

# Discovering Semantic Regularity in Lexical Resources

## Wim Peters

Department of Computer Science  
University of Sheffield  
U.K.  
w.peters@dcs.shef.ac.uk

## Adam Kilgarriff

ITRI  
University of Brighton  
U.K.  
adam@itri.bton.ac.uk

### Abstract

We first describe four varieties of thesaurus: (1) Roget-style, produced to help people find synonyms when they are writing; (2) WordNet and EuroWordNet; (3) thesauruses produced (manually) to support information retrieval systems; and (4) thesauruses produced automatically from corpora. We then contrast thesauruses and dictionaries, and present a small experiment in which we look at polysemy in relation to thesaurus structure. It has sometimes been assumed that different dictionary senses for a word that are close in meaning will be near neighbours in the thesaurus. This hypothesis is explored, using as inputs the hierarchical structure of WordNet 1.5 and a mapping between WordNet senses and the senses of another dictionary. The experiment shows that pairs of 'lexicographically close' meanings are frequently found in different parts of the hierarchy. These pairs often display regular polysemic relations, i.e. they constitute systematic sense combinations that are valid for more than one word.

The second experiment is an attempt to gain more insight into the mechanisms that underlie lexicalized regular polysemy. The hierarchical structure of WordNet is exploited to create a working definition of systematic polysemy and extract regular polysemic patterns at a level of semantic generalization that allows the identification of fine-grained semantic relations between the senses of the words participating in the regular polysemic pattern.

## 1 Introduction

Lexicons, iceberg-like, carry far more than they show. Any dictionary or thesaurus has a dense network of word-meanings implicit within it. Only a small proportion of the relations are explicit. For purposes of lexicography, linguistics and language technology, it is desirable to make the relations explicit. It is desirable for lexicography, so that the editor can make informed and consistent decisions about presenting or indicating particular relations. It is desirable for linguistics simply because the relations are central to the language system. It is desirable for language technology so that inferences that use the relations between word-meanings can be drawn. Thesauruses present groups of related words, so make a suitable starting point for charting the network. In the first part of the paper, we present a taxonomy of thesaurus. We then focus on the special case of polysemy. In both dictionaries and thesauruses, the analysis of a word into its distinct senses is a key aspect of the lexicographer's work, but it is subject to different constraints in the two cases. We discuss the relation between dictionary word senses and thesaurus word senses, and present a small experiment in which we explore whether pairs of meanings of the same word which are close from a lexicographer's perspective are often close in terms of the structure of one particular thesaurus. Part of this work was undertaken in the context of the SENSEVAL exercise in evaluating word sense disambiguation programs. SENSEVAL is briefly described in section 3.

Relations between different senses of the same word are often systematic, as noted by various authors, under various names such as 'animal' - 'food' and 'producer' - 'product'. These systematic sense combinations are known as instances of regular polysemy. In the final part of the paper, we look in greater depth at how links between thesaural categories in the Wordnet network can be made explicit. We present an automatic method for discovering relations of regular polysemy.

## 2 Taxonomy

'Thesaurus' can mean a number of different language resources, useful for a range of different language engineering purposes. We work from an inclusive definition of a thesaurus: "a resource in which words with similar meanings are grouped together". The varieties include at least the following:

- 1 Roget, Macquarie and others, produced, as books, to help writers with word selection
- 2 WordNet and EuroWordNet
- 3 Thesauruses produced manually for use in information retrieval systems
- 4 'Automatic thesauruses', produced by processing corpora, with similarity between words measured (directly or indirectly) by co-occurrence.

There is of course a vast literature on the use of thesauruses in computational linguistics, stretching back to the earliest days of the enterprise when Roget was hand-punched onto cards and the links used for a disambiguation engine (Masterman, 1957), cited in (Wilks et al., p89) and the extensive work of the Sedelows (Sedelow, 1992). Here our references to the literature will be indicative.

### 2.1 Roget-type thesaurus

Landau comments on what he calls the extreme inclusiveness of thesauruses:

Rarely used words, non-English words, names, obsolete and unidiomatic expressions, phrases: all are thrown in together along with common words without any apparent principle of selection. For example, in the fourth edition of Roget's International Thesaurus - one of the best of the conceptually arranged works - we find included under the subheading orator: Demosthenes, Cicero, Franklin D. Roosevelt, Winston Churchill, William Jennings Bryan. Why not Pericles and Billy Graham? When one starts to include types of things, where does one stop? There is actually a list of insects (paragraph 414.36), which is even more of a random sampling than that of orators. Such works are a potpourri of everything the compiler can think of. (Landau, 1989, p.108)

The market for Roget-style thesauruses is distinct from that for dictionaries. They are marketed as aids to help writers choose the appropriate word, and for this the critical consideration is to provide a wide range of possibilities. This is quite unlike the native-speaker dictionary market, where the main purposes are to help with finding meanings for rare words, finding correct spellings, and as an arbiter for word games and family disputes (Summers, 1988), which means that the penalty for the sins sketched by Landau is not great. We may also classify under this heading some items produced for the language learners' market. McArthur's Lexicon (Macquarie, 1986) is, like Roget, organised hierarchically by meaning but restrains itself to a modest 15,000 items, and for each of these it provides a definition and sometimes an example. It is thus quite dictionary-like, and takes a dictionary-like approach to sense division. It frequently has several numbered senses of a word under a thesaural heading, with, for example, twenty-five meanings for *turn* under the thesaural heading **N297: twisting and turning**. The Longman Activator (Summers, 1995) complements it: whereas McArthur first addresses "real world" vocabulary, where the difference between two close items corresponds to their different non-linguistic denotations, the Activator leaves real world vocabulary out entirely, focusing instead on parts of the vocabulary, more often verbal and adjectival, where differences between near-synonyms are usually better explained by differences in circumstances and patterns of use than denotations. Many Activator headwords are phrasal, and the analyses of meanings are designed to help learners select the appropriate English word or phrase from the set that the Activator offers. Like McArthur's Lexicon, the Activator is hierarchically organised and provides definitions and examples.

### 2.1.1 The Macquarie: a case study

Table 1 sets out the entire entry in the Macquarie Thesaurus (Macquarie,1986) constituting section 494, under the heading of NATURE.

NATURE 494		
Part of speech	Sense number	definition
n	1	<b>nature</b> , the great outdoors, the wild, tiger country, waste, wilderness (area); <b>balance of nature</b> , ecosystem.
	2	<b>ecology</b> , autecology, bionomics, natural history, natural science, nature study, physic ( <i>Obs.</i> ), physiography, synecology
	3	<b>naturalist</b> , bionomist, ecologist, physiographer; <b>nature lover</b> , conservationist, greenie
	4	<b>primitive</b> , child of nature, noble savage, savage
adj	5	<b>natural</b> , innate, instinctive, normal, unformed, unschooled; <b>primitive</b> , in a state of nature, feral, native, savage, uncivilised, unlearned.
	6	<b>wild</b> , feral, ladino, tameless, warrigal, wilding ( <i>Archaic</i> ), wildish; <b>undeveloped</b> , rough, trackless, unimproved, untouched, waste.
v	7	<b>go back to nature</b> , escape, go bush, go wild, rough it.
adv	8	<b>naturally</b> , wild; <b>primitively</b> , savagely, wildly; <b>instinctively</b> , by birth, innately.

Table 1: Macquarie Thesaurus, section 494, NATURE

There are no definitions, and the user is left to infer the appropriate senses of words that have several dictionary definitions, such as *nature* and *wild*. Some structure is signalled by the use of part-of-speech labels and numbered subsections. The cross-reference to 'related keywords' can be read as implying a larger semantic framework, but the cross-references here also demonstrate how far semantic affinity may be stretched. One reference is to the section headed BUSH, which includes nouns like *churl* and *wench*, and adjectives like *boorish* and *provincial*. While the semantic connections are salient for typical thesaurus use, they go well beyond straightforward relationships such as synonymy or hyponymy. There are no explicit indications of semantic relationships within or across subsections.

