Decorrelation and Shallow Semantic Patterns for Distributional Clustering of Nouns and Verbs

Yannick Versley
SFB 441
University of Tübingen
versley@sfs.uni-tuebingen.de

Abstract

Distributional approximations to lexical semantics are very useful not only in helping the creation of lexical semantic resources (Kilgariff et al., 2004; Snow et al., 2006), but also when directly applied in tasks that can benefit from large-coverage semantic knowledge such as coreference resolution (Poesio et al., 1998; Gasperin and Vieira, 2004; Versley, 2007), word sense disambiguation (McCarthy et al., 2004) or semantical role labeling (Gordon and Swanson, 2007).

We present a model that is built from Web-based corpora using both shallow patterns for grammatical and semantic relations and a window-based approach, using singular value decomposition to decorrelate the feature space which is otherwise too heavily influenced by the skewed topic distribution of Web corpora.

1 Introduction

It is well-established that human learning of lexical items beyond a certain point is driven by considering the *contexts* in which a word occurs, and it has been confirmed by McDonald and Ramscar (2001) that few occurrences of a word in informative contexts suffice to influence similarity judgements for marginally known words.

Computational models of word semantics based on this assumption are not only attractive for psychological modelling of language, but also for the purposes of automatic text processing, especially for applications where manual ontology construction would be infeasible or overly expensive, or to aid manual construction of lexical resources (cf. Kilgariff et al. 2004).

A common approach (Philips, 1985; Hindle, 1990) is to represent the context a word appears in by the words occurring in that context, weighting more heavily the context elements that co-occur more often than expected for random co-occurrences.

It is possible to group the approaches to use collocate features into two main areas:

- relation-free methods aim to directly use vectors of collocate words as a representation without distinguishing the relation between the target word and its collocates. Thus, related terms such as *doctor*, *hospital* and *treatment* which share many collocates, would be assigned a high similarity value.

- relation-based methods use collocate words together with grammatical relations, so that one noun being a frequent subject and another being a frequent object of a given word would not increase their similarity score – in the hospital example, a context like *the doctor treats the patient* would not contribute to the similarity value of *doctor* and *patient*.

Different methods of extracting word features will pick up different aspects of the denoted concept, from general topic, to sentiment, to ontologically relevant features such as exterior appearance.

In the remainder of this paper, I will start from the hypothesis that basing distributional similarity measures on context elements that are informative (in the sense that they implicitly or explicitly reflect the ontological principles of the targeted taxonomy) is preferable, and, by extension, that explicitly using syntactico-semantic relations yields better results.

2 Experimental Setting

To be useful in real-world tasks, both the size of the vocabulary and the size of the corpus should be large enough, as smaller samples would not contain enough contexts for many of the rarer words. This precludes approaches that rely on large numbers of search engine queries, such as the ones by Markert and Nissim (2005), Almuhareb and Poesio (2005), or Geleijnse and Korst (2006), as achieving significant coverage would necessitate an order of magnitude more effort than the (already very significant) weeks or months of running search engine queries that are necessary for a smaller sample.

On the other hand, the time consumption of full parsing means that approximate methods can be a better fit for processing very large corpora: Curran and Moens (2002) find that the rather large time that full parsing takes (even with a fast parser such as Lin's (1998b) MiniPar) can be reduced by using a reimplementation of Grefenstette's (1992) Sextant system for approximate parsing, which uses a chunker and considers simple neighbourhood relationships between chunks to extract compound, subject and object relations. Since the Sextant reimplementation only uses chunks, it is much faster (by a factor of 27), while the accuracy for the extracted relations is rather close to that of full parsing; Curran also remarks that a simple window-based approach is even faster and can still achieve good quality on large corpora, even though it is inferior to the syntax-based approaches.

In the following, we will explore the use of two large, Web-based datasets, namely UK-WaC (Ferraresi, 2007), as well as Google's n-gram database[1] for unsupervised noun and verb clustering, evaluated on the corresponding datasets proposed by the workshop organisers.

