Distributional Models of Preposition Semantics

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

Timothy Baldwin
CSLI
Stanford University
Stanford, CA 94305 USA
tbaldwin@csli.stanford.edu

Abstract

Prepositions are often considered to have too little semantic content or be too polysemous to warrant a proper semantic description. We first illustrate the suitability of distributional similarity methods for analysing preposition semantics by way of an inter-preposition similarity task, and make the claim that any semantic account of preposition semantics must be partially conditioned on valence. We further apply distributional similarity methods to a particle compositionality task.

Keywords: distributional hypothesis, verb particle construction, preposition semantics

1 Introduction

While nouns, verbs and adjectives have received considerable attention in terms of both lexical semantic language resource development (Ikehara et al. 1991; Mahesh 1996; Fellbaum 1998) and automatic ontology construction (Grefenstette 1994; Lin 1998b; Widdows & Dorow 2002), relatively little work has been done on creating resources for prepositions. Perhaps a large part of the reason for this is that the semantics of a transitive preposition can be bleached and determined largely by the semantics of the head noun it governs (e.g. *at last, on Wednesday, in question*: Pustejovsky (1995)) or its governing verb (e.g. *refer to, talk about*). However, many prepositions also have predicative usages (e.g. *time is up, the cheese is off, flairs are in*), and the semantics of peripheral PPs is determined largely by the preposition (e.g. *from March, in Toulouse, by the gate*). Accordingly, some account of preposition semantics seems unavoidable.

Past research on preposition semantics falls into two basic categories: large-scale symbolic accounts of preposition semantics, and disambiguation of PP sense. The most comprehensive lexical semantic account of English prepositions we are aware of is that of Dorr (1997), who classifies 165 English prepositions into 122 intransitive and 375 transitive senses using lexical conceptual semantics (LCS: Jackendoff (1983)). In a similar vein, Cannesson & Saint-Dizier (2002) developed a formal description of the semantics of 170 French prepositions, paying particular attention to their corpus usage. In contrast to these resource-development efforts, O'Hara & Wiebe (2003) targeted PPs in the context of sense disambiguation task, classifying PP tokens according to their case-role. The relative sparseness of research on preposition semantics can perhaps be explained by the perception that prepositions are both semantically vacuous and distributionally highly promiscuous, and consequently have a very low information content. This is most pronounced in bag-of-words tasks such as information retrieval where prepositions are generally listed in "stop word" lists for exclusion as index terms.

Our interest is in testing the viability of distributional methods in the derivation of a model of preposition semantics, working under the hypothesis that preposition semantics are stable enough

In Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, Toulouse, France, pp. 169-80.

that they can be classified accurately by distributional similarity techniques. Our approach here is based on the distributional hypothesis of Harris (1968) that similar words tend to occur in similar linguistic contexts. This observation has been used to explain various aspects of human language processing, from lexical priming (Lund et al. 1995) to retrieval in analogical reasoning (Ramscar & Yarlett 2003). It has also been employed in a range of natural language processing tasks, including word sense disambiguation (Schütze 1998) and automatic thesaurus construction (Lin 1998a). To our knowledge it has not previously been used to analyse the meaning of closed-class words.

As well as demonstrating the ability of similarity methods to capture intuitive correlations in the semantics of prepositions, we are interested in unearthing semantic anomalies between particles and transitive prepositions and motivating a valence-conditioned classification of English prepositions. **Intransitive prepositions** (Huddleston & Pullum 2002) (which we will interchangeably refer to as **particles**) are valence-saturated and occur most commonly as: (a) components of larger multiword expressions (notably **verb particle constructions**, or VPCs, such as *pick up*, *call in* and *chicken out*), (b) predicates (e.g. *time is* ***up****, flairs are* ***in***) or (c) prenominal modifiers (e.g. *the* ***up*** *escalator,* ***off*** *milk*). **Transitive prepositions**, on the other hand, select for NP complements to form prepositional phrases (PPs, e.g. ***at*** *home,* ***in*** *the end*). The bare term **preposition** is valence-underspecified.

It is relatively easy to find senses which are attested for only intransitive prepositions (e.g. the "hip/in fashion" sense of *in* above) and also uniquely transitive prepositions (e.g. *from*) which by definition do not have intransitive semantics. Of greater interest is the degree of correlation between intransitive and transitive preposition sense according to automatically-derived semantic classifications. That is, we seek to quantify the degree of semantic divergence between intransitive and transitive usages of different prepositions.

