Lexical Substitution as a Framework for Multiword Evaluation

Diana McCarthy

Department of Informatics,
University of Sussex,BN1 9QJ
dianam@sussex.ac.uk

Abstract

In this paper we analyse data from the SemEval lexical substitution task in those cases where the annotators indicated that the target word was part of a phrase before substituting the target with a synonym. We classify the types of phrases that were provided in this way by the annotators in order to evaluate the utility of the method as a means of producing a gold-standard for multiword evaluation. Multiword evaluation is a difficult area because lexical resources are not complete and people's judgments on multiwords vary. Whilst we do not believe lexical substitution is necessarily a panacea for multiword evaluation, we do believe it is a useful methodology because the annotator is focused on the task of substitution. Following the analysis, we make some recommendations which would make the data easier to classify.

1. Introduction

There is a growing interest in "multiwords" in the computational linguistics community owing to their common occurrence in everyday language and the problems that they cause automatic systems. The definition of multiwords provided by Sag et al. (2002) "idiosyncratic interpretations that cross word boundaries (or spaces)" is useful, though it is acknowledged to be a rough guide rather than a precise definition because of the great variety of phenomena encompassed by the term multiword. There are a large number of approaches that aim to detect multiwords automatically, using statistics or linguistics or a mixture of the two, however an outstanding issue is evaluation methodology (Grégoire et al., 2008). Previous work has relied on i) manual scrutiny of the lists output from systems (Lin, 1999; Krenn and Evert, 2001; Blaheta and Johnson, 2001; Piao et al., 2003), ii) comparison with predefined lexical resources (Baldwin and Villavicencio, 2002; Fazly and Stevenson, 2006) and also iii) human judgments of compositionality (Bannard et al., 2003; McCarthy et al., 2003; Venkatapathy and Joshi, 2005). Most of these approaches for evaluation produce useful results by specifically targeting a particular sub-type of multiword expressions, such as verb-particles or verb-objects, where it is easier for humans to make manual judgments on the given type of expression. There are however residual issues which researchers acknowledge because predefined resources are incomplete and manual judgments show low agreement because the notion of multiword is not clear cut. Furthermore, non-compositionality is only one aspect of multiwords since there are non-productive yet compositional phrases for example *frying pan* (Bannard et al., 2003).

This article explores the use of substitution as a methodology for creating multiword data. To do this we examine the dataset created for the English Lexical Substitution task in SemEval (McCarthy and Navigli, 2007) (hereafter referred to as LEXSUB). In this paper, we examine the subset of the LEXSUB dataset where annotators identified that the target was an integral part of a phrase. We wish to see how well suited the annotations are as a gold standard for multiword evaluation.

The LEXSUB task involved a team of 5 annotators who provided substitutes (near synonyms or paraphrases) for target words in sentences. This evaluated the performance of systems on the hybrid task of finding good synonyms, and determining the right meaning (and therefore choice of synonym) in the right context. Multiwords have pretty much been ignored in the tasks from SemEval and its predecessors: SENSEVAL (Kilgarriff and others, 1998), SENSEVAL-2 (Cotton et al., 2001) and SENSEVAL-3 (Mihalcea and Edmonds, 2004). In the WSD tasks, multiwords have usually been manually marked up [1] and , in the event of more than one entry of the same multiword in the same dictionary, systems have then performed WSD just as they do for regular words, without any need to identify multiwords. In the lexical substitution task in SemEval, the annotators had to identify sentences where the target word was "an integral part of a phrase" and what that phrase was. The task was specifically designed in this way to cater for multiwords, indeed multiword detection and identification were evaluated as a

[1] Even in the first SENSEVAL where they were not manually annotated, the list of multiwords was predetermined by the dictionary and identification was a relatively trivial enterprise.

subtask though the main focus of LEXSUB was examining the synonym identification and disambiguation capabilities of systems.

In this paper, we present a classification of the LEXSUB annotations where the target was determined by the annotators to be part of a phrase. The classification was devised to distinguish whether the annotations are due to syntactic paraphrasing necessitated by the act of substitution or whether they are due to collocational, syntactic or semantic idiosyncrasies that are characteristic of multiword expressions. The author manually analysed the annotations according to the various categories of this classification and we examine the results of the classification.

