Measuring MWE Compositionality Using Semantic Annotation

Scott S. L. Piao[1], Paul Rayson[1], Olga Mudraya[2], Andrew Wilson[2] and Roger Garside[1]

[1]Computing Department
Lancaster University
Lancaster, UK
{s.piao, p.rayson, r.garside}@lancaster.ac.uk

[2]Dept. of Linguistics and EL
Lancaster University
Lancaster, UK
{o.mudraya, a.wilson}@lancaster.ac.uk

Abstract

This paper reports on an experiment in which we explore a new approach to the automatic measurement of multi-word expression (MWE) compositionality. We propose an algorithm which ranks MWEs by their compositionality relative to a semantic field taxonomy based on the Lancaster English semantic lexicon (Piao et al., 2005a). The semantic information provided by the lexicon is used for measuring the semantic distance between a MWE and its constituent words. The algorithm is evaluated both on 89 manually ranked MWEs and on McCarthy et al's (2003) manually ranked phrasal verbs. We compared the output of our tool with human judgments using Spearman's rank-order correlation coefficient. Our evaluation shows that the automatic ranking of the majority of our test data (86.52%) has strong to moderate correlation with the manual ranking while wide discrepancy is found for a small number of MWEs. Our algorithm also obtained a correlation of 0.3544 with manual ranking on McCarthy et al's test data, which is comparable or better than most of the measures they tested. This experiment demonstrates that a semantic lexicon can assist in MWE compositionality measurement in addition to statistical algorithms.

1 Introduction

Over the past few years, compositionality and decomposability of MWEs have become important issues in NLP research. Lin (1999) argues that "non-compositional expressions need to be treated differently than other phrases in many statistical or corpus–based NLP methods". Compositionality means that "the meaning of the whole can be strictly predicted from the meaning of the parts" (Manning & Schütze, 2000). On the other hand, decomposability is a metric of the degree to which the meaning of a MWE can be assigned to its parts (Nunberg, 1994; Riehemann, 2001; Sag et al., 2002). These two concepts are closely related. Venkatapathy and Joshi (2005) suggest that "an expression is likely to be relatively more compositional if it is decomposable".

While there exist various definitions for MWEs, they are generally defined as cohesive lexemes that cross word boundaries (Sag et al., 2002; Copestake et al., 2002; Calzolari et al., 2002; Baldwin et al., 2003), which include nominal compounds, phrasal verbs, idioms, collocations etc. Compositionality is a critical criterion cutting across different definitions for extracting and classifying MWEs. While semantics of certain types of MWEs are non-compositional, like idioms "kick the bucket" and "hot dog", some others can have highly compositional semantics like the expressions "traffic light" and "audio tape".

Automatic measurement of MWE compositionality can have a number of applications. One of the often quoted applications is for machine translation (Melamed, 1997; Hwang & Sasaki, 2005), in which non-compositional MWEs need special treatment. For instance, the translation of a highly compositional MWE can possibly be inferred from the translations of its constituent words, whereas it is impossible for non-compositional MWEs, for which we need to identify the translation equivalent for the MWEs as a whole.

In this paper, we explore a new method of automatically estimating the compositionality of MWEs using lexical semantic information, sourced from the Lancaster semantic lexicon (Piao et al., 2005a) that is employed in the USAS[1] tagger (Rayson et al., 2004). This is a

[1] UCREL Semantic Analysis System

Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, pages 2–11,
Sydney, July 2006.

large lexical resource which contains nearly 55,000 single-word entries and over 18,800 MWE entries. In this lexicon, each MWE[2] and the words it contains are mapped to their potential semantic categories using a semantic field taxonomy of 232 categories. An evaluation of lexical coverage on the BNC corpus showed that the lexical coverage of this lexicon reaches 98.49% for modern English (Piao et al., 2004). Such a large-scale semantic lexical resource allows us to examine the semantics of many MWEs and their constituent words conveniently without resorting to large corpus data. Our experiment demonstrates that such a lexical resource provides an additional approach for automatically estimating the compositionality of MWEs.

One may question the necessity of measuring compositionality of manually selected MWEs. The truth is, even if the semantic lexicon under consideration was compiled manually, it does not exclusively consist of non-compositional MWEs like idioms. Built for practical discourse analysis, it contains many MWEs which are highly compositional but depict certain entities or semantic concepts. This research forms part of a larger effort to extend lexical resources for semantic tagging. Techniques are described elsewhere (e.g. Piao et al., 2005b) for finding new candidate MWE from corpora. The next stage of the work is to semi-automatically classify these candidates using an existing semantic field taxonomy and, to assist this task, we need to investigate patterns of compositionality.

