Automatic Extraction of Idioms using Graph Analysis and Asymmetric Lexicosyntactic Patterns

Dominic Widdows
MAYA Design, Inc.
Pittsburgh, Pennsylvania
widdows@maya.com

Beate Dorow
Institute for Natural Language Processing
University of Stuttgart
dorowbe@IMS.Uni-Stuttgart.DE

Abstract

This paper describes a technique for extracting idioms from text. The technique works by finding patterns such as "thrills and spills", whose reversals (such as "spills and thrills") are never encountered.

This method collects not only idioms, but also many phrases that exhibit a strong tendency to occur in one particular order, due apparently to underlying semantic issues. These include hierarchical relationships, gender differences, temporal ordering, and prototype-variant effects.

1 Introduction

Natural language is full of idiomatic and metaphorical uses. However, language resources such as dictionaries and lexical knowledge bases give at best poor coverage of such phenomena. In many cases, knowledge bases will mistakenly 'recognize' a word and this can lead to more harm than good: for example, a typical mistake of blunt logic would be to assume that "somebody let the cat out of the bag" implied that "somebody let some mammal out of some container."

Idiomatic generation of natural language is, if anything, an even greater challenge than idiomatic language understanding. As pointed out decades ago by Fillmore (1967), a complete knowledge of English requires not only an understanding of the semantics of the word *good*, but also an awareness that this special adjective (alone) can occur with the word *any* to construct phrases like *"Is this paper any good at all?"*, and traditional lexical resources were not designed to provide this information. There are many more general examples occur: for example, "the big bad wolf" sounds right and the "the bad big wolf" sounds wrong, even though both versions are syntactically and semantically plausible. Such examples are perhaps 'idiomatic', though we would perhaps not call them 'idioms', since they are compositional and can sometimes be predicted by general pattern of word-ordering.

In general, the goal of manually creating a complete lexicon of idioms and idiomatic usage patterns in any language is unattainable, and automatic extraction and modelling techniques have been developed to fill this ever-evolving need. Firstly, automatically identifying potential idioms and bringing them to the attention of a lexicographer can be used to improve coverage and reduce the time a lexicographer must spend in searching for such examples. Secondly and more ambitiously, the goal of such work is to enable computers to recognize idioms independently so that the inevitable lack of coverage in language resources does not impede their ability to respond intelligently to natural language input.

In attempting a first-pass at this task, the experiments described in this paper proceed as follows. We focus on a particular class of idioms that can be extracted using *lexicosyntactic patterns* (Hearst, 1992), which are fixed patterns in text that suggest that the words occurring in them have some interesting relationship. The patterns we focus on are occurrences of the form "A and/or B", where A and

Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, pages 48–56,
Ann Arbor, June 2005.

B are both nouns. Examples include "football and cricket" and "hue and cry." From this list, we extract those examples for which there is a strong preference on the *ordering* of the participants. For example, we do see the pattern "cricket and football," but rarely if ever encounter the pattern "cry and hue." Using this technique, 4173 potential idioms were extracted. This included a number of both true idioms, and words that have regular semantic relationships but do appear to have interesting orderings on these relationships (such as earlier before later, strong before weak, prototype before variant).

The rest of this paper is organized as follows. Section 2 elaborates on some of the previous works that motivate the techniques we have used. Section 3 describes the precise method used to extract idioms through their asymmetric appearance in a large corpus. Section 4 presents and analyses several classes of results. Section 5 describes the methods attempted to filter these results into pairs of words that are more and less contextually related to one another. These include a statistical method that analyses the original corpus for evidence of semantic relatedness, and a combinatoric method that relies on link-analysis on the resulting graph structure.

2 Previous and Related Work

This section describes previous work in extracting information from text, and inferring semantic or idiomatic properties of words from the information so derived.

