Building a Collocational Semantic Lexicon

David Hardcastle
Open University, Milton Keynes
Birkbeck College, University of London

Abstract

Natural Language Generation (NLG) systems require access to collocational information to help determine lexical choices constrained both by syntactic and semantic concerns. Constructing linguistic resources to support these decisions can be time-consuming whereas, if the information is extracted automatically, data sparsity limits the variety of the output. This paper reports on a method for extracting collocational data from the British National Corpus, and then generalizing it using WordNet to tackle the sparsity problem. The method is evaluated using the lexical choice component of ENIGMA, an NLG system that generates cryptic crossword clues.

1 Introduction

This paper presents the construction of the data sources that support lexical choice for ENIGMA, an NLG system that generates cryptic crossword clues. Lexical choice, or lexicalisation, is an integral part of most NLG systems. The goal is to select "words that adequately express the content that is to be communicated" (Stede 1993:1). In some NLG contexts lexical choice focuses on the relationship between words and the underlying concepts that they represent. For example Reiter et al (2005) address the mapping of non-linguistic data (such as time-series weather data) onto words, and Williams and Reiter (2007) explore the process of mining concept-word mappings from domain-specific corpora. In the case of ENIGMA, the resulting text must satisfy crossword conventions, for which the mappings must run from crossword concepts such as anagrams to words, but it must also appear to be a valid fragment of ordinary English. This latter requirement imposes non-domain specific constraints on lexical choice. While the mappings of domain concepts onto particular words are driven by well-known conventions (for example the adjectives *crazy* or *broken* express the concept of an anagram), the relationship between the resulting lexical units is left open and must be constrained to fit the semantics of the English language. For example, in spoken English one might refer to a *crazy idea* or a *broken anvil*, but would not say *crazy anvil* or *broken idea*.

One could uncover information about the semantic association between terms such as *broken* and *anvil* using a distributional analysis of a corpus, such as the log-likelihood measure proposed by Dunning (1993) or Church and Hanks' (1990) MI-score, but such measures do not provide sufficient information about the syntactic dimension of the association. Instead, ENIGMA relies on adjective-noun, subject-verb and verb-object syntactic dependencies extracted from the British National Corpus (BNC) using a parser or regular expressions (see also Kilgarriff (2004); Lin (1997); Hindle (1990); Zinsmeister et al (2003)). The resulting data is then aggregated into a lexicon of 'collocational semantic' information that informs ENIGMA's lexical choice decisions. Unfortunately, the data in the lexicon is sparse, and this impinges on the range of crossword domain concepts that can be chosen for a given clue and on the lexical variety with which each can be expressed. To combat this sparsity the sets of nouns that participate in each dependency relation are generalized using WordNet, following Small (n.d.).

2 Lexical Choice in ENIGMA

Cryptic crossword clues must operate on two levels to be valid: they must appear to be a fragment of natural language text and they must also set a puzzle to the solver using a prescribed set of keywords and conventions hidden in the text of the clue. For example, consider the clue:

`Marinade a dash of grilled lamb (6)`

The clue appears to be a fragment of text with a simple syntactic structure (a main verb with a noun phrase as its argument) and reasonable semantics. The semantics appear reasonable partly because there is a consistency of domain throughout the clue - ENIGMA can verify this using a scoring system derived from a distributional analysis of the British National Corpus (Hardcastle, 2005) - but also because the semantic relationships between the verb and its object, or the adjective and the noun that it modifies, are semantically valid.

At the level of a puzzle the clue simply informs the solver that the answer is a six letter word that means *marinade* and that is formed of a word or abbreviation meaning *dash* (or a type of *dash*) followed by an anagram of *lamb*, the anagram being indicated by the adjective *grilled.* The answer is *embalm* (*em* followed by *balm*). Lexical choice in ENIGMA is determined through the application of a set of constraint filters, such as the following examples:

- Rubric constraints. The word *balm* can be presented to the solver as an anagram of the word *lamb*.
- Clue constraints. The word *lamb* must be preceded or followed by a keyword indicating an anagram.
- Syntactic constraints. Taking *lamb* as a noun, a valid noun phrase can be generated *x lamb*, where *x* is an adjective.
- Crossword convention mappings. Adjectives that mean anagram include *broken*, *grilled*, *jumbled*, etc.
- Collocational semantic constraints. *Grilled lamb* has a semantic fit, whereas *broken lamb* and *jumbled lamb* do not.

