Contextual Clues to Word-Meaning

ANTOINETTE RENOUF
University of Liverpool

LAURIE BAUER
University of Liverpool and Victoria University of Wellington

1. Introduction

How do native speakers interpret new words when they read or hear them? How much help and what kind of help are they given by the immediate context in which the new word occurs? These are questions which we set out to investigate on the basis of a large corpus of newspaper English. We were working within the framework of the APRIL project, which is a three year, EPSRC-funded research project aimed at classifying the rare words (initially hapax legomena) in text. We had expected to find solid correlations between the type of formation and the type of help provided in the text, but no such results were obtained. However, we were able to classify the types of help provided within the immediate context, and it is this taxonomy which provides the core of the material for this paper.

In a recent paper, Plag (1996: 774) makes the following statement:

> "Linguists who work with native-speaker informants often experience that words or sentences are first rejected by informants because speakers fail to make sense of them, and not because the data violate morphological or syntactic rules of their language. Presented with an appropriate context that provides a possible interpretation, the same informants may readily accept the data presented to them."

This is presented merely as an anecdotal remark, not as a quantified research finding; yet we feel sure that it will strike chords with many morphologists and syntacticians. Studies such as that by Gleitman and Gleitman (1970) illustrate the difficulties that speakers can have in interpreting novel words (in that case compounds) when no context is provided to support an interpretation. It is also clear that difficulties of interpretation can persist even when a context is given: it is not only in academic discourse that phrases such as 'I'm not sure what you mean by X' occur, where X is some word in the immediately preceding discourse. Nevertheless, such problems appear to be much reduced from the level at which they occurred in Gleitman and Gleitman's study. In journalistic writing, it is not possible to know precisely how often newspaper readers have such problems with the novel words they encounter; our own experience in dealing with a particular subset of novel words is that some problems do persist, but they are infrequent. And it must not be forgotten that journalists are typically trained and experienced communicators who will find it against their own interests to fail to transfer meaning to their readership. Pragmatically, it must be assumed that such communicators will provide such assistance with interpretation as they consider necessary to ensure the smooth flow of information, though assistance in the comprehension of coinages intended to amuse may, precisely for that reason, not be forthcoming.

The implication of this is that in at least some cases there will be overt information in the context which will allow readers (in the case of our corpus, but presumably also listeners in the case of spoken language) to interpret novel words. Circumstantial support is provided for this assumption by Baayen and Neijt (1997), who, in their study of the Dutch suffix *-heid* (the equivalent of English *-ness)* in journalistic text, find that, on the whole,

> "hapax legomena are characterized by a higher degree of contextual anchoring than high-frequency words."

'Contextual anchoring' is Baayen and Neijt's term for the phenomenon we are discussing here. Their study is concerned with the different kinds of anchoring associated with the conceptual and referential functions of more and less common words, though not with the issue of novelty as such.

There are two major issues involved in an analysis of the contextualisation of new words. One is how to establish what a new word is from the analyst's point of view. Our definition is a surface textual one: we identify novelty by deeming a word to be a novel usage when it occurs for the first time (at which point it is also a hapax legomenon) in our chronologicallyprocessed data. This approach has the merit of identifying lexical candidates which have an objective claim to novelty.

We can assume that there will not be a total match between our automatically registered new words and those treated as new by the writers of our texts, and this leads us to the second issue: how to establish what a new word is from the writer's and reader's, point of view. Ideally we need to know, in each case, whether the writers (speakers) concerned recognise that the word they are using is novel and to whom it is new. A word might be (a) novel from the reader's (listener's) point of view but not from the writer's (speaker's) or (b) novel for both the reader and the writer. In principle, it is also possible for the writer to coin a neologism which the reader has nevertheless heard before. The writer has to make these judgements about novelty at the time of writing in order to know where interpretative aid is required. This involves guesswork on his/her part, and it is clear that there will be instances where aid is provided unnecessarily (or unnecessarily for the individual reader) and instances where no aid is provided though it would have been useful (for the individual or readers in general). Moreover, in some cases, the writer may not consciously recognise but only feel that a word will be unknown and, thus, subconsciously make contextual adjustments. A further complication is that the daily newspaper journalist, whether of tabloid or broadsheet, assuming regular readership, may assume more knowledge and feel freer to construct words and create puns or to introduce the creations of a third party without providing overt explanation (except perhaps when using new technical terms) to a greater extent than is the case with other kinds of writers.

All this means that we cannot know for sure which individual cases qualify as novel from the perspective of the writer. A question we would like to address is whether and how the contextual anchoring provided for novel words differs according to whether the writer perceives them to be new or not. But in view of the imponderables above, it begins to become clear that there is unlikely to be a particularly straightforward correlation between innovation (however defined) and overt help with interpretation.

This is our starting point in this paper of a discussion of the various kinds of information that are found in texts which may be viewed as providing support in the interpretation of novel vocabulary. We shall first comment on 'overt' (and probably more or less conscious) information, bearing in mind

that this may be of many types, some more overt or more helpful than others; and then catalogue and discuss the 'covert' (and probably unconscious) information that we have found. In the process, we move from obviously linguistic clues to meaning to clues of the metalinguistic kind, such as realworld knowledge. The latter distinctions mirror to some extent those made by Baayen and Neijt (1997) by the terms 'morphological' and 'thematic.'