Several kinds of relationships can be inferred within the NATURE entry. Subsection 2 begins with *ecology* in bold type, followed by words and phrases which are, loosely, synonyms, such as *bionomics* and *nature study*. Subsection 3 begins with *naturalist* in bold type, followed by words such as *bionomist* and *ecologist*. Following this, still within subsection 3 is another bold entry, *nature lover*, followed by *conservationist* and *greenie*. Thus subsection 2 consists of nouns which are more or less synonymous with *ecology* and which can be characterized as 'the study or science of nature'. By contrast, subsection 3 consists of human nouns which could be paraphrased either as 'a person who studies nature' (in the

first set beginning with *naturalist*) or as 'a person who loves or wants to preserve nature' (in the second set beginning with *nature lover*).

In general, the words immediately following a bold entry up to the next semi-colon are (errors apart) synonyms or near-synonyms. Given that thesaurus compilers will probably seek to begin such strings with a relatively general term, some strings may be better characterized as hyponymous rather than synonymous. It is interesting to note that, in subsection 2, *synecology* ('that branch of autecology which deals with the relation between the species or group and its environment') is a hyponym of *autecology* ('that branch of ecology which deals with the individual organism or a single-species population and its environment') which is in turn a hyponym of the boldface term, *ecology*. The first, boldface term is frequently a superordinate rather than synonym of the other members of the group. Many of the semi-colon-separated groups (or synsets, after WordNet) can be related to the headword by simple linguistic operations. These include morphological derivation to change the part of speech (e.g. *nature - natural - naturally*) and predications, whether expressed nominally (*nature - study of nature*) or verbally (*nature - go back to nature*). There are also distinctions of scope or application, such as the difference between adjectives which typically qualify humans (*unschooled*) and those which typically qualify land (*unimproved*). Evaluative meanings are also relevant, as in the distinction between *natural* and the derogatory *primitive* when applied to humans or their behaviour.

Using transparent paraphrases which stay close to the English of the thesaurus itself, the structure of the NATURE entry is made explicit in Table 2. Capitals indicate the general terms used to show the structure; in some instances these simply repeat words or phrases already appearing in the entry, in others they have been introduced to declare a semantic relationship. The words within the section have not been rearranged: the words in capitals serve only to try to make explicit, in an informal way, a semantic arrangement which is already implicit.

<b>NOUNS</b>
NATURE <b>nature</b> , the great outdoors, the wild, tiger country, waste, wilderness (area).
BALANCE OF NATURE <b>balance of nature</b> , ecosystem.
STUDY OF NATURE <b>ecology</b> , autecology, bionomics, natural history, natural science, nature study, physis ( <i>Obs.</i> ), physiography, synecology
PERSON WHO STUDIES NATURE <b>naturalist</b> , bionomist, ecologist, Physiographer
PERSON WHO WANTS TO PRESERVE NATURE <b>nature lover</b> , conservationist, greenie.
PERSON WHO BELONGS TO NATURE <b>primitive</b> , child of nature, noble savage, savage.
<b>ADJECTIVES</b>
NATURAL OF HUMANS <b>natural</b> , innate, instinctive, normal, unformed, unschooled.
NATURAL OF HUMANS DEROGATORILY <b>primitive</b> , in a state of nature, feral, native, savage, uncivilised, unlearned.
NATURAL OF ANIMALS <b>wild</b> , feral, ladino, tameless, warrigal, wilding ( <i>Archaic</i> ), wildish.
NATURAL OF LAND <b>undeveloped</b> , rough, trackless, unimproved, untouched, waste.
<b>VERBS</b>
GO BACK TO NATURE <b>go back to nature</b> , escape, go bush, go wild, rough it.
<b>ADVERBS</b>
NATURALLY <b>naturally</b> , wild
NATURALLY OF HUMANS DEROGATORILY <b>primitively</b> , savagely, wildly
NATURALLY OF HUMANS <b>instinctively</b> , by birth, innately.

Table 2: Macquarie NATURE, with semantic relations

The analysis raises many questions of detail. Some relate to the kinds of informal descriptions we are using. For example, are categories like BALANCE OF NATURE and STUDY OF NATURE adequate? They stay close to the kind of English recorded in the thesaurus, but the structure 'NOUN of NOUN' is notoriously ambiguous as so many different relations can be borne by *of*.

Other questions relate to the compilation of the thesaurus. *Ecology*, *autecology* and *synecology* are clearly not synonyms; is it an error for them to be in the same synset? The adverbs have not been grouped in the same way as their adjective counterparts. Arguably, they should have been. The thesaurus may have been

compiled quickly from a variety of sources without sufficient attention to details, especially as compilers are not obliged to make the semantic organization explicit. On the other hand, it is always dangerous to assume generalized semantic frameworks and structures that do not take account of genuine usage. An example here is the relatively small number of adverbs. It would be foolish to assume that every adjective has a corresponding adverb. There are adjectives such as *feral* and *waste* which clearly belong in this section of the thesaurus (*feral cats, feral instincts, waste ground, waste land*) but which have no cognate adverbial forms and are used adverbially only in very restricted combinations (*go feral, lay waste*). The point is not to attack or defend the arrangement of any particular thesaurus, nor to justify the details of the kind of informal analysis we are suggesting here. Rather, the point is that any thesaurus does imply considerable organization, even if that organization is not clearly presented to the user (and, indeed, may not have been clearly in the minds of the compilers).

## 2.2 WordNet/EuroWordNet

WordNet (Fellbaum, 1998) is a lexical database produced on psycholinguistic principles. It has been developed at Princeton University and, for ten years now, has been freely available. It has been very extensively used in language engineering research. WordNet is an English-language resource: in the recent EU project EuroWordNet (Vossen, 1998) (<http://www.hum.uva.nl/~wn>) and in a number of associated research activities wordnets for several other languages have been developed, on the same basic plan but with further sophistications and with the added benefit of the Inter-Lingual Index, which links synsets of different languages.

The original WordNet (with capitalised N) is principally organised according to synonymy and hyponymy for nouns and verbs<sup>1</sup> and antonymy for adjectives. Each meaning of each word is located in a synset ("synonym set") and synsets stand in hierarchical and other relations to each other. There are around a dozen further lexical relations linking synsets (or, on occasion, word meanings). The database keeps the four open-class parts of speech distinct, so there are almost no relations linking, for example, nouns and verbs. (This is one limitation that EuroWordNet has stepped beyond.) In WordNet 1.5 there are 25 top-level classes for nouns and, for verbs, there is a top-level classification into 15 types (though this is not straightforwardly hierarchical: a synset belonging to one type may have, as its superordinate, a synset of another type.)

Wordnets are thesaurus-like rather than dictionary-like in that their principle mode of organisation is the synset. They guard against Landau's complaint by only allowing specified semantic relations between word-meanings and synsets.

The WordNet 1.5 account of *nature* comprises six meanings, in six synsets. The three that relate to the Macquarie paragraph are shown below.

**1** *nature, wild, natural state, state of nature* - (a wild primitive state untouched by civilization; "he lived in the wild"; "they tried to preserve nature as they found it")

→ *state* - (the way something is with respect to its main attributes; "the current state of knowledge"; "his state of health"; "in a weak financial state")

**2** *universe, nature, creation, world, cosmos, macrocosm* - (everything that exists anywhere; "they study the evolution of the universe")

→ *natural object* -- (an object occurring naturally; not made by man)

**3** *natural phenomenon, nature* -- (all non-artificial phenomena)

→ *phenomenon* -- (any state or process known through the senses rather than by intuition or reasoning)

The entries as presented here comprise the sense number, the synset, a gloss (in brackets, similar to a dictionary definition), and then, following the arrow, the superordinate synset and its gloss. *Nature* is alone in its synset for senses 1, 2 and 5.