Besides a purely window-based approach, which we will present in section 4, we will present an approach that uses shallow patterns to approximate syntactic and semantic relationships, in section 3; even though some of the relations need more processing in different languages (most notably verb arguments, which are nontrivial to identify in languages with free word order such as German or Czech, or between compound parts in languages with synthetic compounds), we can show that this approach is not only computationally relatively inexpensive but also yields high-quality clustering results for verb clustering, where current approaches do not consider semantic relations at all.

[1] Thorsten Brants, Alex Franz (2006): Web 1T 5-gram Version 1, LDC2006T13

2.1 Relational Features for Nouns

Most older approaches to distributional similarity focus on syntactic relations, such as the compound noun, adjective-noun, subject and object relations that Grefenstette (1992) extract from his SEXTANT shallow parser, or the larger set of relations that Lin (1998a) extracts by full parsing.

Clustering words using such ontologically motivated patterns has been used by Evans (2003), who uses hypernymy patterns such as those popularised by Hearst (1992) to cluster named entities, and by Almuhareb and Poesio (2005), who use a pattern inspired by Berland and Charniak's (1999) to cluster nouns by their attributes. Using pattern search on the World Wide Web, Almuhareb and Poesio are able to achieve very good results. Some researchers such as Pantel et al. (2004) use supervised training to learn patterns corresponding to a single relation; going past single ontological relations, Baroni and Lenci (2008) use supervised learning of surface patterns corresponding to relations out of an inventory of 20 relations.

For our experiments, we used a combination of syntactic patterns targeting the same relations as Grefenstette (1992), variants of the hypernymy and meronymy-related patterns popularised by Hearst (1992) and Berland and Charniak (1999), respectively, as well as coordinate structures (X and/or Y); in contrast to Cederberg and Widdows (2003), we use second-order associations (regarding as similar terms which are coordinated with the same feature words) and do not see coordination as an indication for similarity of the conjuncts.

2.2 Relational Features for Verbs

Clustering and classification of verbs in the literature McCarthy (2000); Schulte im Walde and Brew (2002) often makes heavy use of information about argument structure, which is hard to come by without parsing; Stevenson and collaborators (Stevenson and Merlo, 1999; Joanis et al., 2007) use shallower

UK-Wac relation	entropy	purity
nv	0.209	0.818
vn^{-1}	0.244	0.750
jjn^{-1}	0.205	0.773
nn	**0.172**	**0.841**
nn^{-1}	0.218	0.795
cc:and	0.241	0.750
cc:and^{-1}	0.210	0.750
cc:or	0.203	0.767
cc:or^{-1}	0.200	0.795
Y's X	0.566	0.475
Y's X^{-1}	0.336	0.725
X of Y	0.437	0.655
X of Y^{-1}	0.291	0.750
Google n-grams relation	entropy	purity
of the	0.516	0.579
of the^{-1}	0.211	0.818
and other	0.237	0.744
and other^{-1}	0.458	0.632
such as	0.335	0.692
such as^{-1}	0.345	0.675

Table 1: Shallow patterns for nouns

features of which some do not necessitate parsed input, but they concentrate on verbs from three classes and it is not certain whether their features are informative enough for larger clustering tasks.

Schulte im Walde (2008) uses both grammatical relations output by a full parser and and part-of-speech classes co-occurring in a 20 word window to cluster German verbs. Comparing her clustering to gold standard classifications extracted from GermaNet (a German wordnet) and German FrameNet and another gold-standard using classes derived from human associations. She found that the different gold standards preferred different classes of grammatical relations: while GermaNet clustering results were best using subjects of nontransitive verb occurrences, FrameNet results were best when using adverbs, and the human association were best matched using NP and PP dependents on verbs.

In addition to syntactic correlates such as those investigated by Schulte im Walde (2008), we use several patterns targeted at more semantic relations.

