Which statistics reflect semantics? Rethinking synonymy and word similarity

Derrick Higgins
Educational Testing Service

1 Overview

A great deal of work has been done of late on the statistical modeling of word similarity relations (cf.Schütze (1992), Lund and Burgess (1996) Landauer and Dumais (1997), Lin (1998), Turney (2001)). While this has largely been viewed as an engineering task (with the notable exception of much writing on Latent Semantic Analysis (LSA)), the relative success of different approaches to constructing word similarity measures is highly relevant to issues in theoretical semantics and language acquisition.

With this background in mind, this paper has two main aims. First, we will present yet another statistical approach to the calculation of word-similarity scores (LC-IR), which significantly outperforms other methods on standard benchmarks including the 80-question set of TOEFL® synonym test items first employed by Landauer and Dumais (1997).[1] Second, we hope to demonstrate that

- various methods for assessing word similarity are based on fundamentally different assumptions about the statistical properties which synonyms can be expected to display,

- the performance of each method can be taken as a judgment on the validity of these assumptions, and

- whether these predictions regarding the statistical distribution of synonyms in a corpus are borne out ought to be taken into account in any consideration of the acquisition of meaning as part of language, and the mental representation of meaning.

2 Statistical approaches to word similarity

Without indulging in too much of a caricature, we can classify different approaches to statistical estimation of word similarity according to the assumptions which they make about the distribution of synonyms (actually, *plesionyms*; cf. Edmonds and Hirst (2002)). The three main assumptions made by existing word similarity measures are the **topicality** assumption, the **proximity** assumption, and the **parallelism** assumption.

2.1 *Topicality*: LSA et al.

The techniques of Latent Semantic Analysis, Random Indexing, and Lund & Burgess' HAL all collect statistics on the relative frequency with which a word appears "near" other words. Similar words can then be identified as those which have a similar profile of content words which tend to occur near them.

These approaches to word similarity are based on the idea of situating each word in a high-dimensional vector space, so that the similarity between words can be measured as the cosine of the angle between their vectors (or a similar metric). Latent Semantic Analysis (LSA) (Landauer and Dumais 1997) is the most widely cited of these vector-space methods. It involves first constructing a term-by-document matrix based on a training collection, in which each cell of the matrix indicates the number of times a given term occurs in a given document (modulo the term weighting scheme). Given the expectation that similar terms will tend to occur in the same documents, similar terms ought to have similar term vectors in this scheme.

Singular-value decomposition (SVD) is then applied to this matrix, a dimensionality reduction technique which blurs the distinctions between similar terms and improves generalization. Typically, around 300 factors are retained. To illustrate, singular-value decomposition of the term-by-document matrix M produces the three matrices T, S, and C, as indicated in Figure 1. S is a diagonal matrix containing the top 300 singular values of M, and T and C allow term and document vectors, respectively, to be mapped into the reduced space. The product $T \times S \times C$ of these three matrices approximates the original matrix M. Now, in order to find the similarity between any two words, instead of calculating the cosine of the angle between row vectors from M, the vectors are first mapped into the 300-dimensional factor space, and the cosine similarity metric is calculated on these reduced vectors.

Figure 2 shows the most similar words to *ship* in one LSA space, in order to illustrate the sort of word similarity relationships induced using an LSA

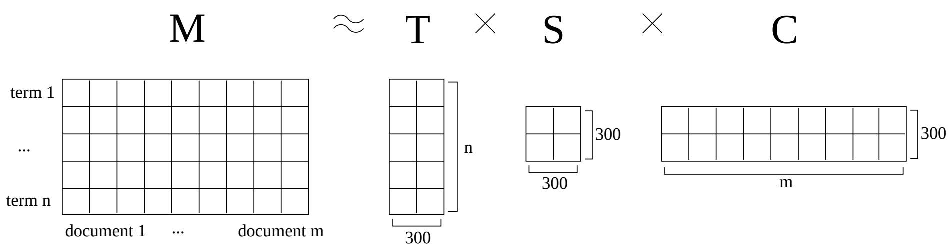

Figure 1: Singular-value decomposition of the term-by-document matrix

Word	**Similarity**	**Word**	**Similarity**
ship	1.00	decks	.82
crew	.90	rigging	.82
aboard	.89	mast	.82
captain	.87	sailors	.80
deck	.85	sails	.80
masts	.85	hull	.78
sailor	.83	ships	.78

Figure 2: Sample similarity scores produced by Latent Semantic Analysis. (Similarity to word *ship*)

analysis.