2 Related Work

In recent years, various approaches have been proposed to the analysis of MWE compositionality. Many of the suggested approaches employ statistical algorithms.

One of the earliest studies in this area was reported by Lin (1999) who assumes that "non-compositional phrases have a significantly different mutual information value than the phrases that are similar to their literal meanings" and proposed to identify non-compositional MWEs in a corpus based on distributional characteristics of MWEs. Bannard et al. (2003) tested techniques using statistical models to infer the meaning of verb-particle constructions (VPCs), focusing on prepositional particles. They tested four methods over four compositional classification tasks, reporting that, on all tasks, at least one of the four methods offers an improvement in precision over the baseline they used.

McCarthy et al. (2003) suggested that compositional phrasal verbs should have similar neighbours as for their simplex verbs. They tested various measures using the nearest neighbours of phrasal verbs and their simplex counterparts, and reported that some of the measures produced results which show significant correlation with human judgments. Baldwin et al. (2003) proposed a LSA-based model for measuring the decomposability of MWEs by examining the similarity between them and their constituent words, with higher similarity indicating the greater decomposability. They evaluated their model on English noun-noun compounds and verb-particles by examining the correlation of the results with similarities and hyponymy values in WordNet. They reported that the LSA technique performs better on the low-frequency items than on more frequent items. Venkatapathy and Joshi (2005) measured relative compositionality of collocations having verb-noun pattern using a SVM (Support Vector Machine) based ranking function. They integrated seven various collocational and contextual features using their ranking function, and evaluated it against manually ranked test data. They reported that the SVM based method produces significantly better results compared to methods based on individual features.

The approaches mentioned above invariably depend on a variety of statistical contextual information extracted from large corpus data. Inevitably, such statistical information can be affected by various uncontrollable "noise", and hence there is a limitation to purely statistical approaches.

In this paper, we contend that a manually compiled semantic lexical resource can have an important part to play in measuring the compositionality of MWEs. While any approach based on a specific lexical resource may lack generality, it can complement purely statistical approaches by importing human expert knowledge into the process. Particularly, if such a resource has a high lexical coverage, which is true in our case, it becomes much more useful for dealing with general English. It should be emphasized that we propose our semantic lexical-based approach not as a substitute for the statistical approaches.

[2] In this lexicon, many MWEs are encoded as templates, such as *driv*_* {Np/P*/J*/R*} mad_JJ*, which represent variational forms of a single MWE, For further details, see Rayson et al., 2004.

Rather we propose it as a potential complement to them.

In the following sections, we describe our experiment and explore this approach to the issue of automatic estimation of MWE compositionality.

3 Measuring MWE compositionality with semantic field information

In this section, we propose an algorithm for automatically measuring MWE compositionality based on the Lancaster semantic lexicon. In this lexicon, the semantic field of each word and MWE is encoded in the form of semantic tags. We contend that the compositionality of a MWE can be estimated by measuring the distance between semantic fields of an MWE and its constituent words based on the semantic field information available from the lexicon.

The lexicon employs a taxonomy containing 21 major semantic fields which are further divided into 232 sub-categories.[3] Tags are designed to denote the semantic fields using letters and digits. For instance, tag *N3.2* denotes the category of {*SIZE*} and Q4.1 denotes {*media: Newspapers*}. Each entry in the lexicon maps a word or MWE to its potential semantic field category/ies. More often than not, a lexical item is mapped to multiple semantic categories, reflecting its potential multiple senses. In such cases, the tags are arranged by the order of likelihood of meanings, with the most prominent one at the head of the list. For example, the word "mass" is mapped to tags *N5*, *N3.5*, *S9*, *S5* and *B2*, which denote its potential semantic fields of {*QUANTITIES*}, {*MEASUREMENT: WEIGHT*}, {*RELIGION AND SUPERNATURAL*}, {*GROUPS AND AFFILIATION*} and {*HEALTH AND DISEASE*}.

The lexicon provides direct access to the semantic field information for large number of MWEs and their constituent words. Furthermore, the lexicon was analysed and classified manually by a team of linguists based on the analysis of corpus data and consultation of printed and electronic corpus-based dictionaries, ensuring a high level of consistency and accuracy of the semantic analysis.