One piece of preliminary evidence which underlines the potential applicability of the distributional hypothesis to prepositions comes from the field of English part-of-speech (POS) tagging. All major POS tagsets[1] prefer to underspecify valence (e.g. there is no tag distinction between intransitive and transitive verbs), with the glaring exception of prepositions which are in all cases partitioned into intransitive and transitive instances. If there were a sharp demarcation in wordform between intransitive and transitive prepositions in English, this finding would perhaps not be surprising. However, a summary analysis of the written component of the British National Corpus (BNC, Burnard (2000)) reveals that while the type overlap between the two classes is only around 8%, the token overlap is roughly 70%. That is, roughly 70% of preposition token instances are potentially ambiguous between an intransitive and transitive usage. Given that taggers are able to deal with this ambiguity, generally using the immediate lexical context of a given preposition token, it would appear that intransitive and transitive usages of a given preposition are to some degree distributionally dissimilar. In this paper, we seek to confirm that this distributional dissimilarity correlates with semantic disparity, and at the same time determine whether semantically-related prepositions are distributionally similar.

The remainder of this paper is structured as follows. Section 2 looks at distributional similarity as a means of modelling simplex preposition semantics. Section 3 describes the application of such techniques in the analysis of verb particle construction compositionality. Finally, we conclude the paper in Section 4.

2 Inter-preposition similarity

We first consider the task of inter-preposition similarity, that is determination of the relative similarity of different preposition pairs. Below, we outline the procedure used to calculate preposition similar-

[1] By which we specifically refer to the International Corpus of English, Penn and various CLAWS tagsets.

$\mathbf{PREP_1}$	$\mathbf{PREP_2}$	
(A) all	(B) intransitive	(C) transitive
.304	.365	.386

Table 1: LCS-based evaluation of verb similarity results

ity, evaluate the method relative to a semantically-grounded preposition lexicon, and then provide a qualitative analysis of the results of the method over the preposition *up*.

2.1 Similarity methods

We took a knowledge-free approach to measuring distributional similarity, based on Latent Semantic Analysis (LSA, Deerwester et al. (1990)). Our technique is very similar to the approach taken to building a "context space" by Schütze (1998) . We measured the frequency of co-occurrence of our target words (the 20,000 most frequent words, with a set of 1000 "content-bearing" words (we used the 51st to the 1050th most frequent words, the 50 most frequent being taken to have extremely low infomation content). A target word was said to co-occur with a content word if that content word occurred within a window of 5 words to either side of it. In order to overcome data sparseness, we used Singular Value Decomposition (SVD) to reduce the dimensionality of the feature space from 1000 to 100. This limits each target word vector to 100 factors which reflect the patterns of association in the matrix, allowing relations to be discovered between target words even if there is not direct match between their context words. We used the various tools in the GTP software package, created at the University of Tennessee,[2] to build these matrices from the co-occurrence data and to perform SVD analysis.

The resulting representation is a 100-feature vector for each target word. Using this we can calculate the similarity between two terms by finding the cosine of the angle between their vectors.

As mentioned above, we distinguish prepositions according to valence, and seek to provide evidence for divergences in transitive and intransitive preposition semantics. This is achieved according to Methods PREP_1 and PREP_2, as detailed below. Here and for the remainder of the paper, we evaluate the methods over the written component of the BNC (90m words).

Method $\mathbf{PREP_1}$

First, we ran the above method over wordforms. With this method, we are thus unable to differentiate intransitive and transitive usages of a given preposition.

Method $\mathbf{PREP_2}$

Second, we ran our method including POS tags from the output of the RASP system (Briscoe & Carroll 2002), i.e. treating each wordform–POS tag pair as a single token. The RASP tagger is based on the CLAWS-4 tagset, and thus offers a fine grained distinction between different kinds of prepositions and particles. In extracting our context space we collapsed the different varieties of prepositions to give us one category for transitive prepositions and one for intransitive prepositions.

