An Empirical Model of Multiword Expression Decomposability

Timothy Baldwin[†], Colin Bannard[‡], Takaaki Tanaka[*] and Dominic Widdows[†]

† CSLI
Stanford University
Stanford CA 94305, USA
{tbaldwin,dwiddows}@csli.stanford.edu

‡ School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
c.j.bannard@ed.ac.uk

∗ Communication Science Labs
NTT Corporation
Kyoto, Japan
takaaki@cslab.kecl.ntt.co.jp

Abstract

This paper presents a construction-inspecific model of multiword expression decomposability based on latent semantic analysis. We use latent semantic analysis to determine the similarity between a multiword expression and its constituent words, and claim that higher similarities indicate greater decomposability. We test the model over English noun-noun compounds and verb-particles, and evaluate its correlation with similarities and hyponymy values in WordNet. Based on mean hyponymy over partitions of data ranked on similarity, we furnish evidence for the calculated similarities being correlated with the semantic relational content of WordNet.

1 Introduction

This paper is concerned with an empirical model of multiword expression decomposability. **Multiword expressions** (MWEs) are defined to be cohesive lexemes that cross word boundaries (Sag et al., 2002; Copestake et al., 2002; Calzolari et al., 2002). They occur in a wide variety of syntactic configurations in different languages (e.g. in the case of English, compound nouns: *post office*, verbal idioms: *pull strings*, verb-particle constructions: *push on*, etc.). **Decomposability** is a description of the degree to which the semantics of an MWE can be ascribed to those of its parts (Riehemann, 2001; Sag et al., 2002). Analysis of the semantic correlation between the constituent parts and whole of an MWE is perhaps more commonly discussed under the banner of compositionality (Nunberg et al., 1994; Lin, 1999). Our claim here is that the semantics of the MWE are deconstructed and the parts coerced into often idiosyncratic interpretations to attain semantic alignment, rather than the other way around. One idiom which illustrates this process is *spill the beans*, where the semantics of `reveal′(secret′)` are decomposed such that *spill* is coerced into the idiosyncratic interpretation of `reveal′` and *beans* into the idiosyncratic interpretation of `secret′`. Given that these senses for *spill* and *beans* are not readily available at the simplex level other than in the context of this particular MWE, it seems fallacious to talk about them composing together to form the semantics of the idiom.

Ideally, we would like to be able to differentiate between three classes of MWEs: non-decomposable, idiosyncratically decomposable and simple decomposable (derived from Nunberg et al.'s sub-classification of idioms (1994)). With **non-decomposable** MWEs (e.g. *kick the bucket*, *shoot the breeze*, *hot dog*), no decompositional analysis is possible, and the MWE is semantically impenetrable. The only syntactic variation that non-decomposable MWEs undergo is verbal inflection (e.g. *kicked the bucket*, *kicks the bucket*) and pronominal reflexivisation (e.g. *wet oneself*, *wet themselves*). **Idiosyncratically decomposable** MWEs (e.g. *spill the beans*, *let the cat out of the bag*, *radar footprint*) are decomposable but coerce their parts into taking semantics unavailable outside the MWE. They undergo a certain degree of syntactic variation (e.g. *the cat was let out of the bag*). Finally, **simple decomposable** MWEs (also known as "institutionalised" MWEs, e.g. *kindle excitement*, *traffic light*) decompose into simplex senses and generally display high syntactic variability. What makes simple decomposable expressions true MWEs rather than productive word combinations is that they tend to block compositional alternates with the expected semantics (termed **anti-collocations** by Pearce (2001b)). For example, *motor car* cannot be rephrased as **engine car* or **motor automobile*. Note that the existence of anti-collocations is also a test for non-decomposable and idiosyncratically decomposable MWEs (e.g. *hot dog* vs. *#warm dog* or *#hot canine*).

Our particular interest in decomposability stems from ongoing work on grammatical means for capturing MWEs. Nunberg et al. (1994) observed that idiosyncratically decomposable MWEs (in particular idioms) undergo much greater syntactic variation than non-decomposable MWEs, and that the variability can be partially predicted from the decompositional analysis. We thus aim to capture the decomposability of MWEs in the grammar and use this to constrain the syntax of MWEs in parsing and generation. Note that it is arguable whether simple decomposable MWEs belong in the grammar proper, or should be described instead as lexical affinities between particular word combinations.