<sup>1</sup> WordNet distinguishes "troponymy", for verbs, from hyponymy for nouns. but both play a similar role in organising the hierarchy (Fellbaum, 1998, p 79-87).

WordNet is a database and a number of other presentation forms are available. All the options cannot be shown here, but the range of semantic relations is still strictly limited, in contrast to the Macquarie thesaurus. If two synsets are not related by one of the small set of semantic relations, then no relation between them will be recorded, however intuitively 'close' in meaning they may be. WordNet suffers from the "tennis problem". It classifies nets and rackets and umpires, but offers no way of associating them all as concepts related to tennis. This is in contrast to the Roget-type strategy, where the connectedness of the words alone supports putting them together. In the Macquarie Thesaurus, the keyword SPORTS has a long list of paragraphs, with headings such as MOVES AND STROKES, SCORING, and subheadings under each of these for the different sports, so a motley collection of tennis terms, including *net* and *umpire* but not *racquet*, can be collected there.

### 2.3 IR-manual thesauruses

In many specialist areas where information retrieval systems are widely used, domain-specific thesauruses have been developed. As with WordNet, the organising principles are synonymy and taxonomy, which make it possible for searches to be broadened or narrowed, and for searches to be matched against documents using synonyms of the search terms. (Baeza-Yates et al., 1999) present the basic relations for IR-manual thesauruses as BT (broader term), NT (narrower term) and RT (related term). IR-manual thesauruses will often also use semantic relations of particular salience in the domain, for example medical thesauruses may include relations such as "located", "prevents" and "diagnoses". A resource such as the Unified Medical Language System (UMLS) is a highly sophisticated object incorporating a very large quantity of medical knowledge and supporting inference of various kinds (EAGLES, 1999). IR-manual thesauruses are domain-specific, and are thereby not the core concern of this paper. The key difference between WordNet and IR-thesauruses is, arguably, that WordNet addresses general language.

### 2.4 Automatic thesauruses

There is a substantial body of work on the automatic generation of thesauruses and related resources from large corpora. Some of this work takes place under the heading of NLP, and some under the heading of Information Retrieval.

The simplest strategy for automatic thesaurus generation is:

**For each content word in the corpus  
for each other content word,  
find how often both occur within k  
words (or characters) of each other.**

If there are  $n$  content words in the corpus, each word can then be represented by a vector of length  $n$ ; similarities between vectors can be computed using any of a variety of similarity measures, and for each word we can identify the most similar words.

Schütze's "word space" (Schütze, 1998) is defined in this way; he then uses a mathematical technique (SVD, singular value decomposition) to reduce the dimensionality of the space. Latent Semantic Indexing (LSI) (Deerwester, 1990) also uses SVD, but applies it to a matrix of counts of words in documents, so words label the rows of the matrix and documents label the columns (or vice versa). (Hindle, 1990), (Lin, 1998), (Grefenstette, 1994a), (Grefenstette, 1994b) all find and count triples of **grammatical-relation, word1, word2** rather than simple unordered word co-occurrences.

Grefenstette (1994a) distinguishes first, second and third order affinities between words. First order affinities are between words that co-occur with each other. Second-order affinities are between the words that co-occur with the same words. Thus words with complementary distributions, such as *tumor* and *tumour*, have no first-order affinity but a marked second-order one, since documents will either contain *tumor* or *tumour* but not both, yet the two will occur in the same kinds of contexts. Spelling variants will be an extreme case of words having a strong second-order affinity yet no first-order affinity. Third order affinities relate to distinct senses of polysemous words: we would like to identify that *bank* }

in one sense, has an affinity with *river*, and in another, an affinity with *business*.

## 2.5 Discussion

Note the parallels between 'looser' and 'tighter' manual and automatic thesauruses. A second-order automatic thesaurus like Lin's, where words are deemed similar to the extent that they occur in the same grammatical relations with the same other words, will tend to give sets of words in the same semantic class. LSI, which treats words as more similar, the more documents they co-occur in, and is thereby closer to a first-order technique, will be more akin to a classification of words according to the domains or 'subject fields' they occur in, and closer to a Roget-style thesaurus. Looser thesauruses such as LSI and Roget-type will be suited to different language engineering uses than tighter ones. For Information Retrieval purposes such as finding related documents, connectedness is of interest irrespective of the semantic relation. For other purposes, such as developing a lexicon with detailed selection restrictions for verbs, the tighter thesauruses (either automatic, or WordNet, or IR-manual) will be appropriate.

## 3 SENSEVAL

There are now many computer programs for automatically determining which sense a word is being used in. One would like to be able to say which were better, which worse, and also which words, or varieties of language, presented particular problems to which programs. To this end, an evaluation exercise, *Senseval*, was organised<sup>2</sup>. It comprised word sense disambiguation (WSD) tasks for English, French and Italian. The exercise is chronicled in a Special Issue of *Computers and the Humanities* (Kilgarriff and Palmer, 2000)<sup>3</sup>. A second SENSEVAL will take place in 2001.

The form of SENSEVAL 1 was a quantitative, competitive evaluation (as in the ARPA evaluations model, see (Hirschman, 1998)). The model is like an exam. The 'examinees' are word sense disambiguation programs; that is, programs which, given a dictionary and a text, assign sense numbers from the dictionary to words in the text. The programs (or their creators) are given texts they have not seen before, and must return their examination scripts comprising the answers their program gives by a given date. The organisers then mark the scripts: they can do this because the disambiguation task has also been done by people, to give a 'gold standard' set of correct answers. Results are then published, compared and discussed at a workshop.

In SENSEVAL-1, there were tasks for English French and Italian, with a total of 25 participating programs. The largest exercise was for English, with 18 programs and a test set of 8448 word occurrences to be disambiguated. The dictionaries that provided the sense inventories were HECTOR (an experimental resource produced in a joint project between Oxford University Press and Digital (Atkins, 1993) for English, Le Petit Larousse Illustré (1995) for French, and Garzanti (1995) for Italian.

One issue that SENSEVAL needed to clarify was the extent to which people would agree on the appropriate sense tag to assign to a corpus instance. If people often disagreed, the validity of the evaluation would be in doubt, as different human taggers would have produced a different gold standard. To this end, all evaluation materials were tagged by two or more people; where they gave different tags, the matter was arbitrated by a third person (for English) or either tag was considered correct (for French and Italian). For English, a small replication study established that, if the whole procedure was repeated, we should expect to get the same tags again 95% of the time.

The best scores achieved by programs were around 77%. All the highest-scoring systems were 'supervised training' systems: that is, they learned how to distinguish senses not directly from the dictionary, but from sentences containing the word to be disambiguated which had previously been disambiguated by a person. Clearly, this raises questions: manually disambiguated training material is not a raw material that is

---

<sup>2</sup> SENSEVAL took place under the auspices of ACL SIGLEX (the Lexicons Special Interest Group of the Association for Computational Linguistics), EURALEX (European Association for Lexicography), ELSNET, and EU Projects SPARKLE and ECRAN.

<sup>3</sup> *Senseval* materials are available at <http://www.itri.bton.ac.uk/events/senseval>

generally available. The issue of making do with very little training material, or using ingenious methods to 'bootstrap' it with minimal human intervention, is being vigorously pursued at several sites.

#### 4 Thesaurus word senses and dictionary word senses

At one level, the difference between a dictionary and a Roget-type thesaurus is one of indexing: the dictionary is organised alphabetically, the thesaurus by meaning or word group. If this were the only difference, a computational environment that offered both indexing possibilities would remove the distinction, and a resource such as WordNet, which offers both options, would be equally dictionary and thesaurus.