[2] `http://www.cs.utk.edu/~lsi/soft.html`

2.2 Quantitative evaluation of inter-preposition similarity

Quantitative evaluation of the two similarity methods was hampered by the fact that there is no established gold-standard resource for preposition semantics to use as a point of comparison. The only large-scale, publicly available resource we are aware of that provides a systematic account of preposition semantics is the LCS-based semantic lexicon of Dorr (1997).[3] Here, each preposition is classified into transitive and intransitive senses, each of which is described in the form of an LCS-based representation such as `(toward Loc (nil 2) (UP Loc (nil 2) (* Thing 6)))`, corresponding to the "up the stairs" sense of *up*$_{\text{TRANS}}$. Resnik & Diab (2000) propose a method for deriving similarities from LCS representations by: (1) decomposing them into feature sets, (2) calculating the information content $I(f)$ of each unit feature f based on the overall feature distribution, and (3) measuring the similarity between two LCS representations according to:

$$sim_{LCS}(e_1, e_2) = \frac{2 \times I(F(e_1)) \cap I(F(e_2))}{I(F(e_1)) + I(F(e_2))} \quad (1)$$

where e_1 and e_2 are lexicon entries, $F(e_i)$ is the decomposed feature set associated with e_i, and $I(F(e_i))$ is the information content of that feature set. Resnik & Diab define the similarity between two words to be the maximum value of $sim_{LCS}(e_1, e_2)$ over the cross product of all lexical entries for the words.

We can evaluate our distributional similarities according to their correlation with these LCS-derived similarities. We determine the correlation for three distinct datasets: (A) preposition similarity according to PREP_1 (with underspecification of valence); (B) particle similarity according to PREP_2; and (C) transitive preposition similarity according to PREP_2. In the case of (A), therefore, we calculate the distributional similarity of prepositions in the absence of POS information, and likewise do not distinguish between intransitive and transitive prepositions in the LCS lexicon. For (B) and (C), on the other hand, we consider only prepositions of fixed transitivity in both the BNC data and LCS lexicon.

For each of the three datasets, we calculated correlation according to Pearson's r, averaged over the nine prepositions *about, down, in, off, on, out, over, through* and *up*. The mean r values are given in Table 1. While the values are not high, the correlations for the intransitive and transitive preposition similarity tasks ((B) and (C), respectively) are higher than that for the (valence-underspecified) preposition similarity task, at a level of statistical significance (based on the two-tailed t-test, $p < .05$). This suggests that our model of preposition semantics is more stable when valence is specified, providing tentative support for the claim that preposition semantics are to some degree conditioned on valence.

Note that these results must be qualified by the observation that Resnik & Diab (2000) found only moderate correlation between the LCS-based similarities and human judgements in a small-scale verb similarity task. Having said this, our main interest is in the relative values and the finding that valence-specified models of distributional similarity are more stable then valence-underspecified models. We leave as an item for future research the determination of how well correlated the LCS-derived similarities are with human judgements on preposition similarity.

2.3 Qualitative evaluation of inter-preposition similarity

We qualitatively evaluate the different models of inter-preposition similarity by presenting in Table 2 the 10 most similar items to *up* based on methods (A), (B) and (C) from Section 2.2 (with respect to (valence-underspecified) *up*, *up*$_{\text{INTRANS}}$ and *up*$_{\text{TRANS}}$, respectively), and also: (D) *up*$_{\text{INTRANS}}$ vs. other intransitive and transitive prepositions according to PREP_2; and (E) *up*$_{\text{TRANS}}$ vs. other intransitive and

