Generative lexicon meets corpus data: the case of non-standard word uses

Adam Kilgarriff
ITRI
University of Brighton

January 20, 1999

Abstract

There are various ways to evaluate the Generative Lexicon (GL). One is to see to what extent it accounts for what we find in text corpora. This has not previously been done, and this chapter presents a first foray. The experiment looks at the "nonstandard" uses of words found in a sample of corpus data: "nonstandard" is defined as not matching a literal reading of any of the word's dictionary definitions. For each nonstandard instance we asked whether it could be analysed using GL strategies. Most cases could not. The chapter discusses in detail a number of non-standard uses and presents a model for their interpretation which draws on large quantities of knowledge about how the word has been used in the past. The knowledge is frequently indeterminate between 'lexical' and 'general', and is usually triggered by collocations rather than a single word in isolation.

1 Introduction

The GL claims to be a general theory of the lexicon. Pustejovsky identifies "the creative uses of words in novel contexts" (Pustejovsky1995, p 1) as one of two central issues which GL addresses, where other formal theories have remained silent. He asserts as a principle that "a clear notion of semantic well-formedness will be necessary in order to characterise a theory of possible word meaning" (em ibid, p 6) and identifies a generative lexicon as a framework in which a core set of word senses is used to generate a larger set, according to a set of generative devices. Most work in the GL tradition has been concerned to identify and formally specify those devices.

This suggests a method for evaluating the theory against corpus data. If GL is a good general theory, then all meanings of all words as found in the corpus will be in principle analysable according to the methods characteristic of GL. A GL analysis of a non-standard meaning of a word takes the word's base meaning and applies one or more of the generative devices to give the non-standard meaning. Given the youth of GL theory, one would not expect all varieties of the devices to be specified, so we would not expect every non-standard word use to be analysable according to one of the meaning-composing operations already discussed in the literature. Nonetheless, it is generally possible to say whether a non-standard use is related to a standard use in a way which would fall under some process of composition, if the catalogue of processes were complete.

2 Polysemy (in the lexicon) and non-standard uses (in the corpus)

The GL makes two sets of predictions, and in this section they are distinguished, the better to focus on the one which is the topic of this chapter.

Pustejovsky introduces the GL with reference to a set of sentence-pairs such as

> The glass **broke**.
> John **broke** the glass.

He then argues that earlier approaches to the lexicon had no option but to treat these two types of uses of *break* as distinct senses, not only for *break* but also for all the other ergative verbs, so missing a generalisation and making the lexicon far bulkier than it need be. By contrast, in the GL, *break* is underspecified for transitivity, no duplication of senses is required, and the lexicon is more compact. This suggests a lexicon-based method for empirical evaluation of the GL: given the set of pairs of two senses of the same word in a pre-existing dictionary, how much of the time can the relation between the two be accounted for by the GL? This is a question of great interest for the GL, and has been investigated at some length in (Buitelaar1997; Buitelaar1998).[1] If it is the predictions of the GL for polysemy which are to be scrutinised, then a lexicon or dictionary is the appropriate object to investigate.

The GL also makes predictions about how words may be used which go beyond anything listed in existing lexicons. A "theory of possible word meaning" will account for novel uses of words. To investigate the coverage of the GL in relation to these, we need to look in a corpus.

3 What is non-standard use?

To test whether GL accounts for novel word uses, we must first identify a set of them. This involves distinguishing standard and non-standard uses of words.

'Standard' and 'non-standard' are loaded terms. A 'standard' case tends to be the kind of case that a particular theory has a vested interest in. Textbook examples of sense extension or logical polysemy simply assert that one use of the word is 'standard', another is 'extended' or similar. For our task, such a relaxed strategy is not viable. Identifying 'standard' or 'central' or 'core' or 'prototypical' uses of a word is an arduous and challenging intellectual task for anyone —linguist or lexicographer— whose job it is to do it in a principled and systematic way.

The standard/non-standard distinction must not be confused with productive uses of language. Consider the use of *see* to mean "understand". There is a substantial literature on the productive or semi-productive process underlying the meaning transfer (Lakoff and Johnson1980; Sweetser1990) yet there is nothing non-standard about the use of *see* in "I see what you

[1] See also the close study of sense pairs in (Kilgarriff1993).

mean". Conversely, "productive" implies a rule, so to assume that all non-standard uses were productive would be to pre-judge the issue that the experiment sets out to test.