2. The data

The source of data for this study is, as noted above, journalistic in nature and consists of the ten years' of text published in the UK daily broadsheet newspaper, the *Independent*, from 1988-1999. The total data over this period amounts to over 360 million word tokens.

The criterion for 'newness' for a given word is, as stated previously, that it has not previously occurred in the data, which is processed chronologically. This is a pragmatic approach, adopted in order to try to automate and accelerate the process of identification. The words occurring for the first time in each quarter over the total period can be extracted by software tools which were developed as sub-routines for the ACRONYM project (Renouf 1996; Collier and Pacey 1997). This project ran from 1994-1997 and produced as its primary output an automated system which identifies semantically related pairs of words using surface features of text. The later stages of the project required that the system become dynamic, or diachronic, in its analysis, identifying changes in the thesaurus across time. At that point, building on the experience of similar activity in an earlier project known as AVIATOR (Renouf 1993; Collier 1993), new tools were developed to process the database chronologically and record first occurrences of words in successive time-chunks of text. This was done by matching each quarter's word index against an accumulating master wordlist, identifying the new items, having recorded these, and adding them to the master wordlist to prevent their being presented again. A further operation recorded typographical and other surface features in the context of each new word in order to assign it to one of several sub-classes of new words including 'ordinary' (i.e., lower-case) words and 'proper' (i.e., a subset of upper-case-initial) names.

Our automated analysis yields a miscellany of candidate neologisms. Some of these are new in the language: new referents to real-world phenomena or concepts, everyday or technical; words created in the making of linguistic jokes; words constructed for the purposes of lexical cohesion. But other first-occurrence words are only new within the context of the ten-year database: rare words; archaisms *(jack-o-lantern),* and revivals *(poll tax; ecu).* Equally, a word when it is recorded as novel may simply be an orthographic or typographical variation: a solid compound previously hyphenated or vice versa (earth-quake), rare plurals, possessives of known words, or errors.

In 1996, a sample of around 11 000 new words-all the first occurrences in the nine million words of running text of the last three months of 1995 *Independent* data-was extracted for manual classification by Renouf and Baayen (1998). They analysed the 8000 lower-case neologisms, among other things assigning them to categories of word formation. In our current study, we focus on their approximately 2940 new compounds and 1820 new derivations: the two commonest classes found, amounting to approximately 4760 new words in total. Though there is also much of interest to be said about the contextualisation of proper nouns, we do not deal with them here.

A study of contextual anchoring requires a definition of context. We settled on a context of thirty words to either side of the keyword, longer than an average sentence, while we accept that help may be found beyond these confines.

3. Methodology

Having selected our output from the Renouf-Baayen analysis, we classified it. We began with an initial set of expectations, based on our linguistic experience, and augmented and modified this progressively in response to our data. The first level of distinction was between 'conscious' and 'unconscious' help. Conscious help is that provided where the writer feels the word is new, and unconscious help is that which is inadvertently provided by the writer or which is otherwise present in the environment. To analyse the conscious help which is available, we needed to identify words which were treated as, and thus presumably considered, new by the given writers.

At a second level of classification, we identified the various types of anchoring which are provided or available in the contexts of new words. In this paper, we examine the main categories of help. Of these, the overt or conscious clues (all typographical or linguistic) are:

- quotation marks
  (other non-linguistic markers of novelty-dashes, commas, brackets are not dealt with in this paper, though they merit study)
- glosses
- introductory and following phrases.

The covert or unconscious help (both linguistic and metalinguistic) falls into the following categories:

- root or base repetition
- exact repetition
- collocation
- semantically-related words
- lexical signals
- parallelism
- lexical field
- pragmatics

Though the various types of contextual clue can occur in isolation, they more often occur in combination, particularly those which are overt or conscious. Multiple anchoring is indicated at points in Section 4 and summarised in Section 5.

It should be noted that we are restricting ourselves to those clues which are directly retrievable from an unparsed text; in particular, we are not considering syntactic clues, such as the use of definite reference and anaphora, which may allow the human reader to deduce that two syntactically different phrases are, nevertheless, supposed to have the same referent, and, thus, to be quasi-synonymous in context.

4. Results

4.1 Overt or conscious help

4.1.1 Quotation marks

Quotation marks are used to mark new words, thereby alerting the reader to a probable need to look to the context to interpret the meaning of a signalled word. In our data, they mark both new individual words (58 cases) and new words within phrases (15 cases), usually noun phrases. The originator of the novel word varies. In most cases, the writer uses quotation marks to report, directly or indirectly, a new word from a third-party source. The third party may have coined the term, as in:

> Greil Marcus, in his great essay on Newman in the book 'Mystery Train', used the phrase 'slot-mouthed', which is so right for the slightly combative tension in Newman's lips as he sings...

or may have employed it as an existing technical term, as in:

> he admitted it might have happened when [the cat] was put in what is alarmingly known as a 'crush-cage', after she resisted his attempts to anaesthetise her in a more conventional manner.