The main technique used in this paper to extract groups of words that are semantically or idiomatically related is a form of lexicosyntactic pattern recognition. Lexicosyntactic patterns were pioneered by Marti Hearst (Hearst, 1992; Hearst and Schütze, 1993) in the early 1990's, to enable the addition of new information to lexical resources such as WordNet (Fellbaum, 1998). The main insight of this sort of work is that certain regular patterns in word-usage can reflect underlying semantic relationships. For example, the phrase "France, Germany, Italy, and other European countries" suggests that *France*, *Germany* and *Italy* are part of the class of *European countries*. Such hierarchical examples are quite sparse, and greater coverage was later attained by Riloff and Shepherd (1997) and Roark and Charniak (1998) in extracting relations not of hierarchy but of *similarity*, by finding conjunctions or co-ordinations such as "cloves, cinammon, and nutmeg" and "cars and trucks." This work was extended by Caraballo (1999), who built classes of related words in this fashion and then reasoned that if a hierarchical relationship could be extracted for *any* member of this class, it could be applied to *all* members of the class. This technique can often mistakenly reason across an ambiguous middle-term, a situation that was improved upon by Cederberg and Widdows (2003), by combining pattern-based extraction with contextual filtering using latent semantic analysis.

Prior work in discovering non-compositional phrases has been carried out by Lin (1999) and Baldwin et al. (2003), who also used LSA to distinguish between compositional and non-compositional verb-particle constructions and noun-noun compounds.

At the same time, work in analyzing idioms and asymmetry within linguistics has become more sophisticated, as discussed by Benor and Levy (2004), and many of the semantic factors underlying our results can be understood from a sophisticated theoretical perspective.

Other motivating and related themes of work for this paper include collocation extraction and example based machine translation. In the work of Smadja (1993) on extracting collocations, preference was given to constructions whose constituents appear in a fixed order, a similar (and more generally implemented) version of our assumption here that asymmetric constructions are more idiomatic than symmetric ones. Recent advances in example-based machine translation (EBMT) have emphasized the fact that examining patterns of language use can significantly improve idiomatic language generation (Carl and Way, 2003).

3 The Symmetric Graph Model as used for Lexical Acquisition and Idiom Extraction

This section of the paper describes the techniques used to extract potentially idiomatic patterns from text, as deduced from previously successful experiments in lexical acquisition.

The main extraction technique is to use lexicosyntactic patterns of the form "A, B and/or C" to find nouns that are linked in some way. For example, consider the following sentence from the British National Corpus (BNC).

> **Ships** laden with **nutmeg**, **cinnamon**, **cloves** or **coriander** once battled the Seven **Seas** to bring **home** their precious **cargo**.

Since the BNC is tagged for parts-of-speech, we know that the words highlighted in bold are nouns. Since the phrase "nutmeg, cinnamon, cloves or coriander" fits the pattern "A, B, C or D", we create nodes for each of these nouns and create links between them all. When applied to the whole of the BNC, these links can be aggregated to form a graph with 99,454 nodes (nouns) and 587,475 links, as described by Widdows and Dorow (2002). This graph was originally used for lexical acquisition, since clusters of words in the graph often map to recognized semantic classes with great accuracy ($> 80\%$, (Widdows and Dorow, 2002)).

However, for the sake of smoothing over sparse data, these results made the assumption that the links between nodes were *symmetric*, rather than *directed*. In other words, when the pattern "A and/or B" was encountered, a link from A to B *and* a link from B to A was introduced. The nature of symmetric and antisymmetric relationships is examined in detail by Widdows (2004). For the purposes of this paper, it suffices to say that the assumption of symmetry (like the assumption of transitivity) is a powerful tool for improving recall in lexical acquisition, but also leads to serious lapses in precision if the directed nature of links is overlooked, especially if symmetrized links are used to infer semantic similarity.

This problem was brought strikingly to our attention by the examples in Figure 1. In spite of appearing to be a circle of related concepts, many of the nouns in this group are not similar at all, and many of the links in this graph are derived from very very different contexts. In Figure 1, *cat* and *mouse* are linked (they are re both animals and the phrase "cat and mouse" is used quite often): but then *mouse* and *keyboard* are also linked because they are both objects used in computing. A *keyboard*, as well as being a typewriter or computer keyboard, is also

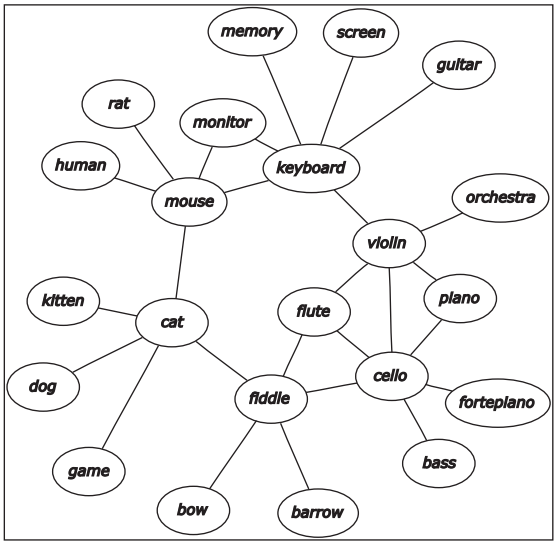

Figure 1: A cluster involving several idiomatic links

used to mean (part of) a musical instrument such as an organ or piano, and *keyboard* is linked to *violin*. A *violin* and a *fiddle* are the same instrument (as often happens with synonyms, they don't appear together often but have many neighbours in common). The unlikely circle is completed (it turns out) because of the phrase from the nursery rhyme

