What's in a name?
The automatic recognition of metonymical location names.

Yves Peirsman
Quantitative Lexicology and Variational Linguistics
University of Leuven, Belgium
yves.peirsman@arts.kuleuven.be

Abstract

The correct identification of metonymies is not normally a problem for most people. For computers, things are different, however. In Natural Language Processing, metonymy recognition is therefore usually addressed with complex algorithms that rely on hundreds of labelled training examples. This paper investigates two approaches to metonymy recognition that dispense with this complexity, albeit in different ways. The first, an unsupervised approach to Word Sense Discrimination, does not require any labelled training instances. The second, Memory-Based Learning, replaces the complexity of current algorithms by a 'lazy' learning phase. While the first approach is often able to identify a metonymical and a literal cluster in the data, it is the second in particular that produces state-of-the-art results.

1 Introduction

In the last few years, metonymy has emerged as an important focus of research in many areas of linguistics. In Cognitive Linguistics, it is often defined as "a cognitive process in which one conceptual entity, the vehicle, provides mental access to another conceptual entity, the target, within the same domain, or idealized cognitive model (ICM)" (Kövecses, 2002, p.145). In example (1), for instance, *China* and *Taiwan* provide mental access to the governments of the respective countries:

(1) *China* has always threatened to use force if *Taiwan* declared independence. (BNC)

This paper is concerned with algorithms that automatically recognize such metonymical country names. These are extremely relevant in Natural Language Processing, since any system that automatically builds semantic representations of utterances needs to be able to recognize and interpret metonymical words.

Early approaches to metonymy recognition, such as Pustejovsky's (1995), identified a word as metonymical when it violated certain selectional restrictions. Indeed, in example (1), *China* and *Taiwan* both violate the restriction that *threaten* and *declare* require an animate subject, and thus have to be interpreted metonymically. This view is present in the psycholinguistic literature, too. Some authors argue that a figurative interpretation of a word typically comes about when all literal interpretations fail; see Gibbs (1994) for an overview. This failure is often due to the violation of selectional restrictions.

However, in psycholinguistics as well as in computational linguistics, this approach has lost much of its appeal. It has become clear to researchers in both fields that many metonymies do not violate any restrictions at all. In *to like Shakespeare*, for instance, there is no explicit linguistic trigger for the metonymical interpretation of *Shakespeare*. Rather, it is our world knowledge that pre-empts a literal reading of the author's name. Examples like this one demonstrate that metonymy recognition should not be based on rigid rules, but rather, on information about the semantic class of the target word and the semantic and grammatical context in which it occurs. In psycholinguistics, this insight (among others) has given rise to theories claiming that a figurative interpretation does not follow the failure of a literal one, but that both processes occur in parallel (Frisson and Pickering, 1999). In computational linguistics, it has led to the development of statistical, corpus-based approaches to metonymy recognition.

This view was first put into computational practice by Markert and Nissim (2002a). Their key to success was the realization that metonymy recognition is a sub-problem of Word Sense Disambiguation (WSD). They found that most metonymies in the same semantic class belong to one of a limited number of metonymical patterns that can be defined a priori. The task of metonymy recognition thus consists of the automatic assignment of one of these readings to a target word. Since all words in the same semantic class may undergo the same semantic shifts, there only has to be one classifier per class (and not per word, as in classic WSD).

In this paper I will be concerned with the automatic identification of metonymical location names. More particularly, I will test two new approaches to metonymy recognition on the basis of Markert and Nissim's (2002b) corpora of 1,000 mixed country names and 1,000 instances of the country name *Hungary*.[1] The most important metonymical patterns in these corpora are `place-for-people`, `place-for-event` and `place-for-product`. In addition, there is a label `mixed` for examples that have two readings, and `othermet` for examples that do not belong to any of the pre-defined metonymical patterns.

On the mixed country data, Nissim and Markert's (2003) classifiers achieved an accuracy of 87%. This was the result of a combination of both grammatical and semantic information. Their grammatical information included the function of a target word and its head. The semantic information, in the form of Dekang Lin's (1998) thesaurus of semantically similar words, allowed the classifier to search the training set for instances whose head was similar, and not just identical, to that of a test instance.

Markert and Nissim's (2002a) and Nissim and Markert's (2003) study is the only one to approach metonymy recognition from a data-driven, statistical perspective. However, it also has a number of disadvantages. First, it requires the annotation of a large number of training and test instances. This compromises its possible application to a wide variety of metonymical patterns across a large number of semantic categories. Second, its algorithms are rather complex. In the training phase, they calculate smoothed probabilities on the basis of a large annotated training corpus and in the test phase, they iteratively search through a thesaurus of semantically similar words. This leads to the question if this complexity is indeed necessary in metonymy recognition.