We cannot be sure whether the signalled new word is actually new to the writer himself/herself except where this is indicated in the context; see below, where the italicised fragments signal the writer's unfamiliarity with the term:

> At the risk of being vulgar, I would like to tell you about an American gen
> tleman who sells *what he calls* 'salon wear' . *I think* for people who work in beauty parlours. He goes by the name of Ivan Bonk and he offers a Velcro
> Body Wrap. . .

Quotation marks rarely (6 cases) surround new words which are otherwise left to be interpreted solely on the basis of the elements which make them up. Most (36 cases) are accompanied in the near environment by actual glosses or full background explanations of the new word, and a further 23 cases are accompanied by some other sort of contextual help.

It is noticeable that there are many new words which are not themselves marked as new but are situated within a quoted stretch of context which may be partly or wholly novel. The context for the word *cartoonland* is one such case:

> The backing singers are in Victorian nighties, with angels' wings on the back. The ears are a multiple reference, [Annie] Lennox says-'to some kind of fantasy, some kind of cartoonland, some kind of naivety, some kind of corporate entity. And a statement about myself, because in a way I might represent-an iconic thing . . . '

Quotation marks are ambiguous. They can indicate novelty, but they can also flag the metalinguistic status of a word, new or otherwise; or the fact that the writer has a particular reaction to a word; or that he anticipates or

wishes to stimulate a reaction in the reader. These functions often overlap, as in the next example, where the word *timbrophilia* is marked linguistically as unknown, metalinguistically as an obsolete term for stamp-collecting, and pragmatically as naughty.

> Maybe it's the arid taste of that word 'philately' which won out against the more logical, and sexier, 'timbrophilia' in the 19th century.

In all, it seems that one can say that where a word is new, the quotation marks play a part in the interpretation of new words by alerting the reader to this novelty.

4.1.2 Introductory and following phrases

Introductory phrases are constructions such as *what he terms, so-called* or *what researchers call,* preceding the neologism. The most frequent verb in this usage is *call.* We find approximately 20 instances of introductory phrases with neologisms in our data.

The typical pattern with neologisms is for the introductory phrase to come immediately before the neologism, which mayor may not also be typographically marked with quotation marks (q.v.), dash(es), or colons, with the neologism being explained in the context. In 10 such cases in our corpus, three receive no further clarification from the context, one of those being deliberately humorous with a very unusual (at least in print) introductory phrase:

> the book is devoted to a hatchet job on the entire pantheon of feminist theorists, which is made the funnier by such incidents as Freely's own experiences of being-er-matronised through Russian cigarette smoke by Marilyn French in the lounge of Claridges.

The standard pattern is illustrated by:

> What all this adds up to is what researchers call metamood-the *ability to recognise our emotions and so have a better chance of handling them in a creative and productive way.*

Where supplementary contextual clarification is provided, this generally takes the form of a direct gloss or explanation.

Introductory phrases can have other functions than introducing neologisms per se: they are also used, for instance, to introduce noun phrases functioning as labels, as in:

> The party is expected to establish a commission in the new year to examine possible answers to the so-called *'West Lothian'* question.

The two types of usage are subtly different, the difference being that where the introductory phrase is used to introduce whole noun phrases, there is no further clarification to be found in the context (although we have instances where meaning is clarified to different degrees):

> He also revealed that Mc Yeltsin has the suitcase containing the *so-called* nuclear-launch button with him.

Following phrases, such as *if there is such a thing* or *if you like,* are rarer (6 instances), less homogeneous in terms of lexis, and appear to serve a different purpose. The primary purpose of these seems to be to draw attention to the novelty of a word (which again mayor may not be marked typographically). The word's interpretation is usually already established in the context by the repetition of the base (q.v.) or occasionally some fuller explanation:

> Snores, like cold cream and curlers, are not sexy. They are an embarrassment to the snorer and a torment to the snoree, *if there is such a thing.*
>
> his campaign is bogged down in embarrassment after accusations of cannabis use among his fellow wrestlers. 'Spliff-gate,' as *it may yet become known,* was triggered by an article in Shukan Gendai, a weekly magazine.

4.1.3 Gloss

Perhaps the most direct help that can be given for the interpretation of a word is a direct gloss, of which there are around 65 in our data. Typically, a direct gloss will immediately follow the word to be glossed, sometimes set off by parenthetical markers: brackets; commas; dashes. A typical example (carrying two instances) is thus:

> We do not know whether they are at increased risk because the drug causes super-ovulation *(the production of many eggs)* or because it is an oncogenic *(cancer causing)* drug, in its own right.

Glosses occasionally precede the new word, as in the two next examples:

> Java allows *extra little programs* (applets) to be downloaded with a Web page, regardless of the platform of the user
>
> In co-operation with Robert Silvey he studied and wrote on *the private worlds of children,* for which MacKeith coined the word 'paracosm.'

These preceding glosses seem to have a rather different function, namely, the introduction of jargon rather than the explication of text. This may explain why the majority of preceding glosses simply provide a superordinate term for the unknown word rather than a straight synonym:

> Why are rocks different colours? The colour of a rock depends on the *minerals* it contains. For example, biotite mica, hornblende and pyroxenes are dark brown, dark green or black, so a rock containing a high percentage of these *minerals* will appear dark or melanocratic.