In our context, we interpret the task of measuring the compositionality of MWEs as examining the distance between the semantic tag of a MWE and the semantic tags of its constituent words. Given a MWE M and its constituent words w_i (i = 1, .., n), the compositionality D can be measured by multiplying the semantic distance SD between M and each of its constituent words w_i. In practice, the square root of the product is used as the score in order to reduce the range of actual D-scores, as shown below:

$$(1) \quad D(M) = \sqrt{\prod_{i=1}^{n} SD(M, w_i)}$$

where D-score ranges between [0, 1], with 1 indicating the strongest compositionality and 0 the weakest compositionality.

In the semantic lexicon, as the semantic information of function words is limited, they are classified into a single grammatical bin (denoted by tag Z5). In our algorithm, they are excluded from the measuring process by using a stop word list. Therefore, only the content constituent words are involved in measuring the compositionality. Although function words may form an important part of many MWEs, such as phrasal verbs, because our algorithm solely relies on semantic field information, we assume they can be ignored.

The semantic distance between a MWE and any of its constituent words is calculated by quantifying the similarity between their semantic field categories. In detail, if the MWE and a constituent word do not share any of the major 21 semantic domains, the SD is assigned a small value λ.[4] If they do, three possible cases are considered:

Case a. If they share the same tag, and the constituent word has only one tag, then SD is one.

Case b. If they share a tag or tags, but the constituent words have multiple candidate tags, then SD is weighted using a variable α based on the position of the matched tag in the candidate list as well as the number of candidate tags.

Case c. If they share a major category, but their tags fall into different sub-categories (denoted by the trailing digits following a letter), SD is further weighted using a

[3] For the complete semantic tagset, see website: http://www.comp.lancs.ac.uk/ucrel/usas/

[4] We avoid using zero here in order to avoid producing semantic distance of zero indiscriminately when any one of the constituent words produces zero distance regardless of other constituent words.

variable β which reflects the difference of the sub-categories.

With respect to weight α, suppose L is the number of candidate tags of the constituent word under consideration, N is the position of the specific tag in the candidate list (the position starts from the top with N=1), then the weight α is calculated as

$$(2)\quad \alpha = \frac{L - N + 1}{L^2},$$

where N=1, 2, …, n and N<=L. Ranging between [1, 0), α takes into account both the location of the matched tag in the candidate tag list and the number of candidate tags. This weight penalises the words having more candidate semantic tags by giving a lower value for their higher degree of ambiguity. As either L or N increases, the α-value decreases.

Regarding the case c), where the tags share the same head letter but different digit codes, i.e. they are from the same major category but in different sub-categories, the weight β is calculated based on the number of sub-categories they share. As we mentioned earlier, a semantic tag consists of an initial letter and some trailing digits divided by points, e.g. *S1.1.2* {*RECIPROCITY*}, *S1.1.3* {*PARTICIPATION*}, *S1.1.4* {*DESERVE*} etc. If we let T_1, T_2 be a pair of semantic tags with the same initial letters, which have k_i and k_j trailing digit codes (denoting the number of sub-division layers) respectively and share n digit codes from the left, or from the top layer, then β is calculated as follows:

$$(3)\quad \beta = \frac{n}{k};$$

$$(4)\quad k = \max(k_i, k_j).$$

where β ranges between (0, 1). In fact, the current USAS taxonomy allows only the maximum three layers of sub-division, therefore β has one of three possible scores: 0.500 (1/2), 0.333 (1/3) and 0.666 (2/3). In order to avoid producing zero scores, if the pair of tags do not share any digit codes except the head letter, then n is given a small value of 0.5.

Combining all of the weighting scores, the semantic distance *SD* in formula (1) is calculated as follows:

$$(5)\quad SD(M, w_i) = \begin{cases} \text{if no tag matches, then } \lambda; \\ \text{if a), then } 1; \\ \text{if b), then } \prod_{i=1}^{n} \alpha_i; \\ \text{if c), then } \prod_{i=1}^{n} \alpha_i \beta_i. \end{cases}$$

where λ is given a small value of 0.001 for our experiment[5].

Some MWEs and single words in the lexicon are assigned with combined semantic categories which are considered to be inseparable, as shown below:

petrol_NN1 station_NN1 *M3/H1*

where the slash means that this MWE falls under the categories of M3 {*VEHICLES AND TRANSPORTS ON LAND*} and H1 {*ARCHITECTURE AND KINDS OF HOUSES AND BUILDINGS*} at the same time. For such cases, criss-cross comparisons between all possible tag pairs are carried out in order to find the optimal match between the tags of the MWE and its constituent words.