Chklovski and Pantel (2004) extract 29,165 pairs of transitive verbs that co-occur with the same subject and object role, using Lin and Pantel's (2001) DIRT (Discovery of Inference Rules from text) approach, and then classify the relation between these verbs into several relations using Web patterns indicating particular relations (*similarity*, *strength*, *antonymy*, *enablement*, and *succession*.

Besides detecting conjunctions of verbs (allowing other words in between, but requiring the part-of-speech tags to match to exclude matches like "see how scared I *was* and *started* to calm me"), and capturing general within-sentence co-occurrence of verbs, we also tried to capture discourse relations more explicitly by limiting to certain discourse markers, such as *that*, *because*, *if*, or *while*.

3 Clustering Results

To determine the weight for an association in the vector calculated for a word, we use the pointwise mutual information value:

$$mi^{+}(w_1, w_2) = \max\left(0, \log \frac{p(X = w_1 | Y = w_2)}{p(X = w_1)}\right)$$

We then use the vectors of mi^{+} values for clustering in CLUTO using repeated bisecting k-means with cosine similarity.[2]

For the nouns, we use a the last noun before a verb as an approximation of subjecthood (`vn`), the next head noun as an approximation for direct objects (`nv`), as well as adjective modifiers (`jjn`), and noun compounds (`nn`) on UK-WaC using the provided lemmas. Using Berland and Charniak's patterns A and B (Y's X, X of Y) on UK-WaC, we found that a surface string search (using Minnen et al.'s (2001) morphological analyser to map word forms to their lemmas) on the Google n-gram dataset gave superior results. We used the same surface string search for Hearst's *X and other Ys* and *Ys such as X* patterns (restricting the "Ys" part to plural nouns to improve the precision). As the Hearst-style patterns are relatively rare, the greater quantity of data from the Google n-grams outweighs the drawback of having no part of speech tagging and only approximate lemmatisation.

Both on UK-WaC and on Google's n-gram dataset, we find a stark asymmetry in the clusterings

[2]Note that the resulting clusters can vary depending on the random initialisation, which means that re-running CLUTO later can result in slightly better or worse clustering.

UK-Wac relation	entropy	purity
nv^{-1}	0.398	0.556
vn	0.441	0.511
rv^{-1}	**0.342**	**0.622**
vi	0.397	0.556
vv	0.423	0.533
vv^{-1}	0.378	0.556
that	0.504	0.467
that^{-1}	0.479	0.489
because	0.584	0.378
because^{-1}	0.577	0.400
if	0.508	0.444
if^{-1}	0.526	0.444
while	**0.477**	**0.511**
while^{-1}	0.502	0.444
by Xing	0.488	0.489
by Xing^{-1}	0.380	0.600
then	0.424	0.533
then^{-1}	0.348	0.600
cc:and	0.278	0.711
cc:and^{-1}	0.329	0.622
cc:or	**0.253**	**0.733**
cc:or^{-1}	0.323	0.667

Table 2: Shallow patterns for verbs

of meronymy patterns, probably due to the fact that parts or attributes provide useful information, but the nouns in the evaluation set are not meaningful parts of other objects.

Considering the verbs, we found that a preceding adverb (rv) provided the most useful information, but other patterns, such as subject-verb (nv), and verb-object (vn), as well as using the following preposition (vi) to approximate the distribution of prepositional modifiers of the verb, give useful results, as much as the following verb (vv), which we used for a very rough approximation of discourse relations. Using verbs linked by subordinate conjunctions such as *if*, *that*, or *because*, performs comparatively poorly, however.

A third group of patterns is inspired by the patterns used by Chklovski and Pantel (2004) to approximate semantic relations between verbs, namely *enablement* relations expressed with gerunds (linking the previous verb with the gerund in sentences such as "Peter *altered* the design by *adding* a green button"), *temporal succession* by relating any verb that is modified by the adverb *then* with its preceding verb, and *broad similarity* by finding pairs of coordinated verbs (i.e., having a coordination between them and marked with the same part-of-speech tag).