But there are further differences. Firstly, most published dictionaries give only limited space to word clusters. Most Roget-type thesauruses do not include definitions, and group words according to implicit rather than explicit semantic categorizations, so the information for reading a resource either way is absent. Secondly, most existing resources have been developed from the one perspective or the other, but not both. When a lexicographer is producing a dictionary entry, the goal is to provide a coherent analysis that separates out the distinct meanings and patterns of use the word has, with each part of the entry making sense in relation to the others. When s/he is producing a thesaurus entry (at least for paper publication), the unit which must appear coherent is the thesaurus entry or word group. Thus where a word has two distinct but closely-related meanings, but the distinction is not salient for other words and the senses both fall in the same thesaurus category, the compiler will not present the word twice in the same thesaurus entry. Without definitions justifying the different senses, the presentation of the same word twice would be confusing to the user. So a dictionary distinction may be lost in the thesaurus. Conversely, a single dictionary meaning is commonly found in more than one section of the thesaurus.

Consider the word *listless* in the Macquarie Dictionary (Macquarie, 1997) and Macquarie Thesaurus. In the dictionary, *listless* has two definitions:

1. feeling no inclination toward or interest in anything.
2. characterised by or indicating such feeling: a listless mood.

The difference is between the adjective describing persons ('they all seemed quite listless') and the adjective applied to certain other nouns ('a listless mood', 'a listless wave of the hand').

In the Thesaurus, *listless* appears in three places, within the sections headed BOREDOM, IDLENESS and APATHY. No-one would claim that these three nouns are strictly synonymous, but they are close, and a user will be happy to find *listless* in any of the three sections. In none of these three sections is there any attempt to differentiate two senses of *listless*, one applying to persons, the other applying to other nouns. A comparable extension of other adjectives, as in 'a tired atmosphere', 'an idle moment', 'a bored glance', is assumed in various parts of the Thesaurus. So the inclusiveness of a thesaurus allows *listless* to be entered under different semantic headings that are not specified in different dictionary senses of the word, while the two dictionary definitions are not distinguished at all in the thesaurus.

Mapping from the senses in one dictionary, to the senses in another, is very often difficult or impossible simply because the lexicographers have chosen to divide up the semantic space in different ways (Stock, 1983, Atkins and Levin, 1991). When thesaurus senses are to be compared with dictionary senses, the likelihood of a clear mapping declines further as the different organisation of the two books imposes different requirements on how the lexicographer should analyse a word's meaning.

#### 5 'Lexicographically Close' and 'Hierarchically Close' Polysemy

We have further explored relations between dictionary and thesaurus word senses as follows.

A central feature of the first three varieties of thesauruses is their hierarchical or network organisation. This offers many benefits for language engineering, including the potential for measuring semantic similarity between two word meanings by finding the length of the shortest path between them across the network. WordNet has been used extensively in this way, with various measures proposed and explored



(see papers in Fellbaum 1998). One interesting possibility is that graph-based metrics can be applied to pairs of meanings of a single word. One might suppose that, where two different meanings of the same word are 'close' in meaning, they will be found in 'close' parts of the thesaurus network. If this were so, it would be useful: for many language engineering purposes, WordNet word senses are often viewed as too fine-grained (as are LDOCE's (LDOCE, 1978) and CIDE's (CIDE, 1995) and other dictionaries' senses). Finer-grained senses produce more ambiguity. Some words which, given coarser-grained senses, would not have been ambiguous, now will be. They also tend to make the disambiguation problem harder, as there will be more meanings to select between. At discussions at the SENSEVAL workshop on evaluating WSD systems (Kilfarriff and Palmer, 2000), numbers of people shared the opinion that a coarser-grained sense inventory was required for Language Engineering, and should be sought for further WSD evaluations. The relation between the grain-size of the sense distinctions, and the thesaurus hierarchy, rests on the assumption that finer-grained sense distinctions correspond to distinctions between items at or near the leaves of the taxonomic tree. Then, looking only at coarser distinctions would correspond to ignoring the distinctions in the hierarchical tree of greater than a certain depth. In many dictionaries, dictionary entries are hierarchical, with subsenses (and sub-subsenses) indicating fine distinctions, and it is tempting to think that the structure of the individual word's dictionary entry will map onto the overall thesaurus hierarchy for the full language. This was the assumption we investigated.

The hypothesis is that 'lexicographically close' word sense are 'hierarchically close'. We define hierarchically close senses as ones that share a superordinate (directly, or at one or two removes) in the thesaurus hierarchy. 'Lexicographically close' is less straightforward to define; lexicographically close senses are those that are often confused, or overlapping, or where a distinction may be made in one dictionary but not in another. We considered various methods for identifying lexicographically close senses, including finding which pairs of senses were often confused where people sense-tagged corpus data, or where a lexicographer was often unsure which of a pair of senses applied in a given corpus instance. We did not have data available for either of these strategies. A third strategy involved mapping between dictionaries. If such a mapping exists, then, where two senses of the first dictionary both map to the same sense of the second dictionary, we say that the two senses of the first dictionary are lexicographically close.

For this strategy, we did have data available. In the course of SENSEVAL a mapping had been produced, for 41 words, by a professional lexicographer, from WordNet senses to HECTOR senses<sup>4</sup>.

## 5.1 Experiment 1

The mapping, for one of the sample words, *excess*, was as follows.

- 1: n: aglut or surplus or toomuch
- 2: n: ott or toomuch
- 3: n: toex
- 4: n: overind

The first of the colon-separated columns is the WordNet sense, the second, the part of speech, and the third, a mnemonic (or series of mnemonics separated by *or*) for the HECTOR senses. Thus WordNet sense 1 maps to HECTOR senses *aglut, surplus or toomuch*.

Wherever two or more WordNet senses mapped to the same HECTOR sense (or there was a non-empty intersection of the two sets of HECTOR senses, as here) the pair or triple of WordNet senses was declared 'lexicographically close'. Here, 1 and 2 are lexicographically close because they share the HECTOR sense *toomuch*. There were 30 such pairs or triples.

Each pair and triple was examined in WordNet, to establish whether the items were hierarchically close. Of the seven adjective pairs, little could be said because adjectives are not organised hierarchically. Of the 9 noun pairs, there is just one pair of 'sisters', sharing a superordinate. One pair meet one further step up the tree, a further two meet several steps up the hierarchy, and the remaining five do not meet at all but are classified under different top nodes. The full set of lexicographically-close noun pairs is presented in

---

<sup>4</sup> The lexicographer was Clare McCauley.

Table 3. Of the 14 verb pairs and triples, five shared a superordinate; the other nine did not, and indeed did not share the same top level category<sup>5</sup>.

Word	Sense	Gloss/synset/top-level category (in italics)	Verdict	WN group
band	1	instrumentalists not including string players	Sisters	Yes
	5	a group of musicians playing popular music for dancing		
behaviour	1	demeanor, trait ( <i>attribute</i> )	Remote	
	3	manner of acting ( <i>act</i> )		
bet	1	money risked ( <i>possession</i> )	Remote	
	2	<i>act</i>		
excess	1	a quantity much larger than needed	Meet high	Yes (with others)
	2	excessiveness, immoderation		
giant	4	very large	Meet at Persons	
	5	abnormal		
onion	1	plant	Remote	Yes (with another)
	3	food		
sack	1	sackful ( <i>measure</i> )	Remote	
	7	bag ( <i>artifact</i> )		
sanction	1	formal and explicit approval ( <i>communication</i> )	Remote	
	3	authorization ( <i>act</i> )		
scrapheap	1	<i>location</i>	Remote	
	2	<i>group</i>		

Table 3: Lexicographically-close noun sense pairs

## 5.2 Related Work

The WordNet database itself has a notion of lexicographically close, or 'grouped' senses<sup>6</sup>. These are 'sisters', 'cousins' and 'twins'. Sisters are two senses of the same word in synsets which share a superordinate (see *band* in Table 3) Cousins are related by one of 88 relations of which some instantiate cases of regular polysemy, such as container/containerful (see *sack*). Twins are synsets with three or more members in common. The creators of WordNet have manually checked all pairs which are lexicographically close according to these criteria, and maintain a list of exceptions to the grouping principles, that is, sense pairs which meet the criteria but which are not lexicographically close. Grouping has only been undertaken for nouns. The last column of Table 3 indicates where a pair was classified as belonging in the same group in WordNet. For *excess* and *onion*, WordNet grouped all senses. For *giant*, WordNet grouped some senses, but not the two which were lexicographically close according to our criteria.