[3] `http://www.umiacs.umd.edu/~bonnie/AZ-preps-English.lcs`

$\mathbf{PREP_1}$		$\mathbf{PREP_2}$							
(A) *up* vs. all		(B) up_{INTRANS} vs. all_{INTRANS}		(C) up_{TRANS} vs. all_{TRANS}		(D) up_{INTRANS} vs. all		(E) up_{TRANS} vs. all	
out	0.996728	out_{INTRANS}	0.994579	$down_{\text{TRANS}}$	0.811990	out_{INTRANS}	0.994579	$down_{\text{TRANS}}$	0.811990
over	0.993389	$down_{\text{INTRANS}}$	0.986999	$along_{\text{TRANS}}$	0.793635	at_{TRANS}	0.992184	$along_{\text{TRANS}}$	0.793635
into	0.992472	on_{INTRANS}	0.979039	$away_{\text{TRANS}}$	0.772063	$into_{\text{TRANS}}$	0.990876	off_{INTRANS}	0.785038
at	0.991531	off_{INTRANS}	0.973902	$around_{\text{TRANS}}$	0.771186	$with_{\text{TRANS}}$	0.990123	$away_{\text{TRANS}}$	0.772063
through	0.990224	in_{INTRANS}	0.965397	$across_{\text{TRANS}}$	0.764267	on_{TRANS}	0.988517	$around_{\text{TRANS}}$	0.771186
with	0.990063	$over_{\text{INTRANS}}$	0.954401	off_{TRANS}	0.762640	$from_{\text{TRANS}}$	0.988196	$down_{\text{INTRANS}}$	0.768646
on	0.989493	$through_{\text{INTRANS}}$	0.941474	$behind_{\text{TRANS}}$	0.762197	to_{TRANS}	0.987719	$across_{\text{TRANS}}$	0.764267
before	0.989433	$about_{\text{INTRANS}}$	0.858488	$like_{\text{TRANS}}$	0.755297	for_{TRANS}	0.987178	off_{TRANS}	0.762640
after	0.988360	$across_{\text{INTRANS}}$	0.711517	$near_{\text{TRANS}}$	0.745630	$down_{\text{INTRANS}}$	0.986999	$behind_{\text{TRANS}}$	0.762197
to	0.988262			$beside_{\text{TRANS}}$	0.737918	in_{TRANS}	0.985999	$like_{\text{TRANS}}$	0.755297
back	0.987985			$past_{\text{TRANS}}$	0.735457	of_{TRANS}	0.984720	up_{INTRANS}	0.748533
about	0.987345			$into_{\text{TRANS}}$	0.726435	$after_{\text{TRANS}}$	0.982437	$near_{\text{TRANS}}$	0.745630
off	0.987056			$until_{\text{TRANS}}$	0.725211	$over_{\text{TRANS}}$	0.980920	$beside_{\text{TRANS}}$	0.737918
from	0.986760			$before_{\text{TRANS}}$	0.722251	$about_{\text{TRANS}}$	0.979719	$past_{\text{TRANS}}$	0.735457
down	0.986703			$inside_{\text{TRANS}}$	0.720766	by_{TRANS}	0.979224	in_{INTRANS}	0.732547
around	0.984672			$over_{\text{TRANS}}$	0.720766	on_{INTRANS}	0.979039	$over_{\text{INTRANS}}$	0.730410
in	0.984560			$after_{\text{TRANS}}$	0.718141	as_{TRANS}	0.978447	out_{INTRANS}	0.728146
by	0.979810			$through_{\text{TRANS}}$	0.711883	off_{INTRANS}	0.973902	$into_{\text{TRANS}}$	0.726435
away	0.979766			$towards_{\text{TRANS}}$	0.710432	$through_{\text{TRANS}}$	0.972225	$until_{\text{TRANS}}$	0.725211
without	0.975238			at_{TRANS}	0.704271	$before_{\text{TRANS}}$	0.972216	$before_{\text{TRANS}}$	0.722251

Table 2: Semantic neighbours of *up* with different transitivities

transitive prepositions according to PREP_2. Note that the most prevalent sense of up_{INTRANS} is the perfective, as in *eat up*.

What is most striking about Table 2 is the disparity of synonyms for the five different system configurations. In class (A), we get a fairly arbitrary set of prepositions due to the lack of valence to condition the semantics of the prepositions. In class (B), on the other hand, out_{INTRANS} is analysed as being very similar to up_{INTRANS}. We suggest that this is because, in addition to the directional sense of *out*, it also has a perfective sense (e.g. *print out*) similar to that of *up* (e.g. *finish up*). That is, the existence of an analogous sense to the two words, specific to intransitive usages, leads to the inflated similarity. In class (C), directional prepositions feature high in the ranking and there is little sign of any non-literal senses having influenced the results (with *out* at rank 35 with a similarity of .665). The results for class (D) overlap with those for (B), but we also get transitive prepositions with temporal senses interspersed amongst them. We suggest this is because of the frequent co-occurrence of temporal PPs with perfectives. Finally, the results for class (E) overlap with those for (C), but we this time get directional intransitive particles interspersed with the directional transitive prepositions. Interestingly, the directional prepositions are generally analogous to *up* in having PATH+GOAL semantics and serving to telicise motion verbs, suggesting that the distributional model is able to capture deep semantic consistencies between the prepositions (Dowty 1991; Jackendoff 1996).

Clearly, this is just one isolated example, but it illustrates the general process of semantic differentiation according to valence. Other prepositions which have differentiated semantics in the intransitive and transitive forms and where an analogous effect was observable were *on* and *out*. We take this as partial evidence for the need to include valence conditioning in any account of preposition semantics.