Noun compounds for nouns and preceding adverbs for verbs already give slightly better clusterings than an approach simply considering words co-occurring in a one-word window (see table 3), with coordination and some of the semantic patterns also yielding results on par with (for nouns) syntactic relations.

4 Window-based approach with decorrelation

As reported by Curran and Moens (2002), a simple cooccurrence-window-based approach, while inferior to approaches based on full or shallow parsing, is amenable to the treatment of much larger data quantities than parsing-based approaches, and indeed, some successful work such as Rapp (2003) or Ravichandran et al. (2005) does not use syntactic information at all.

In this section, we report the results of our approach using window-based cooccurrence on Google's n-gram dataset, using different weighting functions, window sizes, and number of feature words. As a way to minimize the way of uninformative collocates, we simply excluded the 500 most frequent tokens for use as features, using the next most frequent N words (for N in 8k, 24k, 64k, 512k).

Besides the positive mutual information measure introduced earlier, we tried out a simple logarithmic weighting function:

$$Log(w_1, w_2) = \log(1 + C(w_1, w_2))$$

(where $C(w_1, w_2)$ is the raw count for w_1 and w_2 co-occurring in a window), and the entropy-weighted variant used by Rapp (2003):

$$LogEnt(w_1, w_2) = \log(1+C(w_1, w_2))\cdot H(X|Y{=}w_2)$$

This weighting function emphasizes features (i.e., values for w_2) which co-occur with many different target words.

Generally, we found that the window-based approach gave the best results with mutual information weighting (with clustering entropy values for verbs between 0.363, for using 8k features with a

window size of 1 word around the target word, and 0.504, for using 512k features with a window size of 4) than for the other methods (which yielded entropy values between 0.532, for 64k features with a window of 2 words and logarithmic weighting and 0.682, for 8k features with a window size of 4 words and log-entropy weighting). This difference is statistically very significant ($p < 0.0001$ for a paired t-test between mi^+ and Log over combinations of three different window sizes and four different vocabulary sizes).

To see if singular value decomposition would improve the clustering results, we collected co-occurrence vectors for the clustering target verbs in addition to a collection of frequent verbs that we obtained by taking the 2000 most frequent verbs or nouns and eliminating verbs that correspond to very frequent noun forms (e.g., to machine), as well as all non-nouns (e.g. gonna), arriving at a set of 1965 target verbs, and 1413 target nouns, including the items to be clustered. Even though using this larger data set makes it more difficult to experiment with larger feature spaces, we saw the possibility that just using the words from the data set would create an artificial difference from the transformation one would get using SVD in a more realistic setting and the transformation obtained in the experiment.

Using singular value decomposition for dimensionality reduction only seems to have a very small positive effect on results by itself: using mutual information weighting, we get from 0.436 to between 0.408 (for 100 dimensions), with other weighting functions, dimensionality values, or vocabulary sizes perform even worse.

This is in contrast to Rapp (2003), who achieved vastly better results with SVD and log-entropy weighting than without in his experiments using the British National Corpus, and in parallel to the findings of Baroni and Lenci (2008), who found that Rapp's results do not carry over to a web-based corpus such as UK-WaC. Looking at table 4, we find it plausible that the window-based approach tends to pick up topic distinctions instead of semantic regularities, which gives good results on a carefully balanced corpus such as the BNC, but drowns other information when using a Web corpus with a (typically) rather biased topic distribution.[3]