There is comparable work for verbs, but it is more specific in focus and explores interactions with syntax in diathesis alternations (Levin, 1993, Kohl et al., 1998, Dang et al., 1998).

(Peters et al., 1998) had the goal of clustering WordNet 1.5 senses in order to cluster some fine-grained sense distinctions within the Inter-Lingual Index (ILI) of the EuroWordNet database (Vossen, 1998; <http://www.hum.uva.nl/~wn/>). WordNet 1.5 was the starting point for the ILI, but if lexicographically-close sense pairs are present in the ILI, it is always likely that one EuroWordNet for, e.g., Spanish, will link their synset to one of the pair, while a EuroWordnet for, e.g., Dutch, will link to the other of the pair, and then the possibility of linking the Dutch and Spanish words will have been missed. The first goal was to find

<sup>5</sup> Sometimes verb synsets of one top-level category have, as superordinates, synsets of a different top-level category, so there are several ways in which this could be computed. In this dataset there was just one case where a pair which did not share a top level category directly did have superordinates which shared a top level category

<sup>6</sup> WordNet is available from <http://www.cogsci.princeton.edu/~wn/> --see manual page for 'groups'.

lexicographically close senses. Their starting points were sisters, cousins and twins, as above, and also auto-hyponyms -- words with one sense which is a hyponym for another of its senses -- and links with other resources. They gather some evidence that, if the ILI is rationalised through merging close senses, it functions more effectively as a mediator between languages.

Both WordNet and Peters make extensive use of regular polysemy (see section 6). (Buitelaar, 1998) takes this further in CoreLex, a lexical resource derived from WordNet which gives centre stage to the principles of the Generative Lexicon (Pustejovsky, 1995). The principle is similar to WordNet cousins, but a wider range of regular polysemous relations, applying to larger classes of words, is assumed, with the result that the semantic relations holding between pairs in CoreLex are rather loose (see section 6.3.1 for a more detailed discussion).

### 5.3 Discussion

Despite the small size of the sample, the evidence is resounding: it is invalid to assume that lexicographically close senses are hierarchically close.

As the related work indicates, the reason is often that close senses are related by polysemous relation which cut across the hierarchy. High-level categories in the hierarchy often relate to different facets of the same object or event, as where *bet* is either the money risked, or the act of betting. There are classes of cases where the relations are regular. In WordNet, some of these are captured by 'cousins', and EuroWordNet has taken the process one step further, with synsets having multiple parents, in a lattice rather than a hierarchy, and inheriting one 'sense' from each<sup>7</sup>. The cases where lexicographically close senses are close according to some regular criterion have been explored and account for a substantial share of the data<sup>8</sup>. The 'sisters' cases were of two varieties. In cases such as *band*, the denotations of the two senses were clearly different so it was apparent why the meanings, though close, were distinct. However for senses 2 and 5 of verbal *seize*:

2 appropriate, seize, take over, take possession of - (take without permission)

5 capture, seize, take over, conquer - (as of land)

which both have as superordinate 'take, take by force; "Hitler took the Baltic Republics"; "The army took the fort on the hill" it is simply unclear what distinction the lexicographer intended. This was sometimes also the case for pairs which were not close in the hierarchy, as with nominal *sanction*. Where the distinction was not clear in WordNet, the lexicographer producing the WordNet/HECTOR mapping was unlikely to distinguish them as relating to different HECTOR senses.

The conclusion of the experiment is that the hierarchical structure of WordNet alone cannot be used for moving from a fine-grained inventory of word senses to a coarse grained one. 'Hierarchically close' is just one of a number of possible relations between lexicographically close word senses. There are many cases where it is hard to identify the distinction the lexicographer intended, and many which cannot be easily or usefully classified.

Regular polysemy relations account for a substantial number of lexicographically close pairs. The obstacle here is finding and identifying the set of regular polysemy relations, and the Wordnet categories they connect. The next experiment directly addresses this issue.

## 6 Automatic Discovery of Regular Polysemy

In this part of the paper we describe an automatic procedure for finding patterns of regular polysemy in WordNet. The hypothesis is framed the opposite way round to the first experiment; there, the question was

---

<sup>7</sup> WordNet does make limited use of multiple inheritance, but this is more extensively used and more fully worked through in EuroWordNet.

<sup>8</sup> We examined whether any other WordNet relations held between the set of lexicographically close pairs. None did.

"how often are lexicographically-close sense-pairs hierarchically close?" Here, it is "how often are systematically-related sense-pairs, that can be identified algorithmically, lexicographically close?"

## 6.1 Definition of Regular Polysemy

Regular polysemy is defined by a set of word senses that are related in systematic and predictable ways. In principle, a pattern is considered regular if the combinations of related senses are valid for more than one word (see below). In the literature this relatedness is also described in terms of the Aristotelian tropes or figures of speech metonymy and synecdoche (Ross, 1924).

Apresjan (1973) defines regular polysemy as follows:

'Polysemy of the word A with the meaning  $a_i$  and  $a_j$  is called regular if, in the given language, there exists at least one other word B with the meanings  $b_i$   $b_j$ , which are semantically distinguished from each other in exactly the same way as  $a_i$  and  $a_j$  and if  $a_i$  and  $b_i$ ,  $a_j$  and  $b_j$  are nonsynonymous.

Polysemy is called irregular if the semantic distinction between  $a_i$  and  $a_j$  is not exemplified in any other word of the given language.' (Apresjan, 1973)

In other words, regular polysemy applies if at least two words share the same sense combination.

Regular polysemy is Apresjan's term. Closely related phenomena are sense extension (Copestake, 1995) and transfers of meaning (Nunberg, 1996). According to Apresjan, the semantic regularity it displays is a distinctive feature of metonymic transfers. On the other hand, irregular polysemy is more typical of metaphorical transfers which express a semantic similarity relation between two concepts that yields a conceptual mapping, a list of correspondences, between the source and target domains of a metaphor (Lakoff, 1993).

Regular polysemy has mostly been approached from a theoretical perspective. There is a limited set of default relations identified in the literature of which the following are a subset (Ostler and Atkins, 1991, Pustejovsky, 1995). Most of these relations have been arrived at by examination of a limited quantity of linguistic material (texts, dictionaries) or introspection.

1. container/containerful	cup
2. animal/food	lamb, chicken
3. animal/skin	crocodile
4. plant/food	banana
5. product/producer	newspaper, Honda
6. substance/colour	jade, amber
7. object/shape	pyramid
8. language/people	Spanish
9. music/dance	waltz
10. figure/ground	door/window
11. place/people	city, New York

Any discussion of regular polysemy is never far from the lexicon/pragmatics divide. Consider the words *turkey* and *ostrich*. Both may be used to denote the bird, or the food which is the meat of the bird. However the use of *turkey* to mean the food is familiar (to the speaker community of which the authors are part) in a way in which the use of *quail* to mean the food is not. The food reading of *turkey* will be found in dictionaries, the food reading of *ostrich* will not. While both words can mean the food, for *turkey* this is lexicalized, and we shall say that this is a **sense** of the word, whereas for *ostrich*, there is no such sense; that is, there is no such meaning listed in the lexicon and we resort to pragmatic processing in order to interpret *ostrich* as food.