Schütze 1992 and Lund & Burgess 1996 have also produced vector-based methods of assessing word similarity. The primary differences between these methods and LSA are, first, that they use a sliding text window to calculate co-occurrence, rather than requiring that the text be pre-segmented into documents, and second, that they construct a term-by-term matrix instead of a term-by document matrix. In this term-by-term-matrix, each cell represents the co-occurrence of a term with another term within the text window, rather than the occurrence of a term within a document. The methods remain very similar to LSA, however; in each case, a vector is constructed to represent the meaning of a word based on the content words it occurs with, and the similarity between words is calculated as the cosine of the angle between the term vectors.

A slightly different vector-based word similarity metric is Random Indexing (Kanerva et al. 2000; Sahlgren 2001). Sahlgren's application of this method involves first assigning a *label vector* to each word in the vocabulary, an 1800-length sparse vector in which a small number of elements have been randomly set to 1 or -1. The *index vector* for each word is then derived as the sum of the label vectors of all words occurring within a certain distance of the target word in the training corpus (weighted according to their distance from the target word). Sahlgren uses a window size of 2-4 words on each side of the target word. This is similar to the other vector-based approaches mentioned here, but it is more scalable because it does not require a computationally intensive matrix reduction step like SVD.

While the specifics vary between these different approaches to similarity calculation—for example, the proximity required for words to count as "near" one another varies from a distance of 3 words (Random Indexing) to as much as 300 words (LSA)—these approaches are similar enough that we can say they fundamentally depend on the assumption that *similar words tend to have the same neighboring content words*. We will refer to this as the **topicality** assumption, making the inference that synonyms tend to have the same neighbors because they are in passages which are on the same topic.

2.2 ***Proximity*: PMI-IR**

On the other hand, PMI-IR also involves the collection of statistics regarding the relative frequency with which word occur in proximity, but the assumption made regarding how this relates to synonymy is quite different. Instead of the assumption that similar words will occur near the same words, the calculation which forms the core of PMI-IR assumes that similar words will tend to occur near *each other*.

In particular, PMI-IR measures the degree to which words tend to occur near one another using pointwise mutual information, as shown in (1–2) (thus the motivation for the acronym PMI-IR: *pointwise mutual information–information retrieval*).[2] The information retrieval aspect of PMI-IR consists in the fact that web search statistics are used to estimate the frequency with which word pairs appear together. (3) illustrates that the pointwise mutual information is proportional to a measure expressed in terms of the expected counts of words and word pairs in some corpus, and (4) shows how this is estimated using web search statistics. Turney (2001) uses the NEAR operator of the AltaVista search engine, which finds words within a ten-word window of one another, to calculate term co-occurrence.[3]

$$
\begin{aligned}
\mathit{Similarity}_{\mathrm{PMI-IR}}(w_1, w_2) &= \mathrm{PMI}(w_1, w_2) && (1)\\
&= \frac{\mathrm{P}(w_1\ \&\ w_2)}{\mathrm{P}(w_1) \times \mathrm{P}(w_2)} && (2)\\
&\propto \frac{\mathrm{Count}(w_1\ \&\ w_2)}{\mathrm{Count}(w_1) \times \mathrm{Count}(w_2)} && (3)\\
&\approx \frac{\mathrm{Hits}(w_1\ \mathrm{NEAR}\ w_2)}{\mathrm{Hits}(w_1) \times \mathrm{Hits}(w_2)} && (4)
\end{aligned}
$$

The intuitive basis for assuming that similar words will tend to occur near each other is not as clear as the basis for the topicality assumption, but the good results of PMI-IR lend it some empirical credence. We will refer to it as the **proximity** assumption.

2.3 Grammatical *parallelism*

Finally, Dekang Lin's work (Lin 1998; Pantel and Lin 2002) could be said to be based on the **parallelism** assumption: synonyms ought to be found in similar grammatical frames. The primary statistics gathered by Lin's method are the frequencies with which words occur linked by specific grammatical relations with other words. Lin applies a parser to the training corpus to extract triples consisting of two words and the grammatical function by which they are linked, and then constructs an information-theoretic measure on the basis of these triples, which serves as a word similarity score. Since grammatical functions (such as *subject-verb* and *verb-object*) are the basic datum of this method, these scores are based in large part on the selectional properties of verbs.

