Size Matters.
Tight and Loose Context Definitions in English Word Space Models

Yves Peirsman, Kris Heylen and Dirk Geeraerts
Quantitative Lexicology and Variational Linguistics (QLVL)
University of Leuven
Leuven, Belgium
{yves.peirsman|kris.heylen|dirk.geeraerts}@arts.kuleuven.be

Abstract

Word Space Models use distributional similarity between two words as a measure of their semantic similarity or relatedness. This distributional similarity, however, is influenced by the type of context the models take into account. Context definitions range on a continuum from tight to loose, depending on the size of the context window around the target or the order of the context words that are considered. This paper investigates whether two general ways of loosening the context definition — by extending the context size from one to ten words, and by taking into account second-order context words — produce equivalent results. In particular, we will evaluate the performance of the models in terms of their ability (1) to discover semantic word classes and (2) to mirror human associations.

1 Introduction

In recent years, Word Space Models (Landauer and Dumais, 1997; Schütze, 1998; Padó and Lapata, 2007; Baroni et al., 2007) have become the standard NLP answer to any question concerning lexical semantics. Be it query expansion, automated essay rating, thesaurus extraction, word sense disambiguation or question answering, Word Space Models are readily applied to the task at hand. Their success almost makes us forget that the word space approach itself presents us with a number of questions. For instance: what kind of semantic relations are captured by these models? Is it semantic similarity — as between *car* and *truck* — or more topical relatedness — as between *car* and *road*? Moreover, what is the influence of all parameters involved — from the definition of context to the similarity measure used to compare the context vectors of two words? In this paper, we will focus on the precise definition of context that the models use and investigate its effect on the semantic relations that they find.

1.1 Word Space Models

In order to get at the semantic relatedness between two words, word space approaches model their use. They do so by recording in a so-called *context vector* the contextual features that each word co-occurs with in a corpus. For instance, first-order bag-of-word models simply keep track of the context words that appear within a context window of n words around the target (Gale et al., 1994; Levy and Bullinaria, 2001; Bullinaria and Levy, 2007). This implies that two words are similar when they often cooccur with the same context words. The tightest definition of context for bag-of-word models restricts itself to one word to the left and right of the target. Because this restriction may lead to data sparseness, it is often loosened in one of two ways: either the context window is stretched to a higher number of words around the target (Sahlgren, 2006), or the models take into account not the direct context words of the target, but the context words of these context words (Schütze, 1998). In this paper, we will investigate whether these two ways of loosening the context definition have the same influence on the results of the Word Space Models.

Without any doubt, enlarging the context window will change the type of features that the models are based on. With just one word to the left and the right of the target, an English noun will tend to have mostly adjectives and verbs as contextual features, for instance. Most of these context words will moreover be syntactically related to the target. If we extend the window size to five words, say, the noun's context vector will look very different. Not only are other nouns more likely to appear; the majority of words will not be in a direct syntactic relation to the target, but will merely be topically linked to it. We can expect this to have an influence on the type of semantic relatedness that the Word Space Models distinguish.

This effect of context has obviously been noted before. Sahlgren (2006) in particular observes that in the literature, all sorts of context sizes can be found, from fifty words to the left and right of the target (Gale et al., 1994) via fifty words in total (Schütze, 1998) to a mere three words (Dagan et al., 1993). Through a series of experiments, Sahlgren was able to confirm his hypothesis that large context windows tend to model syntagmatic — or topical — relations better, while small context windows are better geared towards paradigmatic — similarity or antonymy — relations. In a similar vein, we investigated the influence of several context definitions on the semantic characteristics of a wide variety of Word Space Models for Dutch (Peirsman et al., 2007; Peirsman, 2008). We found that syntactic models worked best for similarity relations, while first-order bag-of-word approaches modelled human associations better, among other things.