Examining the singular vectors and values we get out of the SVD results, we find that the first few singular values are very large, and the corresponding vectors seem to represent more a topic distinction than a semantic one. Parallel to this, the results for the SVD of log-weighted data is plateauing after the first few singular vectors are added, quite possibly due to the aforementioned drowning of information by the topical distinctions. To relieve this, we altered the size of singular values before clustering, either by taking the square root of the singular values, which has the effect of attenuating the effect of the singular vectors with large values, or by setting all singular values to 1, creating a feature space that has a spherically symmetric data distribution (usually referred to as decorrelation or whitening). As can be seen in figure 1, decorrelation yields clearly superior results, even though they are clearly much noisier, yielding wildly varying results with the addition of just a few more dimensions. For the decorrelated vectors, we find that depending on the other parameters, positive mutual information is either significantly better ($p \approx 0.0001$ for paired t-test over results for different dimension numbers with a window size of 1 and 8k features), or insignificantly worse ($p \approx 0.34$ for a window size of 2 and 24k features). We attribute the fact that the best clustering result for the window-based approach was achieved with log-entropy weighting to the fact that the log and log-entropy based vectors are noisier and have more variance (with respect to number of dimensions), thus possibly yielding artifacts of overfitting the small test data set; however, further research will be necessary to confirm or deny this.

5 Results and Discussion

To get a better clustering than would be possible using single features, we tried combinations of the most promising single features by first normalizing the individual feature vectors by their L_p norm,

[3] Cf. table 4: besides the first two vectors, which seem to identify frequency or content/navigation distinction, the second and third singular vector are clearly influenced by dominant web genres, with a pornography vs. regulatory documents axis for v2 and a Unix/programming vs. newswire documents axis for vector v3.

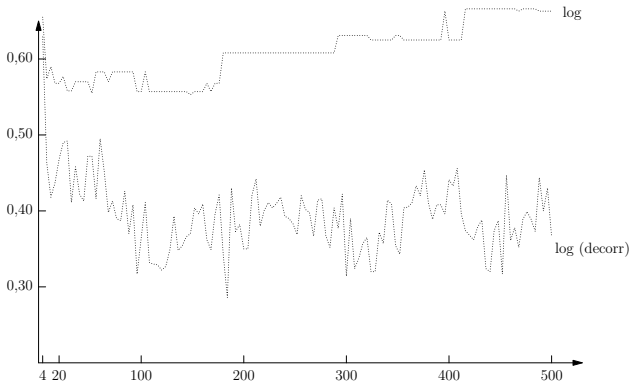

Figure 1: Influence of decorrelation on clustering quality (4-word window, 8k features)

Clustering Entropy in verb clustering vs. number of dimensions; lower is better

noun clustering relation	entropy	purity
win(1), 64k features, mi^+	0.221	0.818
best SVD+decorrelation	0.196	0.795
nn	0.172	0.841
cc:or^{-1}	0.200	0.795
nv^{-1}+jjn^{-1}+and other, 7cl.	**0.034**	**0.977**
verb clustering relation	**entropy**	**purity**
win(1), 1M features, mi^+	0.376	0.600
best SVD+decorrelation	0.280	0.711
rv^{-1}	0.342	0.622
cc:or	0.253	0.733
cc:and+then^{-1}	**0.218**	**0.778**

Table 3: Results overview

for $p = 1.5$.[4] We then concatenate the normalized vectors for the different relations to get the vector used in clustering. As can be seen in table 3, the window-based approach comes near the best results for a single syntax-based pattern, whereas the semantically motivated patterns work better than either syntactic patterns or the window-based approach. The best combinations we found involve several of the semantically motivated patterns and, in the case of nouns, also informative syntactic relations the key seems to be that the different relation focus on complementary aspects of the classes. While the decorrelation-based approach is an improvement over a simpler window-based approach, it does not seem possible to get much larger improvements; however, it should be said that both window size and feature space were constrained due to limitations of the Google n-gram data on one hand and memory limitations on the other.