Figure 3 once again shows the most similar words to *ship*, this time using Dekang Lin's similarity scores. Whereas the LSA space identified words likely to occur in a discourse in which ships are discussed, Lin's method identifies words which are possible substitutes for the word *ship*.

3 LC-IR

Adding to this list of approaches, we present LC-IR (local context–information retrieval), a method for constructing word similarity scores which is inspired by PMI-IR, but which differs in its basic assumptions, and produces significantly better results. LC-IR, like PMI-IR, collects counts from the Web on

Word	**Similarity**	**Word**	**Similarity**
ship	1.00	freighter	.20
vessel	.32	plane	.20
boat	.25	cargo ship	.18
warship	.23	fishing boat	.17
submarine	.22	barge	.17
tanker	.20	helicopter	.17
aircraft	.20	ferry	.16

Figure 3: Sample similarity scores produced by Dekang Lin's information-theoretic method. (Similarity to word *ship*)

how often words occur near one another, but it uses a smaller window size (requiring absolute adjacency).

As shown in (5-7), LC-IR starts from the same assumption made by PMI-IR, namely that a measure of lexical association (pointwise mutual information) ought to be a good predictor of synonymy. In Equation (8), however, we require absolute adjacency for words to be counted as occurring together.

$$
\begin{aligned}
\mathit{Similarity}_{\mathrm{LC-IR}}(w_1, w_2) &= \mathrm{PMI}(w_1, w_2) && (5)\\
&= \frac{\mathrm{P}(w_1 \;\&\; w_2)}{\mathrm{P}(w_1) \times \mathrm{P}(w_2)} && (6)\\
&\propto \frac{\mathrm{Count}(w_1 \;\&\; w_2)}{\mathrm{Count}(w_1) \times \mathrm{Count}(w_2)} && (7)\\
&\approx \frac{\mathrm{Hits}(w_1 \;\textbf{NEXT-TO}\; w_2)}{\mathrm{Hits}(w_1) \times \mathrm{Hits}(w_2)} && (8)\\
&= \frac{\mathrm{Hits}(\text{“}w_1\ w_2\text{”}) + \mathrm{Hits}(\text{“}w_2\ w_1\text{”})}{\mathrm{Hits}(w_1) \times \mathrm{Hits}(w_2)} && (9)
\end{aligned}
$$

At first glance, this would seem to be a minor modification to the basic PMI-IR model, and not one which influences its fundamental assumptions. However, we will show that the small window size is of paramount importance to the model, and almost guarantees that LC-IR will identify synonyms conforming to the **parallelism** assumption, whereas standard PMI-IR is based on the more nebulous **proximity** assumption.

One final modification is necessary in order to complete the description of the LC-IR lexical similarity statistic. As stated in (9), the similarity calculation is very sensitive to collocation effects. Because we sum the number of times w_1 occurs immediately before w_2 and the number of times w_2 occurs immediately before w_1, a high count of either bigram will suffice to produce a high similarity score for the word pair, even if the word bigram produced by reversing the order does not occur at all. This is particularly troublesome when comparing words which belong to different parts of speech, or are ambiguous as to part of speech. For example, if we wish to evaluate the similarity of the words *private* and *practice,* (9) indicates that we should start by summing the frequencies of the bigrams *private practice* and *practice private*. Of course, the former is much more frequent, because of the adjectival sense of *private*, and the fact that *private practice* is a common expression for an individually owned medical or legal office. Unfortunately, the frequency of this collocation could lead to the prediction that *private* is more similar to *practice* than to nouns such as *lieutenant* or *corporal*. To mitigate these collocational effects, we replace the sum in (9) with the **min** function, so that only the less frequent bigram is considered in our calculation:

$$Similarity_{\mathrm{LC-IR}}(w_1, w_2) \quad = \quad \frac{\min(\mathrm{Hits}(\text{“}w_1\ w_2\text{”}), \mathrm{Hits}(\text{“}w_2\ w_1\text{”}))}{\mathrm{Hits}(w_1) \times \mathrm{Hits}(w_2)} \qquad (10)$$

Table 1 compares the performance of LC-IR and a number of other semantic similarity measures on the task of correctly answering multiple-choice synonym test items. The three sets of test items are, first, the 80 questions from the Test of English as a Foreign Language, introduced by Landauer and Dumais (1997), 50 ESL questions used by Turney (2001), and finally a set of 300 items culled from the Reader’s Digest Word Power feature and first used in Jarmasz and Szpakowicz (2003). Each item consists of a stem word and four option words, and the test-taker’s task is to identify which of the four is most nearly synonymous with the stem. Since there are four possible answers for each question, random guessing gives us a baseline performance of 25% accuracy. Partial credit is given in case of a tie, in which a model assigns equal similarity scores to two or more options.