1.2 Research hypothesis

In line with Sahlgren (2006), our research hypothesis is that tight context windows will give better results for semantic similarity, while looser context windows will score higher with respect to more general topical relatedness. 'Loose' here refers to the use of a larger context window or of second-order context words.

We will test this hypothesis through a number of experimental tasks that have been released for the ESSLLI 2008 Lexical Semantics Workshop. First, section 2 will present the setup of our experiments. Section 3 will then discuss three word clustering tasks, in which the Word Space Models are required to discover semantic word classes. In section 4, we will investigate if the models are equally suited to model free associations. Finally, section 5 will wrap up with conclusions and an outlook for future research.

2 Experimental setup

The data for our experiments was the British National Corpus, a 100 million word corpus of British English, drawn from across a wide variety of genres, spoken as well as written. On the basis of this corpus, we constructed fourteen Word Space Models, seven first-order and seven second-order ones. Context size varied from 1 via 2, 3, 4, 5 and 7 to 10 words on either side of the target.

We reduced the dimensionality of the context vectors by treating only the 5,000 most frequent words in the BNC as possible features — a simple, yet popular way of dimensionality reduction (Padó and Lapata, 2007). Although working with all features could still improve performance (Bullinaria and Levy, 2007), we feel confident that cutting off at 5,000 dimensions has no direct influence on the relationships between the models, and the semantic relations they prefer. Semantically empty words in our stop list were ignored, and all words were lemmatized and tagged by their part of speech. In addition, we also used a cut-off that linearly increased with context size. For context size n, with n words on either side of the target word, we only took into account a feature if it occurred at least n times together with the target word. This variable cut-off keeps the number of non-zero cells in the word by feature matrices from exploding for the larger contexts.

The context vectors did not contain the frequency of the features, but rather their point-wise mutual information (PMI) with the target. This measure indicates whether the feature occurs together with the target more or less often than we would expect on the basis of their individual frequencies. Finally, the similarity between two context vectors was operationalized as the cosine of the angle they describe.

3 Task 1: Word Clustering

In Task 1, we tested the ability of our models to discover semantic classes for three types of words: concrete nouns, verbs, and a mixture of concrete and abstract nouns. The data sets and their sources are described on the website of the ESSLLI workshop.[1] The set of concrete nouns consisted of words like *hammer*, *pear* and *owl*, which our models had to cluster into groups corresponding to a number of semantic classes. The output was evaluated at three levels. The most fine-grained class distinctions were those between tools, fruit, birds, etc. — six clusters in total. Next, we checked the models' ability to recognize the differences between artifacts, vegetables and animals. Finally, animals and vegetables had to be combined into one natural category.

The second test set consisted of a mixture of concrete and abstract nouns — *truth* and *temptation* versus *hammer* and *eagle*, for instance. Here, the models were simply required to make the distinction between concrete and abstract — a task they were well capable of, as we will see.

The final test set contained only verbs. Again the models were evaluated several times. At the first stage, with nine clusters, we checked for the distinction between verb classes like *communication* (e.g., *speak*), *mental state* (e.g., *know*) and *body action* (e.g., *eat*). At the second stage, with five clusters, the categories were reduced to the likes of *cognition* and *motion*.

The vectors output by the models were clustered with the *repeated bisections* algorithm implemented in CLUTO (Karypis, 2003). This is a so-called partitional algorithm, which starts with one large cluster that contains all instances, and repeatedly divides one of its clusters in two until the requested number of clusters is reached. The resulting clusters are then evaluated against two measures: entropy and purity.

The entropy of cluster S_r of size n_r is defined as follows:

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \tag{1}$$

Here, q is the number of word classes in the data set, and n_r^i the number of words of class i in cluster r. As always, entropy expresses the *uncertainty* of a cluster — the degree to which it mixes up several categories. The lower the entropy, the better the cluster.