These and other constraints are applied to the data as generation options are explored, and also when the raw data is checked against frames that marshal the output. For ENIGMA to test the collocational semantic constraints in this example it must choose from a list of several hundred adjectives that are keywords for an anagram[1] those that could reasonably modify the noun *lamb*. In other words, ENIGMA must have access to a service that can answer questions such as "Can lamb be grilled?"

3 Extracting Dependency Information

Choices about the fit between pairs or groups of words can be informed by distributional information from corpora analysed using statistical techniques such as MI-score or log-likelihood (Church and Hanks 1990; Dunning 1993), and this data can inform lexical choice in NLG. For example, Smadja and McKeown (1990) extract likely "binary lexical relations" from a corpus using cooccurrence information and statistical analysis and use it to assist lexical choice; Langkilde and Knight (1998) use statistical information about bigrams to support determiner-noun, subject-verb and other collocational lexicalization decisions, and Inkpen and Hirst (2002) use a variety of statistical methods to determine lexical choices between near-synonyms in collocations. However, statistically significant n-grams could themselves be part of some larger n-gram or frame - consider for example *terms-of* or *accident-insurance* from the sample entries presented in Dunning's paper on log-likelihood (1993). Both of these bigrams are statistically sig-

[1] Any adjective that signals a change of state can indicate an anagram.

nificant collocations, but neither can be used independently of a wider collocational context. Also, most of the collocations that prove to be statistically significant will not be fully compositional (see Manning and Schutze 2002:151), meaning that the bigram itself carries more meaning than the sum of its parts. But in an NLG context we don't necessarily want this additional level of meaning that arises from the collocation; we may want to make lexical choice decisions based simply on compositional meaning, regardless of the frequency of the terms. To ensure that the collocations extracted from the corpus can be employed directly in the generated clues the collocation lexicon is derived not from n-grams but from dependency relations.

In the literature a range of different methods are employed to extract dependency relations from text. For example, Smadja and McKeown (1990) perform a statistical analysis of concordance data, Kilgarriff (2004) uses regular expressions for the Sketch Engine, Velardi et al (1991) apply heuristics to chunk the text and then parse those chunks and Hindle (1990), Lin (1997) and Zinsmeister et al (2003) turn to statistical parsers.

For ENIGMA I explored two approaches, first using regular expressions to match dependencies and secondly using the Stanford parser (Klein and Manning 2003) to locate typed dependencies and filter them. In both systems multi-word units were ignored since ENIGMA's data sources are mostly single-word in an effort to tie down the project scope. While the Stanford parser benefits from the efforts of a group of contributors to the code, and has been trained on a large corpus[2] it also computes a large number of complex relationships and dependencies across the whole sentence, most of which are then discarded when the dependencies are filtered. This means that there is a trade-off between the quantity and quality of the data extracted (where the parser out-performs a regular expression engine) and the resource and time required to perform the extraction (where the use of regular expressions is considerably faster), as demonstrated in Table 1 below.

	Relations Extracted	Processing Time
Parser	9,635	5,447 sec
Regex Engine	6,959	2.3 sec

Table 1. The Stanford parser extracted more adjective-noun, subject-verb and verb-object dependency relations from a subcorpus of around 9,000 sentences of the BNC than the regular expression engine, but it required considerably longer to perform the task. To parse all 6.5 million sentences in the BNC the parser would have to run continuously for a number of weeks, whereas the regular expression engine completes the task in a matter of hours.

Furthermore, the regular expression search strings can become quite complex, making the program brittle. It is possible that a broad coverage parser such as MiniPar (Lin 1998) might represent a best of both worlds solution, although at the time of writing this option had not been fully explored.

4 Generalization

Collocational data extracted from corpora is notoriously sparse, since the data relies not just on the frequency of the words in question, but on the frequency of their use in combination. The data relating to dependency relations suffers even more from sparsity than cooccurrence information based on distributional analysis. While almost all occurrences of a word in the corpus have some surrounding context, and

[2] I used the training data distributed with the parser. Although this is based on the WSJ which is US English I did not feel that this would have a noticeable impact on performance on British English text from the BNC when it came to high-level dependency relations such as adjective modifier or direct object relationships.

thus co-occur with some other words, few occurrences may participate in the dependency relations mined from the corpus.