Landauer and Dumais (1997) report an accuracy of 64.4% on the 80-question TOEFL® test using Latent Semantic Analysis, but this is omitted from Table 1 because we do not have corresponding test results for the other data sets. It is also not possible to provide a fair comparison of Dekang Lin’s similarity model, because many of the words used in the test sets were not included in his analysis. See Jarmasz and Szpakowicz (2003) for a partial evaluation of Lin’s model.

Table 1: Comparison of word similarity results across three synonym tests

	TOEFL®	RDWP	ESL	Overall
Baseline	25%	25%	25%	107.5/430 = 25%
Random Indexing	67.5%	36.4%	39.2%	182.8/430 = 42.5%
PMI-IR	80.0%	72.3%	66.0%	314.08/430 = 73.0%
LC-IR	81.3%	74.8%	78.0%	328.33/430 = 76.4%
Roget's Thesaurus	78.8%	74.3%	82.0%	327/430 = 76.0%

The results for Random Indexing, another vector-based semantic approach, slightly exceed those reported using LSA on the TOEFL® data set, the only one for which we have results for LSA. (The Random Indexing results in the table are for our own re-implementation of Sahlgren's method, yielding slightly lower performance than reported in his 2001 paper. We use our results because Sahlgren provides a performance evaluation only on the basis of the 80-question TOEFL® test set.) Table 1 also shows that PMI-IR substantially outperforms Random Indexing, the representative of approaches based on the **topicality** assumption. (Again, the results reported for PMI-IR are based on our own implementation of Turney's (2001) procedure; this time, our results are slightly higher.) Finally, LC-IR shows an improvement over PMI-IR, which is significant at the .05 level.

In fact, the performance of LC-IR in identifying synonyms, as measured by these test sets, is the highest yet recorded, exceeding even the results of systems using lexical resources such as WordNet (Resnik 1995; Hirst and St-Onge 1997; Leacock et al. 1998) and Roget's Thesaurus (Jarmasz and Szpakowicz 2003). For comparison, the final row of Table 1 provides the performance of Jarmasz and Szpakowicz's thesaurus-based system.

Two other systems deserve mention in this summary of performance results on synonym identification. First, Turney, Littman, Bigham, and Shnayder (2003) use a semi-supervised approach to developing an ensemble-based synonym identifier. Their system achieves over 80% accuracy on a test set very similar to the collections described here (and over 97% accuracy on the TOEFL® items). While this result is very encouraging for the engineering task of predicting synonymy, it is not directly comparable to the other systems which we have described, and is not really relevant to the question of what sort of information provides the clearest cue to synonymy in a language-acquisition scenario.

First, this system requires supervised training in order to set the model parameters which govern the importance of each submodel in the ensemble. Second, this model is not purely a corpus-based statistical one; some of the submodules it employs use information from dictionary and thesauri. Both of these conditions are at odds with the situation presented to language learners in the course of lexical acquisition. Typically, there is no supervisory signal which identifies words as synonymous or not, and of course dictionaries and thesauri are not used in the lexical acquisition scenarios we are interested in.

The other paper which reports a high accuracy on this task is (Rapp 2003), which applies singular-value decomposition to a word-by-word matrix of local associations, to produce an LSA-like vector space method of similarity calculation. This paper reports an accuracy over 90% on the TOEFL® synonym test. Given the small size of this test set, though, it is not warranted to extrapolate from this result to the other test sets. This is especially true given the sharp degradation in performance which the similar Random Indexing model shows on the other two test sets (cf. Table 1).

3.1 LC-IR and *parallelism*

We ascribe the good performance of LC-IR across all of these test sets to two main factors. First, LC-IR benefits from the fact that it uses web search statistics, which addresses the problem of data sparsity afflicting other statistical approaches to word similarity. (This is an advantage which it shares with PMI-IR, of course.) Second, LC-IR differs from PMI-IR in that it is based on the **parallelism** assumption regarding the distribution of synonyms in a corpus, rather than the **proximity** assumption.