Purity, next, is the portion of the cluster taken up by the largest class in that cluster:

[1]http://www.wordspace.collocations.de/doku.php/esslli:start

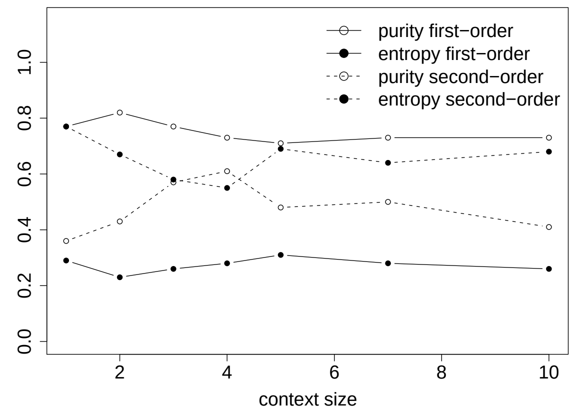

Figure 1: Performance of the Word Space Models in the 6-way concrete noun clustering task.

$$P(S_r) = \frac{1}{n_r}\max_i(n_r^i) \tag{2}$$

The higher the purity of a given cluster, the better. The entropy and purity values of the total solution are simply the sums of the individual cluster scores, weighted according to cluster size.

3.1 Results

By way of example, Figure 1 shows the performance of the models in the 6-way concrete noun clustering task. A number of observations we can make here apply to all results in this section. First, the purity and entropy of the models are almost perfect mirror images of one another. Second, the performance of the first-order models is clearly superior to that of the second-order ones. Purity lies considerably higher; entropy much lower. Third, our expectation that performance would decrease with larger contexts is not fully borne out. For the first-order models, the ideal context size seems to be two words on either side of the target. For the second-order models, it is four. This best second-order approach, however, gives results far lower than the least successful first-order wordspace. In the rest of this section we will therefore focus on the performance of the first-order models only. The results of the second-order approaches were invariably inferior and, because of this lack of quality, often hard to interpret.

Table 1 gives the performance of the first-order

	concrete nouns						concrete – abstract		verbs			
n	6		3		2		2		9		5	
	E	P	E	P	E	P	E	P	E	P	E	P
10	.26	.73	.54	.71	.97	.59	.18	.97	.44	.53	.41	.64
7	.28	.73	**.27**	**.86**	.97	.57	**.00**	**1.0**	**.41**	**.56**	**.39**	**.69**
5	.31	.71	.35	.82	.95	.61	**.00**	**1.0**	**.41**	**.56**	**.39**	**.69**
4	.28	.73	.54	.71	.96	.61	**.00**	**1.0**	.44	.51	**.39**	**.69**
3	.26	.77	.54	.71	.97	.59	**.00**	**1.0**	.42	**.56**	.54	.56
2	**.23**	**.82**	.34	.84	**.55**	**.86**	**.00**	**1.0**	.48	.47	.63	.56
1	.29	.77	.50	.75	.98	.57	**.00**	**1.0**	.42	.53	.51	.60

Table 1: Performance of the first-order Word Space Models in the word clustering tasks.

Word Space Models on the three clustering tasks, for each of the pre-specified numbers of clusters. It is hard to pin down an overall best context size: only the smallest and biggest windows under investigation never gave the best results. Let us first discuss the concrete noun clustering task. Here the systems were evaluated at three steps of their output. Their performance clearly deteriorates with each step. With six clusters, the most successful model is that with context size 2. It gives an average entropy of .23 and an average purity of .82. For the three-way clustering task, however, context size 7 unexpectedly gives the best results. We will see why this happens below. At the final evaluation stage, context size 2 is again distinctly in first position, as the only model that manages to come up with a decent clustering.

The division between concrete and abstract nouns, by contrast, is made much more easily. In fact, six out of seven first-order models are able to perfectly retrieve the two classes in the Gold Standard. The model with context size 10 makes a few mistakes here and there, but still finds a reasonable clustering. The verb clustering task, finally, seems to be of average difficulty. In general, intermediate context windows perform best.