By way of further explanation, the word "brush" as a verb has candidate semantic tags of *B4* {*CLEANING AND PERSONAL CARE*} and *A1.1.1* {*GENERAL ACTION, MAKING*} etc. On the other hand, the phrasal verb "brush down" may fall under either B4 category with the sense of *cleaning* or G2.2 category {*ETHICS*} with the sense of *reprimand*. When we apply our algorithm to it, we get the *D*-score of 1.0000 for the sense of *cleaning*, indicating a high degree of compositionality, whereas we get a low *D*-score of 0.0032 for the sense of *reprimand*, indicating a low degree of compositionality. Note that the word "down" in this MWE is filtered out as it is a functional word.

The above example has only a single constituent content word. In practice, many MWEs have more complex structures than this example. In order to test the performance of our algorithm, we compared its output against human judgments of compositionality, as reported in the following section.

4 Manually Ranking MWEs for Evaluation

In order to evaluate the performance of our tool against human judgment, we prepared a list

[5] As long as λ is small enough, it does not affect the ranking of *D*-scores.

of 89 MWEs[6] and asked human raters to rank them via a website. The list includes six MWEs with multiple senses, and these were treated as separate MWE. The Lancaster MWE lexicon has been compiled manually by expert linguists, therefore we assume that every item in this lexicon is a true MWE, although we acknowledge that some errors may exist.

Following McCarthy et al.'s approach, we asked the human raters to assign each MWE a number ranging between 0 (opaque) and 10 (fully compositional). Both native and non-native speakers are involved, but only the data from native speakers are used in this evaluation. As a result, three groups of raters were involved in the experiment. Group 1 (6 people) rated MWEs with indexes of 1-30, Group 2 (4 people) rated MWEs with indexes of 31-59 and Group 3 (five people) rated MWEs with indexes of 6-89.

In order to test the level of agreement between the raters, we used the procedures provided in the 'irr' package for R (Gamer, 2005). With this tool, the average intraclass correlation coefficient (ICC) was calculated for each group of raters using a two-way agreement model (Shrout & Fleiss, 1979). As a result, all ICCs exceeded 0.7 and were significant at the 95% confidence level, indicating an acceptable level of agreement between raters. For Group 1, the ICC was 0.894 (95% ci = 0.807 < ICC < 0.948), for Group 2 it was 0.9 (95% ci=0.783<ICC<0.956) and for Group 3 it was 0.886 (95% ci = 0.762 < ICC < 0.948).

Based on this test, we conclude that the manual ranking of the MWEs is reliable and is suitable to be used in our evaluation. Source data for the human judgements is available from our website in spreadsheet form[7].

5 Evaluation

In our evaluation, we focused on testing the performance of the *D*-score against human raters' judgment on ranking different MWEs by their degree of compositionality, as well as distinguishing the different degrees of compositionality for each sense in the case of multiple tags.

The first step of the evaluation was to implement the algorithm in a program and run the tool on the 89 test MWEs we prepared. Fig. 1 illustrates the *D*-score distribution in a bar chart. As shown by the chart, the algorithm produces a widely dispersed distribution of *D*-scores across the sample MWEs, ranging from 0.000032 to 1.000000. For example, the tool assigned the score of 1.0 to the *FOOD* sense and 0.001 to the THIEF senses of "tea leaf" successfully distinguishing the different degrees of compositionality of these two senses.

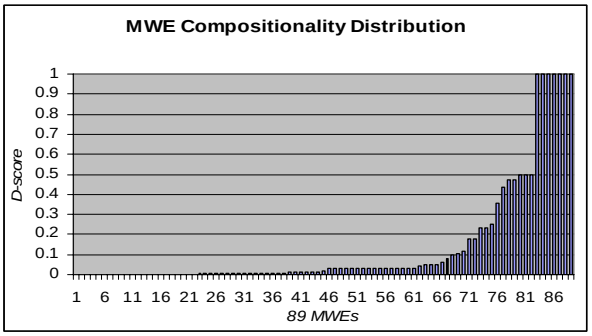

Fig 1: *D*-score distribution across 89 sample MWEs

As shown in Fig. 1, some MWEs share the same scores, reflecting the limitation of the number of ranks that our algorithm can produce as well as the limited amount of semantic information available from a lexicon. Nonetheless, the algorithm produced 45 different scores which ranked the MWEs into 45 groups (see the steps in the figure). Compared to the eleven scores used by the human raters, this provides a fine-grained ranking of the compositionality.