This sparsity problem is mitigated by generalizing the sets of nouns that participate in each recovered relation (for example the set of nouns that are evidenced in the BNC as being *red*) by mapping them into WordNet (Miller 1990) and applying a minimal arc-distance algorithm to group them into sub-trees from which generalizations can be inferred. The use of WordNet as a means to generalize the participants in syntactic relations was first proposed by Michael Small, a fellow student at Birkbeck College, as part of his on-going PhD research into the use of semantic information in a spellchecking system. In Small's implementation all of the senses of each word are mapped into WordNet, and candidate alternatives for possible spelling errors are ranked according to their arc-distance to the terms in WordNet evidenced as participating in the same subject-verb-object relation that are found in the text. In the ENIGMA implementation a coarse-grained sense disambiguation is applied before the data is mapped into WordNet to reduce noise, and the resulting mappings are grouped and generalized into sub-trees that are then filtered for coverage and compiled into a lexicon.

4.1 First-Pass Disambiguation

In addition to reducing noise, sense disambiguation can provide ENIGMA with valuable information about the word sense implied by a particular collocation, information that can add a level of crypticality to the clue. For instance in the example clue above the semantics imply a sense of *dash* meaning a small amount of something, whereas at the level of the puzzle the solver must think of a synonym for *dash* in the sense of a punctuation mark. The use of homographs to mislead the solver is a commonplace feature of the cryptic crosswords found in UK newspapers that form the corpus of target texts for ENIGMA.

When the lexicographers responsible for WordNet tackle a new entry, they first classify it into a broad semantic class known as a lexicographer file number. There are forty-four such classes for WordNet 2.1, of which twenty-nine relate to nouns. ENIGMA performs a first-pass disambiguation using these noun classes as quasi-domains, allowing it to attempt a coarse-grained disambiguation that seeds the allocation of each term in the argument list to a particular WordNet sense. For example, the set of nouns that can be modified by the adjective *grilled* in the BNC includes nouns such as *steak*, *fish*, *bacon* and *bird*. The system spreads the frequency of each collocation (as evidenced in the BNC) over all of the WordNet senses available for the head noun and then aggregates the spread frequencies using the lexicographer file numbers that annotate each word sense, except for proper names and locations which are ignored. These aggregated totals are then normalized using the proportion of WordNet annotated with each lexicographer file number, and result in weights for each file number as shown in Figure 1 below.

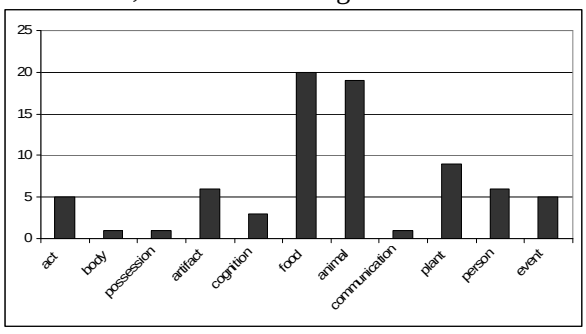

Figure 1. After initial lexicographer file number allocation the entries for *grilled* are spread over a lot of different lexicographer file numbers.

Before these weights are used for disambiguation they are sharpened using a positive feedback algorithm. Each noun in the entry for *grilled* is now allocated to the single synset for which the lexicographer file number weight is the highest and the weights are recalculated. This feedback loop cycles until there are no changes in synset allocations, and the frequency for each collocation has been allocated to a single synset, sharpening the weights as shown below.

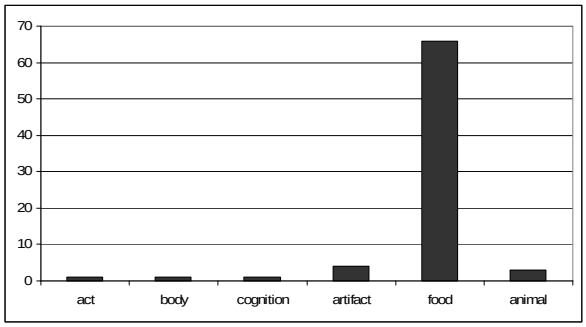

Figure 2. The entries for *grilled* after sharpening, with noise from polysemy filtered out.