The paper is structured as follows. In the next section we describe the LEXSUB task with particular attention to the annotations which involved putative multiwords. In section 3. we describe the classification that we are presenting here and the process of its application. We give the results of the analysis in section 4. followed by a discussion in section 5. In the discussion, we examine the merits and issues with using substitution to create a multiword resource given the methodology of consensus adopted for LEXSUB and we make some future recommendations. We discuss ways in which the various distinctions in our classification might be amenable to automatic detection and we briefly compare the contents of the LEXSUB multiword resource (hereafter LEXSUBMW) with WordNet given that this was used for a baseline system in the task. [2] We conclude in section 6.

2. The Lexical Substitution Task

The LEXSUB task was run as one of 19 semantic evaluation tasks at SemEval 2007 (Agirre et al., 2007). For the LEXSUB task, 2010 sentences were extracted from the Internet Corpus of English (Sharoff, 2006) for a set of 201 target words (nouns, verbs, adjectives and adverbs). Both manual and automatic methods were used for selecting both the words and selecting the sentences (see (McCarthy and Navigli, 2007) for further details). The 5 Annotators were all native English speakers living in the UK; 3 had a linguistics background whilst 2 did not. Each "item" is a target occurrence of a word in a sentence. In the following example item for the target word **post**:

*However, both **posts** include a one-year hand over period and consequently the elections need to be held one year in advance of the end of their terms.*

3 annotators provided *position* as the substitute, 2 provided *job* and 1 provided *role*. [3]

For each item, the annotators are also asked to fill in a box labelled "target word is part of a phrase" (hereafter TWPP) if the word is considered an integral part of a phrase and provide the phrase in that box. The items where the annotators gave such responses is the subset of the LEXSUB dataset that we analyse in this paper. The annotators were given guidelines [4] and were advised as follows:

> If you think the word is actually an integral part of a phrase which appears within the sentence please indicate this in your response by entering the phrase in the box marked *Target word is part of phrase:* and then supplying your substitute in the usual response box.

Examples shown in figure 1 were given and then the annotators were advised that the phrase may appear with intervening words and given the additional example in figure 2.

No formal definition of "multiword" was given. It was anticipated that the annotators could use this box when it is not easy to provide a substitute for the target without considering the phrase because either i) the phrase is lexicalised and the meaning of the target is specialised in the phrase because of this or ii) it is not possible to substitute the word because of syntactic constraints, so the phrase must be considered and a substitute supplied to replace the phrase.

Whilst the second category can cover paraphrases that would not generally be considered multiwords, it will also cover syntactic idiosyncrasies such as the appropriate use of a particle or preposition with a verb. Verb particle constructions and prepositional verbs are usually regarded as multiwords due to some level of syntactic and possibly semantic idiosyncrasy. Furthermore, syntactic information on verb particle constructions is a necessary component of NLP lexicons for both syntactic analysis and generation capabilities.

The TWPP data were converted into a multiword gold-standard (LEXSUBMW) for two reasons. Firstly for a multiword detection and identification subtask of the LEXSUB task and secondly for analysis of system performance on the LEXSUB substitution tasks with and

[2] We use WordNet version 2.1 as this was used for the LEXSUB baseline system.

[3] Note that the annotators were allowed to provide up to 3 substitutes provided that they felt they were all equally as good. Note also that the annotators responses were semi-automatically lemmatised.

[4] The full set of annotator guidelines are available at http://www.informatics.susx.ac.uk/research/nlp/mccarthy/files/instructions.pdf.