The primary issue in our evaluation is the extent to which the automatic ranking of the MWEs correlates with the manual ranking of them. As described in the previous section, we created a list of 89 manually ranked MWEs for this purpose. Since we are mainly interested in the ranks rather than the actual scores, we examined the correlation between the automatic and manual rankings using Spearman's correlation coefficient. (For the full ranking list, see Appendix).

In the manually created list, each MWE was ranked by 3-6 human raters. In order to create a unified single test data of human ranking, we calculated the average of the human ranks for each MWE. For example, if two human raters give ranks 3 and 4 to a MWE, then its rank is (3+4)/2=3.5. Next, the MWEs are sorted by the averaged ranks in descending order to obtain the combined ranks of the MWEs. Finally, we sorted the MWEs by the *D*-score in the same way to obtain a parallel list of automatic ranks. For the calculation of Spearman's correlation coefficient, if *n* MWEs are tied to a score (either *D*-score or the average manual ranks), their ranks were ad-

[6] Selected at random from the Lancaster semantic lexicon.
[7] http://ucrel.lancs.ac.uk/projects/assist/

justed by dividing the sum of their ranks by the number of MWEs involved. Fig. 2 illustrates the correspondence between the adjusted automatic and manual rankings.

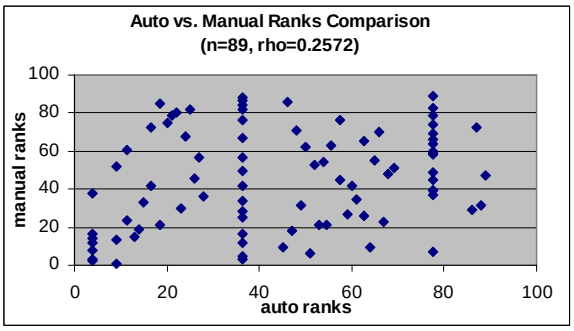

Fig. 2: Scatterplot of automatic *vs.* manual ranking.

As shown in Fig. 2, the overall correlation seems quite weak. In the automatic ranking, quite a few MWEs are tied up to three ranks, illustrated by the vertically aligned points. The precise correlation between the automatic and manual rankings was calculated using the function provided in R for Windows 2.2.1. Spearman's rank correlation (*rho*) for these data was 0.2572 (p=0.01495), indicating a significant though rather weak positive relationship.

In order to find the factors causing this weak correlation, we tested the correlation for those MWEs whose rank differences were less than 20, 30, 40 and 50 respectively. We are interested to find out how many of them fall under each of the categories and which of their features affected the performance of the algorithm. As a result, we found 43, 54, 66 and 77 MWEs fall under these categories respectively, which yield different correlation scores, as shown in Table 1.

numb of MWEs	Percent (%)	Rank diff	*rho*-score	Sig.
43	48.31	<20	0.9149	*P<0.001*
54	60.67	<30	0.8321	*P<0.001*
66	74.16	<40	0.7016	*P<0.001*
77	86.52	<50	0.5084	*P<0.001*
89 (total)	100.00	<=73	0.2572	*P<0.02*

Table 1: Correlation coefficients corresponding different rank differences.

As we expected, the *rho* decreases as the rank difference increases, but all of the four categories containing a total of 77 MWEs (86.52%) show reasonably high correlations, with the minimum score of 0.5084. [8] In particular, 66 of them (74.16%), whose ranking differences are less than 40, demonstrate a strong correlation with rho-score 0.7016, as illustrated by Fig. 3

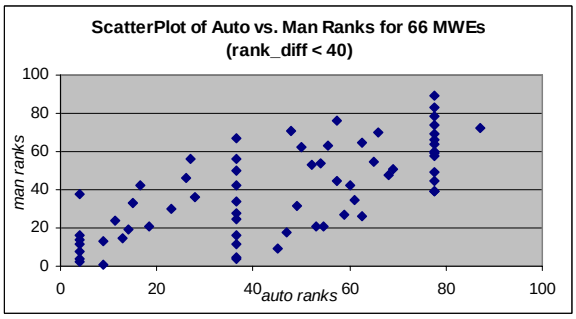

Fig 3: ScatterPlot for 66 MWEs (rank_diff < 40) which shows a strong correlation

Our manual examination shows that the algorithm generally pushes the highly compositional and non-compositional MWEs towards opposite ends of the spectrum of the *D*-score. For example, those assigned with score 1 include "aid worker", "audio tape" and "unemployment figure". On the other hand, MWEs such as "tea leaf" (meaning thief), "kick the bucket" and "hot dog" are given a low score of 0.001. We assume these two groups of MWEs are generally treated as highly compositional and opaque MWEs respectively.