Perhaps because a gloss is used to fulfil a variety of functions, ranging from providing a definition of the meaning of a new word, to the specification of the reference of a word assumed to be known or internally interpretable, the borders of what should count as a gloss are far from clear. We find a number of deviations from the direct gloss: deviations which, in the clear cases, we might wish to establish as separate categories of support for the interpretation but which frequently merge into one another.

For instance, we find clarification or specification by example, with the example being more or less explicitly marked. Consider:

> Kenneth Clarke started the wordification of Tony Blair when he accused him of *'Blairing away. '*

> Though it is claimed that {house-spiders} prey on pests, they seem to prefer just hanging out in fly-free places *(such as the bath).*

> Goldeneye is carefully modelled on the new wave of Clancey-inspired, postcold war techno-thrillers *(there's more than a touch of The Hunt for Red October here).*

> A blizzard of mock-denials *('Nope-hope-never-touch-a-thing, not-me, neveragain')* and then he reflects that he only ever played a concert while tripping on one occasion.

> Persians are available in a bewildering range of colours and patterns: white, black, cream, 'red self' (a single reddish colour throughout), blue, blue-cream, smoke, bicolour, tabby, tortoiseshell, tortoiseshell-and-white, colourpoint *(like a Siamese),* pewter, chocolate, and even lilac.

Note that these examples differ in how much they actually gloss as well as in how overtly the exemplification is provided. The examples of mockdenials, for instance, fail to illustrate the mock-quality of the denials; *The Hunt for Red October* is only indirectly presented as an example of a technothriller. There are also cases of explanation rather then exemplification, as m:

> Whether you actually enjoy this production depends largely on how you feel about David Bark-lones's laffeir-an eccentric method-style performance, *all shuffling feet, blinking eyes and swallowed vowels.*

We also find explanations that are definition-like in being syntactically substitutable with the definiendum:

> Denise says her daughter was always strong-natured, *having tantrums if she didn't get her way,* but otherwise was a very happy, affectionate little girl.

and/or structurally parallel with it:

> Or was she motivated by the common desire to self-improve, *to see where one has gone wrong, and perhaps to make amends?*

These explanations may be quite meandering, separated from the new word and requiring a fair amount of interpretation on the part of the reader,

> even if a large amount of information is given.

> Greensted church is . . . still the oldest wooden building in Europe. Ian Tyers, dendrochronologist at Sheffield University *who dated the Greensted oaks, has checked rival constructions in Scandanavia, and found none dates from before the early 13th century.*

> 'Should of' is *the kind of written form called* 'eye-dialect' *that novelists use to indicate a lower class character, like 'wot' and 'me moother' for 'what' and 'my mother'.*

> I am at a loss to understand why you regard my occasional friendships with young women in their mid-twenties as analogous to my *avuncular affection for children of eight or nine.* I was never entirely serious, as you find me, when I described myself as 'desperate' over the news of a former child-friend's marriage.

4.2 Covert or unconscious help

By covert help, we mean material in the context which may aid in the interpretation of a new hapax, even though it is not obviously provided with that aim in mind. We suspect that this kind of textual anchoring is done unconsciously by the writer, although it is a technique which can be learnt. The result is that the reader (listener) has more interpreting to do to discover the relevance of the information presented. We believe it is an important part of the contextual help available and thus offers an exposition of the process as we understand it.

4.2.1 Root repetition or base repetition

The root of a neologism is the ultimate, unanalysable base, and the base is that part to which the most external affix is added. We find no difference in function or effectiveness between repeating the one or the other, although the remaining context plays a deciding role in this. In some cases, of course, root and base are identical, which contributes to the strategic similarity of the two features.

Examples of root repetition include the following where the repetition precedes the new form:

> 'Day ton Ohio,' said Slobodan Milosevic on being informed of where he had to go. 'I'm not *a priest*, you know.' Some unpriestly concessions to Balkan mores were made.
>
> Isaacs had warned him, when the two met in a corridor, 'If you screw it up, if you betray it, I'll come back and *throttle* you.' To date Grade remains unthrottled.
>
> No, it is not the *hunting* that Labour would ban, but the huntspersons.

In the following examples, the repetition of the root follows the new form (accompanied in the latter context by a gloss):

> OJ outsoaped the *soaps*.
>
> Scientists there reproduced the techniques used by commercial meat-recovery plants which strip *meat* and protein from sheep and cattle carcasses and then
> use steam and solvents to kill any infectious agents.

It seems to us that root or base repetition will be more or probably less useful as an aid to interpretation of the full neologism depending on how many of the following contextual conditions it meets:

- it occurs before the new word;
- it occurs in close proximity to the new word;
- it occurs in an unsubordinated or other simple syntactic environment;
- it is presented in a syntactic framework which highlights a sort of parallelism between the new word and the base repetition;
- it is unaffixed, particularly un-prefixed;
- it is altogether morphologically simpler than the new word, so that it helps to 'unpack' the elements of the new word;
- it is composed of known elements.

An instance where repetition is probably less helpful because it does not meet all these conditions is:

> I don't want to watch a supermodel being a bit risque. I heard in the pilot they did a thing about periods. Real ladettes wouldn't f ing bother talking about that. We've got better things to do. That's not *laddishness* any more than putting condoms on cucumbers and giggling.