As the first step down the path toward an empirical model of decomposability, we focus on demarcating simple decomposable MWEs from idiosyncratically decomposable and non-decomposable MWEs. This is largely equivalent to classifying MWEs as being **endocentric** (i.e., a hyponym of their head) or **exocentric** (i.e., not a hyponym of their head: Haspelmath (2002)).

We attempt to achieve this by looking at the semantic similarity between an MWE and its constituent words, and hypothesising that where the similarity between the constituents of an MWE and the whole is sufficiently high, the MWE must be of simple decomposable type.

The particular similarity method we adopt is latent semantic analysis, or LSA (Deerwester et al., 1990). LSA allows us to calculate the similarity between an arbitrary word pair, offering the advantage of being able to measure the similarity between the MWE and each of its constituent words. For MWEs such as *house boat*, therefore, we can expect to capture the fact that the MWE is highly similar in meaning to both constituent words (i.e. the modifier *house* and head noun *boat*). More importantly, LSA makes no assumptions about the lexical or syntactic composition of the inputs, and thus constitutes a fully construction- and language-inspecific method of modelling decomposability. This has clear advantages over a more conventional supervised classifier-style approach, where training data would have to be customised to a particular language and construction type.

Evaluation is inevitably a difficulty when it comes to the analysis of MWEs, due to the lack of concise consistency checks on what MWEs should and should not be incorporated into dictionaries. While recognising the dangers associated with dictionary-based evaluation, we commit ourselves to this paradigm and focus on searching for appropriate means of demonstrating the correlation between dictionary- and corpus-based similarities.

The remainder of this paper is structured as follows. Section 2 describes past research on MWE compositionality of relevance to this effort. Section 3 provides a basic outline of the resources used in this research, LSA, the MWE extraction methods, and measures used to evaluate our method. Section 4 then provides evaluation of the proposed method, and the paper is concluded with a brief discussion in Section 5.

2 Past research

Although there has been some useful work on compositionality in statistical machine translation (e.g. Melamed (1997)), there has been little work on detecting "non-compositional" (i.e. non-decomposable and idiosyncratically decomposable) items of variable syntactic type in monolingual corpora. One interesting exception is Lin (1999), whose approach is explained as follows:

> The intuitive idea behind the method is that the metaphorical usage of a non-compositional expression causes it to have a different distributional characteristic than expressions that are similar to its literal meaning.

The expressions he uses are taken from a collocation database (Lin, 1998b). These "expressions that are similar to [their] literal meaning" are found by substituting each of the words in the expression with the 10 most similar words according to a corpus derived thesaurus (Lin, 1998a). Lin models the distributional difference as a significant difference in mutual information. Significance here is defined as the absence of overlap between the 95% confidence interval of the mutual information scores. Lin provides some examples that suggest he has identified a successful measure of "compositionality". He offers an evaluation where an item is said to be non-compositional if it occurs in a dictionary of idioms. This produces the unconvincing scores of 15.7% for precision and 13.7% for recall.

We claim that substitution-based tests are useful in demarcating MWEs from productive word combinations (as attested by Pearce (2001a) in a MWE detection task), but not in distinguishing the different classes of decomposability. As observed above, simple decomposable MWEs such as *motor car* fail the substitution test not because of non-decomposability, but because the expression is institutionalised to the point of blocking alternates. Thus, we expect Lin's method to return a wide array of both decomposable and non-decomposable MWEs.

Bannard (2002) focused on distributional techniques for describing the meaning of verb-particle constructions at the level of logical form. The semantic similarity between a multiword expression and its head was used as an indicator of decomposability. The assumption was that if a verb-particle was sufficiently similar to its head verb, then the verb contributed its simplex meaning. It gave empirical backing to this assumption by showing that annotator judgements for verb-particle decomposability correlate significantly with non-expert human judgements on the similarity between a verb-particle construction and its head verb. Bannard et al. (2003) extended this research in looking explicitly at the task of classifying verb-particles as being compositional or not. They successfully combined statistical and distributional techniques (including LSA) with a substitution test in analysing compositionality. McCarthy et al. (2003) also targeted verb-particles for a study on compositionality, and judged compositionality according to the degree of overlap in the N most similar words to the verb-particle and head verb, e.g., to determine compositionality.