However, the algorithm could be improved. A major problem found is that the algorithm punishes longer MWEs which contain function words. For example, "make an appearance" is scored 0.000114 by the algorithm, but when the article "an" is removed, it gets a higher score 0.003608. Similarly, when the preposition "up" is removed from "keep up appearances", it gets 0.014907 compared to the original 0.000471, which would push up their rank much higher. To address this problem, the algorithm needs to be refined to minimise the impact of the function words to the scoring process.

Our analysis also reveals that 12 MWEs with rank differences (between automatic and manual ranking) greater than 50 results in a degraded overall correlation. Table 2 lists these words, in which the higher ranks indicate higher compositionality.

[8] Salkind (2004: 88) suggests that *r*-score ranges 0.4~0.6, 0.6~0.8 and 0.8~1.0 indicate moderate, strong and very strong correlations respectively.

MWE	Sem. Tag[9]	Auto rank	Manual rank
plough into	A9-	53.5	3
Bloody Mary	F2	53.5	2
pillow fight	K6	26	80.5
lollipop lady	M3/S2	70	15
cradle snatcher	S3.2/T3/S2	73.5	17.5
go bananas	X5.2+++	65	8.5
make an appearance	S1.1.3+	2	58.5
keep up appearances	A8/S1.1.1	4	61
sandwich course	P1	69	11.5
go bananas	B2-/X1	68	10
Eskimo roll	M4	71.5	5
in other words	Z4	12.5	83

Table 2: Twelve MWEs having rank differences greater than 50.

Let us take "pillow fight" as an example. The whole expression is given the semantic tag K6, whereas neither "pillow" nor "fight" as individual word is given this tag. In the lexicon, "pillow" is classified as H5 {*FURNITURE AND HOUSEHOLD FITTINGS*} and "fight" is assigned to four semantic categories including S8- {*HINDERING*}, X8+ {*HELPING*}, E3- {*VIOLENT/ANGRY*}, and K5.1 {*SPORTS*}. For this reason, the automatic score of this MWE is as low as 0.003953 on the scale of [0, 1]. On the contrary, human raters judged the meaning of this expression to be fairly transparent, giving it a high score of 8.5 on the scale of [0, 10]. Similar contrasts occurred with the majority of the MWEs with rank differences greater than 50, which are responsible for weakening the overall correlation.

Another interesting case we noticed is the MWE "pass away". This MWE has two major senses in the semantic lexicon L1- {*DIE*} and T2- {*END*} which were ranked separately. Remarkably, they were ranked in the opposite order by human raters and the algorithm. Human raters felt that the sense *DIE* is less idiomatic, or more compositional, than *END*, while the algorithm indicated otherwise. The explanation of this again lies in the semantic classification of the lexicon, where "pass" as a single word contains the sense T2- but not L1-. Consequently, the automatic score for "pass away" with the sense L1- is much lower (0.001) than that with the sense of T2- (0.007071).

In order to evaluate our algorithm in comparison with previous work, we also tested it on the manual ranking list created by McCarthy et al (2003).[10] We found that 79 of the 116 phrasal verbs in that list are included in the Lancaster semantic lexicon. We applied our algorithm on those 79 items to compare the automatic ranks against the average manual ranks using the Spearman's rank correlation coefficient (*rho*). As a result, we obtained *rho*=0.3544 with significance level of *p*=0.001357. This result is comparable with or better than most measures reported by McCarthy et al (2003).

6 Discussion

The algorithm we propose in this paper is different from previous proposed statistical methods in that it employs a semantic lexical resource in which the semantic field information is directly accessible for both MWEs and their constituent words. Often, typical statistical algorithms measure the semantic distance between MWEs and their constituent words by comparing their contexts comprising co-occurrence words in near context extracted from large corpora, such as Baldwin et al's algorithm (2003).

When we consider the definition of the compositionality as the extent to which the meaning of the MWE can be guessed based on that of its constituent words, a semantic lexical resource which maps MWEs and words to their semantic features provides a practical way of measuring the MWE compositionality. The Lancaster semantic lexicon is one such lexical resource which allows us to have direct access to semantic field information of large number of MWE and single words. Our experiment demonstrates the potential value of such semantic lexical resources for the automatic measurement of MWE compositionality. Compared to statistical algorithms which can be affected by a variety of uncontrollable factors, such as size and domain of corpora, etc., an expert-compiled semantic lexical resource can provide much more reliable and "clean" lexical semantic information.