This paper investigates two approaches that each tackle one of these problems. The unsupervised algorithm in section 2 has the intuitive appeal of not requiring any annotated training instances. I will show that it is nevertheless often able to distinguish between two data clusters that correlate with the two target readings. In section 3, I will again take recourse to a supervised learning method, but one that explicitly incorporates a much simpler learning phase than its competitors in the literature — Memory-Based Learning. I will demonstrate that this algorithm of 'lazy learning' gives state-of-the-art results in metonymy recognition. Moreover, although their psychological validity is not a focus of the present investigation, the two studied algorithms have clear links to models of human behaviour.

2 An unsupervised approach to metonymy recognition

2.1 Background

Unsupervised machine learning algorithms do not need any labelled training examples. Instead, the machine itself has to try and group the training instances into a pre-defined number of clusters, which ideally correspond to the implicit target labels. The approach studied here is Schütze's (1998) Word Sense Discrimination, which uses second-order co-occurrence in order to identify clusters of senses.

Schütze's (1998) algorithm first maps all words in the training corpus onto *word vectors*, which contain frequency information about the word's first-order co-occurrents. It then builds a vector representation for each of the contexts of the target by adding up the word vectors of the words in this context. These second-order context vectors get clustered (often after some form of dimensionality reduction), and each of the clusters is assumed to correspond to one of the senses of the target. The classification of a test word, finally, proceeds by assigning it to the cluster whose centroid lies nearest to its context vector. Schütze showed that, with

[1] This data is publicly available and can be downloaded from http://homepages.inf.ed.ac.uk/mnissim/mascara.

about 8,000 training instances on average, this algorithm obtains very promising results.

This unsupervised algorithm is not just attractive from a computational point of view; it is also related to human behaviour. First, it was inspired by Miller and Charles' (1991) observation that humans rely on contextual similarity in order to determine semantic similarity. Schütze (1998) therefore hypothesized that there must be a correlation between contextual similarity and word meaning as well: "a sense is a group of contextually similar occurrences of a word" (Schütze, 1998, p.99). Second, this algorithm lies at the basis of Latent Semantic Analysis (LSA). Although the psycholinguistic merits of LSA are an object of debate, its performance in several language tasks compares well to that of humans (Landauer and Dumais, 1997). Let us therefore investigate if it is able to tackle metonymy recognition as well.

Schütze's (1998) approach has been implemented in the SenseClusters program (Purandare and Pedersen, 2004)[2], which also incorporates some interesting variations on and extensions to the original algorithm. First, Purandare and Pedersen (2004) defend the use of bigram features instead of simple word features. Bigrams are "ordered pairs of words that co-occur within five positions of each other" (Purandare and Pedersen, 2004, p.2) and will be used throughout this paper. Second, they also found that the hybrid algorithm of Repeated Bisections performs better than Schütze's (1998) clustering algorithm — at least for sparse data — so I will use it here, too. Finally, as with all word sense *discrimination* techniques, evaluation proceeds indirectly: SenseClusters automatically finds the alignment of senses and clusters that leads to the fewest misclassifications — the confusion matrix that maximizes the diagonal sum.

2.2 Experiments

On the basis of Markert and Nissim's location corpora, I tested if unsupervised learning can be applied to metonymy recognition. 60% of the instances were used as training data, 40% as test data, and the number of pre-defined clusters was set to two. The experiments were designed with five specific research questions in mind:

- **Does unsupervised clustering work better with one-word sets?**
  Since the unsupervised WSD approach studied here uses lexical features only, I anticipated it to work better with the *Hungary* data than with the mixed country set. After all, we can expect one word to have fewer typical co-occurrences than an entire semantic class, so its contexts may be easier to cluster.

- **Should a stoplist be used?**
  Unsupervised clustering on the basis of co-occurrences usually ignores a number of words that are thought to be uninformative about the reading of the target. Examples of such words are prepositions and extremely frequent verbs (*be*, *give*, *go*, ...). In metonymy recognition, however, these words may be much more useful than in classic WSD. If a location name occurs in a prepositional phrase with *in*, for instance, it is probably used literally. Similarly, verbs such as *give* and *go* determine the interpretation of a possibly metonymical word in contexts like *give sth. to a country* (metonymical) and *go to a country* (literal). Stoplists may therefore be less useful in metonymy recognition.