We are not the first to consider applying LSA to MWEs. Schone and Jurafsky (2001) applied LSA to the analysis of MWEs in the task of MWE discovery, by way of rescoring MWEs extracted from a corpus. The major point of divergence from this research is that Schone and Jurafsky focused specifically on MWE extraction, whereas we are interested in the downstream task of semantically classifying attested MWEs.

3 Resources and Techniques

In this section, we outline the resources used in evaluation, give an informal introduction to the LSA model, sketch how we extracted the MWEs from corpus data, and describe a number of methods for modelling decomposability within a hierarchical lexicon.

3.1 Resources and target MWEs

The particular reference lexicon we use to evaluate our technique is WordNet 1.7 (Miller et al., 1990), due to its public availability, hierarchical structure and wide coverage. Indeed, Schone and Jurafsky (2001) provide evidence that suggests that WordNet is as effective an evaluation resource as the web for MWE detection methods, despite its inherent size limitations and static nature.

Two MWE types that are particularly well represented in WordNet are compound nouns (47,000 entries) and multiword verbs (2,600 entries). Of these, we chose to specifically target two types of MWE: noun-noun (NN) compounds (e.g. *computer network*, *work force*) and verb-particles (e.g. *look on*, *eat up*) due to their frequent occurrence in both decomposable and non-decomposable configurations, and also their disparate syntactic behaviours.

We extracted the NN compounds from the 1996 Wall Street Journal data (WSJ, 31m words), and the verb-particles from the British National Corpus (BNC, 90m words: Burnard (2000)). The WSJ data is more tightly domain-constrained, and thus a more suitable source for NN compounds if we are to expect sentential context to reliably predict the semantics of the compound. The BNC data, on the other hand, contains more colloquial and prosaic texts and is thus a richer source of verb-particles.

3.2 Description of the LSA model

Our goal was to compare the distribution of different compound terms with their constituent words, to see if this indicated similarity of meaning. For this purpose, we used latent semantic analysis (LSA) to build a vector space model in which term-term similarities could be measured.

LSA is a method for representing words as points in a vector space, whereby words which are related in meaning should be represented by points which are near to one another, first developed as a method for improving the vector model for information retrieval (Deerwester et al., 1990). As a technique for measuring similarity between words, LSA has been shown to capture semantic properties, and has been used successfully for recognising synonymy (Landauer and Dumais, 1997), word-sense disambiguation (Schütze, 1998) and for finding correct translations of individual terms (Widdows et al., 2002).

The LSA model we built is similar to that described in (Schütze, 1998). First, 1000 frequent content words (i.e. not on the stoplist)[1] were chosen as "content-bearing words". Using these content-bearing words as column labels, the 50,000 most frequent terms in the corpus were assigned row vectors by counting the number of times they oc-

[1]A "stoplist" is a list of frequent words which have little independent semantic content, such as prepositions and determiners (Baeza-Yates and Ribiero-Neto, 1999, p167).

curred within the same sentence as a content-bearing word. Singular-value decomposition (Deerwester et al., 1990) was then used to reduce the number of dimensions from 1000 to 100. Similarity between two vectors (points) was measured using the cosine of the angle between them, in the same way as the similarity between a query and a document is often measured in information retrieval (Baeza-Yates and Ribiero-Neto, 1999, p28). Effectively, we could use LSA to measure the extent to which two words or MWEs x and y usually occur in similar contexts.

Since the corpora had been tagged with parts-of-speech, we could build syntactic distinctions into the LSA models — instead of just giving a vector for the string *test* we were able to build separate vectors for the nouns, verbs and adjectives *test*. This combination of technologies was also used to good effect by Widdows (2003): an example of the contribution of part-of-speech information to extracting semantic neighbours of the word *fire* is shown in Table 1. As can be seen, the noun *fire* (as in the substance/element) and the verb *fire* (mainly used to mean firing some sort of weapon) are related to quite different areas of meaning. Building a single vector for the string *fire* confuses this distinction — the neighbours of *fire* treated just as a string include words related to both the meaning of *fire* as a noun (more frequent in the BNC) and as a verb. The appropriate granularity of syntactic classifications is an open question for this kind of research: treating all the possible verbs categories as different (e.g. distinguishing infinitive from finite from gerund forms) led to data sparseness, and instead we considered "verb" as a single part-of-speech type.