Laddishness is morphologically more complex than *ladettes* and occurs rather distantly after it. Another example is:

> than to sit around while everyone else is being *polychronic* too. Which is why
> I do not mind that monochronicism (monochronicity?) seems to be becoming something of an obsession.

Polychronic precedes *monochronicism*, but it is complex too and does not help if the internal elements are not known. A similar case is:

> Stop *smouldering*. Your admirers expect a sunny smile to light up your features once matrimonial bliss with Elizabeth is achieved-but the sad truth is that smoulderers, once they have picked up the habit, seem in general reluctant to let it go.

4.2.2 Exact repetition

The exact repetition of a new word does not usually help much in its semantic interpretation. It provides a second chance to see the new word in another context, but it does not prime the reader in the way that repeated roots and bases do (under optimal conditions). Exact repetition does not seem to be employed in the presentation of new words for the purpose of semantic interpretation, but rather with known ones to achieve textual cohesion and meet syntactic requirements. The form may be identical or repeated, plus or minus an apostrophe, as in:

> He concentrated instead on his art; he also tried out a few other inventions, none of which had the ballpoint's success. The *ballpoint* also made a household name of another entrepreneur.

Lemmatised repetition, which occurs rarely in our data, might be argued to help in the understanding of a new word if the repeated form is simpler and executes one layer of interpretation, as in the case of:

> *Psychosurgery* continues to be used in the treatment of severe depression where drugs and ECT have failed and, in the past few years, two additional centres-Cardiff and Dundee-have begun to carry out psychosurgeries. Practitioners emphasise that the techniques used today are light years away from the old

> lobotomies. They believe *psychosurgery* offers a genuine treatment to seriously ill patients.

However, the repetition is perhaps not quite as helpful as it might be, in spite of preceding and following the new hapax, since it is quite distant.

4.2.3 Collocation

4.2.3.1 Strength or fixedness of collocation

The use of collocation in aiding the interpretation of new words is a tricky matter to explain. The paradox is that if the pairing of the collocate with the new word is recognised as an established one, that should indicate that the new word is not new to the reader. So collocation is probably not used by the writer as a primary aid to decoding a new word. In the following instances, the newness may simply be a matter of (non)hyphenation.

> Everyone (. . .) offered complicated **hardluck** *stories* in which the words 'local council', 'postal strikes' and 'it must be a mistake' figured large.
>
> When Leonard Woolf ran the Hogarth Press, he introduced several series of **softbacked**, cheaply priced *books* specifically to reach readers of modest means.

But if we assume that the words *hardluck* or *softbacked* are unknown to the reader, the collocates *stories* and *books* should not in themselves be helpful in decoding.

Conversely, if the collocate is in itself a word known to the reader, it can go some way to aiding the process of interpreting an unknown new word by suggesting one or more semantic or grammatical features pertaining to it. For instance, in the above examples, *stories* and *books* indicate at least that *hardluck* and *softbacked* are their respective attributes. Or supposing for a moment that we did not know the new word *stainedglass*, in:

> the tall ante-room, clad entirely in dark carved wood and lit by **stainedglass** *windows.*

the collocate *window* would allow us to know that it denoted a quality of a window and that its grammatical status was premodifier to a noun; whilst *lit* might indicate that this quality was a material of the kind that allowed light to pass through.

4.2.3.2 Collocates of whole word vs one element

If a known word is the collocate of one element of the new word, whether base, root, or 1st or 2nd element in a compound, it can serve to decode that element, thereby easing the task of interpreting the whole. It is therefore of more use in decoding new words structured as compounds than as deriva
tions. In the following example, a *bone* is literally something to give *a dog;* moreover, *give a dog a bone* is a quotation. *Dog* does not explain *chew*, but it confirms the meaning of the superordinate second element of the new word.

> *Give the dog a* **chewbone** to occupy it outside training times.

Similarly, in the example:

> when you are *disposing* of your **undisposable** *income* on people you either don't know or know only too well.

the collocate *income* decodes the element *disposable* in the new word; whilst the collocating item *pore over* in:

> He doesn't *pore over* **sourcebooks**, he didn't use Dore for Don Quixote; he prefers to let his imagination work untrammelled.

clarifies the superordinate status of *books* in the new word compound, so that we at least know that *sourcebooks* are a type of book. And the collocate *dancers* of *morris* helps to explain the meaning of *morrismen* in:

> A court in Taunton issued an injunction against a group of morris *dancers* after a farmer claimed they were upsetting his goats. When the **morrismen** danced each Sunday lunchtime in the carpark of a local pub, the goats showed clear signs of distress.

4.2.3.3 Rare word vs frequent word collocates

Another paradox arises here. Collocates which are rarer than the new word should be more useful in helping to predict the commoner word element. The collocate *snow* does not strongly predict the new word *softfallen*, in:

> In the gentler form of a new **softfallen** mask of *snow.*

since *snow* collocates with many words, (most significantly: *ice, rain, snow, artificial,* and *heavy).* Whereas the rarer *softfallen*, occurring only once in 360 million words and so having very few collocates, strongly predicts *snow.*

Equally, *imaginary*, in spite of preceding the new word and, thus, being potentially more 'predictive' than a following collocate, should not be a good collocational clue for *playfellow* in:

her cynical New York friend Fergus-turns out at the end to be an *imaginary* **playfellow**.

since *imaginary* collocates with many more words than *playfellow* does, which occurs only once in 360 million words.