3 Verb Particle Compositionality

The second task to which we apply our models of preposition similarity is the analysis of verb-particle construction compositionality. **Verb-particle constructions** (VPCs hereafter) consist of a head verb

and an obligatory intransitive preposition.[4] Examples of VPCs are *put up*, *finish up*, *gun down* and *make out* as used in the following sentences:

(2) Peter put the picture up

(3) Susan finished up her paper

(4) Philip gunned down the intruder

(5) The couple made out

VPCs are relevant to the issue of preposition semantics because they display varying levels of semantic compositionality relative to the simplex semantics of the component verb and particle. Compare, for example, sentences (2) and (5). In (2), the meaning seems to be that Peter *put* the picture somewhere and that as a consequence the picture was *up*. That is, the verb and the particle make independent contributions to the sentence. If we take (5) we see a rather different situation. Neither Barbara nor Simon can be said to have *made* or to be *out*. In (4), by contrast, it is the particle that contributes its simplex meaning and not the verb. As a consequence of Philip's action the intruder is *down*, but since there is no simplex verb *to gun*, we would not say that anyone *gunned* or *was gunned*. In each case, verb and particle compositionality is reflected in the occurrences of predicates corresponding to simplex senses of each in the logical representation of the sentence (see Bannard (2002)).

For the purposes of this paper, we focus solely on particle compositionality.[5] In order to evaluate particle compositionality, we define it to be an entailment relationship between the whole VPC and the particle. With (2), e.g., it is true that *The picture was up* and the entailment holds. For (3), it is not true that either Susan or the paper were up, and the VPC therefore does not entail the particle. We make the assumption that these relationships between the component words of the VPC and the whole are intuitive to non-experts, and aim to use their "entailment" judgements accordingly. This use of entailment in exploring the semantics of verb and preposition combinations was first proposed by Hawkins (2000).

We are taking a rather simplified view of compositionality here. In sentence two above, we might want to argue (in line with much of the linguistic literature) that the particle is making an independent contribution in terms of aspect. However no such approach is fully robust, and for practical NLP purposes we are forced to adopt a rather straightforward definition of compositionality as meaning that the overall semantics of the VPC can be composed from the the simplex semantics of its parts, thereby ignoring construction specific meanings.

Below, we detail the annotation method used to elicit human judgements on particle compositionality, and also two methods for deriving a unique compositionality judgement for each VPC.

3.1 Eliciting human judgements on particle compositionality

Human annotators were asked to annotate a fixed set of VPCs for particle compositionality. In an attempt to normalise the annotators' entailment judgements, we decided upon an experimental setup where the subject is, for each VPC type, presented with a fixed selection of sentential contexts for that VPC. So as to avoid introducing any bias into the experiment through artificially-generated sentences, we chose to extract the sentential contexts from naturally-occurring text, namely the written component of the BNC.

[4] Strictly speaking, the particle can also take the form of an adjective (e.g. *cut short*) or verb (e.g. *let go*), but for the purposes of this paper, we consider intransitive prepositions to be the only type of particle.

[5] But see Bannard (2002) for a discussion of verb compositionality.

	Overall	*Verbs only*	*Particles only*
Agreement	.677	.703	.650
Kappa	.376	.372	.352
% Yes	.575	.655	.495
% No	.393	.319	.467
% Don't Know	.032	.026	.038

Table 3: Summary of judgements for all VPCs

Extraction of the VPCs was based largely on the method of Baldwin & Villavicencio (2002). First, we used a POS tagger and chunker (both built using fnTBL 1.0 (Ngai & Florian 2001)) to (re)tag the BNC. This allowed us to extract VPC tokens through use of: (a) the particle POS in the POS tagged output, for each instance of which we simply then look for the rightmost verb within a fixed window to the left of the particle, and (b) the particle chunk tag in the chunker output, where we similarly locate the rightmost verb associated with each particle chunk occurrence. Finally, we ran a stochastic chunk-based grammar over the chunker output to extend extraction coverage to include mistagged particles and also more reliably determine the valence of the VPC. The token output of these three methods was amalgamated by weighted voting.

The above method extracted 461 distinct VPC types occurring at least 50 times, attested in a total of 110,199 sentences. After partitioning the sentence data by type, we randomly selected 5 sentences for each VPC type. We then randomly selected 40 VPC types (with 5 sentences each) to use in the entailment experiment.

28 participants took part in the experiment, all of whom were native speakers of English. Each participant was presented with 40 sets of 5 sentences, where each of the five sentences contained a particular VPC. The VPC in question was indicated at the top of the screen, and they were asked two questions: (1) whether the VPC implies the verb, and (2) whether the VPC implies the particle. If the VPC was *round up*, e.g., the subject would be asked "Does *round up* imply *round*?" and "Does *round up* imply *up*?", respectively. They were given the option of three responses: "Yes", "No" or "Don't Know". In this paper, we focus exclusively on the particle compositionality results.