As described above, the proximity assumption consists in the prediction that similar words ought to occur near one another; the parallelism assumption predicts that similar words will occur in grammatically parallel constructions. Given that LC-IR and PMI-IR differ primarily in the size of window used to calculate word co-occurrence, it is not immediately clear why they should rely on fundamentally different assumptions about the distribution of synonyms.

The key observation is that AltaVista's indexing format causes punctuation (such as commas) to be ignored in searches. This means that LC-IR tends to rate highly word pairs which often occur in lists of conjoined items (like "$\mathbf{w_1}$, $\mathbf{w_2}$, and w_3") or other equative contexts. Consider, for example, some of the pages rated most highly in an AltaVista search for the word pair **“assistance help”**:

```
1   Federation Sim Fleet - Assistance & Help
            Sim Fleet - A Star Trek sim group on AOL....
2   EPA: Business Gateway:  Assistance, Help and Training
            Environmental Information.  Environmental Assistance....
3   SEAL - Assistance/Help
            SEAL is a free 32-bit GUI for DOS....
```

In each of these results, the pair of words occurs either in a list of items treated in a parallel fashion, or in an implied conjunctive context. In essence, by setting the proximity threshold so low (requiring absolute adjacency between the target words), we are able to isolate word pairs which have a high degree of grammatical parallelism, because the equative uses which we isolate virtually guarantee parallel use of the terms. It stands to reason that a semantic similarity model with a clearer basis in grammatical parallelism should have higher performance than one which prizes word proximity first and foremost. The relationship between synonymy and similar grammatical behavior is intuitively more understandable than a link relating to proximity.

4 Implications for a theory of lexical semantics and acquisition

Our statistical approach to word similarity, which focuses on identifying words with parallel grammatical behavior, produces good results in identifying synonyms, as shown in the previous section. This demonstrates that grammatical parallelism is a strong correlate of semantic similarity. In this section, we will present an argument that grammatical parallelism is also the best candidate for a *cue* used by language learners to identify words as semantically similar or synonymous.

In two different lexical acquisition scenarios, we argue that constructional parallelism could provide a sufficient basis for the acquisition profile actually observed, and that alternative mechanisms based on topicality or proximity are not workable.

4.1 “One-shot” learning

First, we consider the problem of “one-shot” word learning, and argue that this phenomenon is more easily modeled as a special case of learning from parallel word usage than as any corresponding process involving topicality or simple proximity. One-shot word learning, also known as *fast-mapping* (Milostan 1995), is characterized by very rapid lexical acquisition, triggered by a single exposure to a word, or at least a very small number of exposures. This may include learning words through definitions, or through hearing a prototypical instances of word usage which are sufficient to support lexical acquisition. While fast mapping is more typical of adult word learning, it is common to both adults and children.

The idea that a word may be learned from a single prototypical instance was demonstrated by Nelson and Bonvillian (1978) for object names, and by Bates, Bretherton, and Snyder (1978) for action words. To control for the somewhat unnatural experimental setting of these studies, Rice and Woodsmall (1988) investigated children's acquisition of new vocabulary from television viewing, and again found that they were able to learn new words simply from hearing them used a small number of times.

Fast mapping may seem relatively uncontroversial, given its intuitive plausibility. However, this phenomenon has proven very challenging for computational approaches to lexical acquisition (cf. Milostan (1995)). In particular, connectionist approaches such as (Regier 1992), which depend on a long training procedure such as backpropagation of error, have difficulty accounting for such a rapid learning mechanism. A number of modifications to the basic error-driven connectionist learning schemes have been devised in attempts to permit this kind of immediate generalization (Hinton and Plaut 1987; Mikkulainen 1993; Yip and Sussman 1997).

In fact, Latent Semantic Analysis is also a connectionist approach, broadly speaking, albeit not one which requires supervised training. Nevertheless, it does require a computationally expensive training procedure (singular-value decomposition) which makes one-shot word learning hard to model. This is equally true of other vector-based approaches motivated by the topicality assumption, such as Random Indexing; they also require time-consuming procedures for training or dimensionality reduction.

In fact, even if these topicality-based models could be accelerated by some method such as "fast weights" (Hinton and Plaut 1987), they could still not contribute to a realistic model of one-shot word learning, because the data which they use to describe a word is not well-suited to learning in such data-poor situations. The calculation of a word's representation in these models is determined only by the relative frequency with which it occurs near other words in the training corpus. When only a few instances of the target word are available, this kind of loose topical data cannot provide a specific enough semantic representation. (Consider a document containing five new words; LSA would assign all five words the same representation.)