On the other hand, the typical collocates of a fairly high frequency word like *snow*, or *imaginary-of* which the most significant collocates are: *real, conversations, world, line, worlds, friend, friends, characters, entirely, landscape, island, set, journey, town,* and *conversation-fall* into clear semantic groups (the latter representing universal aspects of life such as people, human interaction, locations). It is likely that these will be intuitively known to the reader and help to place the new word.

4.2.3.4 Position of collocates: before or after the new word

To us, the positioning of the collocate does not seem to make any difference to the task of decoding. This may say something about our reading process: we have been reading for phraseology, in overlapping chunks, backwards and forwards-assigning provisional meaning, then revising in the light of subsequent or cumulative information. We expect that the average reader of a new word reads from left to right, then if necessary 'does a double-take' on reaching a new word which has not been decoded by what preceded it. A particular reading strategy may be invoked by the particular type of patteming surrounding the new word.

4.2.4 Semantically-related words

One specific type of semantically related words which may be present in the environment of a new word is a synonym or antonym, sometimes of the whole word, sometimes of just one of the elements in the word.

In our data, synonyms can be helpful, though they need not be. For example, the presence of a synonymous prefix *'in(delicate)'* in:

Not to mention the totally unearnest and indelicate men who make Tube travel

a nightmare, morning (groping), noon (ogling) and night (vomiting).

does not help the interpretation of *unearnest* beyond highlighting its negativity. On the other hand, the full-word synonymy in the example:

'I usually grow various forms *ofjacial hair* for a role. For myself, I'm *unmoustached. Moustache-less.'* He is getting into his stride here. 'Demoustachioed.'

Kline is sharp and articulate, with a penchant for wordplay.

does provide help, in juxtaposing *unmoustached, moustache-less,* and *de-moustachioed.* But the hyponymy of *facial hair* and *moustache* is also helpful, as is the repetition of the base *moustache.* Synonymy in itself becomes crucial to the comprehension of a new expression when there is metalinguistic commentary on the synonymy. Consider its clarifying role in:

(...) a text notable also for its author's exhausting *synonymisation* of the female pudenda *(honey pot, lambpot, cream-box,* fur-pie, *thennal pudding,* etc., etc., etc. pretty well ad infinitum).

The presence of synonymous elements can be rather more helpful where the intended interpretation of the new word is too dependent on metaphor to be easily retrievable out of context. This is the case for the new word *fur-pie* in the example above. In the example below, the metaphor *bestraddled* is explained by paraphrase:

He has (...) **bestraddled** the often mutually uncomprehending worlds of the scholar and the practitioner in international affairs, and of the churches and Whitehall. He has *made connections* (. . .) *between* research and policy, politics and ethics, religion and diplomacy; and *brought together* for serious discussion and social fellowship people with so little apparently in common.

Antonyms function less frequently than synonyms as contextual anchors in our data, and they do not appear to contribute greatly to the interpretation of a new word. An important factor is whether they are morphological or lexical. Morphological antonyms are more helpful than lexical antonyms in being easily recognisable and interpretable, like root or base repetition. Take, for example:

Doctors also recognise that it is not necessary for a woman to have an *orgasm* to get pregnant, since **anorgasmic** women (women who have never had an *orgasm)* have become mothers.

Assuming for a moment that we do not know the derivation *anorgasmic* but do know its root, the presence of the latter would be helpful. But morphological antonyms are very rare in our data.

The fact that, in our data, antonyms generally follow the new word in whose interpretation they might play a role, could be a partial reason for their limited helpfulness. However, even where the antonym (in the next example, of one element of a compound) precedes, it is not necessarily particularly helpful. In this example:

> How many people realise that cocoa beans vary in quality and that the beans from the criollo tree are *delicate* compared to the **crude-ftavoured** forastero which provides the bulk of the world's chocolate?

the semantic connection is obscure, perhaps also because *delicate* is ellipted from *delicate-flavoured*, a parallelism with the new compound which would have alerted us better to the semantic connection. Yet in cases where there is no ellipsis and a preceding antonym, the context is still not immediately
helpful:

> It made me wonder why we have the Order of the Bath. Was it an attempt to distinguish our *fragrant* selves from the slurry-covered enemy across the Channel, perhaps? I also wonder whether young Mr Major could foster his attempts at 'democratising' honours by inventing a new one more in keeping with the times.

This illustrates that antonymy in text is usually unconventional and often phrasal, both potential obstacles to interpretation. Even eliminating the second of these, the information may not be of much benefit, as we see in the next
two examples:

> As walkers and hikers don kagouls to cope with the onset of *sodden* autumn/winter conditions, Nike has been preparing state-of-the-art equipment to keep their lower extremities defiantly **unmoist**.
> His first band 'The Frantic Elevators' was an altogether **scratchier** affair than the *super-slick* 'Simply Red', though.

Although these may be prime examples of the thesaurus in operation
in text, the antonymous pairs *soddenlunmoist* and *scratchier/super-slick* are hardly core components of the mental lexicon, and so perhaps do not jump out at the reader.

