VLIS

Using only two parameters –  $A_{v \rightarrow n}$  and  $R_{v \rightarrow n}$  – gives the highest f-measure ( $\pm 14\%$ ), with a precision and recall of about 17% and about 12% respectively. Adding parameter  $R_{n \rightarrow v}$  increases precision but degrades recall, and this tendency continues when adding both parameters  $A_{n \rightarrow v}$  and  $R_{n \rightarrow v}$ . In all cases, a higher

threshold increases precision but degrades recall. When using a high threshold for all parameters, the algorithm is able to reach a precision of  $\pm 38\%$ , but recall is low ( $\pm 4\%$ ).

The lexical fixedness score is able to reach an f-measure of  $\pm 12\%$  (using a threshold of 3.00). These scores show the best performance that we have reached using lexical fixedness.

#### 4.1.2 Human evaluation

The evaluation procedure described above was applied fully automatically by comparing the output of our method to two existing Dutch lexical databases. We are aware of the fact that the automated annotation process may introduce some errors. There may be extracted expressions wrongly labeled as true MWEs but also extracted expressions erroneously labeled as false MWEs. Furthermore, it is known that the used lexical databases are static resources that are likely to miss actual MWEs found in large corpora. This is either because the lexical resources are incomplete, or because the MWEs were not included due to a different understanding of the concept of MWE. With this motivation, we set up a human evaluation experiment. From the dataset that produced the best f-measure ( $A_{v \rightarrow n} = .10$  and  $R_{v \rightarrow n} = .80$ ), 200 expressions were semi-randomly selected. To assess the performance of our method across different frequency ranges, we selected 100 high frequent MWE candidates (frequency  $\geq 100$ ) and 100 low frequent ones (frequency  $< 100$ ).

Three human judges were asked to label the expressions as MWE or as non-MWE. The judges were asked to always provide an answer. To investigate if the rankings from the 3 judges agreed, we employed the Kappa statistic (Cohen 1960). We obtained an average pairwise interannotator agreement of  $\kappa = .60$ , showing a reasonable correlation between the judges.

The scores assigned by the judges differed severely with regard to frequency range. In the high frequency range, our method was given an average precision of 33.00%. In the low frequency range, precision dropped down to 6.67%. In §4.2.2, the results of our human evaluation are evaluated more extensively.

## 4.2 Qualitative evaluation

In this section, we elaborate upon advantages and disadvantages of our semantics-based MWE extraction algorithm by examining the output of the procedure, and looking at the characteristics of the correct MWEs found and the errors made by the algorithm.

### 4.2.1 Advantages of the method

First of all, our algorithm is able to filter out grammatical collocations that cause problems in traditional MWE extraction paradigms. Two examples are given in (5) and (6).

- (5) benoemen tot minister, secretaris-generaal  
 appoint to minister, secretary-general  
*appoint s.o. {minister, secretary-general}*
- (6) voldoen aan eisen, voorwaarden  
 meet to demands, conditions  
*meet the {demands, conditions}*

In traditional MWE extraction algorithms, based on collocations, highly frequent expressions like the ones in (5) and (6) often get classified as a MWE, even though they are fully compositional. Such algorithms correctly identify a strong lexical affinity between two component words (*voldoen, aan*), which make up a grammatical collocation; however, they fail to capture the fact that the noun may be filled in by a semantic class of nouns. Our algorithm filters out those expressions, because semantic similarity is taken into account.

Our quantitative evaluation shows that the algorithm reaches the best results (i.e. the highest f-measures) when only two parameters ( $A_{v \rightarrow n}$  and  $R_{v \rightarrow n}$ ) are taken into account. But upon closer inspection of the output, we have noticed that  $A_{n \rightarrow v}$  and  $R_{n \rightarrow v}$  are often able to filter out non-MWEs like the expressions b in (7) and (8).