The current experiment calls for an operational definition of '(non-)standard'. The only possibility the author is aware of is to use an existing dictionary. We then classify any corpus occurrence which fits a dictionary definition of the word as 'standard', and misfits are classified as 'non-standard'.

This may seem unpalatable. Dictionaries are imperfect artifacts, produced for particular markets, under time constraints, by teams of people some of whom are more skilled than others. Any two dictionaries will differ in the meanings they say words have at innumerable points. All of this sets them at a great distance from the theoretical realm GL inhabits and makes them seem clumsy tools for evaluating the theory.

It also may be objected that a dictionary is too detailed, or too coarse-grained, for the current exercise. It may be objected that a dictionary is too detailed because it specifies, as separate senses, those productive uses that the GL would explain as the outcome of generative processes. But this objection misses the mark because, as discussed above, the experiment aims to look at non-standard uses, not the polysemy question. The divide between the two issues will be re-drawn by the selection of a dictionary but both questions remain, and the non-standard uses, according to a particular dictionary, still remain a valid dataset regarding which we can ask, "are they accounted for by the GL?"

A dictionary may be too coarse-grained because it sweeps two uses into a single sense where it is the achievement of the GL to explain the difference between the two readings. Thus *enjoy* can take a verb phrase ("enjoy doing something") or a noun phrase denoting an event ("enjoy the party") or a noun phrase denoting a physical object with an associated telic reading ("enjoy the paper"), and the GL analysis demonstrates how the third is implicit in the first, given the appropriate lexical entries and coercion mechanisms. But dictionaries do not specify the three distinct readings as separate senses. In general, it may frequently be the case that the grain-size assumed in GL work is too fine to be spotted by strategies using dictionaries. A different methodology would be required to investigate how many phenomena there were in a corpus sample that **were** susceptible to GL analysis.

Many of the distinctions that the GL provides analyses for will fall through the net of the dictionaries' senses. However that does not invalidate the ones that are caught by the net, as a suitable dataset for the experiment. Some alternations that have received GL analyses do also give rise to distinct senses in dictionaries. For example the 'container' and 'containee' readings of *cup* are each assigned their own sense in (LDOCE1995). Moreover, if the GL is a general theory of the lexicon, it should account for novelty whether or not the novelty was closely related to existing GL analyses.

Close reading of definitions from a published dictionary does not provide an ideal method for distinguishing standard from non-standard uses of words. However, the method has no fundamental flaws, and there is no better method available.

4 Experimental design

The design was as follows:

- take a sample of words
- take a set of corpus instances for each
- choose a dictionary
- sense-tag
- identify **mismatches** to dictionary senses
- determine whether they fit the GL model

The materials used for the experiment were available from another project. This was SENSEVAL (Kilgarriff and PalmerForthcoming), an evaluation exercise for Word Sense Disambiguation programs, which needed a set of correctly disambiguated corpus instances to evaluate against. The HECTOR lexical database (Atkins1993) was used. It comprises a dictionary of several hundred words and a corpus in which all the occurrences of those words have been manually sense-tagged by professional lexicographers.

For Pilot SENSEVAL, the corpus instances were tagged twice more (again by professional lexicographers), and where the taggers disagreed the data was sent to an arbiter. The taggings thereby attained were 95% replicable (Kilgarriff1999; Kilgarriff and PalmerForthcoming).[2]

Sample of words

In most GL work, words to be studied have been hand-selected according to the interests and hypotheses of the researcher. For a study such as this (and indeed any study which explores the viability of the GL as a general theory of the lexicon) it is essential to approach sampling more systematically. A random sample of the words available in the HECTOR dictionary was used.[3] The words investigated were *modest, disability, steering, seize, sack*

[2] We are most grateful to Oxford University press for permission to use the Hector database, and to the UK EPSRC for the grant which supported the manual re-tagging of the data.

[3] The approach to sampling of which this was a degenerate version is described in detail in (Kilgarriff1998a).

(noun), *sack* (verb), *onion, rabbit*, also *handbag* (taken from a different dataset).

Sample of corpus instances

The HECTOR corpus is a 20-million word corpus comprising mainly journalism, books and magazines. It was a pilot for the British National Corpus, and some of the data is shared with the BNC. Around two hundred corpus instances per word were randomly selected from all the HECTOR data available. The exact number of corpus lines per word varied according to the BNC frequency of the word, its level of polysemy, and the number of its corpus lines which turned out to be personal names, of the wrong word class, or otherwise anomalous. There were usually two sentences of context available, the sentence containing the word and the preceding one, but occasionally more and occasionally less, depending on the structures available in HECTOR.

