8.1 Introduction

In the previous chapters, we have shown that it is possible to automatically extract a database of semantically similar words. In this chapter, we will explain how such a database of semantically similar words can in turn be used to automatically extract multi-word expressions (mws) from text.

mws are expressions whose linguistic behaviour is not predictable from the linguistic behaviour of their component words. Baldwin (2006) characterizes the idiosyncratic behavior of mws 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 mws the most is semantic non-compositionality. mws are typically non-compositional. As a consequence, it is not possible to replace the content words of a mwe by semantically similar words. Take for example the expressions in (1) and (2):

(1)  
   a. break the vase  
   b. break the cup  
   c. break the dish

---

1This chapter presents joint work with Begoña Villada Moirón. The research presented in this chapter has been published as Van de Cruys and Villada Moirón (2007a) and Van de Cruys and Villada Moirón (2007b).
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 et al., 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 et al., 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 (1998a), McCarthy, Keller, and Carroll (2003) measure the semantic similarity between expressions (phrasal verbs) as a whole and their component words. 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) correlate poorly 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 \textit{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. Separately, 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 \textit{mwe}s. 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 \textit{mwe}s, for instance, when coding empirical attributes of the candidates. Similar to Lin (1999), McCarthy, Keller, and Carroll (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.
8.3 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 MWES extraction can be done in a fully automated way. A number of statistical measures are developed that try to capture the MWES’s non-compositional bond between a verb-preposition combination and its noun by comparing the particular noun of an MWES candidate to other semantically related nouns.

8.3.1 Data extraction

The MWES 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 with 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 MWES candidate. 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 §8.3.3.

8.3.2 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 similarity techniques. A syntax-based model is created, similar to the models described in the evaluation part of this dissertation, using the Twente Nieuws Corpus. The model contains 10k nouns by 100k dependency relations, and is weighted with pointwise mutual information. The noun vectors are then clustered into 1,000 clusters using a k-means clustering algorithm (MacQueen, 1967) using 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)).

---

2The lowest frequency verb-preposition combination (with regard to the 10,000 nouns) appears 3 times.

3A detailed description of distributional similarity techniques is given in chapter 2.
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.

8.3.3 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.

Verb preference

The first two measures, $A_{v\rightarrow n}$ (equation 8.2) and $R_{v\rightarrow n}$ (equation 8.3), formalize the ‘unique’ preference of the verb for the noun, i.e. the preference of a particular verb $v$ to combine with a particular noun $n$ (from a cluster $C$ of semantically related nouns). Equation 8.1 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 8.2. Equation 8.2 models the actual preference of the verb for the noun.

\[
S_v = \sum_n p(n|v) \log \frac{p(n|v)}{p(n)} \tag{8.1}
\]

\[
A_{v\rightarrow n} = \frac{p(n|v)}{S_v} \log \frac{p(n|v)}{p(n)} \tag{8.2}
\]

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

Equation 8.3 gives the ratio of the verb preference for a particular noun, compared to the other nouns that are present in the cluster.

\[^4\text{We will use ‘verb’ to designate a prepositional verb, i.e. a combination of a verb and a preposition.}\]
\[ R_{v \rightarrow n} = \frac{A_{v \rightarrow n}}{\sum_{n' \in C} A_{v \rightarrow n'}} \]  

(8.3)

where \( C \) is the cluster in which the noun \( n \) appears. 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, since the noun of the expression cannot be substituted with a semantically similar noun.

**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 8.4 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 8.5, which models the preference of the noun for the verb.

\[ S_n = \sum_v p(v|n) \log \frac{p(v|n)}{p(v)} \]  

(8.4)

\[ A_{n \rightarrow v} = \frac{p(v|n) \log \frac{p(v|n)}{p(v)}}{S_n} \]  

(8.5)

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

Equation 8.6 gives the ratio of noun preference for a particular verb, compared to the other nouns that are present in the cluster.

\[ R_{n \rightarrow v} = \frac{A_{n \rightarrow v}}{\sum_{n' \in C} A_{n' \rightarrow v}} \]  

(8.6)

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.

**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).\(^5\) The lexical fixedness is computed following equation 8.7

\[
\text{fixedness}_{\text{lex}}(v, n) = \frac{\text{pmi}(v, n) - \text{pmi}}{s}
\]  

where \(\text{pmi}\) 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 value for \(m\) varies.

**8.3.4 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)  
\[\begin{array}{l}
a. \text{in de smaak vallen} \\
\quad \text{in the taste fall} \\
\quad \text{to be appreciated} \\
b. \text{in de put vallen} \\
\quad \text{in the well fall} \\
\quad \text{to fall down the well} \\
\end{array}\]

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)  
\[\begin{array}{l}
\end{array}\]

---

\(^5\)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.
Next, the various measures described in §8.3.3 are applied. Resulting scores are given in tables 8.1 and 8.2.

<table>
<thead>
<tr>
<th>MWE Candidate</th>
<th>$A_{v\rightarrow n}$</th>
<th>$R_{v\rightarrow n}$</th>
<th>$A_{n\rightarrow v}$</th>
<th>$R_{n\rightarrow v}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>val# in smaak</td>
<td>.12</td>
<td>1.00</td>
<td>.04</td>
<td>1.00</td>
</tr>
<tr>
<td>val# in geur</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>val# in zicht</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

Table 8.1: Scores for MWE candidate *in de smaak vallen* and other nouns in the same cluster

Table 8.1 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 for the noun smaak. $R_{v\rightarrow n}$ shows that there is 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, and this preference is unique (1.00).