3.3 MWE extraction methods

NN compounds were extracted from the WSJ by first tagging the data with fnTBL 1.0 (Ngai and Florian, 2001) and then simply taking noun bigrams (adjoined on both sides by non-nouns to assure the bigram is not part of a larger compound nominal). Out of these, we selected those compounds that are listed in WordNet, resulting in 5,405 NN compound types (208,000 tokens).

Extraction of the verb-particles was considerably more involved, and drew on the method of Baldwin and Villavicencio (2002). Essentially, we used a POS tagger and chunker (both built using fnTBL 1.0 (Ngai and Florian, 2001)) to first (re)tag the BNC. This allowed us to extract verb-particle tokens through use of the particle POS and chunk tags returned by the two systems. This produces high-precision, but relatively low-recall results, so we performed the additional step of running a chunk-based grammar over the chunker output to detect candidate mistagged particles. In the case that a noun phrase followed the particle candidate, we performed attachment disambiguation to determine the transitivity of the particle candidate. These three methods produced three distinct sets of verb-particle tokens, which we carried out weighted voting over to determine the final set of verb-particle tokens. A total of 461 verb-particles attested in WordNet were extracted (160,765 tokens).

For both the NN compound and verb-particle data, we replaced each token occurrence with a single-word POS-tagged token to feed into the LSA model.

3.4 Techniques for evaluating correlation with WordNet

In order to evaluate our approach, we employed the lexical relations as defined in the WordNet lexical hierarchy (Miller et al., 1990). WordNet groups words into sets with similar meaning (known as "synsets"), e.g. {*car, auto, automobile, machine, motorcar* } . These are organised into a hierarchy employing multiple inheritance. The hierarchy is structured according to different principles for each of nouns, verbs, adjectives and adverbs. The nouns are arranged according to hyponymy or `ISA` relations, e.g. a *car* is a kind of *automobile*. The verbs are arranged according to troponym or "manner-of" relations, where *murder* is a manner of killing, so *kill* immediately dominates *murder* in the hierarchy.

We used WordNet for evaluation by way of looking at: (a) hyponymy, and (b) semantic distance.

Hyponymy provides the most immediate way of evaluating decomposability. With simple decomposable MWEs, we can expect the constituents (and particularly the head) to be hypernyms (ancestor nodes) or synonyms of the MWE. That is, simple decomposable MWEs are generally endocentric, although there are some exceptions to this generalisation such as *vice president* arguably not being a hyponym of *president*. No hyponymy relation holds with non-decomposable or idiosyncratically decomposable MWEs (i.e., they are exocentric), as even if the semantics of the head noun can be determined through decomposition, by definition this will not correspond to a simplex sense of the word.

We deal with polysemy of the constituent words and/or MWE by simply looking for the existence of a sense of the constituent words which

fire (string only)		fire_nn1		fire_vvi	
fire	1.000000	fire_nn1	1.000000	fire_vvi	1.000000
flames	0.709939	flames_nn2	0.700575	guns_nn2	0.663820
smoke	0.680601	smoke_nn1	0.696028	firing_vvg	0.537778
blaze	0.668504	brigade_nn1	0.589625	cannon_nn0	0.523442
firemen	0.627065	fires_nn2	0.584643	gun_nn1	0.484106
fires	0.617494	firemen_nn2	0.567170	fired_vvd	0.478572
explosion	0.572138	explosion_nn1	0.551594	detectors_nn2	0.477025
burning	0.559897	destroyed_vvn	0.547631	artillery_nn1	0.469173
destroyed	0.558699	burning_aj0	0.533586	attack_vvb	0.468767
brigade	0.532248	blaze_nn1	0.529126	firing_nn1	0.459000
arson	0.528909	arson_nn1	0.522844	volley_nn1	0.458717
accidental	0.519310	alarms_nn2	0.512332	trained_vvn	0.447797
chimney	0.489577	destroyed_vvd	0.512130	enemy_nn1	0.445523
blast	0.488617	burning_vvg	0.502052	alert_aj0	0.443610
guns	0.487226	burnt_vvn	0.500864	shoot_vvi	0.443308
damaged	0.484897	blast_nn1	0.498635	defenders_nn2	0.438886

Table 1: Semantic neighbours of *fire* with different parts-of-speech. The scores are cosine similarities

subsumes a sense of the MWE. The function $\mathrm{hyponym}(word_i, mwe)$ thus returns a value of 1 if some sense of $word_i$ subsumes a sense of mwe, and a value of 0 otherwise.