Dictionary

The HECTOR dictionary was produced in tandem with the sense-tagging of the HECTOR corpus, so the HECTOR dictionary entries are probably more closely tied to the corpus evidence than any published dictionary. Only a sample of several hundred entries were prepared, and they were never polished and double-checked for publication. The entries include more examples than standard dictionaries, and provide more explicit information on lexico-grammatical patterning.

Sense-tagging

The basic task was to assign each corpus instance for a word to one (or more) of the meanings in the HECTOR dictionary entry for that word. The task had been done once already prior to SENSEVAL and was done twice more for SENSEVAL. The options available to the taggers were:

- simple assignment of one sense
- more than one sense, eg "1 or 2"
- sense plus suffix: suffixes were:
    - P for proper-name use, eg "Peter Rabbit"
    - A or N for adjectival or nominal use of a sense that wasn't standardly adjectival/nominal
    - M for metaphorical or metonymic use
    - X for other exploitations of the sense
    - ? for awkward and unclear cases

- T, P, U for Typographical errors, Proper names (where the use is not also a regular use of the word – cf. the P suffix) or Unassignable

Some words were easy and quick, others hard and slow. The average time taken was one minute per citation.

Identify mismatches

For this experiment, it was necessary to identify all those cases which were not covered by literal readings of dictionary entries. We took all those instances where there was anything less than complete agreement by all three taggers on a single, simple sense (eg, without suffixes) and re-examined them. That is, all those cases where there was any disagreement, or where there were suffixes, or where there were disjunctive answers, were re-analysed. This cast the net wide, and in some cases over half the data was re-examined. Each of these cases was then classified as standard or non-standard by the author.

GL?

For the non-standard cases, the author then also assessed whether a GL-style analysis might plausibly apply.

5 Examples

Different words behaved in different ways, and in this section we make some comments on each of the words in turn. A number of corpus citations are provided, as that serves to demonstrate the nature of the exercise and the sensitivities required for the analysis of non-standard word use.

The numbers in brackets following each word give, first, the number of corpus instances that were re-examined specifically for this exercise, and second, the complete sample size for the word.

modest (164/270)

The HECTOR lexicographers had split the meaning of *modest* between nine senses, in contrast to 3 (CIDE1995), 4 (LDOCE1995) or 5 (COBUILD1995) in other dictionaries. There was a high degree of overlap, and the sense distinctions could not be drawn sharply. (This supports findings in other exercises that this is characteristic of adjectives: they can be assigned to a wide range of nouns, sometimes more literally, sometimes less so, but it is the meaning of the modified noun which determines the sense of the adjective. Where the nouns do not fall into neat categories, nor will the adjective senses.)

Faced with this indeterminacy, the taggers often gave disjunctive or different answers. But in none of the 164 cases of non-agreement was it appropriate to classify the corpus instance as a non-standard use of the word.

disability (29/160)

HECTOR distinguished two senses, one "medical", for physical or mental disabilities, the other for anything non-medical. However the non-medical, residual sense was marked "chiefly legal". This seems a lexicographic error, as most of the non-medical instances in the corpus were not legal either: the lexicographer should identify the technical, legal sense as distinct, and provide a third, residual sense for instances such as:

> Satie may well have lacked accomplishment, but like all major artists he managed to turn his **disabilities** to account.

The cases re-examined were either of this type, or names such as "Disability Alliance", which were clearly both the medical sense, and (parts of) names.

steering (16/177)

There were two senses in HECTOR:

- the activity eg. *his steering was careless* vs.
- the mechanism eg. *they overhauled the steering*

These are metonymically related. Most of the sixteen re-examined corpus instances were simple cases of underspecification, e.g.

> it has the Peugeot's steering feel

One more complex case was:

> After nearly fifty years [as a bus driver] Mr. Hannis stepped down from behind the steering wheel

This is of interest because it makes passing reference to the idiomatic reading of *behind the steering wheel* in which it means "to be the driving force behind (an organisation)". Had Mr. Hannis's occupation been not bus-driving but managerial, the instance would have been clearly idiomatic. As it is, the sentence carries traces of both the literal and idiomatic readings.

seize (53/259)

HECTOR gives 10 'senses' for *seize* (excluding phrasal verbs). On closer inspection, it would seem that these 'senses' are better interpreted as features, as they are not exclusive and frequently co-occur. The HECTOR labels for the first five senses are GRAB, HOSTAGE, CONFISCATE, POSSESS/INVADE, OPPORTUNITY. These are all aspects of the meaning of *seize* which might or might not be evident in a particular instance. Most of the re-examined cases were ones where more than one feature was salient, and the taggers had given two senses and/or given different ones. In

> [He] slipped out of the hands of the two [gangsters] who had seized him

both HOSTAGE and GRAB are salient. In

> Bruges Group Tory MPs claimed victory last night after seizing all the top places on the backbench European affairs committee

both OPPORTUNITY and POSSESS/INVADE are present. (OPPORTUNITY is implicitly present in a high proportion of instances: replacing *seize* by *seize the opportunity (to take)* would not, in most cases, change the meaning.)