Using the lexicographer file numbers as stand-in domain annotation provides an axis orthogonal to the hyponymy hierarchy of WordNet, and also provides a granularity that is sufficiently coarse-grained that a simple pooling algorithm can succeed (see Ciaramita et al (2003) who also use lexicographer file numbers as coarse-grained sense markers). It might be interesting to investigate using the finer grained WordNet Domains described in Magnini et al (2002), although these are annotated against an earlier version of WordNet[3] and so considerable mapping effort would most likely be required.

The graph in Figure 2 also demonstrates how first-pass disambiguation reduces noise. Many of the nouns that are found to be modified by *grilled* in the BNC have a sense that means the flesh of an animal, but can also mean the animal itself (such as *lamb*, *fish*, *chicken* or *sardine*). In a more general language processing context, blurring this distinction may be helpful, as it would help the application to deal with figurative language, or unusual collocates such as *grilled dog*. For generation, however, using positive feedback to reduce polysemy reduces noise and mitigates against the generation of peculiar sounding collocations.

4.2 Generalization by Hyponymy

This coarse-grained disambiguation associates each noun for an entry such as *grilled* with a specific WordNet sense, and therefore a specific WordNet synset. The next step in the process is to use the hyponymy structure of WordNet to construct sub-trees. A sub-tree is composed of synsets that share a common cohypernym that is within a parameterized arc-distance from all members of the sub-tree. The sub-tree is then analysed for coverage by counting every synset (in other words all senses, regardless of the first pass disambiguation) of each member of the set that is a hyponym of the cohypernym root and dividing this number by the total number of nodes in WordNet that are hyponyms of that root. Frequency is not used here as coverage is measuring the representativity of the evidence for the generalization that will be made if the sub-tree is used. In particular, this prevents very large sub-trees of WordNet being selected on the basis of a small number of synsets with low depth.

For example, in the set for *broken* the sub-tree under the root node *adornment* is retained, as there is direct evidence in the BNC for twenty of its seventy-five hyponyms. By contrast the sub-tree under the root

[3] Version 1.7 rather than version 2.1.

node *cognition* is discarded as only twenty of its nearly four thousand hyponyms is directly evidenced as being *broken* in the BNC.

The crossword application uses a tight arc-distance threshold (3) and a high threshold for coverage (20%). In an information retrieval context these parameters could be relaxed, allowing wider generalizations to be made on the basis of less evidence. For ENIGMA a cautious approach is more appropriate, since any member of the generalization could be used in a clue, even if a directly evidenced alternative exists, to maximize lexical variety in generation. It is important to note that some members of the set may now have multiple synsets represented in the generalization, or may include a synset that was not the dominant sense that emerged from the sharpening phase of first pass disambiguation. This allows the system to handle collocations such as *to throw a ball* where the whole collocation can be read polysemously, and also to handle occasions where multiple senses of a noun could reasonably be modified by an adjective, or governed by a verb (consider *large bank*, for example).

Finally the generalized data is written to a 'collocational semantic lexicon' listing against each headword the indices of the cohypernym roots of each sub-tree. The lexical choice component can use this lexicon to determine the semantics of a proposed relation by checking to see if any synset for the proposed noun is a hyponym of any of the root nodes listed against the headword. The lexicon also contains all of the remaining nouns that were evidenced in the BNC but have not been allocated to a sub-tree, in other words that are not part of any generalization. Collocations with unusually high log-likelihood (Dunning 1993) are flagged[4]; these are likely to be non-compositional collocations, and so will add weight to a clue's ranking for idiomaticity if used.

4.3 Sample Output

As an example of the output of the system, the entry for the adjective *red* includes the following (among many others):

Generalizations	vegetable, coat, furniture, merchandise, flower, injury
Evidenced	alligator, blister, belly, phosphorus, stone, flame, sauce, crescent*, admiral*, squirrel*, label*, meat*

Table 2. Sample entry in the collocational semantic lexicon for the adjective *red*. The generalizations are nodes in WordNet, the lexicon asserts that any hyponym of these nodes can be said to be *red*. The other entries were evidenced in the BNC but did not form part of any generalization. Those marked with asterisks are flagged as plausible non-compositional collocations.