Of course, this observation holds equally for a semantic similarity metric based on the proximity assumption, such as PMI-IR. When the goal is to learn a word's meaning based on a single observation, the data will simply be too sparse to use such a method. (Given a new word, we cannot estimate an association statistic such as pointwise mutual information, because we have only observed it co-occurring with a handful of other words.)

The problem is not as acute, however, when we turn to models based on the parallelism assumption. The crucial point is that the primary datum for parallelism-based word similarity is a linguistic construction whose attestation is often definitive, whereas the requisite data for the other approaches is typically much too sparse for one-shot learning. Of course, the exact method used in LC-IR is not available in the case of one-shot word learning; language learners cannot consult the web to look for words which occur in parallel contexts with the target word. However, the grammatical construction in which a word is used, including the other lexical heads with which it is associated, may well be sufficient data to identify the word's meaning at once. It is just this sort of parallelism of grammatical relations on which LC-IR is based.

4.2 Vocabulary learning by children

Second, we consider the more general issue of children's gradual acquisition of vocabulary, addressed by Landauer and Dumais (1997). In this domain as well, we argue that parallelism is a better cue to word similarity than either topicality or proximity, in part due to the nature of the primary linguistic data with which the child is confronted. In addition to data sparsity issues, which are in play here as well as in adult word learning, such approaches would find it difficult to deal with language data which consists of relatively short utterances, and can be lacking in topical coherence.

To illustrate this fact, consider the discourse in Figure 4.2, an excerpt from the CHILDES database (MacWhinney 2000) of childrens' conversational interactions. The discourse rapidly shifts from a recap of a trip the child took recently, to a discussion of when one should and should not have chewing gum, to an exchange about the child's shirt, and finally to a request to play with a dog. See Spooren and Sanders (in prep.) for an analysis of children's acquisition of discourse coherence relations.

This is a far cry from the kind of strong topical coherence required of a collection used to train an LSA space. In fact, the best sort of text to use for training an LSA space for general use is an encyclopedia, precisely because each article is so sharply focused on a particular subject. In much of the language to which children are exposed, then, loose topical connections cannot provide a strong cue for word learning.

5 Conclusion

This paper has presented a statistical model of word similarity, LC-IR, which is based on web search statistics regarding the frequency with which words appear adjacent to one another. This metric achieves a higher level of overall performance on three major synonym identification test sets than any other purely statistical system has reported.

We have argued that this system's performance gains can in part be attributed to the fact that it is based on fundamentally different assumptions about the distribution of synonyms in a text than other models. In particular, this model assumes that words which are similar in meaning will occur in the same grammatical frames (the parallelism assumption). Other models are based on the idea that similar words ought to occur near the same set of other words (the topicality assumption), or that they ought to occur near those words which are most similar to them (the proximity assumption). The parallelism assumption has the best support among the three, both on the grounds of empirical results, and on the basis of theoretical considerations of language acquisition.

Acknowledgements

I would like to thank the conference organizers for providing an open forum for discussion, and my ETS colleagues for their helpful comments on an earlier draft of this paper. Any opinions expressed here are those of the author, and not necessarily of Educational Testing Service.

Notes

[1] The current version of the TOEFL® test does not include synonym items.

[2] Terra and Clarke (2003) compare other statistical measures of word association on the synonym identification task, such as log-likelihood ratio and chi-square, but none of the other measures which they investigate performs better than pointwise mutual information.

```
*MOT: where did you meet Minoru ?
*CHI: where ?
*MOT: we came to your school the other day # didn’t we ?
*MOT: didn’t he take you somewhere ?
*CHI: where ?
*MOT: where did he take you ?
*CHI: where did he take me ?
*MOT: where did we go with him # remember ?
*MOT: we came to pick you up at school ?
*CHI: we go to a doughnut store .
*MOT: and what did you do there ?
*CHI: eat a doughnut .
*MOT: and what else ?
*CHI: we didn’t have gum .
*MOT: no .
*MOT: why didn’t we have gum ?
*CHI: cause +...
*CHI: that was dessert .
*MOT: and it was breakfast .
*CHI: yeah [= yes] .
*MOT: does one eat gum after breakfast ?
*CHI: no .
*MOT: let me see that .
*MOT: oh # that looks nice on you # Nina .
*MOT: do you like that shirt ?
*MOT: what’s this ?
*CHI: a pocket .
*CHI: I wanna play +...
*CHI: this big honey doggy .
*MOT: he’s cute .
*CHI: I wanna play with him .
```