As with any corpus-based approach to lexical semantics, our study of VPCs is hampered by polysemy, e.g. *carry out*$_{\text{TRANS}}$ in the *execute* and *transport out (from a location)* senses.[6] Rather than intervene to customise example sentences to a prescribed sense, we accepted whatever composition of senses random sampling produced. Participants were advised that if they felt more that one meaning was present in a set of sentences, they should base their decision on the sense that had the greatest number of occurrences in the set.

The experiment was conducted remotely over the Web, using the experimental software package WebExp (Corley et al. 2000). Experimental sessions lasted approximately 20 minutes and were self-paced. The order in which the forty sets of sentences were presented was randomised by the software.

3.2 Conversion into gold-standard compositionality data

In order to measure how difficult the task is, we performed a pairwise analysis of the agreement between our 28 participants. The overall mean agreement was .655, with a kappa score (Carletta 1996) of .329. An initial analysis showed that two participants strongly disagreed with the other

[6]The effects of polysemy were compounded by not having any reliable method for determining valence. We consider that simply partitioning VPC items into intransitive and transitive usages would reduce polysemy signifi cantly.

VPC	*Particle*	*Particle compositional?*		
		Yes	*No*	*Don't Know*
get down	down	14	10	2
move off	off	19	7	0
throw out	out	15	10	1
pay off	off	16	8	2
lift out	out	26	0	0
roll back	back	14	12	0
dig up	up	18	7	1
lie down	down	25	1	0
wear on	on	3	22	1
fall off	off	25	1	0
move out	out	26	0	0
hand out	out	19	7	0
seek out	out	15	11	0
sell off	off	16	9	1
trail off	off	10	16	0
stay up	up	21	5	0
go down	down	22	3	1
hang out	out	25	1	0
get back	back	19	6	1
throw in	in	13	12	1
put off	off	5	19	2
shake off	off	15	11	0
step off	off	26	0	0
give off	off	21	5	0
carry away	away	6	18	2
throw back	back	21	4	1
pull off	off	13	6	7
carry out	out	0	25	1
brighten up	up	16	10	0
map out	out	10	16	0
slow down	down	19	7	0
sort out	out	11	15	0
bite off	off	16	8	2
add up	up	19	6	1
mark out	out	14	12	0
lay out	out	10	14	2
catch up	up	7	18	1
run up	up	13	10	3
stick out	out	15	11	0
play down	down	6	20	0

Table 4: Participant entailment judgements

participants, achieving a mean pairwise kappa score of less than .100. We decided therefore to remove these from the set before proceeding, resulting in a final mean agreement of .688. The overall results for the remaining 26 participants can be seen in Table 3. The kappa score over these 26 participants (.376) is classed as fair (0.2–0.4) and approaching moderate (0.4–0.6) according to Altman (1991).

3.3 Computational models of particle compositionality

Having created our gold-standard data, we set about implementing some statistical techniques for automatic analysis. In this, we use the VPC tokens with sentential contexts extracted from the BNC as reported in Section 3.1, i.e. a superset of the data used to annotate the VPCs.

The following sections describe four methods for modelling VPC compositionality, each of which is tested over the particle compositionality classification task. The results for each method are given in Table 5. Here, the baseline is the score obtained when we assign the majority class (particle compositional) to all items. Each method is evaluated in terms of (classification) accuracy, precision,

	Accuracy	*Precision*	*Recall*	*F-score*
Majority class	.750	.750	1.000	.857
VPC_1	.425	**.818**	.300	.442
VPC_2	.425	**.818**	.300	.442
VPC_3	.425	**.769**	.333	.480
VPC_4	.725	.758	.833	.793

Table 5: Results for the four methods over the particle compositionality classification task

recall and F-score ($\beta = 1$), and all values which exceed the baseline are indicated in **boldface**. Note the difference between accuracy and precision: accuracy is the ratio $\frac{TP+TN}{TP+FP+FN+TN}$ whereas precision is the ratio $\frac{TP}{TP+TN}$ (where TP = no. true positives, FP = no. false positives, FN = no. false negatives and TN = no. true negatives). That is, accuracy records the relative number of correct classifications by the classifier (whether positive or negative exemplars), whereas precision records the relative success of the classifier at positively-classifying data instances.