Lexical semantics may seem a politically neutral territory but this is not always so. Consider

> ... examine charges that Israeli soldiers were intimidating local residents. Al-Haq, a human rights organisation on the West Bank, charged soldiers with non-registration of property seized, assault and tearing up identity cards.

If your sympathies are with the Israelis, this is CONFISC. If they are with the Palestinians, it is POSSESS/INVADE.

Research of this kind cannot readily be done by anyone who is not a native speaker, and it is also as well if the researcher is from the same culture as the intended readership. Consider

> Tolley drove uppishly at a half-volley and was caught at short mid-wicket; Lord, cutting without due care, was **seized** at gully off Tim Wren

Enquiries of people who are not British[4] are met with blank stares. The context is of course cricket, and what happened was that the ball was caught. (The object of *seize* is *Lord*, who hit the ball that was caught. The relation between the ball and Lord is metonymic. The cricket use of *catch X* where X is a player is a distinct sense in LDOCE3 and CIDE, the complication here being that the verb is not *catch* but *seize*.)

[4] Nor from the West Indies, the Indian sub-continent, Australia, New Zealand, or South Africa, one might suppose.

sack/v (5/178)

Four of the five re-examined cases were errors. The fifth,

> And Labour MP, Mr Bruce George, has called for the firm to be **sacked** from duty at Prince Andrew's £5 million home at Sunningwell Park near Windsor.

is non-standard because the CEASE-EMPLOYING meaning of *sack* is specified in HECTOR as taking a person as its direct object. Here, the object is the company. This is an instance that the GL could in principle account for.

sack/n (7/82)

The instances re-examined were typing errors, two instances of *sack race*,[5] one instance with insufficient context to determine the sense, and one non-standard use based on a metaphor:

> Santa Claus Ridley pulled another doubtful gift from his sack.

(Ridley is a British politician.)

handbag (30/715)

The *handbag* data has a different origin: the British National Corpus. It was analysed as part of a different study, reported in (Kilgarriff1998b) with goals similar to the current exercise. Thirty non-standard instances were found, comprising metaphors, handbag-as-Thatcher's-symbol, handbags-as-weapons, the idiom *dance round your handbag* and exploitations of the idiom in the sublanguage of nightclubs, where *handbag* denotes a music genre. There was just one instance which potentially supported a GL analysis:

> She moved from handbags through gifts to the flower shop

(*Handbags* denotes the handbag department of a department store.)

onion (34/214)

The lexical entry distinguishes PLANT onions from VEGETABLE onions, and ten of the re-examined cases bridged that distinction, eg:

> Plant the sets two inches apart in rows 10 inches apart to produce a good yield of medium-sized onions.

[5] A kind of race in which the contestants stand in a sack, which they hold around their waist, and hop; usually encountered at school sports days and village fêtes.

There was a simile and a metaphor, in which a speeding tennis ball is likened to an onion. Other anomalies included instances in which onion and derivatives were being used as medicine, as a decorative feature, and for dying. In each case, neither the PLANT nor the VEGETABLE sense was more applicable than the other.

In

> It's not all frogs legs and strings of onions in the South of France

we have a cliché of Frenchness rather than a vegetable, and in

> In Leicestershire, machine drivers have their own names for river plants, such as 'water onions' for the true bulrush

the occurrence belongs to a sublanguage and is signalled as such.

For purposes of counting numbers, just the tennis ball metaphor and the 'water onions' were counted as non-standard, though clearly other decisions could have been made.

rabbit (52/224)

This was the most fecund of the words. First, the word enters into a large number of names, and these accounted for half the instances re-examined. There were:

- Rabbit (Winnie the Pooh's friend)
- Peter Rabbit
- Crusader Rabbit
- Brer Rabbit
- (Who framed) Roger Rabbit
- The White Rabbit
- Care For Your Rabbit (book title)
- Super Rabbit:

  > Now Oxfordshire grain growers are facing a new enemy, the Super Rabbit. Super Rabbit is different from anything ever seen before in the county because he seems pretty well indestructible.