Whether or not a particular reading for a word is lexicalized depends on the frequency of occurrence of the sense, and the level of convention that the similarity or contiguity relation has attained in language use. The consolidation of regular polysemy into attested sense distinctions shifts and changes, reflecting various degrees of lexicalization. Some theoretical work (e.g. Copestake, 1996) views regular polysemy as a set of

processes represented within the lexicon by means of lexical rules, while other work takes the pragmatic tack, representing the productive processes outside the lexicon as pragmatically defined meaning transfers (e.g. Nunberg, 1996). The boundary between lexical and pragmatic explanations of regular polysemy is determined by synchronic analysis of language in use on the one hand and lexicographic practice on the other. The latter represents the historical picture of the development of a word's senses and is heavily influenced by the type of resource it wants to produce. Dictionary users and size determine the choice between lumping and splitting word senses (Evens, 1988, Kilgarriff, 1997), and no matter how fine-grained a dictionary's inventory of senses is, it can never be exhaustive, nor can it fully cover the dynamic meaning potential of language.

## 6.2 Regular Polysemy in WordNet

In general, dictionaries are not primarily directed towards encoding semantic regularities across the lexicon, but towards distinguishing senses that enable the user to interpret texts. WordNet with its dictionary-based sense distinctions forms no exception; the semantic encoding of WordNet is not aimed at the implementation of explicit patterns of regular polysemy. However, a number of semantic regularities have been captured by the so-called 'cousin' relation mentioned in section 5.2, which is used to group concepts according to similarity of meaning. The relation is computed from a table of 88 concept pairs in WordNet1.5, and 220 in WordNet1.6, the latest version of the thesaurus. All their hyponymic concepts, i.e. all concepts that occur below them in the hierarchy, are assumed to be involved in the semantic relation. Exceptions are listed in a separate file.

Although several regularities have been observed, there is no explicit description of the semantic relations covered by the concept pairs. Close inspection shows that these pairs implicitly signal a semantic relation which is not always metonymic. While the relations are a useful starting set, neither the relations nor the exceptions form a complete set

## 6.3 Detecting Regular Polysemy in WordNet

We have operationalised Apresjan's definition of regular polysemy by exploiting WordNet's hierarchical structure: wherever there are two or more words with senses in one part of the hierarchy which also have senses in another part of the hierarchy, then we have a candidate pattern of regular polysemy. We have restricted our experiments to cases where the related meanings are of the same syntactic class (nouns), and we have explored various ways of defining "the part of the hierarchy". We would expect the procedure to produce a high level of noise, where words A and B both have senses in both parts I and J of the hierarchy but there is no great similarity in the relations between the meaning pairs. We would expect more noise when we look at larger parts of the hierarchy. We would also not expect the procedure to discover all regular polysemy relations.

### 6.3.1 Step 1: Combinations of High-Level Concepts

In our investigation, we examined a number of combinations of WordNet1.5 unique beginners (i.e. the top level nodes of individual branches within the WordNet taxonomy) that share the same word form as a hyponym at any level in the hierarchy. Examples of unique beginners are:

<b>artefact-1</b>	a man-made object
<b>group-1</b>	any number of entities (members) considered as a unit

The initial hypothesis was that combinations of these unique beginners yield regular polysemous sense distinction patterns at a very general level of description. Previous work has been performed by Buitelaar (1998) who extracted 126 classes consisting of combinations of 35 high-level concepts in the WordNet hierarchy, including WordNet's 24 unique beginners. These classes display various types of regular polysemy from WordNet1.5.

We have examined pair-wise combinations of unique beginners. The examples below lists a small subset of the words that occur as hyponyms of both **artefact** and **group**:

**institution-2** a building or complex of buildings where an organization for the promotion of some cause is situated

**institution-1** an organization founded for a specific purpose

**guard-3** a device designed to prevent injury

**guard-5** a group of men who escort and protect some important person

**menagerie-2** the facility where wild animals are housed for exhibition

**menagerie-1** a collection of live animals for study or display

**shower-1** a plumbing fixture that sprays water over you

**shower-5** a party held by friends to present gifts to a person

**type-6** a small block of metal bearing a raised character on one end; produces a printed character when inked and pressed on paper; “he dropped a case of type so they made him pick them up”

**type-3** (biol) the taxonomic group whose characteristics are used to define the next higher taxon

**unit-6** a combination of interrelated interacting elements designed to work as a coherent entity

**unit-2** an organization regarded as part of a larger social group

Further investigation of the above examples shows some disadvantages of restricting regular polysemy to combinations of high level concepts. It leads to inappropriate instantiations of a pattern, as in the case of *shower* and *type*, where there is no systematic relation between the two senses. Whereas there seems to be no meaning relation between the various senses of *type*, the relation between the two senses of *shower* is an isolated case of metaphorical transfer. It also leads to the creation of groups that contain words that are *not semantically similar*. For instance, *menagerie* and *guard* seem to display a ‘facility/collection’ and a ‘group/device’ alternation respectively. Using high level concepts for the characterization of regular polysemy can block the identification of *subgroups* that are semantically more coherent.

### 6.3.2 Step 2: Combinations of More Specific Concepts

The next step was to identify combinations of hypernyms at a more specific level, giving rise to semantically more coherent groups. We selected pair-wise combinations of nodes in the WordNet hierarchy that were more specific than the unique beginners but still general enough to cover a substantial set of words. The nodes we considered met the following criteria:

1. Node pairs must function as a hypernym of at least three words having at least one sense in both unique beginner (UB) branches. This is in fact a stronger constraint than the one expressed in Apresjan’s definition cited above where two words already constitute a regular polysemic pattern.
2. The distance of the word senses to the hypernym in terms of node traversal in the WordNet taxonomy is never greater than 4. This distance has mainly been chosen to limit processing time and volume of output data.

Figure 1 is a graphical representation of the two WordNet branches headed by the unique beginners ‘artefact’ and ‘action’ covering the regular polysemic pattern music – dance. The words are all within the maximum distance to the conceptual signposts ‘music’ and ‘dance’, but not at the same level in the hierarchy.