Figure 4: Sample mother-child interaction from the CHILDES database

3 At the time of this writing, AltaVista no longer supports the NEAR operator in its searches. While it is possible to approximate this sort of search using the Google search engine (using, for example, the Google API Proximity Search at `http://staggernation.com/cgi-bin/gaps.cgi`), this workaround is less than ideal. Unfortunately, corpus linguistics work which depends on web statistics is at the mercy of those few services with the resources to index such a large collection.

References

Bates, E., I. Bretherton, and L. Snyder
1978 Acquisition of a novel concept at 20 months. In *From First Words to Grammar: Individual Differences and Dissociable Mechanisms*, pp. 124–134. Cambridge University Press, Cambridge, NY.

Edmonds, Philip and Graeme Hirst
2002 Near-synonymy and lexical choice. *Computational Linguistics*, 28(2): 105–144.

Hinton, G. E. and D. C. Plaut
1987 Using fast weights to deblur old memories. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society*, pp. 177–186. Seattle, WA.

Hirst, Graeme and David St-Onge
1997 Lexical chains as representation of context for the detection and correction of malapropisms. In C. Fellbaum, (ed.), *WordNet: An electronic lexical database and some of its applications*. The MIT Press, Cambridge, MA.

Jarmasz, Mario and Stan Szpakowicz
2003 Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pp. 212–219. Borovets, Bulgaria.

Kanerva, P., J. Kristoferson, and A. Holst
2000 Random indexing of text samples for latent semantic analysis. In L. R. Gleitman and A. K. Josh, (eds.), *Proc. 22nd Annual Conference of the Cognitive Science Society*.

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

Leacock, Claudia, Martin Chodorow, and George Miller
1998 Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1): 147–165.

Lin, Dekang
1998 An information-theoretic definition of similarity. In *Proc. 15th International Conference on Machine Learning*, pp. 296–304. Morgan Kaufmann, San Francisco, CA.

Lund, Kevin and Curt Burgess
1996 Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research Methods, Instruments and Computers*, 28(2): 203–208.

MacWhinney, Brian
2000 *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ. Third edition.

Mikkulainen, Risto
1993 *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. MIT Press.

Milostan, Jeanne
1995 *Connectionist Modeling of the Fast Mapping Phenomenon*. UCSD ms.

Nelson, K. E. and J. D. Bonvillian
1978 Early semantic development: Conceptual growth and related processes between 2 and 4 1/2 years of age. In K. E. Nelson, (ed.), *Children's Language*, volume 1, pp. 467–556. Gardner Press, New York.

Pantel, Patrick and Dekang Lin
2002 Document clustering with committees. In *Proceedings of SIGIR02*. Tampere, Finland.

Rapp, Reinhard
2003 Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of Machine Translation Summit IX*.

Regier, Terry
1992 *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization*. Ph.D. thesis, University of California, Berkeley.

Resnik, Philip
1995 Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pp. 448–453.

Rice, M. L. and L. Woodsmall
1988 Lessons from television: Children's word learning when viewing. *Child Development*, 59: 420–429.

Sahlgren, Magnus
2001 Vector based semantic analysis: Representing word meanings based on random labels. In *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*. Helsinki, Finland.

Sch¨utze, Hinrich
1992 Dimensions of meaning. In *Proceedings of Supercomputing '92, Minneapolis.*, pp. 787–796.

Spooren, W. and T. Sanders
in prep. What does children's discourse tell us about the nature of coherence relations?

Terra, Egidio L. and Charles L. A. Clarke
2003 Frequency estimates for statistical word similarity measures. In Marti Hearst and Mari Ostendorf, (eds.), *HLT-NAACL 2003: Main Proceedings*, pp. 244–251. Association for Computational Linguistics, Edmonton, Alberta, Canada.

Turney, Peter D.
2001 Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, pp. 491–450.

Turney, Peter D., Michael Littman, Jeffrey Bigham, and Victor Shnayder
2003 Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pp. 482–489. Borovets, Bulgaria.

Yip, Kenneth and Gerald Jay Sussman
1997 Sparse representations for fast, one-shot learning. In *AAAI/IAAI 1997*, pp. 521–527.