Semantics-based Multiword Expression Extraction

Tim Van de Cruys and Begoña Villada Moirón
Alfa Informatica, University of Groningen
Oude Kijk in 't Jatstraat 26
9712 EK Groningen, The Netherlands
{T.Van.de.Cruys|M.B.Villada.Moiron}@rug.nl

Abstract

This paper describes a fully unsupervised and automated method for large-scale extraction of multiword expressions (MWEs) from large corpora. The method aims at capturing 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. Our approach has been tested on Dutch, and automatically evaluated using Dutch lexical resources.

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". Some MWEs show productive morphology and/or syntactic flexibility. Therefore, these two aspects are not sufficient conditions to discriminate actual MWEs from productive expressions. Nonetheless, the mentioned characteristics are useful indicators to distinguish literal and idiomatic expressions (Fazly and Stevenson, 2006).

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 noun of a MWE by semantically related nouns. 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-a) is fully compositional. Therefore, *vase* can easily be replaced with semantically related nouns such as *cup* and *dish*. Expression (2-a), on the contrary, is non-compositional; *ice* cannot be replaced with semantically related words, such as *snow* and *hail* without loss of the original meaning.

Due to the idiosyncratic behavior, current proposals argue that MWEs need to be described in the lexicon (Sag et al., 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.

Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pages 25–32,
Prague, June 2007.

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 et al., 2006), WordNet (Pearce, 2001), automatically generated thesauri (Lin, 1999; McCarthy et al., 2003; Fazly and Stevenson, 2006), 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 (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, standard association measures 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 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 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 we believe the method can be easily 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.[1] 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 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.[2] This way, a frequency matrix of 10K nouns by 100K dependency relations is constructed. The cell frequencies are replaced by pointwise mutual information scores (Church et al., 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.

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, 4 measures are used, the latter two being the symmetric counterpart of the former two.

The first two measures, $A_{v \to n}$ (equation 2) and $R_{v \to n}$ (equation 3), formalize the unique preference of the verb[3] 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.

$$S_v = \sum_n p(n \mid v) \, log \frac{p(n \mid v)}{p(n)} \tag{1}$$

$$A_{v \to n} = \frac{p(n \mid v) \, log \frac{p(n|v)}{p(n)}}{S_v} \tag{2}$$

When $p(n|v)$ is 0, $A_{v \to n}$ 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.

$$R_{v \to n} = \frac{A_{v \to n}}{\sum_{n' \epsilon C} A_{v \to n'}} \tag{3}$$

When $R_{v \to n}$ is more or less equally divided among the different nouns in the cluster, there is no

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

[2] e.g. dependency relations that qualify *apple* might be 'object of *eat*' and 'adjective *red*'. This gives us dependency triples like $< apple, obj, eat >$.

[3] We will use 'verb' to designate a prepositional verb, i.e. a combination of a verb and a preposition.

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 $R_{v \to n}$ value approaches 1 are likely to be non-compositional expressions.

In the latter two measures, $A_{n \to v}$ and $R_{n \to v}$, the direction of preference is changed: equations 4 and 5 are the symmetric counterparts of equations 2 and 3. Instead of the preference of the verb for the noun, the preference of the noun for the verb is modelled. Except for the change of preference direction, the characteristics of the former and the latter two measures are the same.

$$A_{n \to v} = \frac{p(v \mid n) \; log \frac{p(v|n)}{p(v)}}{S_n} \qquad (4)$$

$$R_{n \to v} = \frac{A_{n \to v}}{\sum_{n' \epsilon C} A_{n' \to v}} \qquad (5)$$

Note that, despite their symmetry, 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 number 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.

3.4 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', *hiaat* 'gap', *hol* 'cave', *kloof* 'gap', *krater* 'crater', *kuil* 'hole', *lacune* 'lacuna', *leemte* 'gap', *valkuil* 'pitfall'

Next, the various measures described in section 3.3 are applied. Resulting scores are given in tables 1 and 2.