- (7) a. op toneel verschijnen  
 on stage appear  
*to appear*
- b. op toneel zingen  
 on stage sing  
*to sing on the stage*
- (8) a. in geheugen liggen  
 in memory lie  
*be in memory*
- b. in ziekenhuis liggen  
 in hospital lie  
*lie in the hospital*

When only taking into account the first two measures (a unique preference of the verb for the noun), the expressions in b do not get filtered out. It is only when the two other measures (a unique preference of the noun for the verb) are taken into account that they are filtered out – either because the preference of the noun for the verb is very low, or the noun preference for the verb is more evenly distributed among the cluster. The b expressions, which are non-MWEs, result from the combination of a verb with a highly frequent PP. These PPs are typically locative, directional or predicative PPs, that may combine with numerous verbs.

Also, expressions like the ones in (9), where the fixedness of the expression lies not so much in the verb-noun combination, but more in the PP part (*naar school, naar huis*) are filtered out by the latter two measures. These preposition-noun combinations seem to be institutionalized PPs, so-called determinerless PPs

(Baldwin, Beavers, van der Beek, Bond, Flickinger and Sag 2006).

- (9) a. naar school willen  
to school want  
*want to go to school*
- b. naar huis willen  
to home want  
*want to go home*

#### 4.2.2 Errors of the method

In this section, we give an exhaustive list of the errors made by our algorithm, and quantitatively evaluate the importance of each error category.

1. First of all, our algorithm highly depends on the quality of the noun clustering. If a noun appears in a cluster with unrelated words, the measures will overrate the semantic uniqueness of the expressions in which the noun appears.
2. Syntax might play an important role. Sometimes, there are syntactic restrictions between the preposition and the noun. A noun like *pagina* ‘page’ can only appear with the preposition *op* ‘on’, as in *lees op pagina* ‘read on page’. Other, semantically related nouns, such as *hoofdstuk* ‘chapter’, prefer *in* ‘in’. Due to these restrictions, the measures will again overrate the semantic uniqueness of the expression.
3. We found many expressions in which the fixedness of the expression lies not so much in the combination of the verb and the prepositional phrase, but rather in the prepositional phrase itself (*naar school*, *naar huis*). Note, however, that our two latter measures were able to filter out many of those expressions (as noted in §4.2.1). But in our error evaluation, we used the result that yields the highest f-measure (and does not take the latter measures into account).
4. Our hard clustering method does not take polysemous nouns into account. A noun can only occur in one cluster, ignoring other possible meanings. *Schaal*, for example, means ‘dish’ as well as ‘scale’. In our clustering, it only appears in a cluster of dish-related nouns. Therefore, expressions like *maak gebruik op [grote] schaal* ‘make use of [sth.] on a [large] scale’, receive again overrated measures of semantic uniqueness, because the ‘scale’ sense of the noun is compared to nouns related to the ‘dish’ sense.
5. Related to the previous error category is the fact that certain nouns – although occurring in a perfectly sound cluster – possess a semantic feature or characteristic that distinguishes them from the other nouns in the cluster, and causes the verb to uniquely prefer that particular noun. An example of this kind of error is the expression *eet in restaurant* ‘eat in a restaurant’, which

is perfectly compositional. But due to the fact that the noun *restaurant* ends up in a cluster with nouns such as *bar* ‘bar’, *café* ‘bar’,  *kroeg* ‘pub’, *winkel* ‘shop’, *hotel* ‘hotel’ – which are places where one is less likely to eat – the fixedness of the expression is overestimated.

6. The effectiveness of our method is highly dependent on the corpus distribution. Sometimes, expressions that would be effective counterweights for the erroneous classification of compositional expressions as MWE just are not found in the corpus. This might be either due to sparseness of the data, or due to the specific nature of the corpus itself. Examples are *sluit wegens verbouwing* ‘close due to alteration’, with cluster members such as *restauratie* ‘restoration’ and *renovatie* ‘renovation’, and *uit van emotie* ‘express emotion’, with cluster members such as *agressie* ‘aggression’, *irritatie* ‘irritation’, *ongeduld* ‘impatience’. Expressions such as *sluit wegens renovatie* or *uit van irritatie* are perfectly possible, but are not (sufficiently) attested in the corpus. Therefore, the compositional forms which are attested in the corpus are overestimated as MWE.
7. Finally, misclassifications may be caused by parsing errors or other technical issues.