- Sumo Rabbit and His Inescapable Trap of Doom (song title)

Rabbit also brought the issue of representations to the fore. HECTOR included a "toy" sense of rabbit, which might seem an innocent choice. However the data included

> Some people learn by watching videos of the great players, Borg, McEnroe, Navratilova and Evert. I thought it would be fun to make Monica an animated film of a rabbit playing tennis set to music, and this was a success.

> It contains three drawings of Cecily Parsley, the rabbit innkeeper, a hand-painted Christmas card and two amateurish Lake District views.

> Playboy was described as a pleasure-primer, its symbol was a rabbit and its bait was the Playmate of the month, the girl who was unfolded in the centre wearing a staple through her navel but not much else.

> Marie Holmes as the nervy Piglet, John Nolan as the garrulous Owl, and Judy Eden as the troublesome Rabbit all perform competently, and Anne Belton, as Kanga, fusses in matronly fashion over young Roo (Jonathan Eden) and the other animals.

To try to unpick just this last example: there was a toy rabbit, belonging to Christopher Robin, called Rabbit. Christopher Robin's father, A. A. Milne, wrote stories about Rabbit in which he imputed to it some TOY- and some ANIMAL-properties. The books of the stories were published and became popular and now have been turned into a play so a person (Judy Eden) now 'pretends' to be this TOY-ANIMAL individual.

Rabbit also supports a number of conventionalised metaphors and collocations with both literal and metaphorical meanings: *frightened rabbits, froze like rabbits, running like rabbits*, rabbit *holes, hutches* and *warrens* all occurred in the data. Only *rabbit warren* was explicitly mentioned in HECTOR.

There are several instances that allude to magicians pulling rabbits out of their hats:

> The violins waved and swayed like cornstalks in the wind. The drummer, white haired, might have been a conjuror drawing **rabbits** from his instrument's interior.

This is a distinct sense in the HECTOR entry, so the instances are allocated to it and correspondingly classified as standard.

6 Results

Of 2276 corpus instances examined, there were 390 where the lexicographers had not all agreed on the same unique tag in the first pass. Of these, on closer examination 41 instances were found to be non-standard word uses. Thus just under 2% of the corpus instances were non-standard.

Of these, just two, or 5%, were plausible candidates for GL treatment.

The quantitative results are presented in Table 1.

7 Discussion

The exercise puts the spotlight on the dictionary as much as on the words. Many readers will have granted the argument of Section 3 that a published

Word	Sample	Re-ex	NS	GL
modest	270	164	0	0
disability	160	29	0	0
steering	177	16	0	0
seize	259	53	0	0
sack/n	178	5	1	1
sack/v	82	7	1	0
onion	214	34	2	0
rabbit	224	52	7	0
handbag	712	30	30	1
TOTALS	2276	390	41	2

Table 1: Experimental results, showing, for each word, the size of the dataset (Sample), the number of instances re-examined (Re-ex), the number of those which were classified as non-standard uses (NS) and the number of those which were plausibly accounted for by GL analyses (GL).

dictionary had to be used for this exercise, but may now feel this argument must have been flawed and that there must be a more tolerable strategy than working to the vagaries of one particular dictionary. The author can only agree that it would be nice if there were one.[6]

41 of the 2276 instances in the dataset were identified as non-standard, and just two of these — the "handbag department" use of *handbags* and the use of verbal *sack* with a company rather than an individual as object — were identified as candidates for GL-style analysis. As is evident from the examples, another analyst would probably not have arrived at identical figures, but they would, in all likelihood, have pointed to the same conclusion: GL analyses will only ever account for a small proportion of non-standard word uses.

[6] Some GL literature (eg (Copestake and Briscoe1996)) points to co-predication and related ambiguity tests as a way of identifying the distinct senses. The proposal is explored at some length in (Kilgarriff1998b). It suffers from numerous drawbacks. First, there is simply no inventory of senses available, which has been developed according to these criteria. Second, different speakers very often disagree on the acceptability of the test sentences. Thirdly, the relation between evidence from co-predication tests and the pre-theoretical notion of a word sense is far from clear. (Cruse1986) argues that the tests are criterial for the notion of a distinct sense, but his methods are not based on corpus or lexicographic evidence or systematic sampling of the lexicon and bear no relation to lexicographic practice. (Geeraerts1993) presents a critique of the logic of the tests. Experiments to explore the relation between linguists' ambiguity tests and lexicographers' polysemy judgements are currently underway.