The resulting clusters generally seem rather sensible, although they sometimes incorporate distinctions that are slightly different from those in the gold standard: in many clusterings, the class of birds and ground animals are split according to a different pattern, e.g. domestic and wild animals. Some other divisions are very consistently found in all clusterings: Even in the best clustering, artifacts are split into a container-like group including *bottle, bowl, cup* and others, and a handle-like artifact group including *chisel, hammer, screwdriver*, and fruits and vegetables are merged into one group unless the number of clusters is increased to seven. *chicken* also seems to be consistently misclustered as a cooking ingredient rather than an animal.

For the verbs, the communication verbs are split into the non-directive verbs *read*, *speak* and *talk*, which are clustered with two mental state verbs which are less action-focused, *know* and *remember*, as well as *listen*, which the gold standard categorizes as a body sense verb, whereas the more directive communication verbs *request* and *suggest* are clustered together with the more action-focused mental state verbs *check* and *evaluate*, and *repair*, which the gold standard groups with the state change verbs (*break*, *die*, *kill*).

6 Outlook

We presented two approaches for using distributional statistics extracted from large Web-based corpora to cluster nouns and verbs: one using shallow patterns to extract syntactically and semantically motivated relations, and the other using a small window size together with Google's n-gram dataset, showing how manipulating the SVD-transformed representation helps overcome problems that are due to the skewed topic distribution of Web corpora. We also showed how multiple relations can be combined to arrive at high-quality clusterings that are better

[4] The Lebesgue norm $L_p = \left(\sum |x_i|^p\right)^{1/p}$ has the euclidean norm L_2 as a special case. For $1 \leq p < 2$, the L_p-norm is larger than the euclidean norm if there are multiple non-zero values in a vector; we think that normalizing by the $L_{1.5}$ norm rather than L_2 norm has the beneficial effect of slightly emphasizing relations with a smaller feature space.

v0: $\lambda = 56595$		v1: $\lambda = 2043.5$		v2: $\lambda = 2028.7$		v3: $\lambda = 1760.5$	
fundraise	*0.0000	ensure	*-9999.99	f–ck	a–s	configure	src
exhilarate	*Reserved	determine	*Verzeichnis	suck	p–ssy	filter	header
socialize	*Advertise	process	*-99	*amend	*pursuant	*accuse	*father
pend	*Cart	identify	*-999	*comply	*Agreement	*murder	*whom

Table 4: Singular vectors for the largest singular values (8k features, 4-word window)
Most important target verbs (left) and features (right), starred words have a negative weight in the vector. Some explicit words in vector 2 have been redacted by replacing middle letters with a dash.

noun clusters		
banana	cat	bottle[1]
cherry	cow	bowl[1]
pear	dog	kettle[1]
pineapple	elephant	pencil[1]
chisel[1]	lion	pen[1]
hammer[1]	pig	spoon[1]
knife[1]	snail	telephone[1]
scissors[1]	turtle	
screwdriver[1]		
duck	*chicken*	boat
eagle	corn	car
owl	lettuce	helicopter
peacock	mushroom	motorcycle
penguin	onion	ship
swan	potato	truck
verb clusters		
breathe	drive	carry
cry	fly	pull
drink	ride	push
eat	run	send
	walk	
acquire	break	feel
buy	destroy	look
lend	die	notice
pay	kill	smell
sell	*fall*	*smile*
check[2]	know[2]	arrive
evaluate[2]	remember[2]	enter
repair	*listen*	leave
request[3]	read[3]	rise
suggest[3]	speak[3]	*move*
	talk[3]	*forget*

Table 5: Resulting verb and noun clusters
(Each cluster is one column. Italicized items are the only members of their class in the cluster)

than would be possible using either single relations or the best results achieved using the window-based approach.

Several open questions remain for future research: One would be the use of supervised learning approaches to perform automatic weighting and/or acquisition of patterns. The other one would be a question of how these approaches can be scaled up to the size needed for real-world applications. While the most important issue for the window-based approach is the use of Singular Value Decomposition, which scales poorly with both the size of the dataset due to nonlinear growth of computation time as well as memory consumption, the relation-based approach may suffer from data sparsity when considering rare words, especially using the rarer semantic relations; however, an approach like the ones by Snow et al. (2006) or Baroni and Lenci (2008) that is able to learn patterns from supervised training data may solve this problem at least partially.