- **Are smaller context windows better than large ones?**
  Markert and Nissim (2002a) discovered that, with co-occurrence features, the reduction of window sizes from 10 to about 3 led to a radical improvement in precision (from 25% to above 50%) and recall (from 4% to above 20%). Schütze's (1998) original algorithm, however, used context windows of 25 words on either side of the target.

- **Does Singular Value Decomposition result in better performance?**[3]
  Schütze (1998) found that his algorithm performs better with SVD than without. SVD is said to abstract away from word dimensions, and to discover topical dimensions instead. This helps tackle vocabulary issues such as synonymy and polysemy, and moreover addresses data sparseness. However, as Markert and Nissim (2002a) argue, the sense distinctions between the literal and metonymical meanings of a word are not of a topical

[2] This software package is freely available and can be downloaded from http://senseclusters.sourceforge.net.

[3] With SVD, I set the number of dimensions to 300, as in Purandare & Pedersen (2004).

	+LL, +SVD		+LL, -SVD		-LL, +SVD		-LL, -SVD	
context	Acc	F	Acc	F	Acc	F	Acc	F
20	62.70	37.84**	73.78	11.01	60.54	28.43	54.72	26.24
15	55.95	34.54*	51.08	34.18	55.68	30.51	56.49	27.15
12	58.92	38.71**	60.54	39.67**	65.14	18.87	66.22	18.30
10	55.68	36.92**	59.46	41.41**	61.35	20.99	65.95	20.25
7	54.32	35.25**	66.76	32.79**	59.73	26.60	65.14	25.43
5	66.76	38.81**	52.97	33.08	54.32	28.09	67.03	29.07
3	58.38	37.40**	70.54	14.17	61.62	37.17**	61.62	36.04**

Table 1: Results on the mixed country data of four algorithms with varying context sizes and without a stoplist.

+LL	:	statistical feature selection
-LL	:	frequency-based feature selection
+SVD	:	dimensionality reduction with SVD
-SVD	:	no dimensionality reduction
**	:	F-score is significantly better than random assignment of data to clusters ($p < 0.05$)
*	:	difference between F-score and random assignment approaches significance ($p < 0.10$)

nature. Word dimensions may thus lead to better performance.

- **Should features be selected on the basis of a statistical test?**[4]
  Purandare and Pedersen (2004) used a log-likelihood test to select their features, probably because of the intuition that "candidate words whose occurrence depends on whether the ambiguous word occurs will be indicative of one of the senses of the ambiguous word and hence useful for disambiguation" (Schütze, 1998, p.102). Schütze, in contrast, found that statistical selection is outperformed by frequency-based selection when SVD is not used.

Like Nissim and Markert (2003), I used four measures to evaluate the experimental results: precision, recall and F-score for the metonymical category, and overall accuracy. They are defined in the following way:

- Overall accuracy is the total number of instances that is classified correctly.
- Precision for the metonymical category is the percentage of metonymical labels that the classifier assigns correctly.
- Recall for the metonymical category is the percentage of metonymies that the classifier recognizes.
- F-score is the harmonic mean between precision and recall:

$$F = \frac{2 \times P \times R}{P + R} \tag{2}$$

Let us use the confidence matrix below to illustrate these measures:

	LIT	MET
LIT	208	86
MET	37	39

If the rows represent the correct labels and the columns the labels returned by the classifier, we get the following results:

$$Acc = \frac{208 + 39}{208 + 86 + 37 + 39} = 66.76\% \tag{3}$$

$$P = \frac{39}{39 + 86} = 31.20\% \tag{4}$$

$$R = \frac{39}{39 + 37} = 51.32\% \tag{5}$$

$$F = \frac{2 \times 31.20\% \times 51.32\%}{31.20\% + 51.32\%} = 38.81\% \tag{6}$$

In engineering terms, a WSD system is only useful when its accuracy beats the so-called majority baseline. This is the accuracy of a system that simply gives the same, most frequent, label to all test instances. Such a classifier reaches an accuracy of 79.46% on the test corpus of mixed country names and of 77.35% on the test corpus with instances of *Hungary*.

[4] I again followed Purandare & Pedersen (2004) by selecting bigrams with a log-likelihood score of 3.841 or more.