3.2 Error analysis

Let us now take a closer look at the results. Again we start with the concrete noun subtask. At a first level, the models were required to distinguish between six possible classes. Broadly speaking all models here have the same three difficulties: (1) they are often not able to distinguish between vegetables and fruit, (2) they confuse some of the ground animals with birds, and (3) the tools are scattered among several clusters. Context size 1 makes a separate category for *screwdriver*, *chisel*, *knife* and *scissors*, for instance. The larger context sizes tend to put *spoon*, *bowl*, *cup* and *bottle* in a separate cluster, sometimes together with a number of animals or kinds of fruit. At the later stages, a hard core of artifacts seems to be easily grouped together, but the natural kinds (animals and fruit or vegetables) are still much harder to identify. Here and there a *kitchen* cluster that combines several types of tools, fruit and vegetables might be discerned instead of the Gold Standard grouping, but this is obviously open to interpretation.

The good performance of context size 2 in semantic similarity tasks has been observed before (Sahlgren, 2006). This is no doubt due to the fact that it combines useful information from a number of sources: a noun's adjectives, verbs of which it is the subject, and those of which it fills the object position. This last source of information is often absent from context size 1, at least when the noun is preceded by an article.

With three clusters, we observed that context size 7 suddenly outperforms this seemingly ideal configuration. This actually appears to be a question of chance. The main reason is that with six clusters, the model with context size 7 splits the ground animals and the birds evenly over two clusters. Because of their similarity, these are merged correctly afterwards. Context size 2 gives a far better classification early on, but at the next stage, it recovers less well from its few mistakes than context 7 does. It thus looks like the high performance of context 7 may partly be an artifact of the data set. Overall, context size 2 still seems the best choice for a classification task of concrete nouns.

Let us now turn to the verb clustering task. At the lowest level, the models were asked to produce nine clusters. The models with intermediate context sizes performed best, although the differences are small. This might be due to the fact that verb clustering benefits from information from a large number of arguments to the verb: subjects and objects as well as prepositional and adverbial phrases. Note that verb classification seems harder than the noun clustering tasks. The boundaries between the classes are indeed more subtle and fuzzy here. Differences, for instance, between *change location* (as in *move*), *motion direction* (as in *arrive*) or *motion manner* (as in *run*) are often too small to discover on a distributional basis.

In this analysis, we regularly mentioned syntactically related words as interesting sources of semantic information. We can therefore expect a model that takes into account these syntactic relations (Padó and Lapata, 2007; Peirsman, 2008) to outperform the simple bag-of-word approaches in these tasks. For the time being, such a model is outside the scope of our investigation, however.

4 Task 2: Free Associations

Of course, two words can also be related across semantic classes. *Doctor* is linked to *hospital*, for instance, even though the former refers to a human being and the latter to a building. Similarly, *car* and *drive* are associated, despite the fact they belong to different parts of speech. In this second task, we will try and investigate the degree to which our models are able to capture this type of semantic relatedness, by comparing their nearest neighbours for a target word with the results from a psycholinguistic experiment in which people were asked to give an association for each cue word they were presented with.

Both training and test sets consist of a number of *cue word – association* pairs. All words occurred in at least fifty BNC documents. It was now the general task of our Word Space Models to automatically find the associate for each cue word. This differs considerably from the previous task: whereas word clustering requires the Word Space Models only to consider the words in the test set, now they have to compare the targets with a far larger set of words. We chose to use the 10,000 most frequent words in the BNC as potential associates, including semantically empty words, plus those associates in the test set that did not survive the cut-off at 10,000 words. Even though the words in the training and test set were not tagged for part of speech, our Word Space Models did take these tags into account. Each cue word therefore automatically received its most frequent part of speech in the BNC.