In order to get a better view of the errors of the method, we manually classified the expressions that were evaluated as non-MWE by our judges. Each expression was assigned to one of the error categories described above. Overall results, and results for high and low frequency expressions are given.

	overall (%)	high freq. (%)	low freq. (%)
1 erroneous clustering	3.6	3.8	3.4
2 specific preposition	6.4	15.4	1.1
3 PP fixedness	26.4	21.2	29.5
4 polysemous word	15.7	13.5	17.0
5 specific semantic feature	22.9	30.8	18.2
6 corpus distribution	21.4	13.5	26.1
7 parsing/other	3.6	1.9	4.5

Table 4: Quantitative error evaluation

Misclassifications due to erroneous clustering or parsing errors only constitute a small part of the errors. Also, misclassifications due to syntactic restrictions (specific prepositions) are responsible for only a small part of the errors. More important are misclassifications due to fixedness in the PP, or due to polysemy or specific semantic features of the nouns. The former might be remedied by a more effective use of our measures  $A_{n \rightarrow v}$  and  $R_{n \rightarrow v}$ , the latter by taking on a soft clustering approach. Finally, there are quite some errors due to the specific

distribution of MWEs in the corpus. These errors are more common in the low frequency range. Clearly, our method is highly dependent on the corpus that is used, and it should be sufficiently large in order to adequately classify less frequent MWEs.

#### 4.2.3 MWE fuzziness

A last observation to mention is that the status of certain expressions extracted with our method is unclear. Expressions such as *vraag met klem* ‘ask with emphasis’ or *ga over tot orde [van de dag]* ‘pass to the order [of the day]’ seem to be on the border of compositionality vs. non-compositionality, and therefore cannot be adequately qualified as MWE or non-MWE. This observation is confirmed by the conflicting views the three judges showed when assessing these kind of expressions.

## 5 Conclusions and further work

Our algorithm based on non-compositionality explores a new approach aimed at large-scale MWE extraction. Using only two parameters,  $A_{v \rightarrow n}$  and  $R_{v \rightarrow n}$ , yields the highest f-measure. Using the two other parameters,  $A_{n \rightarrow v}$  and  $R_{n \rightarrow v}$ , increases precision but degrades recall. Due to the formalization of the intuition of non-compositionality (using an automatic noun clustering), our algorithm is able to rule out various expressions that are coined MWEs by traditional algorithms.

Note that our algorithm has taken on a purely semantics-based approach. ‘Syntactic fixedness’ of the expressions is not taken into account. Combining our semantics-based approach with other MWE extraction methods that take into account different features might improve the results significantly.

We conclude with some issues saved for future work. First of all, we would like to combine our semantics-based method with other methods that are used to find MWEs (especially syntax-based methods), and implement the method in general classification models (decision tree classifier and maximum entropy model). This includes the use of a more principled (machine learning) framework in order to establish the optimal threshold values, and the use of appropriate median values and confidence intervals in order to model the different levels within a continuum of compositionality.

Next, we would like to investigate a number of topics to improve on our semantics-based method. First of all, using the top  $k$  similar nouns for a certain noun – instead of the cluster in which a noun appears – might be more beneficial to get a grasp of the compositionality of MWE candidates. Also, making use of a verb clustering in addition to the noun clustering might also help in determining the non-compositionality of expressions. Disambiguating among the various senses of nouns should also be a useful improvement. Furthermore, we would like to generalize our method to other syntactic patterns (e.g. verb object combinations), and test the approach for English.

We believe that our method provides a genuine and successful approach to

get a grasp of the non-compositionality of MWES in a fully automated way. We also believe that it is one of the first methods able to extract MWES based on non-compositionality on a large scale, and that traditional MWE extraction algorithms will benefit from taking this non-compositionality into account.

### Acknowledgements

This research was carried out as part of the IRME STEVIN research project. We would like to thank our three human judges (Nicole Grégoire, Jori Mur, Gertjan van Noord) and the two anonymous reviewers for their helpful comments on an earlier version of this paper.