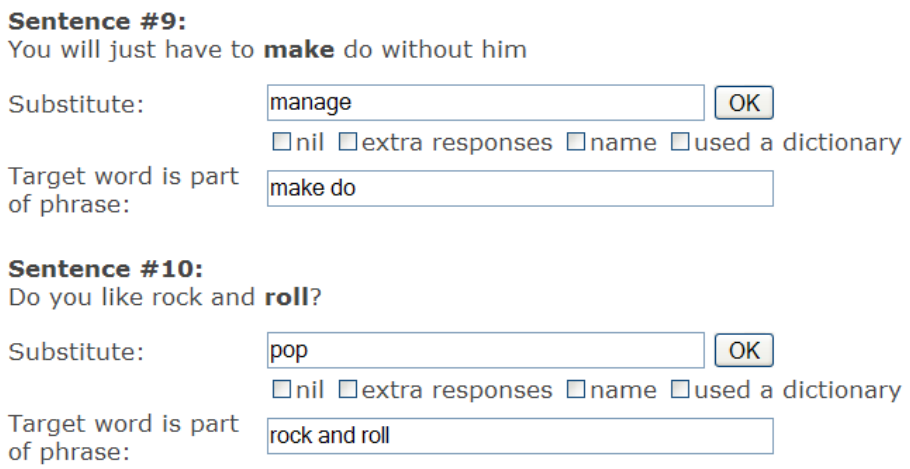

Figure 1: Examples for TWPP box

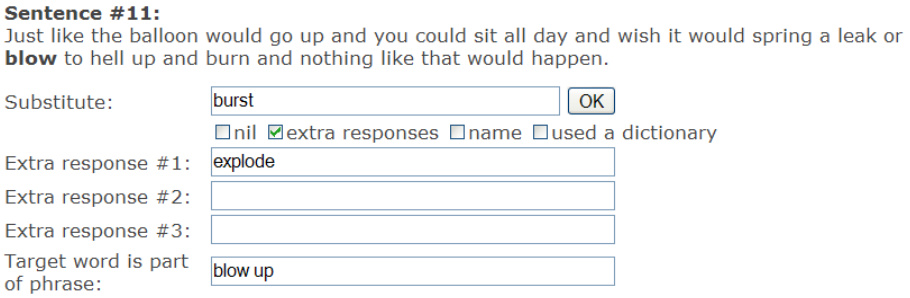

Figure 2: Further Examples for TWPP box: with intervening words

without the subset of items judged to be multiwords (i.e. those items in LEXSUBMW).

Whilst there were 5 annotators, the TWPP box was optional and for many items there was not a TWPP response. Furthermore, for those items where there was a TWPP response, not all the annotators provided one and the annotators did not always agree on the phrase. The construction of LEXSUBMW from the TWPP responses was performed as follows. The TWPP phrases provided by the annotators were semi-automatically lemmatised. For each item (target word in the context of a sentence) a multiword was entered in the gold-standard if there was a majority vote for the same form of the multiword (after lemmatisation) and there were at least two annotators who stipulated the same phrase (we hereafter refer to this constraint as MAJORITY>=*2*). Thus although 1 annotator provided *which way* and 1 provided *which way it would go* for the same item, this item was rejected because annotators do not agree on the actual phrase. From the 1710 sentences released as test data (300 of the 2010 sentences were used as trial data) 282 items were identified by at least one annotator as TWPP. Using the MAJORITY>=*2* criterion resulted in 130 sentences with such a consensus which were entered in LEXSUBMW. Inter-annotator agreement figures and evaluation of WordNet as a baseline system is given in (McCarthy and Navigli, 2007). In the following section we provide a classification of all 282 TWPP responses and analyse the annotators' TWPP responses in terms of this classification.

3. Analysis of the TWPP Responses

Our classification is designed to help us distinguish cases when the TWPP responses are used for paraphrasing to retain grammaticality after substitution compared with cases when the target is a part of a genuine multiword expression with collocational preference or stronger semantic idiosyncracies. We also wished to distinguish compositional collocations from non-compositional multiwords. We do this by assuming that if the phrase has idiosyncratic properties then the substitute will replace the entire phrase. This is true for semantically opaque constructions where semantic interpretation is not easily derived from the constituent words and also, to a lesser extent, for the less semantically transparent collocations.

Verb particles form a large and easily identifiable set of multiword constructions so we classified the verb particle constructions separately from other types. We specified whether the substitute was also a verb particle construction or not. If the substitute was a verb

particle construction this might include the same or a different particle, compared to the TWPP response being substituted. From examining the data, it seems that a substitute verb particle construction is usually more compositional compared to cases where the substitute was a single word. We provide examples with the classification below.