For each of the cue words in the test set, we had the Word Space Models recover the 100 nearest neighbours, in the same way as described in section 2. Since this is an unsupervised approach, we ignored the training set and worked on the test set only. The performance of the models was expressed by the average rank of the association in the list of 100 nearest neighbours to the respective cue word. If the association did not appear in this list, it was automatically given rank 101. Obviously, the lower the score of the model, the better it is able to capture the type of semantic relatedness this task represents.

We also added a different type of algorithm to the experiment. Since we expected syntagmatic relations to play an important role in human associations, we investigated if simple co-occurrence statistics allow us to model the data better than the more advanced Word Space Models. We therefore computed the log-likelihood statistic between each cue word and all potential associates, within a context window of n words to the left and right of the cue. We then simply selected the 100 words with the highest log-likelihood scores.

4.1 Results

The results for the investigated models are presented in Figure 2. Because of the cut-off values, the coverage of our models was not always 100%. Context size 10, for instance, fails to come up with nearest neighbours for 7% of the words in the experiment. This is due to a slight inconsistency between our data and the Gold Standard. While we used a lemmatized version of the BNC, the words in the Gold Standard were not always lemmatized to the same base. A good example is *prepared*: in the lemmatized BNC, this is generally reduced to *prepare/VV*,

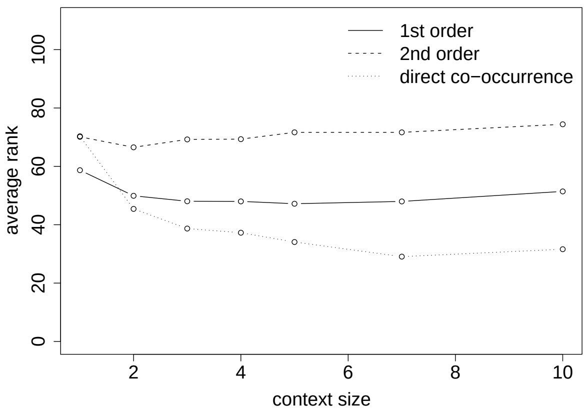

Figure 2: Performance of the Word Space Models in the free association task: average rank of association.

so that *prepared* as an adjective occurs very infrequently. If a cue word was not covered, the example automatically received rank 101.

A Friedman test confirms that there is indeed a statistical influence of the type of model on performance. Interestingly, the direct co-occurrence statistics clearly outperform the Word Space Models. When they take into account seven words to the left and right of the cue, they find the desired association at rank 30, on average. By contrast, the best first-order model (context size 5) only gives this association at a mean rank of 47, and the best second-order model performs even worse, with an average rank of 66.5 for context size 2. Moreover, the performance of the different context sizes seems to contradict our initial research hypothesis, which claimed that tight contexts should score better in the clustering task, while looser context windows should compare more favourably to free association norms. Tests for multiple comparisons after a Friedman test showed significant differences between the three types of models in the association task, but hardly any significant differences between the several context sizes. A detailed error analysis, however, adds some subtlety to this first impression.

4.2 Error analysis

To fully appreciate the outcome of the present task, we need to look at the results of the models in more detail. After all, semantically associated words come in many flavours. Some words may be associated to their cue because they are semantically similar, others because they are part of the same conceptual *frame*, still others because they represent typical collocations. This may explain the relatively low average ranks in Figure 2: each model could have its own preference for a specific type of association. It is therefore interesting to have a closer look at the precise associations that are recovered successfully by the different models.

Table 2 compares the results of the first-order model with context size 1 to those of the first-order model with context size 10. For both these models, it shows the twenty cue–association pairs with the highest gain in ranks, as compared to the other model. For instance, with a context size of 1, the associate of *hard* (*soft*) shows up 78 ranks higher than with a context size of 10. This last model, however, was able to recover the associate of *wave* (*sea*) at rank four — the first does not find it.