### References

- Baldwin, T.(2006), Compositionality and Multiword Expressions: Six of One, Half a Dozen of the Other?, Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties.
- Baldwin, T., Bannard, C., Tanaka, T. and Widdows, D.(2003), An Empirical Model of Multiword Expressions Decomposability, *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 89–96.
- Baldwin, T., Beavers, J., van der Beek, L., Bond, F., Flickinger, D. and Sag, I.(2006), *In search of a systematic treatment of Determinerless PPs*, Computational Linguistics Dimensions of Syntax and Semantics of Prepositions, Kluwer Academic, pp. 163–180.
- Church, K., Gale, W., Hanks, P. and Hindle, D.(1991), Using statistics in lexical analysis, in U. Zernik (ed.), *Lexical Acquisition: Exploiting On-line resources to build a lexicon*, Lawrence Erlbaum Associates, New Jersey, pp. 115–164.
- Cohen, J.(1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.
- Fazly, A. and Stevenson, S.(2006), Automatically constructing a lexicon of verb phrase idiomatic combinations, *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy.
- Katz, G. and Giesbrecht, E.(2006), Automatic identification of non-compositional multi-word expressions using Latent Semantic Analysis, *Proc. of the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, pp. 12–19.
- Lin, D.(1998), Automatic retrieval and clustering of similar words, *Proceedings of COLING/ACL 98*, Montreal, Canada.
- Lin, D.(1999), Automatic identification of non-compositional phrases, *Proceedings of ACL-99*, University of Maryland, pp. 317–324.
- MacQueen, J. B.(1967), Some methods for classification and analysis of mul-

- tivariate observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, pp. 281–297.
- Martin, W. and Maks, I.(2005), *Referentie Bestand Nederlands. Documentatie*.
- McCarthy, D., Keller, B. and Carroll, J.(2003), Detecting a Continuum of Compositionality in Phrasal Verbs, *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.
- Ordelman, R.(2002), Twente Nieuws Corpus (TwNC). Parlevink Language Technology Group. University of Twente.
- Pearce, D.(2001), Synonymy in collocation extraction, *WordNet and Other lexical resources: applications, extensions & customizations (NAACL 2001)*, Carnegie Mellon University, Pittsburgh, pp. 41–46.
- Piao, S., Rayson, P., Mudraya, O., Wilson, A. and Garside, R.(2006), Measuring MWE compositionality using semantic annotation, *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Association for Computational Linguistics, Sydney, Australia, pp. 2–11.
- Resnik, P.(1993), *Selection and Information: A Class-Based Approach to Lexical Relationships*, PhD Thesis, University of Pennsylvania.
- Resnik, P.(1996), Selectional constraints: An information-theoretic model and its computational realization, *Cognition* **61**, 127–159.
- Sag, I., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D.(2002), Multiword Expressions: a pain in the neck for NLP, *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, pp. 1–15.
- van der Plas, L. and Bouma, G.(2005), Syntactic contexts for finding semantically similar words, *Computational Linguistics in the Netherlands 2004. Selected Papers from the Fifteenth CLIN Meeting* pp. 173–184.
- van Noord, G.(2006), At Last Parsing Is Now Operational, in P. Mertens, C. Fairon, A. Dister and P. Watrin (eds), *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, Leuven, pp. 20–42.
- Venkatapathy, S. and Joshi, A.(2005), Measuring the relative compositionality of verb-noun collocations by integrating features, *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, pp. 899–906.
- Villada Moirón, B. and Tiedemann, J.(2006), Identifying idiomatic expressions using automatic word-alignment, *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context*, Trento, Italy, pp. 33–40.
- Weeds, J.(2003), *Measures and Applications of Lexical Distributional Similarity*, PhD Thesis, University of Sussex.
- Wu, Z. and Palmer, M.(1994), Verb semantics and lexical selection, *32nd. Annual Meeting of the Association for Computational Linguistics*, New Mexico State University, Las Cruces, New Mexico, pp. 133–138.