	+LL, +SVD		+LL, -SVD		-LL, +SVD		-LL, -SVD	
context	Acc	F	Acc	F	Acc	F	Acc	F
20	58.52	35.06*	73.28	14.63	57.51	32.39	57.00	34.75
15	54.96	36.10*	60.05	33.19	54.20	34.31	57.25	35.38*
12	53.18	38.67**	54.71	32.06	57.76	34.65	54.96	36.10*
10	55.47	34.46	55.47	33.96	56.23	32.81	55.22	32.31
7	51.91	35.93	51.91	24.70	51.91	35.93	65.90	33.00**
5	54.20	21.74	67.18	36.45**	63.87	35.45**	63.36	28.71
3	65.14	33.82**	64.89	35.51**	57.00	35.25*	59.80	36.80**

Table 2: Results on the Hungary data of four algorithms with varying context sizes and without a stoplist.

	+LL, +SVD		+LL, -SVD		-LL, +SVD		-LL, -SVD	
context	Acc	F	Acc	F	Acc	F	Acc	F
20	67.51	15.89	67.94	16.00	57.76	36.64**	61.07	35.98**
15	70.74	10.85	70.23	12.03	58.52	36.58**	64.89	34.91**
12	66.92	30.11*	72.01	11.29	64.89	35.51**	64.89	34.29**
10	63.87	29.00	61.83	27.88	63.87	29.00	63.87	29.00
7	67.18	27.93	62.60	27.59	64.38	29.29	64.38	29.29
5	67.18	29.51	67.43	29.67*	67.18	29.51	66.16	28.11
3	68.70	30.51**	68.70	30.51**	68.19	28.57	68.19	28.57

Table 3: Results on the Hungary data of four algorithms with varying context sizes and with a stoplist.

2.3 Experimental results

Compared to this majority baseline, the results of the unsupervised approach fall below the mark. None of the accuracy values in tables 1, 2 and 3 lies above this baseline. With baselines of almost 80%, however, this result comes as no surprise. Moreover, the classifier's failure to beat the majority baseline does not necessarily mean that it is unable to identify a 'metonymical' and a 'literal' cluster in the data. This ability should be investigated with a χ^2-test instead, which helps us determine if there is a correlation between a test instance's cluster on the one hand and its label on the other. If we compare the results with this χ^2-baseline, it emerges that in many cases, the identified clusters indeed significantly correlate with the reading of the target words. The default (+LL +SVD) algorithm, for instance, typically identifies a metonymical and a literal cluster in the mixed country data (table 1). It also becomes clear that the best algorithms are not those with the highest accuracy values. After all, an accuracy close to the baseline often results from the identification of one huge 'literal' cluster that covers most metonymies as well.

Let us now evaluate the algorithms with respect to the five research questions I mentioned above. First, a comparison between the results on the mixed country data in table 1 and the *Hungary* data in table 2 shows that the former are more consistent than the latter. The (+LL +SVD) algorithm in particular is very successful on the country data. There is thus no sign of the anticipated difficulty with sets of mixed target words.

Second, when the algorithm is applied to the set of mixed country names, it should not use a stoplist. Not a single time did the resulting clusters correlate significantly with the target labels — the results were therefore not included here. A possible reason may be that the useful co-occurrences in this data tend to be words on the stoplist, but it should be studied more carefully if this is indeed the case.

On the *Hungary* data, the use of a stoplist has a different effect. Overall success rate remains more or less the same (although F-scores with a stoplist are slightly lower on average), but the results display a different pattern. Broadly speaking, a stoplist is most beneficial when feature selection proceeds on the basis of frequency and when large contexts are used. Smaller contexts are more successful without a stoplist. There is a logic to this: as I observed above, stoplist words may be informative about the reading of a possibly metonymical word, but their usefulness increases when they are closer to the target. If *go* occurs within three words of a country name, it may point towards a literal reading; if it occurs within a context of twenty words, it is less likely to do so. This explains why stoplists work best in combination with bigger contexts.

Overall, the influence of context is hard to determine. Small windows of three words on either side of the target are generally most successful, but the context size that should be chosen depends on other characteristics of the algorithm. The same is true for dimensionality reduction and statistical feature selection. In general, the anticipated negative effects of dimensionality reduction were not observed, and frequency-based feature selection clearly benefited algorithms with a stoplist on the *Hungary* data. However, the algorithms should be applied to more data sets in order to investigate the precise effect of these factors.