3.1. The Classification

We designed the classification to answer the following questions:

1. Is the TWPP response a verb particle construction?
   (a) Is the substitute also a verb particle construction?
   (b) Is the substitute a word (not a verb particle construction) which replaces the verb particle construction?
2. If not a verb particle construction, is the phrase sufficiently semantically opaque such that the substitute must replace the whole phrase?
3. If not, does the phrase have a strong collocational preference between the words and
   (a) does the substitute replace the whole phrase (even though the meaning of the individual words in the phrase is relatively transparent compared to 2)?
   (b) does the substitute only replace the target word within the phrase (whilst there is a collocational preference for the phrase, the construction is less fixed compared to 3a)?
4. Has the phrase been used simply for ease of paraphrasing because the substitute that the annotator feels is optimal cannot replace the target on its own because of grammatical constraints?

Cases where the substitute covered the whole phrase were felt to be either more likely to be semantically idiosyncratic (1b,2,3a) OR due to grammatical constraints where the annotator was using the response box for ease of paraphrasing (4 and 1a to a certain extent). 3b is more compositional compared to 3a but does have a certain degree of specialised interpretation. These criteria demonstrate various levels of semantic idiosyncracy and help us to distinguish cases which were simply collocational from those where the constituent words no longer retained their original semantics. The following categories were used to answer the questions above and we also devised a miscellaneous category which we also describe. This was used for problematic responses and the subcategories within the miscellaneous category were devised from manual inspection of the TWPP response where the other categories were not appropriate. In the examples below, the target word is shown in boldface and if there is a substitute, this is shown on the right hand side of the → arrow. [5]

syn a syntactic paraphrase e.g. ***letters*** *were* → *correspondence was*, ***earlier*** *than* → *before*

VP1 verb + particle (adverb or preposition) where the substitute is also a verb + particle construction e.g. ***charge*** *with* → *accuse of*

VP2 verb + particle where the substitute is a verb without a particle e.g. ***pass*** *away* → *disappear*

sem the meaning of the target is specialised in the phrase e.g. ***bull*** *market*

colloc1 the target word is substituted with another word that means something similar to the original target and the other part of the phrase should be retained with the substitute e.g. ***civil*** *law* → *non-criminal*

colloc2 the entire phrase would be replaced as a whole with the substitute, but unlike the **sem** category, the meaning of the target word is reflected in the meaning of the phrase e.g. *critical* ***mass*** → *crucial level*

miscellaneous err errors from the annotators e.g. the phrase box was marked with a single word or the paraphrase retained the original target word rather than providing a substitute

quote e.g. *education, education, education* [source: Tony Blair]

intensifier e.g. *very* ***special*** → *exceptional*

names e.g. ***mad*** *cow disease*

measure e.g. *5 3/4 pounds* → *11.27 kilograms*

The author manually inspected all the TWPP responses from each annotator, alongside the substitutes provided for that item and categorised the TWPP according to these categories. [6] We hypothesised that

[5] Note that there is not always a substitute as the annotators were allowed to supply a NIL response if they could not think of a good substitute for the item.

[6] The division is not always clear and sometimes several categories were applicable. For the analysis we used the category deemed the most appropriate. Ideally we would have several people assign categories to the annotations and determine how reliably these categories can be assigned. We leave that for future work.

TWPPs which were not strong collocations or lexicalised multiword expressions were much less likely to be identified in the same way by the majority of annotators.

3.2. Assigning the Categories

The categories are assigned to each non-empty annotator TWPP response for an item. Then for each item we use the category only if all annotators' responses to the TWPP field for this item are of the same category. That is, we discard items from our analysis where we assigned different categories to the TWPP annotations. There are 11 cases of disagreement between **VP1** and **VP2** verdicts (some substitutes include particles whilst others don't) and 3 between **colloc1** and **colloc2**. There were only 3 other cases of disagreement.

4. Results of the Analysis

In table 1 we show for each category the number of items without disagreement, i.e. where all TWPPs for the item have the same category. We also show for these items the average number of annotators identifying TWPP for each item in the category (#ann per item), the average number of annotators that agree on the exact phrase (# agreeing on form) for each item and the number of items meeting the MAJORITY>=*2* criteria specified for the multiword task (and therefore appearing in LEXSUBMW). From this analysis, we find most consensus on semantically anomalous phrases (**sem**) and verb particles (**VP**) [7] (particularly those which are paraphrased without a particle (**VP2**)). This demonstrates that the MAJORITY>=*2* constraint ensures a higher proportion of semantically idiosyncractic multiwords in the LEXSUBMW. There is also considerable consensus for the miscellaneous **quote** and for the proper names (**name**). These are both fixed phrases which could be identified as such in a lexicon. The **colloc** category where the meaning of the target word is reflected in the phrase has less consensus because the semantic idiosyncrasy is less apparent. This consensus is less for the more transparent **colloc1** than **colloc2** (because in the latter the phrase is substituted as a whole).