The evidence points to the similarity between the lexicographer's task, when s/he classifies the word's meaning into distinct senses, and the analyst's when s/he classifies instances as standard or non-standard. The lexicographer asks him/herself, "is this pattern of usage sufficiently distinct from other uses, and well-enough embedded in the common knowledge of speakers to count as a distinct sense?" The analyst asks him/herself, "is this instance sufficiently distinct from the listed senses to count as non-standard?" Both face the same confounding factors: metaphors, at word-, phrase-, sentence- or even discourse-level; uses of words in names and in sublanguage expressions; underspecification and overlap between meanings; word combinations which mean roughly what one would expect if the meaning of the whole were simply the sum of the meanings of the parts, but which carry some additional connotation.

7.1 Lexicon or pragmatics?

For many of the non-standard instances, an appropriate model must contain both particular knowledge about some non-standard interpretation, and reasoning to make the non-standard interpretation fit the current context. The 'particular knowledge' can be lexical, non-lexical, or indeterminate. Consider

> Alpine France is dominated by new brutalist architecture: stacked rabbit hutches reaching into the sky ...

In this case the particular knowledge, shared by most native speakers, is that

- 'rabbit hutch' is a collocation
- rabbit hutches are small boxes
- to call a human residence a rabbit hutch is to imply that it is uncomfortably small

The first time one hears a building, office, flat or room referred to as a rabbit hutch, some general-purpose interpretation process (which may well be conscious) is needed.[7] But thereafter, the BUILDING reading is familiar. Future encounters will make reference to earlier ones. This can be seen as the **general** knowledge that buildings and rooms, when small and cramped, are like rabbits' residences, or as the **lexical** knowledge that

[7] As ever, there are further complexities. *Hutch* and *warren* are both rabbit-residence words which are also used pejoratively to imply that buildings etc. are cramped. A speaker who is familiar with this use of *warren* but not of *hutch* may well, in their first encounter with this use of *hutch*, interpret by analogy with *warren* rather than interpreting from scratch (whatever that may mean).

hutch or *rabbit hutch* can describe buildings and rooms, with a connotation of 'cramped'.

It is the compound *rabbit hutch* rather than *hutch* alone that triggers the non-standard reading. Setting the figurative use aside, *rabbit hutch* is a regular, compositional compound and there is little reason for specifying it in a dictionary. Hutches are, typically, for housing rabbits so, here again, the knowledge about the likely co-occurrence of the words can be seen as general or lexical. (The intonation contour implies it is stored in the mental lexicon.)

That hutches are small boxes is also indeterminate between lexical and general knowledge. It can be seen as the definition of *hutch*, hence lexical, or as based on familiarity with pet rabbit residences, hence general.

To bring all this knowledge to bear in the current context requires an act of visual imagination: to see an alpine resort as a stack of rabbit hutches.

A different sort of non-standard use is:

> Santa Claus Ridley pulled another doubtful gift from his sack.

Here, the required knowledge is that Santa Claus has gifts in a sack which he gives out and this is a cause for rejoicing. There is less that is obviously lexical in this case, though gifts and sacks play a role in defining the social construct, 'Santa', and it is the co-occurrence of *Santa Claus*, *gifts* and *sack* which triggers the figurative interpretation.

As with *rabbit hutch*, the figure is not fresh. We have previously encountered ironic attributions of "Santa Claus" or "Father Christmas" to people who are giving things away. Interpretation is eased by this familiarity.

In the current context, Ridley is mapped to Santa Claus, and his sack to the package of policies or similar.

These examples have been used to illustrate three themes that apply to almost all the non-standard uses encountered:

1. Non-standard uses generally build on similar uses, as previously encountered
2. It is usually a familiar combination of words which triggers the non-standard interpretation
3. The knowledge of the previously-encountered uses of the words is very often indeterminate between "lexical" and "general".

Any theory which relies on a distinction between general and lexical knowledge will founder.

7.2 Lexicon size

The lexicon is rife with generalisation. From generalisations about transitive verbs, to the generalisation that *hutch* and *warren* are both rabbit residences, they permeate it, and the facts about a word that cannot usefully be viewed as an instance of a generalisation are vastly outnumbered by those that can. GL aims to capture generalisations about words.