Method VPC_1

We decided to gain a sense of the start-of-the-art on the task by reimplementing the substitution-based technique described in Lin (1999) over VPCs. Lin's method is based on the premise that non-compositional items have markedly different distributional characteristics to expressions derived through synonym substitution over the original word composition. Lin took his multiword items from a pre-generated collocation database (Lin 1998b). For each collocation, Lin substituted each of the component words with a word with a similar meaning. The list of similar meanings was obtained by taking the 10 most similar words according to a corpus-derived thesaurus, the construction of which is described in Lin (1998b). For each item that resulted from substitution we found the mutual information (I), taking a collocation to consist of three events: the type of dependency relationship (A), the head lexical item (B), and the modifier (C). The mutual information then is the logarithm of the ratio between the probability of the collocation (where the probability space is all possible collocation triples) and the probability of the head, type and modifier occurring together, taking the head and modifier to be independent given the type:

$$I(A,B,C) = log\frac{P(A,B,C)}{P(B|A)P(C|A)P(A)} \tag{6}$$

Lin's initial observation, given these scores, is that "a phrase is probably non-compositional if such substitutions are not found in the collocation database, or their mutual information values are significantly different from that of the phrase" (p. 319).

In our implementation we replaced Lin's collocations with our VPCs, treating the relationship between a verb and a particle as a kind of grammatical relation. The thesaurus used by Lin has generously been made available online. However this is not fully adequate for our purposes since it includes only verbs, nouns and adjectives/adverbs. We therefore replicated the approach described in (Lin 1998a) to build the thesaurus, using BNC data and including prepositions.

Method VPC_2

VPC_2 is very similar to VPC_1, except that instead of using Lin's method, we derived our thesaurus using the knowledge-free distributional similarity method described in Section 2.1. For each term we then sorted all of the other particles in descending order of similarity, which gave us the thesaurus for

use in substitution. As with the Lin method, we performed substitutions by taking the 10 most similar items for the particle of each VPC.

Method VPC_3

Instead of assuming that an item formed by substitution should have a similar mutual information score to the original item, VPC_3 bases its compositionality judgement on the distributional similarity of original expression and word-substituted derivative expressions. The same method of substitution is used, with each component being replaced by each of its 10 nearest neighbours according to the LSA-based similarity measure described above. We judge a VPC item to be compositional if an expression formed by substitution occurs among the nearest 100 verb-particle items to the original, and failing this, we judge it to be non-compositional. We experimented with a number of cut-off points for identifying semantically similar items, and found that a value of 100 gave the best results.

Method VPC_4

VPC_4 takes a different method from the preceding three methods, in that it doesn't employ substitution. The underlying intuition is that identifying the degree of distributional similarity between a VPC and its particle might be a useful feature in distinguishing a compositional from a non-compositional VPC. That is, if a VPC can be shown to be semantically similar to its particle, then this could be a good indicator that that particle contributes simplex semantics. We again used the LSA-based semantic similarity measure for this purpose. We performed a pairwise comparison of all VPCs with all particles, obtaining cosine similarity scores for each pair.

In order to build a classifier for making compositionality decisions, we again used a neighbour-based approach with a cut-off. We said that a particle was contributing meaning to a VPC if it occurred in the 20 most similar items to the VPC. We tried out a range of different cut-offs for each item and found that this value gave the best results.

3.4 Particle compositionality results

The results in Table 5 show that all other than VPC_4 offer an improvement in precision over the baseline. VPC_1 and VPC_2 performed with relatively high precision but low recall. In contrast, VPC_4 outperformed the other methods in terms of both recall and F-score (nearly levelling with the baseline in F-score), but precision was slightly down.

Based on these results, we can conclude that it is, to a surprising degree, possible to estimate the particle compositionality of a given VPC through distributional similarity of the VPC with its particle in simplex prepositional form. This provides support for our hypothesis that preposition semantics are well-defined enough to be captured by distributional similarity techniques.

4 Conclusion

We have illustrated how distributional similarity methods can be used to successfully calculate inter-preposition similarity, and provided evidence for the valence-dependence of preposition semantics. More generally, we have furnished counter-evidence to the claim that prepositions are ill-suited to distributional similarity methods, in the form of the inter-preposition similarity task and also solid results over a particle compositionality classification task. Our hope is that this research will open the way to research on automatically-derived preposition thesauri to act as the catalyst in the development of preposition ontologies.

There is scope for this research to be extended in the direction of empirically-grounded evaluation of inter-preposition similarity, perhaps using human judgements as for the particle compositional-

ity task. We are also interested in the impact of dependency data on the semantic classification of prepositions. These are left as items for future research.

Acknowledgements

We would like to thank John Beavers, Aline Villavicencio and the two anonymous reviewers for their valuable input on this research. Timothy Baldwin is 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. Colin Bannard is supported by ESRC Grant PTA-030-2002-01740.