5. Discussion

In the analysis we have seen that more lexicalised and semantically idiosyncratic TWPP phrases are more likely to appear as entries in the muliword gold standard. This demonstrates that the MAJORITY>=*2* constraint is a reasonable one which ensured that most entries in the multiword gold-standard were lexicalised multiword expressions. In this section we examine the types of TWPP responses which present as genuine multiwords yet did not make it into the LEXSUBMW (false negatives) and those TWPP that did get entered which are not genuine multiwords (false positives). From our analysis, we provide recommendations for using lexical substitution as a way of finding multiword expressions. We discuss possible ways that one might distinguish the different categories in our analysis automatically. We also discuss briefly the findings from the task as to the overlap between the LEXSUBMW entries and the multiwords in WordNet.

5.1. Issues with determining if a annotator response should be in the LEXSUBMW

The MAJORITY>=*2* constraint was intended to ensure that entries in the LEXSUBMW were more likely to be multiword expressions. To a certain extent this worked as we demonstrated in the last column of table 1 which shows that entries meeting this criteria did tend to be genuine multiwords or fixed expressions (**sem**, **colloc**, particularly **colloc2**, **VP** particularly **VP2**), quotes (**quote**) or proper nouns (**name**), and were less likely to be syntactic paraphrases (**syn**). We examined the cases where we assigned a category of **sem**, **colloc2**, **VP2** and where the item did not meet the MAJORITY>=*2* criterion for the LEXSUBMW and also those cases where the category was **syn** and yet the item did satisfy the MAJORITY>=*2* criteria.

5.1.1. False Negatives

The cases where a genuine multiword response did not satisfy the criteria seem due to several reasons. Often, the expression was sufficiently compositional that whilst the expression might be considered a multiword, it would be possible to substitute the target in context and retain the meaning of the phrase.

For example (as before the target word is shown in boldface):

VP2: ***bring*** *back, move* ***forward***
colloc2: *draw to a* ***close****, free* ***range***
sem: ***throw*** *into sharp relief,* ***skip*** *a beat*

The MAJORITY>=*2* constraint sometimes gave rise to a genuine multiword being omitted from LEXSUBMW, for example we had the following number of TWPP responses for the various forms (numbers of responses for each form are shown in brackets):

comes *to* (2) *it* ***comes*** *to* (2) ***come*** *to* (1)
forward *looking* (1) *be* ***forward*** *looking* (1)
bring *closer to* (1) ***bring*** *closer* (1)

[7] We use **colloc** to refer to **colloc1** and **colloc2**, and **VP** to refer to both **VP1** and **VP2**.

catg	size	#ann per item	# agreeing on form	in LEXSUBMW (% of catg)
syn	44	1.70	1.61	15 (34.09)
colloc1	65	1.14	1.14	9 (13.85)
colloc2	36	1.97	1.72	15 (41.67)
VP1	11	1.45	1.45	5 (45.45)
VP2	49	2.29	2.27	32 (65.31)
sem	46	2.54	2.26	34 (73.91)
misc:err	4	1.00	1.00	0 (0.00)
misc:quote	1	2.00	2.00	1 (100.00)
misc:inten	1	1.00	1.00	0 (0.00)
misc:name	6	3.00	2.33	5 (83.33)
misc:meas	3	1.00	1.67	0 (0.00)

Table 1: Analysis of the classification of items

The problem is due to the lack of an agreed canonical form. Possibly this could be resolved in many cases by linguists scrutinising these cases, or automatically by using a lemmatiser and resolving any difference to the longer or shorter form provided that one is a subsequence of the other. Another option would be to use all choices from the annotators when constructing the gold-standard, with weights on each form, rather than applying the MAJORITY>=*2* constraint. This would also allow some credit where currently the phrase is a false negative because it is sufficiently compositional for substitution of the target without consideration of the entire phrase, as in the examples in the first paragraph in this subsection(5.1.1.) and for the majority of **colloc1** expressions.