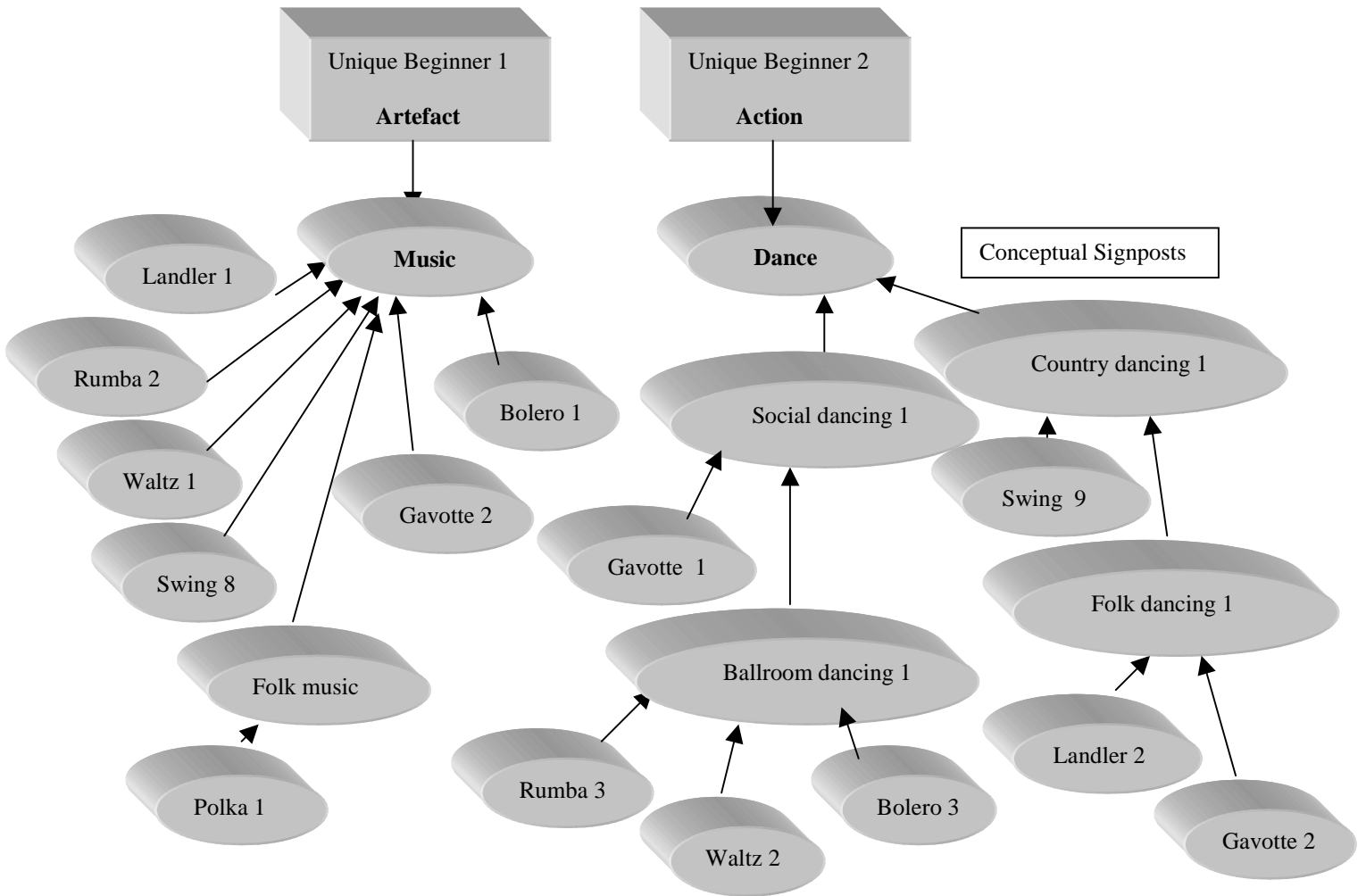


Figure 1: The systematic polysemic pattern **music - dance**

A manual examination of a number of UB combinations yielded the following subgroups. Table 4 to 7 below list regular polysemic patterns under different UB combinations.

Relation	Examples	No of words covered
holdfast – control	lock, clasp, clench	5
picture – painting	etching, engraving, fresco	5
music – dance	waltz, rumba, bolero	20
communication system – broadcast	radio, television, wireless	3
path – travelling	crossing, walk, promenade	4

Table 4: UB pair Artefact - Action

<b>Relation</b>	<b>Examples</b>	<b>No of words covered</b>
publication-publisher	paper, newspaper, magazine	3
musical composition – group of singers	trio, quartet, suite	16
building – institution/association	school, chamber, court	15
package/container – collection	parcel, bundle, pack	5
music – arrangement/formation	line, arrangement, chorus	3
construction – body of people	house, body, camp	5

Table 5: UB pair Artefact – Group

<b>Relation</b>	<b>Examples</b>	<b>No of words covered</b>
supporting structure – theory	framework, foundation, base	5
musical theme – idea	theme, motif, strain	4
concrete obstruction - abstract obstruction	barrier, roadblock, hurdle	8
device - something having influence over somebody	bait, lure, support	3
artistic work – idea	design, motif, pattern	3
product – ability	art, innovation, invention	3
concrete representation/creation - mental representation	figure, model, scene	8

Table 6: UB pair Artefact – Cognition

<b>Relation</b>	<b>Examples</b>	<b>No of words covered</b>
musical composition – amount of items of the same kind	trio, quartet, potpourri	9
long thin implement - unit of measure	pole, rod, yard	3
container - quantity	barrel, firkin, kettle	33

Table 7: UB pair Artefact - Measure

## 6.4 Discussion

### 6.4.1 Evaluation

We evaluated around 100 of the concept pairs that the automatic procedure delivered, by examining the word-sense-pairs and assessing whether the same relation held between each pair, i.e., whether it was a case of regular polysemy.

The concept pairs that were evaluated as valid show a low percentage of unsuccessful combinations. Around 10% of the members extracted on the basis of a particular conceptual signpost pair were considered invalid instantiations of that pattern. The observed regular polysemic patterns correspond in a number of cases to relations attested in the literature (see list in section 6.1). This is particularly true for those sets containing a relatively large number of members, such as ‘music’ – ‘dance’ and ‘container’ – ‘quantity’.



#### 6.4.2 Further work

The methodology described in this paper has several limitations. First, only pair-wise combinations of unique beginners have been taken into account. Combinations of three or more are, in contrast to Buitelaar's classification, only observable in an indirect way, by examining the overlap of sets that have members in common. For example, table 5 and 7 link musical composition, group of singers and amount of items of the same kind (as expressed by words such as 'trio' and 'quartet').

The second limitation is that it is difficult to automate the fine-tuning of the extraction of suitable conceptual pairs derived automatically on the basis of the criteria described above. Several other criteria may be used for this purpose such as the level of the conceptual pair in the hierarchy. In this case there is a trade-off between generality and specificity: the more specific the conceptual pair, the more defined the semantic relation that holds between the concepts in the pair, but also the smaller the set, and it may well happen that a meaningful set is cut up into several overly specific subsets. The best trade-off might be found by determining the most specific subsumer with the highest information content (Resnik, 1999) from the candidates produced by the technique described above.

One final limitation is that imposing a node traversal limit of 4 between hyponymic concepts and conceptual pairs may be too much dependent on the assumption of a balanced hierarchy. In fact, the hierarchies in WordNet are far from balanced, and therefore expansion further downward may yield more candidates.

#### 6.4.3 Productivity

It seems plausible to generalise the regular polysemic patterns over all members of the semantic class captured by the pattern, and postulate new, derived senses for words that only occur in the base sense of a pattern. This would lead to a significant systematic expansion of the semantic coverage of WordNet. For instance, any word with a 'container' sense could automatically be assigned a potential 'quantity' sense if that sense is lacking in WordNet. For example, this would be true for *amphora* and *parcel*. The underlying assumption here is that there are many words sharing the same meaning extension potential other than the words whose list of senses in the dictionary explicitly reflects the regular polysemy patterns. This offers the possibility to enrich sense descriptions in the resource and license dynamic activation of regular polysemic patterns. However, there are several issues that need further attention.

Firstly, we need to determine the base sense of a particular conceptual signpost pair. For example, where words that occur in an animal sense can theoretically occur in a food sense, the reverse implication does not hold. In other words, the animal sense is the primary or base sense of this combination.

Furthermore, activation of the postulated sense might never happen because of lexical blocking such as usage of 'pork' instead of 'pig', pragmatic considerations such as poisonous food substances, lexical preference for derivational forms or the improbability of the context in which the postulated sense can be activated.

#### 6.4.4 Metaphoric Transformations

Some of the unique beginner combinations reflect a metaphoric instead of a metonymic relation (see table 6). Although these are generally not considered productive, the results from the evaluation indicate a dependency on a specific unique beginner pair, *artefact-cognition*. We seem to have stumbled on a metaphorical domain where properties of artificial structures are mapped onto cognitive skills.