4.4 Evaluation

I performed a task-based evaluation of the lexical choice component of ENIGMA using a forced choice questionnaire to test the collocations chosen by the system for a set of sixty adjective-noun pairs generated for nouns known to be anagrams of other words. In each case the adjective was chosen by the system as an apposite indicator of an anagram, but was accompanied by two control adjectives selected at random from the pool of anagram keyword adjectives not thought to be apposite in the particular case. Sub-

[4] The system measures $-2\log\lambda$ as in Dunning's paper but we cannot just use χ^2 significance as the members of the set were extracted on the basis of a syntactic relation and so independence does not apply. Instead a top slice is taken, since these could plausibly be non-compositional collocations so it is useful for the system to flag them.

jects were asked to choose the adjective-noun pairing that they imagined they would be most likely to encounter in spoken English.

```
mixed/ordered/modified spice
broken/corrected/blended anvil
awkward/varied/untrue teenager
```

Figure 3. Some sample adjective-noun choices presented to the subjects. One of the adjectives in each set of three alternatives was selected by the system as apposite for the noun, the other two were chosen at random. In this sample the adjective chosen by the system is on the left in each case, but for the questionnaire the ordering was randomized.

Thirty subjects participated in the experiment. The adjective chosen by the system matched the subjects' choice with $p<0.01$ in fifty-one out of sixty, or eighty-five percent, of the choices presented. Overall, agreement between subjects was very high: on average 80% of subjects chose the leading adjective for each entry, equivalent to $p<0.0001$, and only three of the sixty entries did not have a clear winner with a proportion significant at $p<0.01$. This implies that although there were seven negative results, in which the chosen adjective differed from ENIGMA's selection, there may be other circumstantial factors that made these alternatives seem appealing in their own right.

A full description of the evaluation experiment including results and further discussion is presented in a separate technical note.

5 Discussion

To address the issue of data sparsity the dependency relations extracted from the BNC are generalized using WordNet, as described above. This implies that some isomorphism exists between the hyponomy hierarchies defined in WordNet, and the domain of nouns that can be modified by particular adjectives, or be the subjects or objects of particular verbs, an implication that is implicitly supported by Lin (1997) for whom WordNet functions as a point of comparison in evaluating a machine-generated thesaurus based on a collocational similarity measure. On the other hand, Kilgarriff (1997) proposes that word senses amount to clusters of collocations that are large and distinct enough to be salient, for some purpose or in some context. Similarly, Hindle (1990) presents "an approach to classifying English words according to the predicate-argument structures they show in a corpus of text", as opposed to a static classification in a dictionary or thesaurus. Rather than sharing some isomorphism with WordNet, it could be argued that senses grouped according to their role as participants in relationships such as adjective-noun or subject-verb will belong to many different groupings depending on register, domain, context and other factors. Looking at the data in more detail, there are many examples of situations in which collocations evidenced in the BNC do not map straightforwardly onto WordNet groupings, for a number of different reasons.

- Continuous data. Colours are often cited as an example of meaning that does not translate across cultural boundaries. Since the colour spectrum is continuous all colour distinctions are arbitrary in nature, and although some objects that share the same colour may also share other features, this need not be the case.
- Synecdoche. One might expect that the entry for *broken* would include a generalization about limbs, or parts of the skeleton. In practice the BNC lists some actual bones that are broken but also includes loci such as *ankle*, *shoulder*, *finger* or *leg* that are in a different part of WordNet.
- Figurative speech. Many of the collocations that could not be generalized are idiomatic or figurative in nature, for example *red mist*, *broken heart* or *new potato*. Being non-compositional in meaning they cannot tell the system anything that can be generalized. Captured as single collocations they provide useful data to the system, but during disambiguation and generalization they simply introduce noise.

- Sub-domain vocabulary. In addition to *eggs* we find that *goals* and *passages* can also be *scrambled* in the BNC. This occurs because the BNC includes sports coverage, an idiolect with its own peculiar grammar and its own bespoke collocations. Sub-domain usages such as these defy attempts to systematise collocational relationships.
- Predicate polysemy. In this paper I only try to resolve polysemy at the level of the arguments, where their grouping within WordNet can support disambiguation. There is no data with which to disambiguate the predicates, but for some entries the different sub-trees represent not just different groupings within some shared overall sense but quite distinct senses. Consider for example *broken vase*, *broken beam*, *broken leg*, and *broken computer*.
- WordNet senses. There is only one entry for *kidney* in WordNet, and that is as an organ. This prevents the collocation *grilled kidney* from being included in the generalization about grilled food.
- WordNet topology. The WordNet topology is very uneven, this means that constraints such as arc-distance have a different impact in different parts of the structure. For example, the synset *Irish_water_spaniel* is five edges away from the synset for *dog*, too far to be included in the subtree. However most modifiers that apply to dogs will likely apply to Irish water spaniels too. Conversely the synsets *bleach* and *deus_ex_machina* have an arc-distance of three, but probably rather less in common when it comes to adjective modifiers.