However, we do not suggest that algorithms based on semantic lexical resources can substitute corpus-based statistical algorithms. Rather, we suggest it as a complement to existing statistical algorithms. As the errors of our algorithm

[9] Semantic tags occurring in Table 2: A8 (seem), A9 (giving possession), B2 (health and disease), F2 (drink), K6 (children's games and toys), M3 (land transport), M4 (swimming), P1 (education), S1.1.1 (social actions), S1.1.3 (participation), S2 (people), S3.2 (relationship), T3 (time: age), X1 (psychological actions), X5.2 (excited), Z4 (discourse bin)

[10] This list is available at website: http://mwe.stanford.edu/resources/

reveal, the semantic information provided by the lexicon alone may not be rich enough for a very fine-grained distinction of MWE compositionality. In order to obtain better results, this algorithm needs to be combined with statistical techniques.

A limitation of our approach is language-dependency. In order to port our algorithm to languages other than English, one needs to build similar semantic lexicon in those languages. However, similar semantic lexical resources are already under construction for some other languages, including Finnish and Russian (Löfberg et al., 2005; Sharoff et al., 2006), which will allow us to port our algorithm to those languages.

7 Conclusion

In this paper, we explored an algorithm based on a semantic lexicon for automatically measuring the compositionality of MWEs. In our evaluation, the output of this algorithm showed moderate correlation with a manual ranking. We claim that semantic lexical resources provide another approach for automatically measuring MWE compositionality in addition to the existing statistical algorithms. Although our results are not yet conclusive due to the moderate scale of the test data, our evaluation demonstrates the potential of lexicon-based approaches for the task of compositional analysis. We foresee, by combining our approach with statistical algorithms, that further improvement can be expected.

8 Acknowledgement

The work reported in this paper was carried out within the UK-EPSRC-funded ASSIST Project (Ref. EP/C004574).

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An Empirical Model of Multiword Expression Compositionality. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89-96, Sapporo, Japan.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proc. of the ACL2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo.

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proc. of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands.

Ann Copestake, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan A. Sag, and Dan Flickinger. 2002. Multiword expressions: Linguistic precision and reusability. In *Proc. of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1941–1947, Las Palmas, Canary Islands.

Matthias Gamer. 2005. The irr Package: Various Coefficients of Interrater Reliability and Agreement. Version 0.61 of 11 October 2005. Available from: cran.r-project.org/src/contrib/Descriptions/irr.html

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proc. of the 37th Annual Meeting of the ACL*, pages 317–324, College Park, USA.

Laura Löfberg, Scott Piao, Paul Rayson, Jukka-Pekka Juntunen, Asko Nykänen, and Krista Varantola. 2005. A semantic tagger for the Finnish language. In *Proc. of the Corpus Linguistics 2005 conference*, Birmingham, UK.

Christopher D. Manning and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan.

Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proc. of the 2nd Conference on Empirical Methods in Natural Language Processing* , Providence, USA.

Geoffrey Nunberg, Ivan A. Sag, and Tom Wasow. 1994. Idioms. *Language*, 70: 491–538.

Scott S.L. Piao, Paul Rayson, Dawn Archer and Tony McEnery. 2004. Evaluating Lexical Resources for a Semantic Tagger. In *Proc. of LREC-04*, pages 499–502, Lisbon, Portugal.

Scott S.L. Piao, Dawn Archer, Olga Mudraya, Paul Rayson, Roger Garside, Tony McEnery and Andrew Wilson. 2005a. A Large Semantic Lexicon for Corpus Annotation. In *Proc. of the Corpus Linguistics Conference 2005*, Birmingham, UK.

Scott S.L. Piao., Paul Rayson, Dawn Archer, Tony McEnery. 2005b. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language*, 19, 4: 378–397.

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL Semantic Analysis System. In *Proc. of LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks*, pages 7–12, Lisbon, Portugal.

Susanne Riehemann. 2001. *A Constructional Approach to Idioms and Word Formation*. Ph.D. thesis, Stanford University, Stanford.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Neil J. Salkind. 2004. Statistics for People Who Hate Statistics. Sage: Thousand Oakes, US.