> Hey diddle diddle,
> The cat and the fiddle,
> The cow jumped over the moon;

It became clear from examples such as these that idiomatic links, like ambiguous words, were a serious problem when using the graph model for lexical acquisition. However, with ambiguous words, this obstacle has been gradually turned into an opportunity, since we have also developed ways to used the apparent flaws in the model to detect which words are ambiguous in the first place (Widdows, 2004, Ch 4). It is now proposed that we can take the same opportunity for certain idioms: that is, to use the properties of the graph model to work out which links arise from idiomatic usage rather than semantic similarity.

3.1 Idiom Extraction by Recognizing Asymmetric Patterns

The link between the *cat* and *fiddle* nodes in Figure 1 arises from the phrase "the cat and the fiddle."

Table 1: Sample of asymmetric pairs extracted from the BNC.

First word	Second word
highway	byway
cod	haddock
composer	conductor
wood	charcoal
element	compound
assault	battery
north	south
rock	roll
god	goddess
porgy	bess
middle	class
war	aftermath
god	hero
metal	alloy
salt	pepper
mustard	cress
stocking	suspender
bits	bobs
stimulus	response
committee	subcommittee
continent	ocean

However, no corpus examples were ever found of the converse phrase, "the fiddle and the cat." In cases like these, it may be concluded that placing a *symmetric* link between these two nodes is a mistake. Instead, a *directed* link may be more appropriate.

We therefore formed the hypothesis that if the phrase "A and/or B" occurs frequently in a corpus, but the phrase "B and/or A" is absent, then the link between A and B should be attributed to idiomatic usage rather than semantic similarity.

The next step was to rebuild, finding those relationships that have a strong preference for occurring in a fixed order. Sure enough, several British English idioms were extracted in this way. However, several other kinds of relationships were extracted as well, as shown in the sample in Table 1.[1]

After extracting these pairs, groups of them were gathered together into *directed subgraphs*.[2] Some of these directed subgraphs are reporduced in the analysis in the following section.

[1]The sample chosen here was selected by the authors to be representative of some of the main types of results. The complete list can be found at `http://infomap.stanford.edu/graphs/idioms.html`.

[2]These can be viewed at `http://infomap.stanford.edu/graphs/directed_graphs.html`

4 Analysis of Results

The experimental results include representatives of several types of asymmetric relationships, including the following broad categories.

'True' Idioms

There are many results that display genuinely idiomatic constructions. By this, we mean phrases that have an explicitly lexicalized nature that a native speaker may be expected to recognize as having a special reference or significance. Examples include the following:

thrills and spills
bread and circuses
Punch and Judy
Porgy and Bess
lies and statistics
cat and fiddle
bow and arrow
skull and crossbones

This category is quite loosely defined. It includes

1. historic quotations such as "lies, damned lies and statistics"[3] and "bread and circuses."[4]
2. titles of well-known works.
3. colloquialisms.
4. groups of objects that have become fixed nominals in their own right.

All of these types share the common property that any NLP system that encounters such groups, in order to behave correctly, should recognize, generate, or translate them as phrases rather than words.

Hierarchical Relationships

Many of the asymmetric relationships follow some pattern that may be described as roughly hierarchical. A cluster of examples from two domains is shown in Figure 2. In chess, a rook outranks a bishop, and the phrase "rook and bishop" is encountered much more often than the phrase "bishop and

[3]Attributed to Benjamin Disraeli, certainly popularized by Mark Twain.

[4]A translation of "panem et circenses," from the Roman satirist Juvenal, 1st century AD.