5.1.2. False Positives

In these cases, there was a majority vote from two or more annotators for the same form however, the phrase was identified for ease of paraphrasing. For example:

rather than, earlier than, full of, instead of, on our side, told me (**syn**)

Most of these seem to be collocational grammatical constraints between the target and a closed class grammatical function word. This might be useful information for a computational system, however, these sorts of expressions are not usually considered to be multiwords. One way of distinguishing these might be to discount TWPP where the target word is identified with one closed class grammatical word and where the TWPP is not a verb particle construction. [8]

[8]We could ask annotators to make a distinction between when they are using the TWPP response for ease of paraphrasing and when they believe the phrase is functioning as a single unit/word. This however might be difficult for annotators to judge and so we would prefer to make this distinction automatically.

5.2. Partitioning the Categories

An outstanding question is whether we might be able to partition these TWPP responses to the various categories automatically. The syntactic paraphrasing issues might be detected by ruling out any items with a MAJORITY>=*2* which have the target word with one function word but are not verb particle constructions.

The distinction between collocation and semantic is based on semantic transparency and might perhaps be done automatically by comparing semantics of constituent words with that of the phrase (Katz and Giesbrecht, 2006).

The distinction between whether the substitute replaces the entire phrase, i.e. that between **colloc1** and **colloc2** would best be done in future by asking the annotators to stipulate this in a box. Indeed, it might be preferable to get annotators to focus on only using the TWPP box where they really need to substitute for the entire phrase. This would remove the **colloc1** from the gold-standard and would make the goal of the task simpler for annotators and systems. It might be easier to draw a boundary this way between what is a multiword and what isn't according to the substitutability test, but as we have seen, even semantically idiosyncratic multiwords can have some semantic transparency and flexibility with regard to lexical substitutes. For this reason we would recommend keeping the TWPP as it is (to indicate where the target is an integral part of a phrase) but when entering a TWPP response we would recommend requiring the annotators to stipulate in another box whether their substitute replaces this entire phrase, or just the target word.

5.3. Overlap with WordNet

One future area for research is to what extent the multiword expressions captured in the lexical substitution framework coincide with those found in hand-crafted lexical resources. This work has begun with the LEX-

SUB multiword subtask where the authors provided a baseline using WordNet to examine a 5 word window around the target word (2 words either side) and determine if there was a multiword expression from WordNet in that window. 1 participant (Zhao et al., 2007) also developed a similar system. The precision for finding that there was a multiword (using LEXSUBMW as the gold-standard) on the items which were found in WordNet using the 5-gram window was 43.6% approx and the recall against all the items with an entry in LEXSUBMW was 36.9%. These figures show that there is a lot of room for improvement but one clear impediment is that whichever resource is used will have gaps. Furthermore, there were many entries in WordNet that were not found by the annotators since they are relatively compositional and productive collocations for example *civil law, civil war, civil authority, compost heap, garbage heap, phone number, phone bill*. If these had been found by annotators they would have been categorised here as **colloc1**. For identification of the exact form, precision of the items identified by the WordNet baseline was 40% whereas recall against all in LEXSUBMW was 33.9%. Sometimes the issue is the canonical form. For example, in WordNet there is a multiword *look forward* whereas the majority of annotators voted for the form *look **forward** to*. Sometimes it was simply that the 5-gram window used in the baseline did not allow for multiwords where the target word appeared with more than two words to the right or left for example, ***pull** out all the stops*. If we calculate identification of the correct form only where WordNet did find a multiword AND there was a multiword entry in LEXSUBMW then we get 91.7% precision. The canonical form is much less of a problem compared with the fact that WordNet and other resources will record common collocations which are quite productive and compositional and will not necessarily be picked up by annotators on a substitution task. Furthermore, there are many responses from annotators which are not found in WordNet. The **syn** category are a common example (e.g. *earlier than, rather than*) but also because WordNet has some stored without a space e.g. *goalpost* and *scrapbook*, and, like all lexical resources, it has some omissions e.g. *pass muster*. [9]

[9]Please note that we report here the performance of the WordNet baseline used at SemEval however, performance of a WordNet baseline can be expected to be better than reported due to minor bugs in that baseline system. This can be seen by the slightly better performance of the participant system (Zhao et al., 2007) which adopted a similar approach. We report results for the official baseline as that was reported at the time of the competition, but we only report omissions from WordNet that are not due to bugs in the LEXSUB MW baseline.