Serge Sharoff, Bogdan Babych, Paul Rayson, Olga Mudraya and Scott Piao. 2006. ASSIST: Automated semantic assistance for translators. Proceedings of EACL 2006, pages 139–142, Trento, Italy.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin (2), 420–428.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 899–906, Vancouver, Canada.

Appendix: Manual *vs.* Automatic Ranks of Sample MWEs

The table below shows the human and automatic rankings of 89 sample MWEs. The MWEs are sorted in ascending order by manual average ranks. The top items are supposed to be the most compositional ones. For example, according to the manual ranking, facial expression is the most compositional MWE while tea leaf is the most opaque one. This table also shows that some MWEs are tied up with the same ranks. For the definitions of the full semantic tagset, see website http://www.comp.lancs.ac.uk/ucrel/usas/.

MWE Tag	Sem tag	Man rank	Auto. rank
facial expression	B1	1	9
aid worker	S8/S2	2	4
audio tape	K3	3.5	4
leisure activities	K1	3.5	36.5
advance warning	T4/Q2.2	5	36.5
living space	H2	6	51
in other words	Z4	7	77.5
unemployment figures	I3.1/N5	8	4
camera angle	Q4.3	9.5	45
pillow fight	K6	9.5	64
youth club	S5/T3	11.5	4
petrol station	M3/H1	11.5	36.5
palm tree	L3	13	9
rule book	G2.1/Q4.1	14	4
ball boy	K5.1/S2.2	15	13
goal keeper	K5.1/S2	16.5	4
kick in	E3-	16.5	36.5
ventilation shaft	H2	18	47
directory enquiries	Q1.3	19	14
phone box	Q1.3/H1	21	18.5
lose balance	M1	21	53
bend the rules	A1.7	21	54.5
big nose	X7/X2.4	23	67
quantity control	N5/A1.7	24	11.5
act of God	S9	25	36.5
air bag	A15/M3	26	62.5
mind stretching	A12	27	59
plain clothes	B5	28	36.5
keep up appearances	A8/S1.1.1	29	86
examining board	P1	30	23
open mind	X6	31.5	49
make an appearance	S1.1.3+	31.5	88
cable television	Q4.3	33	15
king size	N3.2	34	36.5
action point	X7	35	61
keep tight rein on	A1.7	36	28
noughts and crosses	K5.2	37	77.5
tea leaf	L3/F2	38	4
single minded	X5.1	39.5	77.5
window dressing	I2.2	39.5	77.5
street girl	G1.2/S5	42	36.5
just over the horizon	S3.2/S2.1	42	60
pressure group	T1.1.3	42	16.5
air proof	O4.1	44.5	57.5
heart of gold	S1.2.2	44.5	77.5
lose heart	X5.2	46	26
food for thought	X2.1/X5.1	47	89
play part	S8	48	68
look down on	S1.2.3	49	77.5
arm twisting	Q2.2	50	36.5
take into account	A1.8	51	69
kidney bean	F1	52	9
come alive	A3+	53	52
break new ground	T3/T2	54	54
make up to	S1.1.2	55	65
by virtue of	C1	56.5	36.5
snap shot	A2.2	56.5	27
pass away	L1-	58	77.5
long face	E4.1	59	77.5
bossy boots	S1.2.3/S2	60	77.5
plough into	M1/A1.1.2	61	11.5
kick in	T2+	62	50
animal magnetism	S1.2	63	55.5
sixth former	P1/S2	64	77.5
pull the strings	S7.1	65	62.5
couch potato	A1.1.1/S2	66	77.5
think tank	S5/X2.1	67	36.5
come alive	X5.2+	68	24
hot dog	F1	69	77.5
cheap shot	G2.2-/Q2.2	70	66

rock and roll	K2	71	48
bright as a button	S3.2/T3/S2	72.5	87
cradle snatcher	X9.1+	72.5	16.5
alpha wave	B1	74	77.5
lollipop lady	M3/S2	75	20
pass away	X5.2+	76.5	57.5
plough into	T2-	76.5	36.5
piece of cake	P1	78.5	77.5
sandwich course	A12	78.5	21
go bananas	B2-/X1	80	22
go bananas	X5.2+++	81.5	36.5
go bananas	E3-	81.5	25
kick the bucket	L1	83	77.5
on the wagon	F2	84	36.5
Eskimo roll	M4	85	18.5
acid house	K2	86	46
plough into	A9-	87	36.5
Bloody Mary	F2	88	36.5
tea leaf	G2.1-/S2mf	89	77.5