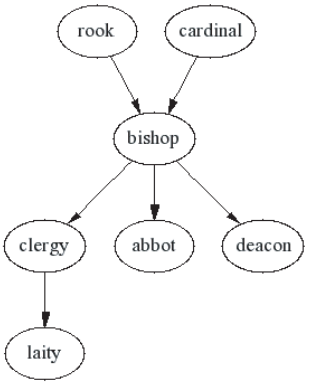

Figure 2: Asymmetric relationships in the chess and church hierarchies

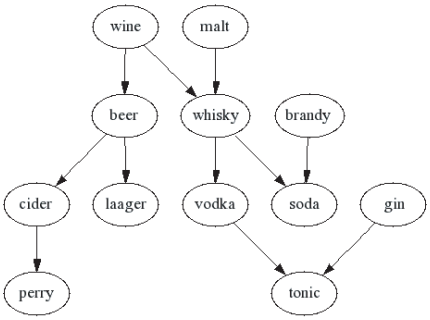

Figure 3: Different beverages, showing their directed relationships

rook." In the church, a cardinal outranks a bishop, a bishop outranks most of the rest of the clergy, and the clergy (in some senses) outrank the laity.

Sometimes these relationships coincide with figure / ground and agent / patient distinctions. Examples of this kind, as well as "clergy and laity", include "landlord and tenant", "employer and employee", "teacher and pupil", and "driver and passengers". An interesting exception is "passengers and crew", for which we have no semantic explanation.

Pedigree and potency appear to be two other dimensions that can be used to establish the directedness of an idiomatic construction. For example, Figure 3 shows that alcoholic drinks normally appear before their cocktail mixers, but that wine outranks some stronger drinks.

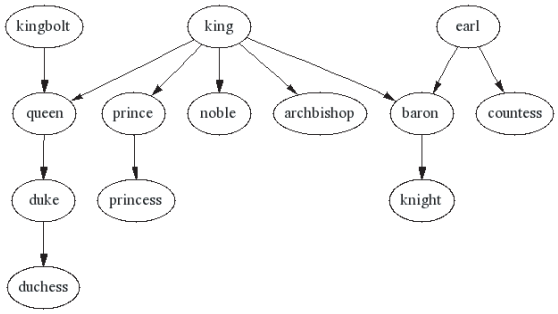

Figure 4: Hierarchical relationships between aristocrats, some of which appear to be gender based

Gender Asymmetry

The relationship between corresponding concepts of different genders also appear to be heavily biased towards appearing in one direction. Many of these relationships are shown in Figure 4. This shows that, in cases where one class outranks another, the higher class appears first, but if the classes are identical, then the male version tends to appear before the female. This pattern is repeated in many pairs of words such as "host and hostess", "god and goddess", etc. One exception appears to be in parenting relationships, where female precedes male, as in "mother and father", "mum and dad", "grandma and grandpa".

Temporal Ordering

If one word refers to an event that precedes another temporally or logically, it almost always appears first. The examples in Table 2 were extracted by our experiment. It has been pointed out that for cyclical events, it is perfectly possible that the order of these pairs may be reversed (e.g., "late night and early morning"), though the data we extracted from the BNC showed strong tendencies in the directions given.

A directed subgraph showing many events in human lives in shown in Figure 5.

Prototype precedes Variant

In cases where one participant is regarded as a 'pure' substance and the other is a variant or mixture, the pure substance tends to come first. These occur particularly in scientific writing, examples including "element and compound", "atoms and

Table 2: Pairs of events that have a strong tendency to occur in asymmetric patterns.

Before	After
spring	autumn
morning	afternoon
morning	evening
evening	night
morning	night
beginning	end
question	answer
shampoo	conditioner
marriage	divorce
arrival	departure
eggs	larvae

molecules", "metals and alloys". Also, we see "apples and pears", "apple and plums", and "apples and oranges", suggesting that an apple is a prototypical fruit (in agreement with some of the results of prototype theory; see Rosch (1975)).

Another possible version of this tendency is that core precedes periphery, which may also account for asymmetric ordering of food items such as "fish and chips", "bangers and mash", "tea and coffee" (in the British National Corpus, at least!) In some cases such as "meat and vegetables", a hierarchical or figure / ground distinction may also be argued.

Mistaken extractions

Our preliminary inspection has shown that the extraction technique finds comparatively few genuine mistakes, and the reader is encouraged to follow the links provided to check this claim. However, there are some genuine errors, most of which could be avoided with more sophisticated preprocessing.

To improve recall in our initial lexical acquisition experiments, we chose to strip off modifiers and to stem plural forms to singular forms, so that "apples and green pears" would give a link between *apple* and *pear*.