In short, although the investigated unsupervised algorithms never beat the majority baseline for Markert and Nissim's (2002b) data, they are often able to identify two clusters of data that correlate with the two possible readings. This is true for the set with one target word as well as for the set with mixed country names. In general, the algorithms that incorporate both statistical feature selection and Singular Value Decomposition lead to the best results, except for the *Hungary* data when no stoplist is used. In this last case, statistical feature selection is best dropped and a large context window should be chosen.

3 Memory-based metonymy recognition

3.1 Background

Memory-Based Learning (MBL), which is implemented in the TiMBL classifier (Daelemans et al., 2004)[5] rests on the hypothesis that people interpret new examples of a phenomenon by comparing them to "stored representations of earlier experiences" (Daelemans et al., 2004, p.19). It is thus related to Case-Based reasoning, which holds that "[r]eference to previous similar situations is often *necessary* to deal with the complexities of novel situations" (Kolodner, 1993, p.5). As a result of this learning hypothesis, an MBL classifier such as TiMBL eschews the formulation of complex rules or the computation of probabilities during its training phase. Instead it remembers all training vectors and gives a test vector the most frequent label of the most similar training vectors.

TiMBL implements a number of MBL algorithms. In my experiments, the so-called IB1-IG algorithm (Daelemans and Van den Bosch, 1992) proved most successful. It computes the distance between two vectors X and Y by adding up the weighted distances δ between their corresponding feature values, as in equation (7):

$$\Delta(X, Y) = \sum_{i=1}^{n} w_i \delta(x_i, y_i) \tag{7}$$

By default, TiMBL determines the weights for each feature on the basis of the feature's Information Gain (the increase in information that the knowledge of that feature's value brings with it) and the number of values that the feature can have. The precise equations are discussed in Daelemans et al. (2004) and need not concern us any further here.

3.2 Experiments

I again applied this IB1-IG algorithm to Markert and Nissim's (2002b) location corpora. In order to make my results as comparable as possible to Markert and Nissim's (2002a) and Nissim and Markert's (2003), I made two changes in the evaluation process. First, evaluation was now performed with 10-fold cross-validation. Second, in the calculation of accuracy, I made a distinction between the several metonymical labels, so that a misclassification within the metonymical category was penalized as well.

I conducted two rounds of experiments. The first used only grammatical features: the grammatical function of the word (subj, obj, iobj, pp, gen, premod, passive subj, other), its head, the presence of a second head, and the second head (if present). Such features can be expected to identify metonymies with a high precision, but since metonymies may have a wide variety of heads, performance will likely suffer from data sparseness (Nissim and Markert, 2003). I therefore conducted a second round of experiments, in which I added semantic information to the feature sets, in the form of the WordNet hypernym synsets of the head's first sense.

WordNet is a machine-readable lexical database that, among other things, structures English verbs, nouns and adjectives in a hierarchy of so-called "synonym sets" or synsets (Fellbaum, 1998). Each word belongs to such a group of synonyms, and each synset "is related to its immediately more general and more specific synsets via direct hypernym and hyponym relations" (Jurafsky and Martin, 2000, p.605). *Fear*, for instance, belongs to the synset *fear, fearfulness, fright*, which has *emotion* as its most immediate, and *psychological fea-*

[5] This software package is freely available and can be downloaded from http://ilk.uvt.nl/software.html.

	Acc	P	R	F
TiMBL	86.6%	80.2%	49.5%	61.2%
N&M	87.0%	81.4%	51.0%	62.7%

Table 4: Results for the mixed country data.
TiMBL: TiMBL's results
N&M: Nissim and Markert's (2003) results

ture as its highest hypernym. This tree structure of synsets thus corresponds to a hierarchy of semantic classes that can be used to add semantic knowledge to a metonymy recognition system.

My experiments investigated a few constellations of semantic features. The simplest of these used the highest hypernym synset of the head's first sense as an extra feature. A second approach added to the feature vector the head's highest hypernym synsets, with a maximum of ten. If the head did not have 10 hypernyms, its own synset would fill the remaining features. The result of this last approach is that the MBL classifier first looks for heads within the same synset as the test head. If it does not find a word that shares all hypernyms with the test instance, it gradually climbs the synset hierarchy until it finds the training instances that share as many hypernyms as possible. Obviously, this approach is able to make more fine-grained semantic distinctions than the previous one.