Interestingly, the nature of the associations for which the models display the highest difference in ranks, varies from one model to the other. The model with context size 1 tends to score comparably well on associations that are semantically similar to their target word. Many are (near-)synonyms, like *rapidly* and *quickly* or *astonishment* and *surprise*, others are antonyms, like *hard* and *soft* or *new* and *old*, while still others are in a IS-A relationship, like *cormorant* and *bird*. The associations for which the larger context window scores far better are generally of a completely different type. Here semantic similarity forms the exception. Most associations are topically related to their target word, either because they belong to the same conceptual *frame*, as with *reflection* and *mirror* or *spend* and *money*, or because they are typical collocates of their target word, like *twentieth* and *century* or *damsel* and *distress*. Of course, no clear line exists between the two categories, since frame-related words will often be collocates of each other.

This contrast is even more outspoken when we compare the first-order model with context size 1 to the best direct co-occurrence model. Among the association pairs recovered by the latter but not by the former are *wizard–oz*, *salvation–army* and *trafalgar–square*. This type of syntagmatic relatedness is indeed seldom modelled by the word spaces.

strengths of context size 1						strengths of context size 10					
cue	asso	diff	cue	asso	diff	cue	asso	diff	cue	asso	diff
melancholy	sad	100	glucose	sugar	63	sill	window	100	damsel	distress	97
rapidly	quickly	98	fund	money	61	riding	horse	100	leash	dog	96
plasma	blood	95	suspend	hang	61	reflection	mirror	100	consultant	doctor	95
astonishment	surprise	91	adequate	enough	54	nigger	black	100	pram	baby	94
joyful	happy	83	levi	jeans	49	hoof	horse	100	barrel	beer	94
hard	soft	78	sugar	sweet	46	holster	gun	100	twentieth	century	91
cormorant	bird	76	din	noise	44	dump	rubbish	100	handler	dog	90
new	old	70	no	yes	42	spend	money	98	scissors	cut	80
combat	fight	69	tumour	brain	39	bidder	auction	98	deck	ship	75
wrath	anger	64	weary	tired	33	wave	sea	97	suicide	death	72

Table 2: Top twenty cue words and associations for which either the first-order model with context size 1 or that with context size 10 scored better than the other.

Finally, when we put the first-order and second-order models with context size 1 side to side, it becomes more difficult to discern a clear pattern. Despite the fact that second-order context words are another way of loosening the definition of context, the second-order model with context size 1 still appears to have a preference for semantic similarity. In fact, word pairs like *companion–friend* and *chat–talk* are better covered here. As Figure 2 suggested, second-order models thus seem to follow the behaviour of the first-order approaches, even though they are consistently less successful.

Our findings so far are confirmed when we look at the parts of speech of the words that are recovered as nearest neighbours to a given cue word. Table 2 showed that for the smallest context window, these nearest neighbours tend to belong to the same part of speech as their cues. This does not hold for the models with larger context sizes. In fact, the table suggests that these sometimes even find nearest neighbours that typically appear as an argument of their cue. Nice examples are *dump–rubbish* or *spend–money*. We therefore calculated for each model the proportion of single nearest neighbours with the same part of speech as their cue. The results are given in Figure 3. It can clearly be seen that, as the context grows larger, the Word Space Models tend to find more neighbours with different parts of speech. For the first-order model with context size 1, 83% of the nearest neighbours have the same part of speech as their cue; for the model with context size 10, this figure has dropped to 58%. The second-order Word Space Models follow the behaviour of the first-order ones here. Not surprisingly, the algorithm that chooses associations on the basis of their log-likelihood score with the target shows the reverse pattern. The larger the co-occurrence span, the higher the chance of finding a word with the same part of speech.