<table>
<thead>
<tr>
<th>MWE Candidate</th>
<th>$A_{v\rightarrow n}$</th>
<th>$R_{v\rightarrow n}$</th>
<th>$A_{n\rightarrow v}$</th>
<th>$R_{n\rightarrow v}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>val# in put</td>
<td>.00</td>
<td>.05</td>
<td>.00</td>
<td>.05</td>
</tr>
<tr>
<td>val# in kuil</td>
<td>.01</td>
<td>.11</td>
<td>.02</td>
<td>.37</td>
</tr>
<tr>
<td>val# in kloof</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.03</td>
</tr>
<tr>
<td>val# in gat</td>
<td>.04</td>
<td>.71</td>
<td>.01</td>
<td>.24</td>
</tr>
</tbody>
</table>

Table 8.2: Scores for MWE candidate *in de put vallen* and other nouns in the same cluster

Table 8.2 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 8.1. $A_{v\rightarrow n}$ – 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).

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.

### 8.4 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.

#### 8.4.1 Quantitative evaluation

**Automatic evaluation**

The **MWEs** that are extracted with the fully unsupervised method described above are automatically evaluated by comparing them 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**.

---

6Note 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’).
The first six rows of table 8.3 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.

<table>
<thead>
<tr>
<th>$A_{v \rightarrow n}$</th>
<th>$R_{v \rightarrow n}$</th>
<th>$A_{n \rightarrow v}$</th>
<th>$R_{n \rightarrow v}$</th>
<th>n</th>
<th>precision (%)</th>
<th>recall (%)</th>
<th>f-measure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>.10</td>
<td>.80</td>
<td>–</td>
<td>–</td>
<td>3175</td>
<td>16.09</td>
<td>13.11</td>
<td>14.45</td>
</tr>
<tr>
<td>.10</td>
<td>.90</td>
<td>–</td>
<td>–</td>
<td>2655</td>
<td>17.59</td>
<td>11.98</td>
<td>14.25</td>
</tr>
<tr>
<td>.10</td>
<td>.80</td>
<td>–</td>
<td>.80</td>
<td>2225</td>
<td>19.19</td>
<td>10.95</td>
<td>13.95</td>
</tr>
<tr>
<td>.10</td>
<td>.90</td>
<td>–</td>
<td>.90</td>
<td>1870</td>
<td>20.70</td>
<td>9.93</td>
<td>13.42</td>
</tr>
<tr>
<td>.10</td>
<td>.80</td>
<td>.01</td>
<td>.80</td>
<td>1859</td>
<td>20.33</td>
<td>9.69</td>
<td>13.13</td>
</tr>
<tr>
<td>.20</td>
<td>.99</td>
<td>.05</td>
<td>.99</td>
<td>404</td>
<td>38.12</td>
<td>3.95</td>
<td>7.16</td>
</tr>
<tr>
<td>fixedness_{lex}(v, n)</td>
<td>.50</td>
<td>.50</td>
<td>.50</td>
<td>3899</td>
<td>15.14</td>
<td>9.92</td>
<td>11.99</td>
</tr>
</tbody>
</table>

Table 8.3: Evaluation results compared to RBN & VLIS

Using only two parameters – $A_{v \rightarrow n}$ and $R_{v \rightarrow n}$ – gives the highest f-measure (± 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 ± 38%, but recall is low (± 4%).

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

**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, as well as extracted expressions erroneously labeled as false MWEs. Furthermore, it is known that the lexical databases used 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 mwe\textsubscript{s} 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 ($A_{v-n} = .10$ and $R_{v-n} = .80$), 200 expressions were randomly selected. To assess the performance of our method across different frequency ranges, we selected 100 highly frequent mwe\textsubscript{s} candidates (frequency $\geq 100$) and 100 less 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%. Below, the results of our human evaluation are evaluated more extensively.

8.4.2 Qualitative evaluation

In this section, we elaborate upon advantages and disadvantages of our semantics-based mwe\textsubscript{e} extraction algorithm by examining the output of the procedure, and looking at the characteristics of the correct mwe\textsubscript{e}s found and the errors made by the algorithm.

Advantages of the method

First of all, our algorithm is able to filter out grammatical collocations that cause problems in traditional mwe\textsubscript{e} 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\textsubscript{e} extraction algorithms, based on collocations, highly frequent expressions like the ones in (5) and (6) often get classified as a mwe\textsubscript{e}, 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. Thesepps are typically locative, directional or predicativepps, 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 institutionalizedpps, so-called determinerlesspps (Baldwin et al., 2006).

(9) a. naar school willen
   to school want
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 explained above). 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 wegen 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* ‘agression’, *irritatie* ‘irritation’, *ongeduld* ‘impatience’. Expressions such as *sluit wegen 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.

<table>
<thead>
<tr>
<th>Overall (%)</th>
<th>High freq. (%)</th>
<th>Low freq. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 erroneous clustering</td>
<td>3.6</td>
<td>3.8</td>
</tr>
<tr>
<td>2 specific preposition</td>
<td>6.4</td>
<td>15.4</td>
</tr>
<tr>
<td>3 PP fixedness</td>
<td>26.4</td>
<td>21.2</td>
</tr>
<tr>
<td>4 polysemous word</td>
<td>15.7</td>
<td>13.5</td>
</tr>
<tr>
<td>5 specific semantic feature</td>
<td>22.9</td>
<td>30.8</td>
</tr>
<tr>
<td>6 corpus distribution</td>
<td>21.4</td>
<td>13.5</td>
</tr>
<tr>
<td>7 parsing/other</td>
<td>3.6</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Table 8.4: 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 mwe 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 mwe.

**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.

### 8.5 Conclusion

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