What's in a thesaurus?

Adam Kilgarriff[*] **and Colin Yallop** [†]

[*] ITRI, University of Brighton, UK
adam@itri.bton.ac.uk
[†] Macquarie University, Sydney
cyallop@ling.mq.edu.au

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.

In the first part of the paper, we present different varieties of thesaurus. In the second part, we contrast thesaurus word senses with dictionary word senses and present a small experiment in which we explore whether 'lexicographically close' meanings are often close in the WordNet network.

1. 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:

Roget Roget, Macquarie and others, produced, as books, to help writers with word selection

WordNet WordNet and EuroWordNet

IR-manual Thesauruses produced manually for use in information retrieval systems

Automatic '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 linguistic, 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., 1996, p 89)) and the extensive work of the Sedelows (Sedelow and Sedelow, 1992). Here our references to the literature will be indicative.

1.1. The Macquarie: a 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 Rogets 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.

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

		NATURE 494
n.	**1.**	**nature**, the great outdoors, the wild, tiger country, waste, wilderness (area); **balance of nature**, ecosystem.
	2.	**ecology**, autecology, bionomics, natural history, natural science, nature study, physic (*Obs.*), physiography, synecology.
	3.	**naturalist**, bionomist, ecologist, physiographer; **nature lover**, conservationist, greenie.
	4.	**primitive,** child of nature, noble savage, savage.
adj.	**5.**	**natural**, innate, instinctive, normal, unformed, unschooled; **primitive**, in a state of nature, feral, native, savage, uncivilised, unlearned.
	6.	**wild**, feral, ladino, tameless, warrigal, wilding (*Archaic*), wildish; **undeveloped**, rough, trackless, unimproved, untouched, waste.
v.	**7.**	**go back to nature**, escape, go bush, go wild, rough it.
adv.	**8.**	**naturally**, wild; **primitively**, savagely, wildly; **instinctively,** by birth, innately.
		Related Keywords: THE BUSH 91; FLORA 260; FAUNA 261.

Table 1: Macquarie Thesaurus, section 494, NATURE

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.

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

Table 2: Macquarie NATURE, with semantic relations

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.

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).

1.2. 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 (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[1] 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 superordinate synset and its gloss. *Nature* is alone in its synset for senses 1, 2 and 5.

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.

1.3. 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 and Ribeiro-Neto, 1999) present the basic relations for IR-manual thesauruses as BT

[1] WordNet distinguishes "troponymy", for verbs, from hyponymy for nouns. but both play a similar role in organising the hierarchy (Fellbaum, 1998, p 79–87).

(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.

1.4. 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 et al., 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) and 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*.

1.5. 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.

2. 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.

3. '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 (Kilgarriff 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 subsubsenses) 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 lexicographical 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.[2]

3.1. Experiment

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.

[2] The lexicographer was Clare McCauley. HECTOR is an experimental lexicon produced in a joint project between Oxford University Press and Digital, see (Atkins, 1993).

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 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.[3]

3.2. Related work

The WordNet database itself has a notion of lexicographically close, or ‘grouped’ senses.[4] 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 105 regular polysemy relations, such as container/containerful (see *sack*). Twins are synsets with three or more members in common. WordNet manually checks all pairs which are lexicographically close according to these criteria, and maintains 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 was classifed 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 remove some fine-grained sense distinctions from the Inter-Lingual Index (ILI) of the EuroWordnet database. 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 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. 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 *et al.* make extensive use of regular polysemy. (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.

3.3. 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.[5] 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.[6]

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 largely negative: the hierarchical structure of WordNet 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. Regular polysemy relations account for a further set of lexicographically close pairs. There are

[3] 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.

[4] WordNet is available from http://www.cogsci.princeton.edu/~wn/ –see manual page for ‘groups’.

[5] WordNet does make limited use of multiple inheritance, but this is more extensively used and more fully worked through in EuroWordNet.

[6] We examined whether any other WordNet relations held between the set of lexicographically close pairs. None did.

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 (*attribute*)	Remote	
	3	manner of acting *act*		
bet	1	money risked (*possession*)	Remote	
	2	*act*		
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 (*measure*)	Remote	Yes
	7	bag (*artifact*)		
sanction	1	formal and explicit approval; *communication*	Remote	
	3	authorization; *act*		
scrapheap	1	*location*	Remote	
	2	*group*		

Table 3: Lexicographically-close noun sense pairs

many cases where it is hard to identify the distinction the lexicographer intended, and many which cannot be easily or usefully classified.

On a more positive note, there was a degree of congruence between lexicographic closeness and the work on grouping senses automatically undertaken in WordNet and by Peters *et al.* in EuroWordNet, and the potential for exploiting that work for the next SENSEVAL exercise will be explored further.

Acknowledgments

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

4. References

Atkins, B. T. S. and Beth Levin, 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.

Atkins, Sue, 1993. Tools for computer-aided corpus lexicography: the Hector project. *Acta Linguistica Hungarica*, 41:5–72.

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto, 1999. *Modern Information Retrieval*. ACM Press and Addison Wesley.

Buitelaar, Paul, 1998. CORELEX*: Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University.

CIDE, 1995. *Cambridge International Dictionary of English*. CUP, Cambridge, England.

Dang, Hao Trang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig, 1998. Investigating regular sense extensions based on intersective Levin classes. In *Proc. COLING-ACL*. Montreal.

Deerwester, S, S. Dumais, G. Furnas, Thomas Landauer, and R. Harshman, 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.

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, Gregory, 1994b. *Explorations in Automatic thesaurus discovery*. Dordrecht: Kluwer.

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

Kilgarriff, Adam and Martha Palmer, 2000. Guest editors, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs. *Computers and the Humanities*.

Kohl, Karen, Douglas Jones, Robert Berwick, and Naoyuki Nomura, 1998. Representign verb alternations in WordNet. In Christiane Fellbaum (ed.), *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press, pages 155–178.

Landau, Sidney, 1989. *Dictionaries: the Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

LDOCE, 1978. *Longman Dictionary of Contemporary English*. Edited by Paul Proctor. Harlow.

Levin, Beth, 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, Margaret, 1957. The thesaurus in syntax and semantics. *Mechanical Translation*, 4:1–2.

Peters, Wim, Ivonne Peters, and Piek Vossen, 1998. Automatic sense clustering in EuroWordNet. In *Proc. First Intnl Conf on Language Resources and Evaluation*. Granada, Spain.

Pustejovsky, James, 1995. *The Generative Lexicon*. Cambridge, Mass.: MIT Press.

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

Sedelow, Sally and Walter Sedelow, 1992. Recent model-based and model-related studies of a large-scale lexical resource (Roget's Thesaurus). In *Proc. 15th COLING*. Nantes.

Stock, Penelope F., 1983. Polysemy. In *Proc. Exeter Lexicography Conference*.

Summers, Della, 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.

Wilks, Yorick, Brian M. Slator, and Louise Guthrie, 1996. *Electric words: dictionaries, computers and meanings*. Cambridge, Mass.: MIT Press.