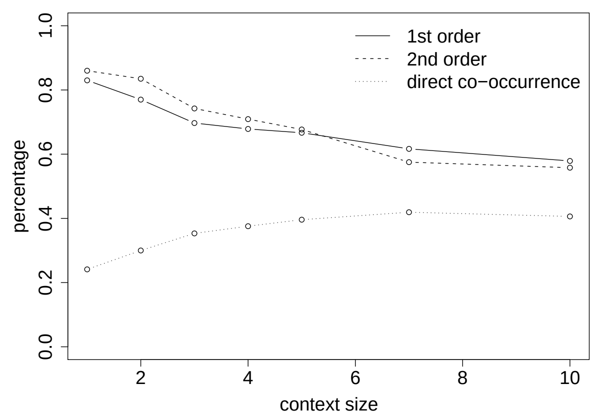

Figure 3: Percentage of nearest neighbours with same tag as their cue word in the free association task.

Overall, our findings demonstrate that human associations are a mixed bag of semantic similarity and topical relatedness. Models with small contexts better recover the former, those with large contexts have a preference for the latter.

5 Conclusions and future work

Word Space Models have become an indispensible tool in many computational-linguistic applications. Yet, the NLP community is only slowly gaining insight in the type of semantic information that gets modelled by these approaches, and how this information is influenced by the way the models operationalize the vague notion of context. While it has previously been shown that first-order bag-of-word models with small context sizes tend to best capture semantic similarity (Sahlgren, 2006), this paper is the first to compare two ways of loosening this context definition. In particular, we contrasted larger first-order context windows with second-order context models, which model the meaning of a word in terms of the context words of its context words, and evaluated them through two series of experiments.

Our findings can now be summarized as follows. (1) Overall, second-order bag-of-word models are inferior to their first-order competitors. Switching to second-order co-occurrence moreover does not lead to an increased preference for syntagmatic relations. (2) With respect to semantic similarity, a context window of size 2 gives the best results for noun clustering. For verbs, the context is better stretched to 4-7 words to the left and right of the target word. (3) Even though there is only a minor impact of context size on the overall performance in the free association task, small contexts display a preference for semantic similarity, while large contexts model syntagmatic relations better. However, the Word Space Models here are clearly outperformed by direct co-occurrence statistics.

Obviously, the Word Space Models under investigation allow for much more variation than we have been able to explore here. Syntactic models, for instance, certainly deserve further investigation, as in our papers on Dutch (Peirsman, 2008). Moreover, the question still remains why the second-order contexts, despite their poor performance generally, did score extremely well on a number of examples. Is this coincidental, or could there be a pattern to this set of cases? Either way, the intriguing variation across the results from the different Word Space Models justifies further research in the precise relationship between distributional and semantic relatedness.

References

Marco Baroni, Alessandro Lenci, and Luca Onnis. 2007. ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the ACL 2007 Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 49–56, East Stroudsburg, PA.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st International Conference on Association for Computational Linguistics (ACL-1993)*, pages 164–171, Columbus, OH.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1994. Discrimination decisions for 100,000-dimensional spaces. In A. Zampoli, N. Calzolari, and M. Palmer, editors, *Current Issues in Computational Linguistics: In Honour of Don Walker*, pages 429–450. Kluwer Academic Publishers.

George Karypis. 2003. CLUTO. A clustering toolkit. Technical Report 02-017, University of Minnesota, Department of Computer Science, Minneapolis, MN.

Thomas. K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Joseph P. Levy and John A. Bullinaria. 2001. Learning lexical properties from word usage patterns: Which context words should be used. In R.F. French and J.P. Sougne, editors, *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop*, pages 273–282. London: Springer.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Yves Peirsman, Kris Heylen, and Dirk Speelman. 2007. Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts. In *Proceedings of the CoSMO workshop, held in conjunction with CONTEXT-07*, pages 9–16, Roskilde, Denmark.

Yves Peirsman. 2008. Word space models of semantic similarity and relatedness. In *Proceedings of the 13th ESSLLI Student Session*, pages 143–152, Hamburg, Germany.

Magnus Sahlgren. 2006. *The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

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