However, in many cases this is a mistake, because the bracketing should not be of the form "A and (B C)," but of the form "(A and B) C." Using part-of-speech tags alone, we cannot recover this information. One example is the phrase "hardware and software vendors," from which we obtain a link between *hardware* and *vendors*, instead of a link between *hardware* and *software*. A fuller degree of syntactic analysis would improve this situation. For extracting semantic relationships,

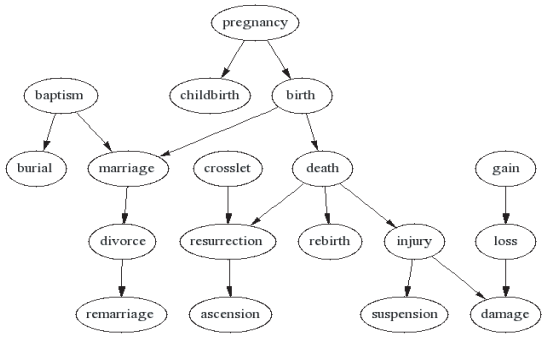

Figure 5: Directed graph showing that life-events are usually ordered temporally when they occur together

Cederberg and Widdows (2003) demonstrated that nounphrase chunking does this work very satisfactorily, while being much more tractable than full parsing.

The mistaken pair *middle* and *class* shown in Table 1 is another of these mistakes, arising from phrases such as "middle and upper class" and "middle and working class." These examples could be avoided simply by more accurate part-of-speech tagging (since the word "middle" should have been tagged as an adjective in these examples).

This concludes our preliminary analysis of results.

5 Filtering using Latent Semantic Analysis and Combinatoric Analysis

From the results in the previous section, the following points are clear.

1. It is possible to extract many accurate examples of asymmetric constructions, that would be necessary knowledge for generation of natural-sounding language.

2. Some of the pairs extracted are examples of general semantic patterns, others are examples of genuinely idiomatic phrases.

Even for semantically predictable phrases, the fact that the words occur in fixed patterns can be very useful for the purposes of disambiguation, as demonstrated by (Yarowsky, 1995). However, it

would be useful to be able to tell which of the asymmetric patterns extracted by our experiments correspond to semantically regular phrases which happen to have a conventional ordering preference, and which phrases correspond to genuine idioms. This final section demonstrates two techniques for performing this filtering task, which show promising results for improving our classification, though should not yet be considered as reliable.

5.1 Filtering using Latent Semantic Analysis

Latent semantic analysis or LSA (Landauer and Dumais, 1997) is by now a tried and tested technique for determining semantic similarity between words by analyzing large corpus (Widdows, 2004, Ch 6). Because of this, LSA can be used to determine whether a pair of words is likely to participate in a regular semantic relationship, even though LSA may not contribute specific information regarding the *nature* of the relationship. However, once a relationship is expected, LSA can be used to predict whether this relationship is used in contexts that are typical uses of the words in question, or whether these uses appear to be anomalies such as rare senses or idioms. This technique was used successfully by (Cederberg and Widdows, 2003) to improve the accuracy of hyponymy extraction. It follows that it should be useful to tell the difference between regularly related words and idiomatically related words.

To test this hypothesis, we used an LSA model built from the BNC using the Infomap NLP software.[5] This was used to measure the LSA similarity between the words in each of the pairs extracted by the techniques in Section 4. In cases where a word was too infrequent to appear in the LSA model, we used 'folding in,' which assigns a word-vector 'on the fly' by adding together the vectors of any surrounding words of a target word that are in the model.

The results are shown in Table 3. The hypothesis is that words whose occurrence is purely idiomatic would have a low LSA similarity score, because they are otherwise not closely related. However, this hypothesis does not seem to have been confirmed, partly due to the effects of overall frequency. For example, the word *Porgy* only occurs in the phrase "Porgy and Bess," and the word *bobs* almost always occurs in the phrase "bits and bobs." A more effective filtering technique would need to normalize to account for these effects. However, there are some good results: for example, the low score between *assault* and *battery* reflects the fact that this usage, though compositional, is a rare meaning of the word *battery*, and the same argument can be made for *element* and *compound*. Thus LSA might be a better guide for recognizing rarity in meaning of individual words than it is for idiomaticity of phrases.