Sense relations can cross grammatical boundaries. For synonyms, the obscurity of the semantic link illustrated below seems partly attributable to that:

> He's also a dab hand with *amateurs.* He showed me how to carry my skis without poking someone's eye out, warm up, sidestep, snowplough-all the dull **beginnery** stuff-with minimum fuss.

Though this does not appear to hinder the immediate perception of semantic links between morphological antonyms, it does with lexical antonyms.

Overall, we find that semantically related words, whether synonyms or antonyms, do little more than confirm the lexical domain within which the text is operating. This confirmation can be helpful, but is, perhaps surprisingly, rarely the crucial piece of information which leads to the interpretation of the new word.

4.2.5 Lexical signals

The recognition of semantic relations in text can be helped by their being signalled by certain accompanying lexical and lexico-syntactic patterns (Renouf, forthcoming). For instance, in the earlier example:

> How many people realise that cocoa beans vary in quality and that the beans from the criollo tree are *delicate* compared to the crude-ftavoured forastero which provides the bulk of the world's chocolate?

there are two signals, *vary in* and *compared to.* In our data, this sentence happens to represent the typical lexico-grammatical pattern in which *vary in* occurs, that is, following a superordinate plural noun and preceding two contrasted co-hyponymic plural nouns. The reader may be subconsciously aware of this, and thus able to deduce that a contrast is likely to be understood between *delicate* and *crude-flavoured,* in a way which compensates for the lack of parallelism mentioned in Section 4.2.4.

The two words signalled as being sense-related may, however, be an unconventional pairing, as in:

> The *anti-aircraft fire* was meagre compared to the *fireworks of 1991.*

Lexical signals in these cases may be taken as a sign of some pragmatic restructuring (Renouf, *ibid.)* and the existence of a semantic relationship, without it necessarily being clear what the precise semantic relationship is. This makes them useful, but not powerful, predictors of semantic content.

4.2.6 Parallelism

By parallelism, we mean the juxtaposition in text of two similar lexico-grammatical strings. Parallelism itself does not explain the meaning of a new word, but it draws attention to the presence of a clue in its environment. It takes several forms as part of the formulation of a syntactically-substitutable definition or gloss (see Section 4.1.3) or in conjunction with the repetition of the root or base of the new word (see Section 4.2.1). It can be quite complex, as in:

> 'Lucky' Jim Dixon, the young lecturer in History, is the stranger at the High Table, happier *with* the bottle of beer and the blonde *than with* academic or

> artistic gatherings, easier *with* his own common sense voice *than with* professorial **high-speak**.

In the above example, there is a rather complex network of relations set up. Information about the semantic domain and the topic is already provided in the opening of the first clause. Then a complex NP, *academic or artistic gatherings*, contrasts lexico-grammatically with *the bottle of beer and the blonde.* Then the new word *high-speak*, together with its pre-modifier *professorial*, is similarly placed in semantic contrast with *his own common sense voice*, and this second complex NP is placed in grammatical parallel with the first. Out of this will come, for some, the realisation that *high-speak* means 'pretentious academic style of verbal expression.'

> He has (...) **bestraddled** the often mutually uncomprehending worlds of the scholar and the practitioner in international affairs, and of the churches and Whitehall. He has *made connections* (. . .) *between* research and policy, politics and ethics, religion and diplomacy; and *brought together* for serious discussion and social fellowship *people with so little apparently in common.*

Here, there are no really significant collocates; the new word is a metaphor, possibly a blend of *bestride* and *straddle*, but its meaning is inducible from the parallelism in structure between *He has* (. . .) *bestraddled* and *He has made connections* (. . .) *between.* This is reinforced by colligational indicators: *bestraddled* is a transitive verb (he has VERB+ed NP), and the NP consists of two ('mutual' and also the pattern NP1I2 and NP3/4) conflicting ('mutually uncomprehending') worlds.

4.2.7 Lexical field

By 'lexical field,' we mean a series of words relating to the same topic area. We employ the term loosely, and it will overlap with related concepts such as domain, genre, and semantic prosody. The lexical field is not likely to be used deliberately to indicate the meaning of a new word, but it will give the reader a general idea of its area of meaning and will reduce the possible interpretations for a polysemous word. For instance, for the new derivation *gartery* in:

> Joe owns Agent Provocateur, a shop ... which sells all kinds of fluffy, frilly, gartery, girly things, a cornucopia of sibiliant sauciness: satin basques, silk knickers, seamed stockings and shiny black bras that lend breasts a conical shape.

circumstantial evidence is offered that it is (an attribute of) an item of underwear worn by females. In this case, the fact that the reader is left by the writer to induce the meaning from the context creates a probably intentional amusing effect. Meanwhile, with the compound *lumpsucker*, in:

> The seals' menu is extremely varied. Mr Strong recorded about 40 species of fish-from little black goby and ugly **lumpsucker** to octopus and squid.
> Salmon and sea trout-of concern to river fishermen-hardly registered.

the surrounding lexis goes quite far in indicating that it means an ugly-looking species of saltwater fish eaten by seals.