Given an appropriate inheritance framework, once a generalisation has been captured, it need only be stated once, and inherited: it does not need to be stated at every word where it applies. So a strategy for capturing generalisations, coupled with inheritance, will tend to make the lexicon smaller: it will take less bytes to express the same set of facts. GL is associated with a compact lexicon, in this sense.

But a compact, or smaller, lexicon should not be confused with a small lexicon. The examples above just begin to indicate how much knowledge of previously encountered language a speaker has at his or her disposal. Almost all the non-standard instances in the dataset call on some knowledge which we may not think of as part of the meaning of the word and which the HECTOR lexicographer did not put in the HECTOR dictionary, yet which is directly linked to previous occasions on which we have heard the word used. The sample was around 200 citations each per word: had far more been data examined, far more items of knowledge would have been found to be required for the full interpretation of the speaker's meaning.[8] The sample took in just nine words. There are tens or even hundreds of thousands of words in an adult vocabulary. The quantity of information is immense. A compact lexicon will be smaller than it would otherwise be — but still immense.

7.3 Quotations

Speakers recognise large numbers of poems, speeches, songs, jokes and other quotations. Often, the knowledge required for interpreting a non-standard instance relates to a quotation. One of the words studied in SENSEVAL was *bury*. The *bury* data included three variants of Shakespeare's "I come to bury Caesar not to praise him", as in:

> [Steffi] Graf will not be there to praise the American but to bury her . . .[9]

We know and recognise vast numbers of quotations. (I suspect most of us could recognise, if not reproduce, snatches from most top ten pop songs from our teenage years.) Without them, many non-standard word uses

[8] The issue of what should count as an interpretation, or, worse, a **full** interpretation leads into heady waters, see eg (Eco1992). We hope that a pre-theoretical intuition of what it is for a reader or hearer to grasp what the author or speaker meant will be adequate for current purposes.

[9] For further details on the *Caesar* cases, and a discussion of other related issues in the SENSEVAL data, see (Krishnamurthy and Nicholls1999, forthcoming).

are not fully interpretable. This may or may not be considered lexical knowledge. Much will, and much will not be widely shared in a speaker community: the more narrowly the speaker community is defined, the more will be shared. Many dictionaries include quotations, both for their role in the word's history and for their potential to shed light on otherwise incomprehensible uses (CIDE1995).

An intriguing analogy is with the memory-based learning (MBL) approach to machine learning. In MBL all instances are retained and a new instance is classified according to the familiar instances which it most resembles. The approach has recently been shown to be well-suited to a range of natural language leaning tasks (Daelemans, van der Bosch, and Zavrelto appear). In MBL, where numbers of instances are similar, they will contribute to future classifications jointly, so do not appear to have roles as individual recollections in memory. Exceptional instances, by contrast, play an explicit role in classification when a new instance matches. Correspondingly, for standard word uses, we do not think in terms of individual remembered occurrences at all. For instances with a touch of idiosyncrasy, like Mr. Hannis's "fifty years [...] behind the steering wheel", or strings of onions as a cliché of Frenchness, we probably do not but might. And for "not to praise but to bury" cases we do.

A proposal in the literature which informs this discussion is (Hanks1994). Hanks talks about word meaning in terms of 'norms and exploitations'. A word has its normal uses, and much of the time speakers simply proceed according to the norms. The norm for the word is its semantic capital, or meaning potential. But it is always open to language users to exploit the potential, carrying just a strand across to some new setting. The evidence encountered in the current experiment would suggest an addendum to Hanks's account: it is very often the exploitations which have become familiar in a speech community which serve as launching points for further exploitations.

In the 1995 book, Pustejovsky reviews recent work by Nunberg, and Asher and Lascarides, and draws the moral that:

> polysemy is not a monolithic phenomenon. Rather, it is the result of both compositional operations in the semantics, such as coercion and co-composition, and of contextual effects, such as the structure of rhetorical relations in discourse and pragmatic constraints on co-reference. (p 236)

Our evidence endorses this weaker view of the role of generative devices, and adds that a prominent role in the analysis should be taken by extensive knowledge of how words have deviated from their norms before.

8 Conclusion

We have described an experiment in which the merits of GL as a general theory of the lexicon, which accounts for non-standard uses of words, were scrutinised. The experiment looked at the non-standard uses of words found in a sample of corpus data, and asked whether they could be analysed using GL strategies. The finding was that most of the time, they could not.

This by no means undermines GL analyses for the kinds of cases discussed in the GL literature. Rather, it points to the heterogeneity of the lexicon and of the processes underlying interpretation: GL is a theory for some lexical phenomena, not all.