Acknowledgements I am grateful to the two anonymous reviewers for helpful comments on an earlier version of this paper. The research reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of Collaborative Research Centre (Sonderforschungsbereich) 441 "Linguistic Data Structures".

References

Almuhareb, A. and Poesio, M. (2005). Finding concept attributes in the web. In *Proc. of the Corpus Linguistics Conference*.

Baroni, M. and Lenci, A. (2008). Concepts and word spaces. *Italian Journal of Linguistics*, to appear.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of ACL-1999*.

Cederberg, S. and Widdows, D. (2003). Using LSA and noun coordination information to improve the preci-

sion and recall of automatic hyponymy extraction. In *Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*.

Chklovski, T. and Pantel, P. (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *Proc. EMNLP 2004*.

Curran, J. and Moens, M. (2002). Scaling context space. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Evans, R. (2003). A framework for named entity recognition in the open domain. In *RANLP 2003*.

Ferraresi, A. (2007). Building a very large corpus of English obtained by Web crawling: ukWaC. Master's thesis, Università di Bologna.

Gasperin, C. and Vieira, R. (2004). Using word similarity lists for resolving indirect anaphora. In *ACL'04 workshop on reference resolution and its applications*.

Geleijnse, G. and Korst, J. (2006). Learning effective surface text patterns for information extraction. In *Proceedings of the EACL 2006 workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

Gordon, A. S. and Swanson, R. (2007). Generalizing semantic role annotations across syntactically similar verbs. In *ACL 2007*.

Grefenstette, G. (1992). Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis. In *ACL Student Session 1992*.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING 92)*.

Hindle, D. (1990). Noun classification from predicate argument structures. In *Proceedings of the 28th annual meeting of the Association for Computational Linguistics*.

Joanis, E., Stevenson, S., and James, D. (2007). A general feature space for automatic verb classification. *Natural Language Engineering*, forthcoming:1–31.

Kilgariff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The sketch engine. In *EURALEX 2004*.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proc. CoLing/ACL 1998*.

Lin, D. (1998b). Dependency-based evaluation of Minipar. In *Workshop on the Evaluation of Parsing Systems*.

Lin, D. and Pantel, P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Markert, K. and Nissim, M. (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–402.

McCarthy, D. (2000). Using semantic prefernces to identify verbal participation in role switching alternations. In *Proc. NAACL 2000*.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant word senses in untagged text. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279, Morristown, NJ, USA. Association for Computational Linguistics.

McDonald, S. and Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proc. 23rd Annual Conference of the Cognitive Society*.

Minnen, G., Caroll, J., and Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards terascale knowledge acquisition. In *Proc. Coling 2004*.

Philips, M. (1985). *Aspects of Text Structure: An investigation of the lexical organization of text*. Elsevier, Amsterdam.

Poesio, M., Schulte im Walde, S., and Brew, C. (1998). Lexical clustering and definite description interpretation. In *AAAI Spring Symposium on Learning for Discourse*.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proc. Ninth Machine Translation Summit*.

Ravichandran, D., Pantel, P., and Hovy, E. (2005). Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In *Proc. ACL 2005*.

Schulte im Walde, S. (2008). Human associations and the choice of features for semantic verb classification. *Research on Language and Computation*, to appear.

Schulte im Walde, S. and Brew, C. (2002). Inducing german semantic verb classes from purely syntactic subcategorization information. In *Proc. ACL 2002*.

Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL 2006*.

Stevenson, S. and Merlo, P. (1999). Automatic verb classification using grammatical features. In *Proc. EACL 1999*.

Versley, Y. (2007). Antecedent selection techniques for high-recall coreference resolution. In *Proc. EMNLP 2007*.