Different types of metaphorical transfer can be identified, such as 'theory as a supporting structure for thought' and 'a melodic theme as pervasive idea'. The fact that this domain is particularly productive for metaphoric extension does seem to indicate that extensions like these can behave more or less regularly as soon as the a correspondence between two domains has been established.

## 6.5 Conclusion

Mining a linguistic resource with semantic information in search for regularities in lexicalization patterns can give insight into:

- the level of encoded semantic regularity in the resource;
- the orientation of lexicographic practise towards this phenomenon;
- the number and nature of lexicalized metonymic patterns;
- the semantic potential of words that guides their dynamic interpretation when used in new contexts;
- the cognitive processes that underlie metonymic sense extensions / metaphorical transformations;
- the taxonomic balance of the resource in terms of chain length and choice of concept nodes.

The resulting enriched lexical knowledge will be useful for a variety of linguistic, lexicographic and language engineering purposes, including providing grouped word senses which are more suited to many language engineering tasks.

## 7 Acknowledgments

Part of this work was instigated under a Macquarie University Research Grant Visiting Fellowship to Kilgarriff for October 1999, the EuroWordNet project (LE-4003) and was further supported by the UK Engineering and Physical Sciences Research Council under grant M54971 (Wasps) and M73521 (Malt). It benefited from discussions with Robert Dale (Language Technology Group, Macquarie Univ), and Richard Tardif and James Lambert (Macquarie Dictionary).

## References

Apresjan, J. (1973).

*Regular Polysemy*

In: *Linguistics* 142: 5-32

Atkins, Sue (1993)

*Tools for computer-aided corpus lexicography: the Hector project.*

*Acta Linguistica Hunga-ica* 41:5-72.

Atkins, B. T. S. and Levin, B. (1991)

*Admitting impediments.*

In Uri Zernik (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon.*

Hillsdale, New Jersey: Lawrence Erlbaum, pages 233-262.

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (1999)

*Modern Information Retrieval.*

ACM Press and Addison Wesley .

Buitelaar, P (1998)

*Corelex: Systematic Polysemy and Underspecification,*

Ph.D., Department of Computer Science, Brandeis University, Boston, U.S.A.

CIDE (1995)

*Cambridge International Dictionary of English.*

CUP, Cambridge, England.

Copestake, A. and Briscoe, T. (1996), *Semi-productive Polysemy and Sense Extension*

In: Pustejovsky, J. and Boguraev, B. (Eds.), *Lexical Semantics, the Problem of Polysemy,*

Clarendon Press, Oxford

Dang, H.T., Kipper, K., Palmer, M. and Rosenzweig, J. (1998)  
Investigating Regular Sense Extensions Based on Intersective Levin Classes.  
In *Proc. COLING-ACL*. Montreal.

Deerwester, S., Dumais, S., Furnes, G., Landauer, T. and Harshman, R. (1990)  
*Indexing by Latent Semantic Analysis*.  
Journal of the American Society for Information Science 41 (16): 391\407.

EAGLES (1999)  
Preliminary recommendations on semantic  
encoding. Technical report, EAGLES Lexicons Interest Group.

Evens, M.W. (ed.) (1988).  
Relational Models of the Lexicon: Representing Knowledge in Semantic Networks,  
Cambridge, CUP

Fellbaum, Christiane (ed.) (1998)  
*WordNet: An Electronic Lexical Database*.  
Cambridge, Mass.: MIT Press.

Grefenstette, Gregory (1994a)  
Corpus-derived first-, second- and third-order word affinities.  
In *Proc. Euralex*.  
Amsterdam.

Grefenstette, G. (1994b)  
Explorations in Automatic thesaurus discovery.  
Dordrecht: Kluwer.

Hindle, D. (1990)  
Noun classification from predicate-argument structures.  
In *ACL Proceedings, 28th Annual Meeting*. Pittsburgh.

Hirschman, L. (1998)  
*The Evolution of Evaluation: lessons from the Message Understanding Conferences*.  
Computer Speech and Language 12 (4): 281\307.

Kilgarriff, A. (1997),  
*I Don't Believe in Word Senses*,  
Special Issue on Word Sense Disambiguation. *Computers and the Humanities*

Kilgarriff, A. and Palmer, M. (2000) Guest editors,  
Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs. *Computers and the Humanities*.

Kohl, K., Jones, D., Berwick, R. and Nomura, N. (1998)  
*Representing Verb Alternations in WordNet*.  
In Christiane Fellbaum (ed.), *WordNet: An Electronic Lexical Database*.  
Cambridge, Mass.: MIT Press,  
pages 155–178

Lakoff, George. 1993. *The Contemporary Theory of Metaphor*  
In: Andrew Ortony (ed.), *Metaphor and Thought*, second ed. New York:  
Cambridge University Press, pp. 202-251

LDOCE (1978)

*Longman Dictionary of Contemporary English.*  
Edited by Paul Proctor

Levin, B. (1993)

*English Verb Classes and Alternations.*  
University of Chicago Press

Lin, Dekang (1998)

*Automatic retrieval and clustering of similar words.*  
In COLING-ACL. Montreal.

Macquarie, (1986). *Macquarie Thesaurus.*

*Edited by J. R. L. Bernard (First published 1984).* Sydney.

Macquarie, (1997). *Macquarie Dictionary, 3rd Edition.*

*Editor in Chief Arthur Delbridge.* Sydney.

Masterman, M., (1957). The thesaurus in syntax and semantics. *Mechanical Translation*, 4:1-2.

McArthur, T. (1981)

Longman Lexicon of Contemporary English  
Longman. London and Harlow.

Miller G.A, Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. (1990),

*Introduction to WordNet: An On-line Lexical Database,*

In: *International Journal of Lexicography*, Vol. 3, No.4, 235-244.

Nunberg, G. (1996), *Transfers of Meaning,*

In: Pustejovsky, J. and Boguraev, B. (Eds.), *Lexical Semantics, the Problem of Polysemy,*  
Clarendon Press, Oxford

Ostler, N. and Atkins, B. (1991), Predictable Meaning Shift: Some Linguistic properties of Lexical Implication Rules,

In: Pustejovsky, J., Bergler, S. (Eds.), *Lexical Semantics and Knowledge Representation, ACL SIGLEX Workshop Berkeley, California*

Peters, W., Peters, I. and Vossen, P. (1998).

Automatic sense clustering in EuroWordNet.

In *Proc. First Intl Conf on Language Resources and Evaluation.*

Granada, Spain

Peters, W. and Peters, I. (2000)

Lexicalised Systematic Polysemy in WordNet

In *Proc. Second Intl Conf on Language Resources and Evaluation*

Athens, Greece

Pustejovsky, J. (1995). *The Generative Lexicon,*

MIT Press, Cambridge MA, U.S.A.

Resnik, P. (1999)

*Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language*

*Journal of Artificial Intelligence Research* 11, pp.95-130

Ross, W.D. (ed.) (1924), *The Works of Aristotle*,  
Oxford Clarendon Press

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

Stock, P. (1983)  
*Polysemy*.  
In Proc. Exeter Lexicography Conference.

Summers, D. (1988)  
*The role of dictionaries in language learning*.  
In R. A. Carter and M. McCarthy (eds.), *Vocabulary and Language Teaching*.  
London: Longman, pages 111–125.

Summers, D. (1995)  
*The Longman Language Activator*. Harlow: Longman.

Vossen, P.  
1998. EuroWordNet: Building a Multilingual Database with Wordnets for European Languages.  
In: K. Choukri, D. Fry, M. Nilsson (eds), *The ELRA Newsletter*, Vol3, n1, 19998. ISSN: 1026-8200.

Wilks, Y., Slator, B.M. and Guthrie, L. (1996)  
*Electric words: dictionaries, computers and meanings*.  
Cambridge