MWE candidate	$A_{v \to n}$	$R_{v \to n}$	$A_{n \to v}$	$R_{n \to v}$
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.

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

$A_{n \to v}$ and $R_{n \to v}$ 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	$A_{v \to n}$	$R_{v \to n}$	$A_{n \to v}$	$R_{n \to v}$
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 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. $A_{v \to n}$ – 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, $R_{v \to n}$ 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).[4]

The other two scores show again a similar tendency; $A_{n \to v}$ – the preference of the noun for the verb – is low in all cases, and when all nouns in the cluster are considered ($R_{n \to v}$), 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.

4 Results & Evaluation

4.1 Quantitative evaluation

In this section, we quantitatively evaluate our method, and compare it to the lexical and syntactic fixedness measures proposed by Fazly and Stevenson (2006). More information about Fazly and Stevenson's measures can be found in their paper.

The potential MWEs that are extracted with the fully unsupervised method described above and with Fazly and Stevenson's (2006) method (FS from here onwards) are automatically evaluated by comparing the extracted list to handcrafted MWE 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. Evaluation scores 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 $A_{v \to n}$, $R_{v \to n}$, $A_{n \to v}$ and $R_{n \to v}$, together with the number of candidates found (n). The last 3 rows show the highest values we were able to reach by using FS's fixedness scores.

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

Lexical fixedness reaches an f-measure of ± 12% (threshold of 3.00). These scores show the best performance that we reached using lexical fixedness. Following FS, we evaluated the syntactic fixedness scores of expressions falling above a frequency cutoff. Since our corpus is much larger than that used by FS, a frequency cutoff of 50 was chosen. The precision, recall and f-measure of the syntactic fixedness measure (shown on table 3) are ± 10%, 41% and 16% respectively, showing worse precision than our method but much better recall and f-measure. As shown by FS, syntactic fixedness performs better than lexical fixedness; $Fixedness_{overall}$ improves on the syntactic fixedness results and also reaches better overall performance than our method.

The compared methods show a different behavior. FS's method favours high recall whereas our method prefers the best trade-off between precision and recall. We wish to highlight that our method reaches better precision than FS's method while handling many low frequency candidates (minimum frequency is 3); this makes our method preferable in some NLP tasks. It is possible that the two methods are capturing different properties of MWEs; in future work, we want to analyse whether the expressions extracted by the two methods differ.

[4] The 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').

parameters					precision	recall	f-measure
$A_{v \to n}$	$R_{v \to n}$	$A_{n \to v}$	$R_{n \to v}$	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
$Fixedness_{syn}(v, n)$ 50				15,630	10.20	40.90	16.33
$Fixedness_{overall}(v, n)$ 50				7819	13.73	27.54	18.33

Table 3: Evaluation results compared to RBN & VLIS

4.2 Qualitative evaluation

Next, 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 MWEs found and the errors made by the algorithm.

First of all, our algorithm is able to filter out grammatical collocations that cause problems in traditional MWE extraction paradigms. An example is given in (5).

(5) 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) 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 between nouns that fill in the object slot is taken into account.

Our quantitative evaluation shows that the algorithm reaches the best results (i.e. the highest f-measures) when using only two parameters ($A_{v \to n}$ and $R_{v \to n}$). Upon closer inspection of the output, we noticed that $A_{n \to v}$ and $R_{n \to v}$ are often able to filter out non-MWEs like the expressions b in (6) and (7).

(6) a. verschijnen op toneel
appear on stage
to appear

b. zingen op toneel
sing on stage
to sing on the stage

(7) a. lig in geheugen
lie in memory
be in memory

b. lig in ziekenhuis
lie in hospital
lie in the hospital

It is only when the two other measures (a unique preference of the noun for the verb) are taken into account that the b expressions are filtered out – either because the noun preference for the verb is very low, or because it 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 (8), where the fixedness of the expression lies not so much in the verb-noun combination, but more in the noun 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.