A more proactive means of utilising the WordNet hierarchy is to derive a **semantic distance** based on analysis of the relative location of senses in WordNet. Budanitsky and Hirst (2001) evaluated the performance of five different methods that measure the semantic distance between words in the WordNet Hierarchy, which Patwardhan et al. (2003) have then implemented and made available for general use as the Perl package `distance-0.11`.[2] We focused in particular on the following three measures, the first two of which are based on information theoretic principles, and the third on sense topology:

- Resnik (1995) combined WordNet with corpus statistics. He defines the similarity between two words as the information content of the lowest superordinate in the hierarchy, defining the information content of a concept c (where a concept is the WordNet class containing the word) to be the negative of its log likelihood. This is calculated over a corpus of text.

- Lin (1998c) also employs the idea of corpus-derived information content, and defines the similarity between two concepts in the following way:

$$\mathrm{sim}(C_1, C_2) = \frac{2 \log P(C_0)}{\log P(C_1) + \log P(C_2)} \quad (1)$$

  where C_0 is the lowest class in the hierarchy that subsumes both classes.

- Hirst and St-Onge (1998) use a system of "relations" of different strength to determine the similarity of word senses, conditioned on the type, direction and relative distance of edges separating them.

The Patwardhan et al. (2003) implementation that we used calculates the information values from SemCor, a semantically tagged subset of the Brown corpus. Note that the first two similarity measures operate over nouns only, while the last can be applied to any word class.

The similarity measures described above calculate the similarity between a pair of senses. In the case that a given constituent word and/or MWE occur with more than one sense, we calculate a similarity for sense pairing between them, and average over them to produce a consolidated similarity value.

4 Evaluation

LSA was used to build models in which MWEs could be compared with their constituent words. Two models were built, one from the WSJ corpus (indexing NN compounds) and one from the BNC (indexing verb-particles). After removing stopwords, the 50,000 most frequent terms were indexed in each model. From the WSJ, these 50,000 terms included 1,710 NN compounds (with corpus frequency of at least 13) and from the BNC, 461 verb-particles (with corpus frequency of at least 49).

We used these models to compare different words, and to find their neighbours. For example, the neighbours of the simplex verb *cut* and the verb-particles *cut out* and *cut off* (from the BNC model) are shown in Table 2. As can be seen, several of the neighbours of *cut out* are from similar semantic areas as those of *cut*, whereas those of *cut off* are quite different.

[2] `http://www.d.umn.edu/~tpederse/distance.html`

cut (verb)		cut out (verb)		cut off (verb)	
cut_verb	1.000000	cut out_verb	1.000000	cut off_verb	1.000000
trim_verb	0.529886	fondant_nn	0.516956	knot_nn	0.448871
slash_verb	0.522370	fondant_jj	0.501266	choke_verb	0.440587
cut_nns	0.520345	strip_nns	0.475293	vigorously_rb	0.438071
cut_nn	0.502100	piece_nns	0.449555	suck_verb	0.413003
reduce_verb	0.465364	roll_nnp	0.440769	crush_verb	0.412301
cut out_verb	0.433465	stick_jj	0.434082	ministry_nn	0.408702
pull_verb	0.431929	cut_verb	0.433465	glycerol_nn	0.395148
fall_verb	0.426111	icing_nn	0.432307	tap_verb	0.383932
hook_verb	0.419564	piece_nn	0.418780	shake_verb	0.381581
recycle_verb	0.413206	paste_nn	0.416581	jerk_verb	0.381284
project_verb	0.401246	tip_nn	0.413603	put down_verb	0.380368
recycled_jj	0.396315	hole_nns	0.412813	circumference_nn	0.378097
prune_verb	0.395656	straw_nn	0.411617	jn_nnp	0.375634
pare_verb	0.394991	hook_nn	0.402947	pump_verb	0.373984
tie_verb	0.392964	strip_nn	0.399974	nell_nnp	0.373768

Table 2: Semantic neighbours of the verbs *cut*, *cut out*, and *cut off*.

Construction	***Method***	***Pearson***	R^2
NN compound	Resnik	.108	.012
	Lin	.101	.010
	HSO	.072	.005
verb-particle	HSO	.255	.065

Table 3: Correlation between LSA and WordNet similarities

This reflects the fact that in most of its instances the verb *cut off* is used to mean "forcibly isolate".