3.3 Experimental results

The experiments with grammatical information showed that TiMBL is able to replicate Nissim and Markert's (2003) results. The obtained accuracy and F-scores for the mixed country names in table 4 are almost identical to Nissim and Markert's figures. The results for the *Hungary* data in table 5 lie slightly lower, but again mirror Nissim and Markert's figures closely (Katja Markert, personal communication). This is all the more promising since my results were reached without any semantic information. Remember that Nissim and Markert's algorithm, in contrast, used Dekang Lin's (1998) clusters of semantically similar words in order to deal with data sparseness. Memory-Based Learning does not appear to need this semantic information to arrive at state-of-the-art performance. Instead, it tackles possible data sparseness by its automatic back-off to the grammatical role if the target's head is not found among the training data.

Acc	P	R	F
84.7%	80.4%	51.9%	63.1%

Table 5: Results for the *Hungary* data.

Of course, the grammatical role of a target word is often not sufficient for determining its literal or metonymical status. Therefore my second round of experiments investigated if performance can still be improved by the addition of semantic information. This does not appear to be the case. Although F-scores for the metonymical category tended to increase slightly (as a result of higher recall values), the system's accuracy hardly changed. In order to check if this was due to the automatic selection of the head's first WordNet sense, I manually disambiguated all heads in the data. This showed that the first WordNet sense was indeed often incorrect, but the selection of the correct sense did not improve performance. The reason for the failure of WordNet information to give higher results must thus be found elsewhere. A first possible explanation is the mismatch between WordNet's synsets and our semantic labels. Many synsets cover such a wide variety of words that they allow for several readings of the target, while others are too specific to make generalization possible. A second possible explanation is the predominance of prepositional heads in the data, for which extra semantic information is useless.

In short, the experiments above demonstrate convincingly that Memory-Based Learning is a simple but robust approach to metonymy recognition. This simplicity is a major asset, and is in stark contrast to the competing approaches to metonymy recognition in the literature. It should be studied, however, if there are other features that can further increase the classifier's performance. Attachment information is one such source of information that certainly deserves further attention.

4 Conclusions

This paper has investigated two computational approaches to metonymy recognition that both in their own way are less complex than their competitors in the literature. The unsupervised algorithm in section 2 does not need any labelled training data; the supervised algorithm of Memory-Based Learning incorporates an extremely simple learning phase. Both approaches moreover have a clear relation to models of human behaviour.

Schütze's (1998) approach is related to LSA, a model whose output correlates with human performance on a number of language tasks. Memory-Based Learning is akin to Case-Based Reasoning, which holds that people approach a problem by comparing it to similar instances in their memory.

Rather than presenting a psycholinguistic critique of these approaches, this paper has investigated their ability to recognize metonymical location names. Not surprisingly, it was shown that the unsupervised approach is not yet a good basis for a robust metonymy recognition system. Nevertheless, it was often able to distinguish two clusters in the data that correlate with the literal and metonymical readings. It is striking that this is also the case for a set of mixed target words from the same category — a type of data set that, to my knowledge, this algorithm had not yet been applied to. Memory-Based Learning, finally, proved to be a reliable way of recognizing metonymical words. Although this approach is much simpler than many competing algorithms, it produced state-of-the-art results, even without semantic information.

Acknowledgements

I would like to thank Mirella Lapata, Dirk Geeraerts and Dirk Speelman for their feedback on this project. I am also very grateful to Katja Markert and Malvina Nissim for their helpful information about their research.

References

W. Daelemans and A. Van den Bosch. 1992. Generalisation performance of backpropagation learning on a syllabification task. In M. F. J. Drossaers and A. Nijholt, editors, *Proceedings of TWLT3: Connectionism and Natural Language Processing*, pages 27–37, Enschede, The Netherlands.

W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 2004. TiMBL: Tilburg Memory-Based Learner. Technical report, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

S. Frisson and M. J. Pickering. 1999. The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25:1366–1383.

R. W. Jr. Gibbs. 1994. *The Poetics of Mind. Figurative Thought, Language and Understanding*. Cambridge: Cambridge University Press.

D. Jurafsky and J. H. Martin. 2000. *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall.

J. Kolodner. 1993. *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann Publishers.

Z. Kövecses. 2002. *Metaphor: A Practical Introduction*. Oxford: Oxford University Press.

T. K. Landauer and S. 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.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, Madison, USA.

K. Markert and M. Nissim. 2002a. Metonymy resolution as a classification task. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, USA.

K. Markert and M. Nissim. 2002b. Towards a corpus annotated for metonymies: the case of location names. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.

G. A. Miller and W. G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

M. Nissim and K. Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, Sapporo, Japan.

A. Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, Boston, USA.

J. Pustejovsky. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

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