6. Conclusions

In this paper we present analysis of a multiword resource that has been produced as a consequence of a lexical substitution task. The analysis performed here investigates the type of multiword expression found in the data by the annotators who were focused on the task of substitution but asked to identify cases where the word was an integral part of a phrase. We examine the responses of the annotators to determine the proportion of verb particles, collocations and semantically idiosyncratic phrases as well as those which seem to be due to syntactic/paraphrasing constraints. We show that the MAJORITY$>=2$ constraint on entries appearing in the gold-standard does indeed ensure that a higher percentage of genuine multiwords appear in the resource than would be if we included all responses, however since there are some genuine multiwords which do not get entered because they fail this constraint, it may be worth entering all responses but with a weighting on the number of annotators selecting that response.

By and large we see many of the multiwords found during the lexical substitution exercise appearing in WordNet, however there are differences with many legitimate multiwords not present in one or other resource. Typically, those in WordNet and not in LEXSUB are often semantically transparent collocations that do not necessitate identification for substitution purposes because the meaning of the constituent words is retained in the expression. There are some expressions found in LEXSUBMW and not in WordNet which are due to the paraphrasing/substitution nature of the LEXSUB task. We believe these can perhaps be weeded out by using various patterns such as if the target word consists of the target word preceded or followed by a closed class word where the phrase is not a verb particle construction.

We do not believe that our approach is a panacea for multiword evaluation, however we do feel it is useful because the annotators are focused on the substitution task and are not asked to make difficult judgments on whether a phrase is a multiword or not. Instead the focus is whether the target word is substitutable in its own right, or whether there are semantic or syntactic peculiarities of the phrasal context that need to be considered.

For the future, we would also recommend instructing annotators to indicate when filling in the TWPP box whether the substitute replaces the whole phrase (as with **VP1**, **VP2**, **colloc2** and **sem**) or just the target

word.

7. Acknowledgements

This work was funded by a UK Royal Society Dorothy Hodgkin Fellowship. We would like to thank Roberto Navigli for useful comments on this paper.

8. References

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, June.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002)*, pages 98–104, Taipei, Taiwan.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL Workshop on multiword expressions: analysis, acquisition and treatment*, pages 65–72.

Don Blaheta and Mark Johnson. 2001. Unsupervised learning of multi-word verbs. In *Proceedings of the ACL Workshop on Collocations*, pages 54–60, Toulouse, France.

Scott Cotton, Phil Edmonds, Adam Kilgarriff, and Martha Palmer. 2001. SENSEVAL-2. http://www.sle.sharp.co.uk/senseval2/.

Afsaneh Fazly and Suzanne 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)*, pages 337–344, Trento, Italy, April.

Nicole Grégoire, Brigitte Krenn, and Stefan Evert, editors. 2008. *Proceedings of the LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 08)*, Marrakech, Morroco.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the ACL Workshop on multiword expressions: identifying and exploiting Underlying Properties*, pages 12–19.

Adam Kilgarriff et al. 1998. SENSEVAL - evaluating word sense disambiguation systems. http://www.itri.brighton.ac.uk/events/senseval/proceedings.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, pages 317–324, Univeristy of Maryland, College Park, Maryland.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 03 Workshop: Multiword expressions: analysis, acquisition and treatment*, pages 73–80.

Rada Mihalcea and Phil Edmonds, editors. 2004. *Proceedings* SENSEVAL-3 *Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Barcelona, Spain.

Scott Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL Workshop on multiword expressions: analysis, acquisition and treatment*, pages 49–56.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan 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 (CICLing 2002)*, pages 1–15, Mexico City, Mexico.

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features. In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, pages 899–906, Vancouver, B.C., Canada.

Shiqi Zhao, Lin Zhao, Yu Zhang, Ting Liu, and Sheng Li. 2007. HIT: Web based scoring method for english lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 173–176, Prague, Czech Republic, June. Association for Computational Linguistics.