In order to measure this effect quantitatively, we can simply take the cosine similarities between these verbs, finding that $\mathrm{sim}(\textit{cut, cut out}) = 0.433$ and $\mathrm{sim}(\textit{cut, cut off}) = 0.183$ from which we infer directly that, relative to the sense of *cut*, *cut out* is a clearer case of a simple decomposable MWE than *cut off*.

4.1 Statistical analysis

In order to get an initial feel for how well the LSA-based similarities for MWEs and their head words correlate with the WordNet-based similarities over those same word pairs, we did a linear regression and Pearson's correlation analysis of the paired data (i.e. the pairing $\langle \mathrm{sim}_{\mathrm{LSA}}(word_i, mwe), \mathrm{sim}_{\mathrm{WN}}(word_i, mwe)\rangle$ for each WordNet similarity measure $\mathrm{sim}_{\mathrm{WN}}$). For both tests, values closer to 0 indicate random distribution of the data, whereas values closer to 1 indicate a strong correlation. The correlation results for NN compounds and verb-particles are presented in Table 3, where R^2 refers to the output of the linear regression test and *HSO* refers to Hirst and St-Onge similarity measure. In the case of NN compounds, the correlation with LSA is very low for all tests, that is LSA is unable to reproduce the relative similarity values derived from WordNet with any reliability. With verb-particles, correlation is notably higher than for NN compounds,[3] but still at a low level.

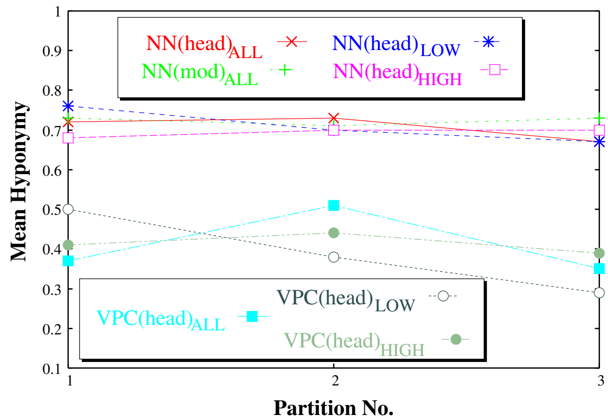

Figure 1: Hyponymy correlation

Based on these results, LSA would appear to correlate poorly with WordNet-based similarities. However, our main interest is not in similarity per se, but how reflective LSA similarities are of the decomposability of the MWE in question. While taking note of the low correlation with WordNet similarities, therefore, we move straight on to look at the hyponymy test.

4.2 Hyponymy-based analysis

We next turn to analysis of correlation between LSA similarities and hyponymy values. Our expectation is that for constituent word–MWE pairs with higher LSA similarities, there is a greater likelihood of the MWE being a hyponym of the constituent word. We test this hypothesis by ranking the constituent word–MWE pairs in decreasing order of LSA similarity,

[3] Recall that HSO is the only similarity measure which operates over verbs.

and partitioning the ranking up into m partitions of equal size. We then calculate the average number of hyponyms per partition. If our hypothesis is correct, the earlier partitions (with higher LSA similarities) will have higher occurrences of hyponyms than the latter partitions.

Figure 1 presents the mean hyponymy values across partitions of the NN compound data and verb-particle data, with m set to 3 in each case. For the NN compounds, we derive two separate rankings, based on the similarity between the head noun and NN compound (*NN(head)*) and the modifier noun and the NN compound (*NN(mod)*). In the case of the verb-particle data, WordNet has no classification of prepositions or particles, so we can only calculate the similarity between the head verb and verb-particle (*VPC(head)*). Looking to the curves for these three rankings, we see that they are all fairly flat, nondescript curves. If we partition the data up into low- and high-frequency MWEs, as defined by a threshold of 100 corpus occurrences, we find that the graphs for the low-frequency data (*NN(head)*$_{\mathrm{LOW}}$ and *VPC(head)*$_{\mathrm{LOW}}$) are both monotonically decreasing, whereas those for high-frequency data (*NN(head)*$_{\mathrm{HIGH}}$ and *VPC(head)*$_{\mathrm{HIGH}}$) are more haphazard in nature. Our hypothesis of lesser instances of hyponymy for lower similarities is thus supported for low-frequency items but not for high-frequency items, suggesting that LSA similarities are more brittle over high-frequency items for this particular task. The results for the low-frequency items are particularly encouraging given that the LSA-based similarities were found to correlate poorly with WordNet-derived similarities. The results for *NN(mod)* are more erratic for both low- and high-frequency terms, that is the modifier noun is not as strong a predictor of decomposability as the head noun. This is partially supported by the statistics on the relative occurrence of NN compounds in WordNet subsumed by their head noun (71.4%) as compared to NN compounds subsumed by their modifier (13.7%).