6 Conclusion

This paper describes the construction of a collocational semantic lexicon that is used by an NLG system to resolve lexicalization options. The lexicon represents the domain of objects that can participate in a syntactic relationship (such as adjective-noun) with a given word (such as the adjective *grilled*). Rather than rely on a distributional analysis the data is resourced from a corpus by extracting relations, which are then generalized to combat sparsity.

The evaluation shows that this lexicon provides ENIGMA with the information it requires to make semantically valid lexical choices. However, the process of pinning down slippery semantic categories onto a static classification sometimes results in semantic assertions that do not hold in practice. For example, when the system generalizes the nouns evidenced to be red in the BNC using WordNet this results in an assertion in the lexicon that all flowers can be red, which is not the case. It is possible that a larger classification system, such as a set of ontologies, might provide a better resource for generalization, but for the reasons set out above a trade off is likely to persist between the data gain needed to assure lexical variety and some resulting simplification of the fuzzy categories that the system aims to capture.

Acknowledgement

The idea of using WordNet to generalize semantic relations was first proposed and extensively explored by Michael Small, a fellow PhD student at Birkbeck College, in his, as yet unpublished, research into the use of semantics to improve the performance of a spellchecker. I am grateful for his cooperation in the production of the paper.

References

Giovanni Pezzulo Bernardo Magnini, Carlo Strapparava and Alfio Gliozzo. 2002. Using domain information for word sense disambiguation. In *Proceedings of Senseval-2 Workshop, Association of Computational Linguistics*, page 111-115.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

David Hardcastle. 2005. Using the distributional hypothesis to derive cooccurrence scores from the British National Corpus. In *Proceedings of Corpus Linguistics*. Birmingham, UK.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *Meeting of the Association for Computational Linguistics*, pages 268–275.

D. Inkpen and G. Hirst. 2002. Acquiring collocations for lexical choice between near synonyms.

Adam Kilgarriff. 1997. I don't believe in word senses.

Adam Kilgarriff. 2004. The sketch engine. In *Proceedings of Euralex*, pages 105–116.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, Morristown, NJ, USA.

Irene Langkilde and Kevin Knight. 1998. The practical value of N-grams in derivation. In Eduard Hovy, editor, *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 248–255. Association for Computational Linguistics, New Brunswick, New Jersey.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Meeting of the Association for Computational Linguistics*, pages 64–71.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *The 17th International Conference on Computational Linguistics*, pages 768–774.

Christopher D. Manning and Hinrich Schutze. 2002. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Thomas Hofmann Massimiliano Ciaramita and Mark Johnson. 2003. Hierarchical semantic classification: Word sense disambiguation with world knowledge. In *The 18th International Joint Conference on Artificial Intelligence*.

George A. Miller. 1990. Wordnet: a lexical database for English. *Commun. ACM*, 38(11):39–41.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computergenerated weather forecasts. *Artif. Intell.*, 167(1-2):137–169.

Frank A. Smadja and Kathleen R. McKeown. 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 252–259, Pittsburgh, PA. Association for Computational Linguistics.

Michael Small. n.d. *Ongoing Phd Thesis*. Ph.D. thesis, Birkbeck College, London.

Manfred Stede. 1993. Lexical choice criteria in language generation. In *Proceedings of the 6th Conference of the European Chapter of the ACL (EACL)*, Utrecht.

Paola Velardi, Michela Fasolo, and Maria Teresa Pazienza. 1991. How to encode semantic knowledge: a method for meaning representation and computer aided acquisition. *Comput. Linguist.*, 17(2):153–170.

Sandra Williams and Ehud Reiter. Forthcoming. Generating basic skills report for low-skilled readers. *Submitted to Natural Language Engineering Journal, 2007*

Heike Zinsmeister and Ulrich Heid. 2003. Significant triples: Adjective+noun+verb combinations. In *Proceedings of Complex 2003*, Budapest, Hungary.