(8) a. naar school willen
to school want
want to go to school

b. naar huis willen
to home want
want to go home

We will now look at some errors made by our algorithm. 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.

Secondly, 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 noun (*pagina* in the example).

Finally, our hard clustering method does not take polysemous nouns into account. A noun may 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 Conclusions and further work

Our algorithm based on non-compositionality explores a new approach aimed at large-scale MWE extraction. Using only two parameters, $A_{v \to n}$ and $R_{v \to n}$, yields the highest f-measure. Using the two other parameters, $A_{n \to v}$ and $R_{n \to v}$, 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 extraction techniques such as the syntactic fixedness measure proposed by Fazly and Stevenson (2006) 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.

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 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.

One final issue is the realization of a manual evaluation of our semantics-based algorithm, by having human judges decide whether a MWE candidate found by our algorithm is an actual MWE. Our automated evaluation framework is error-prone due to mistakes and incompleteness of our resources. During qualitative evaluation, we found many actual MWEs found by our algorithm, that were not considered correct by our resources (e.g. *[iemand] in de gordijnen jagen* 'to drive s.o. mad', *op het [verkeerde] paard gokken* 'back the wrong horse', *[de kat] uit de boom kijken* 'wait to see which way the wind blows', *uit het [goede] hout gesneden* 'be a trustworthy person'). Conversely, there were also questionable MWE candidates that were described as actual MWEs in our evaluation resources (*val op woensdag* 'fall on a wednesday', *neem als voorzitter* 'take as chairperson', *ruik naar haring* 'smell like herring', *ben voor [. . .] procent* 'to be . . . percent'). A manual evaluation could overcome these difficulties.

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 research program IRME STEVIN project. We would also like to thank Gertjan van Noord and the two anonymous reviewers for their helpful comments on an earlier version of this paper.

References

T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. An Empirical Model of Multiword Expressions Decomposability. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

T. Baldwin. 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, July.

K. Church, W. Gale, P. Hanks, and D. Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-line resources to build a lexicon*, pages 115–164. Lawrence Erlbaum Associates, New Jersey.

A. Fazly and S. Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy.

G. Katz and E. Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using Latent Semantic Analysis. In *Proc. of the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of* COLING/ACL *98*, Montreal, Canada.

D. Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, pages 317–324. University of Maryland.

J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley. University of California Press.

W. Martin and I. Maks, 2005. *Referentie Bestand Nederlands. Documentatie*, April.

D. McCarthy, B. Keller, and J. Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

R.J.F. Ordelman. 2002. Twente Nieuws Corpus (TwNC), August. Parlevink Language Techonology Group. University of Twente.

D. Pearce. 2001. Synonymy in collocation extraction. In *WordNet and Other lexical resources: applications, extensions & customizations (NAACL 2001)*, pages 41–46, Pittsburgh. Carnegie Mellon University.

S. Piao, P. Rayson, O. Mudraya, A. Wilson, and R. Garside. 2006. Measuring mwe compositionality using semantic annotation. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 2–11, Sydney, Australia. Association for Computational Linguistics.

P. Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD Thesis, University of Pennsylvania.

P. Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword Expressions: a pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Mexico City, Mexico.

L. van der Plas and G. Bouma. 2005. Syntactic contexts for finding semantically similar words. *Computational Linguistics in the Netherlands 2004. Selected Papers from the Fifteenth CLIN Meeting*, pages 173–184.

G. van Noord. 2006. At Last Parsing Is Now Operational. In P. Mertens, C. Fairon, A. Dister, and P. Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42, Leuven.

S. Venkatapathy and A. Joshi. 2005. Measuring the relative compositionality of verb-noun collocations by integrating features. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 899–906, Vancouver.

B. Villada Moirón and J. Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context"*, pages 33–40, Trento, Italy.

J. Weeds. 2003. *Measures and Applications of Lexical Distributional Similarity*. PhD Thesis, University of Sussex.

Z. Wu and M. Palmer. 1994. Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133–138, New Mexico State University, Las Cruces, New Mexico.