In an ideal world, we would hope that the values for mean hyponymy were nearly 1 for the first partition and nearly 0 for the last. Naturally, this presumes perfect correlation of the LSA similarities with decomposability, but classificational inconsistencies in WordNet also work against us. For example, *vice chairman* is an immediate hyponym of both *chairman* and *president*, but *vice president* is not a hyponym of *president*. According to LSA, however, $\mathrm{sim}(\textit{chairman}, \textit{vice chairman}) = .508$ and $\mathrm{sim}(\textit{president}, \textit{vice president}) = .551$.

It remains to be determined why LSA should perform better over low-frequency items, although the higher polysemy of high-frequency items is one potential cause. We intend to further investigate this matter in future research.

5 Discussion

While evaluation pointed to a moderate correlation between LSA similarities and occurrences of hyponymy, we have yet to answer the question of exactly where the cutoffs between simple decomposable, idiosyncratically decomposable and non-decomposable MWEs lie. While it would be possible to set arbitrary thresholds to artificially partition up the space of MWEs based on LSA similarity (or alternatively use statistical tests to derive confidence intervals for similarity values), we feel that more work needs to be done in establishing exactly what different LSA similarities for different MWE–constituent word combinations mean.

One area in which we plan to extend this research is the analysis of MWEs in languages other than English. Because of LSA's independence from linguistic constraints, it is equally applicable to all languages, assuming there is some way of segmenting inputs into constituent words.

To summarise, we have proposed a construction-inspecific empirical model of MWE decomposability, based on latent semantic analysis. We evaluated the method over English NN compounds and verb-particles, and showed it to correlate moderately with WordNet-based hyponymy values.

Acknowledgements

This material is partly based upon work supported by the National Science Foundation under Grant No. BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. We would like to thank the anonymous reviewers for their valuable input on this research.

References

Ricardo Baeza-Yates and Berthier Ribiero-Neto. 1999. *Modern Information Retrieval*. Addison Wesley / ACM press.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In

Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment. (this volume).

Colin Bannard. 2002. Statistical techniques for automatically inferring the semantics of verb-particle constructions. *LinGO Working Paper No. 2002-06*.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on Wordnet and Other Lexical Resources, Second meeting of the NAACL*, Pittsburgh, USA.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

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 *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1934–40, 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 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1941–7, Las Palmas, Canary Islands.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Martin Haspelmath. 2002. *Understanding Morphology*. Arnold Publishers.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropism. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–32. MIT Press, Cambridge, USA.

Thomas Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition. *Psychological Review*, 104(2):211–240.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*.

Dekang Lin. 1998b. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*.

Dekang Lin. 1998c. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proc. of the 37th Annual Meeting of the ACL*, pages 317–24, College Park, USA.

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*. (this volume).

I. 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 (EMNLP-97)*, Providence, USA.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–44.

Grace Ngai and Radu Florian. 2001. Transformation-based learning in the fast lane. In *Proc. of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2001)*, pages 40–7, Pittsburgh, USA.

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

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proc. of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*, Mexico City, Mexico.

Darren Pearce. 2001a. Synonymy in collocation extraction. In *Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, USA.

Darren Pearce. 2001b. Using conceptual similarity for collocation extraction. In *Proc. of the 4th UK Special Interest Group for Computational Linguistics (CLUK4)*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

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

Patrick Schone and Dan Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 100–108.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Dominic Widdows, Beate Dorow, and Chiu-Ki Chan. 2002. Using parallel corpora to enrich multilingual lexical resources. In *Third International Conference on Language Resources and Evaluation*, pages 240–245, Las Palmas, Spain, May.

Dominic Widdows. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proc. of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*. (to appear).