[5] Freely available from `http://infomap-nlp.sourceforge.net/`

Table 3: Ordering of results from semantically similar to semantically dissimilar using LSA

Word pair	LSA similarity
north south	0.931
middle class	0.834
porgy bess	0.766
war aftermath	0.676
salt pepper	0.672
bits bobs	0.671
mustard cress	0.603
composer conductor	0.588
cod haddock	0.565
metal alloy	0.509
highway byway	0.480
committee subcommittee	0.479
god goddess	0.456
rock roll	0.398
continent ocean	0.300
wood charcoal	0.273
stimulus response	0.261
stocking suspender	0.177
god hero	0.115
element compound	0.044
assault battery	-0.068

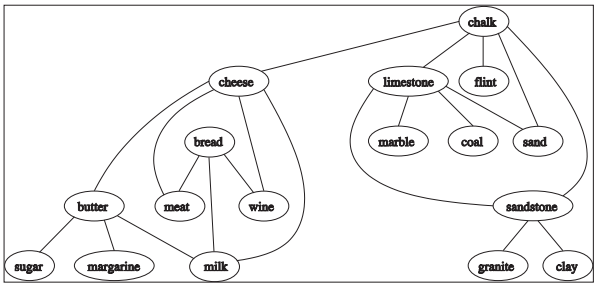

Figure 6: Nodes in the original symmetric graph in the vicinity of *chalk* and *cheese*

5.2 Link analysis

Another technique for determining whether a link is idiomatic or not is to check whether it connects two

areas of meaning that are otherwise unconnected. A hallmark example of this phenomenon is the "chalk and cheese" example shown in Figure 6. [6] Note that none of the other members of the rock-types clusters is linked to any of the other foodstuffs. We may be tempted to conclude that the single link between these clusters is an idiomatic phenomenon. This technique shows promise, but has yet to be explored in detail.

6 Conclusions and Further Work

It is possible to extract asymmetric constructions from text, some of which correspond to idioms which are indecomposable (in the sense that their meaning cannot be decomposed into a combination of the meanings of their constituent words).

Many other phrases were extracted which exhibit a typical directionality that follows from underlying semantic principles. While these are sometimes not defined as 'idioms' (because they are still composable), knowledge of their asymmetric behaviour is necessary for a system to generate natural language utterances that would sound 'idiomatic' to native speakers.

While all of this information is useful for correctly interpreting and generating natural language, further work is necessary to distinguish accurately between these different categories. The first step in this process will be to manually classify the results, and evaluate the performance of different classification techniques to see if they can reliably identify different types of idiom, and also distinguish these cases from false positives that were mistakenly extracted. Once some of these techniques have been evaluated, we will be in a better position to broaden our techniques by turning to larger corpora such as the Web.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

Sarah Bunin Benor and Roger Levy. 2004. The chicken or the egg? a probabilistic analysis of english binomials. `http://www.stanford.edu/~rog/papers/binomials.pdf`.

Sharon Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *37th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pages 120–126.

M Carl and A Way, editors. 2003. *Recent Advances in Example-Based Machine Translation*. Kluwer.

Scott Cederberg and Dominic Widdows. 2003. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Conference on Natural Language Learning (CoNNL)*, Edmonton, Canada.

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

Charles J. Fillmore. 1967. The grammar of hitting and breaking. In R. Jacobs, editor, *In Readings in English: Transformational Grammar*, pages 120–133.

Marti Hearst and Hinrich Schütze. 1993. Customizing a lexicon to better suit a computational task. In *ACL SIGLEX Workshop*, Columbus, Ohio.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*, Nantes, France.

Thomas Landauer and Susan Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition. *Psychological Review*, 104(2):211–240.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *ACL:1999*, pages 317–324.

Ellen Riloff and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124. Association for Computational Linguistics, Somerset, New Jersey.

Brian Roark and Eugene Charniak. 1998. Noun-phrase co-occurence statistics for semi-automatic semantic lexicon construction. In *COLING-ACL*, pages 1110–1116.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192–233.

[6] "Chalk and cheese" is a widespread idiom in British English, used to contrast two very different objects, e.g. "They are as different as chalk and cheese." A roughly corresponding (though more predictable) phrase in American English might be "They are as different as night and day."

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *19th International Conference on Computational Linguistics*, pages 1093–1099, Taipei, Taiwan, August.

Dominic Widdows. 2004. *Geometry and Meaning*. CSLI publications, Stanford, California.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.