A model of the interpretation of non-standard word uses was sketched in which speakers and hearers have access to large quantities of knowledge of how the word (and its near-synonyms) has been used in the past. The knowledge is frequently indeterminate between 'lexical' and 'general', and is usually triggered by collocations rather than a single word in isolation.

There are numerous disputes in linguistics which circle around the question of storage or computation: is the structure recalled from memory, or computed afresh each time it is encountered.[10] The GL is a theory of the lexicon which gives the starring role to computation. The evidence from this experiment is that, while complex computations are undoubtedly required, so too is a very substantial repository of specific knowledge about each word, the kinds of settings it normally occurs in, and the various ways in which those norms have been exploited in the past.

[10] A preliminary version of this chapter was presented at a conference entitle "Storage and Computation in Linguistics", (in Utrecht, the Netherlands, October 1998).

References

Atkins1993 Atkins, Sue. 1993. Tools for computer-aided lexicography: the Hector project. In *Papers in Computational Lexicography: COMPLEX '93*, Budapest.

Buitelaar1997 Buitelaar, Paul. 1997. A lexicon for underspecified semantic tagging. In Marc Light, editor, *Tagging Text with Lexical Semantics: Why, What and How?*, pages 25–33, Washington, April. SIGLEX (Lexicon Special Interest Group) of the ACL.

Buitelaar1998 Buitelaar, Paul. 1998. CoreLex*: Systematic Polysemy and Underspecification.* Ph.D. thesis, Brandeis University.

CIDE1995 CIDE, 1995. *Cambridge International Dictionary of English.* CUP, Cambridge, England.

COBUILD1995 COBUILD, 1995. *The Collins COBUILD English Language Dictionary. 2nd Edition.* Edited by John McH. Sinclair *et al.* London.

Copestake and Briscoe1996 Copestake, Ann A. and Edward J. Briscoe. 1996. Semi-productive polysemy and sense extension. In James Pustejovsky and Branimir Boguraev, editors, *Lexical Semantics: The Problem of Polysemy.* Oxford Univerity Press, Oxford, pages 15–68.

Cruse1986 Cruse, D. A. 1986. *Lexical Semantics.* CUP, Cambridge, England.

Daelemans, van der Bosch, and Zavrelto appear Daelemans, Walter, Anton van der Bosch, and Jakub Zavrel. to appear. Forgetting exceptions is harmful in language learning. *Machine Learning, Special Issue on Natural Language Learning.*

Eco1992 Eco, Umberto. 1992. *Interpretation and Overinterpretation.* Cambridge University Press, Cambridge.

Geeraerts1993 Geeraerts, Dirk. 1993. Vagueness's puzzles, polysemy's vagueness. *Cognitive Linguistics*, 4(3):223–272.

Hanks1994 Hanks, Patrick. 1994. Linguistic norms and pragmatic exploitations or, why lexicographers need prototype theory, and vice versa. In Ferenc Kiefer, Gabor Kiss, and Julia Pajzs, editors, *Papers in Computational Lexicography: COMPLEX '94*, pages 89–113, Budapest.

Kilgarriff1993 Kilgarriff, Adam. 1993. Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26(1–2):365–387.

Kilgarriff1998a Kilgarriff, Adam. 1998a. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language*, forthcoming. Special Issue on Evaluation of Speech and Language Technology, edited by R. Gaizauskas.

Kilgarriff1998b Kilgarriff, Adam. 1998b. 'I don't believe in word senses'. *Computers and the Humanities*, 31(2):91–113.

Kilgarriff1999 Kilgarriff, Adam. 1999. "95% replicability for manual word sense tagging. In *Proc. EACL, submitted*.

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

Krishnamurthy and Nicholls1999, forthcoming Krishnamurthy, Ramesh and Diane Nicholls. 1999, forthcoming. Peeling an onion: the lexicographers' experience of manual sense-tagging. *Computers and the Humanities.* Special Issue on SENSEVAL, edited by Adam Kilgarriff and Martha Palmer.

Lakoff and Johnson1980 Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By.* University of Chicago Press.

LDOCE1995 LDOCE, 1995. *Longman Dictionary of Contemporary English, 3rd Edition.* Edited by Della Summers. Harlow.

Pustejovsky1995 Pustejovsky, James. 1995. *The Generative Lexicon.* MIT Press, Cambridge, Mass.

Sweetser1990 Sweetser, Eve. 1990. *From etymology to pragmatics : metaphorical and cultural aspects of semantic structure.* CUP, Cambridge, England.