References

Altman, Douglas G.: 1991, *Practical Statistics for Medical Research*, Chapman and Hall.

Baldwin, Timothy & 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.

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

Briscoe, Ted & John Carroll: 2002, 'Robust accurate statistical annotation of general text', in *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, pp. 1499–1504.

Burnard, Lou: 2000, '*User Reference Guide for the British National Corpus*', Tech. rep., Oxford University Computing Services.

Cannesson, Emmanuelle & Patrick Saint-Dizier: 2002, 'Defining and representing preposition senses: A preliminary analysis', in *Proc. of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, USA, pp. 25–31.

Carletta, Jean: 1996, 'Assessing agreement on classification tasks: the kappa statistic', *Computational Linguistics*, **22**(2): 249–254.

Corley, Martin, Frank Keller & Christoph Scheepers: 2000, '*Conducting psychological experiments over the world wide web*', Unpublished manuscript, University of Edinburgh and Saarland University.

Deerwester, Scott, Susan Dumais, George Furnas, Thomas Landauer & Richard Harshman: 1990, 'Indexing by latent semantic analysis', *Journal of the American society for information science*, **41**(6): 391–407.

Dorr, Bonnie J.: 1997, 'Large-scale dictionary construction for foreign language tutoring and interlingual machine translation', *Machine Translation*, **12**(4): 271–322.

Dowty, David R.: 1991, 'Thematic proto-roles and argument selection', *Language*, **67**(3): 547–619.

Fellbaum, Christiane, ed.: 1998, *WordNet: An Electronic Lexical Database*, Cambridge, USA: MIT Press.

Grefenstette, Gregory: 1994, *Explorations in Automatic Thesaurus Extractions*, Kluwer Academic Publishers.

Harris, Zellig: 1968, *Mathematical Structures of Language*, New York, USA: Wiley.

Hawkins, John A.: 2000, 'The relative order of preposition phrases in English: Going beyond manner – place – time', *Language Variation and Change*, **11**: 231–266.

Huddleston, Rodney & Geoffrey K. Pullum: 2002, *The Cambridge Grammar of the English Language*, Cambridge, UK: Cambridge University Press.

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo & Hiromi Nakaiwa: 1991, 'Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–', in *Proc. of the Third Machine Translation Summit (MT Summit III)*, Washington DC, USA, pp. 101–106.

Jackendoff, Ray: 1983, *Semantics and Cognition*, Cambridge, USA: MIT Press.

Jackendoff, Ray: 1996, 'The proper treatment of measuring out, telicity and perhaps event quantification in English', *Natural Language and Linguistic Theory*, **14**: 305–54.

Lin, Dekang: 1998a, 'Automatic retrieval and clustering of similar words', in *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, Montreal, Canada.

Lin, Dekang: 1998b, 'Extracting collocations from text corpora', in *Proc. of the COLING-ACL'98 Workshop on Computational Terminology*, Montreal, Canada.

Lin, Dekang: 1999, 'Automatic identification of non-compositional phrases', in *Proceedings of the 37th Annual Meeting of the ACL*, College Park, USA, pp. 317–24.

Lund, Kevin, Curt Burgess & Ruth Ann Atchley: 1995, 'Semantic and associative priming in high-dimensional semantic space', in *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, Pittsburgh, USA, pp. 660–5.

Mahesh, Kavi: 1996, '*Ontology Development for Machine Translation: Ideology and Methodology*', Tech. Rep. MCCS-96-292, Computing Research Laboratory, NMSU.

Ngai, Grace & 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)*, Pittsburgh, USA, pp. 40–7.

O'Hara, Tom & Janyce Wiebe: 2003, 'Preposition semantic classification via Treebank and FrameNet', in *Proc. of the 7th Conference on Natural Language Learning (CoNLL-2003)*, Edmonton, Canada, pp. 79–86.

Pustejovsky, James: 1995, *The Generative Lexicon*, Cambridge, USA: MIT Press.

Ramscar, Michael & Dan Yarlett: 2003, 'Semantic grounding in models of analogy: An environmental approach', *Cognitive Science*, (27): 41–71.

Resnik, Philip & Mona Diab: 2000, 'Measuring verb similarity', in *Proc. of the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000)*, pp. 399–404.

Sch¨utze, Hinrich: 1998, 'Automatic word sense discrimination', *Computational Linguistics*, **24**(1): 97–123.

Widdows, Dominic & Beate Dorow: 2002, 'A graph model for unsupervised lexical acquisition', in *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, pp. 1093–9.