4.2.8 Pragmatics

Perhaps the loosest type of help is that which we have labelled 'pragmatic'; the kind of help that is not present linguistically in the text and which is thus only marginally to be included as a 'gloss' at all. This seems to occur most often in cases where the writer does not allow for the word being new to the reader. So it is probably not provided deliberately but is available to the reader who recognises the real-world import of a particular expression or name (often brand name) present in the context. Consider, for instance:

> Having invented the flavoured *crisp* and so knocked *Smiths* with their little blue **salt-bags** off their perch, *Golden Wonder* did wonderfully well.

where *salt-bag* is interpreted in the light of *crisp* and *Smiths* for those of us who recall the tiny twists of blue paper containing salt in the packets of potato-chips made by the UK company of that name. 1 Or the following:

> The amount of *BSE* around is absolutely enormous and as the quantity goes up so does the *risk.* I'm a happy **beef-eater** but I will never eat processed meat.

where *beef-eater* has to be reinterpreted away from its established meaning ('guard at the Tower of London') to a new literal interpretation guided by *BSE,* which has to be interpreted as something which affects beef. Finally, consider the following, where *malt-seller* is interpreted partly in the light of a number of brand names for malt whiskies and names of distillers:

> Malts were barely sold outside the north of Scotland until 1964, when *William Grant* & *Sons* launched *Glenfiddich* to a wider audience. It is still the world's biggest **malt-seller**, although many others have piled into the market since. *United Distillers,* owned by Guinness, has capitalised on consumer desire for experimentation with its range of Classic Malts *Lagavulin, Talisker, Dalwhinnie, Cragganmore, Oban* and *Glenkinchie-sold* in full bottles or a miniature pack.

We see, therefore, that there is a cline of semantic and pragmatic support, from a directly synonymous phrase to a hint of the area in which an interpretation should be sought, but that a precise subcategorisation of this cline may be neither desirable nor possible.

5. Multiple clues to interpretation

In the last section, we referred to the fact that many of the contextual clues are used, or occur, in combination. Perhaps this is best demonstrated by a small sample of neologisms which have been coded for types of contextual help.

Table 1. *Coded entries (edited) from the Quotation Marks file*

Key to Codes (extract)

base	base form element	qm	quotation marks
el	following phrase	rep	exact repetition
follphr	gloss in context	sem	semantic equivalence/relationship
gl	introductory phrase	sen	in another sentence
introphr	lexical field	syn +/-	synonym
lex	parenthetical	1st/2nd	feature follows/precedes
par			first/second

Word: *blotterhound*
qm gl- par

> around and try to pick up the scent on the other side. He's been training four otterhounds since April this year, and is now starting on a bloodhound-otterhound cross-a '**blotterhound**' called Bowman. Out on moorland on the edge of the New Forest, Cautious and Grayling go lolloping through treacherous-looking bogs, and plunge into a ditch of brackish water, basking in

Word: *fruit-only*
qm I stellexsen-l

> version of freshly baked scones stuffed with clotted cream and strawberry jam Wafer-thin, organic brown flour, low-salt bran scones with a scraping of soya-based sour cream and a smidgin of '**fruit-only**', no-sugar jam. And for all the effort taken to avoid feeling high, low or giddy, food mood bores always end up looking spooky anyway. Ever noticed healthy eaters always look

Word: *neo-mercantilist* qm
introphr gl+

> same time the country found itself by force of events in the international limelight. Using its new-found respectability, it latched on to what Professor Vale calls economic pragmatism, or the '**neo-mercantilist**' model of international affairs, which views the world as being driven only by economic issues: trade, industry, etc. The Foreign Ministry spent most of its energies using Mr Mandela's reputation

Word: *re-ratting*
qm follphr gl+ baserep-2

> to imagine him defecting again, this time perhaps to something exotically millenarian The converts themselves may strenuously deny it, but conversion, with its attendant passions, becomes a habit. 'Ratting' and '**re-ratting**', as Churchill described his repeated journeys across the Commons, is the modem disease. It is the nomadism of modern man; the realisation in the guts that one's true spiritual home.

Word: *specs-ist*
qm basesynsen+ 1 basesemsen-l

> self-regulating and un-enslaved-which was what the sexual revolution was about, after all. Letter: Four-eyes strike back (By MS FRANCES GILYEAD. Sir: Shame on Vicky Ward for her thoughtless '**specs-ist**' remarks about those beautiful Bostridge boys (Diary, 10th October). Glasses do not automatically make the wearer undesirable nor pitiable. For proof of this fact, Ms Ward could take a quick

As can be seen from the examples above, we have conducted a fairly extensive codification and established that certain contextual features tend to co-occur. The overt or conscious types of anchoring are the ones most likely to cluster. Further detailed statistical analysis of our markup might lead to new insights.

6. Discussion

Although Aitchison and Lewis (1995), in a discussion of the treatment of the word *wimp* in British newspapers, find that

> "over 80% of *wimp-word* tokens contain information on their meaning in the immediate surrounding text."

(even when it has stopped being so new!), our findings are that texts are not generally as helpful as that. The journalistic nature of the data is likely to be a factor. We began by hypothesising that journalists wish partly to amuse and assume knowledge of their style in an established readership, and for both these reasons, they would be likely to introduce new words without overt explanation. Looking to our classes of overt contextual help as specified in Section 3, only around 70 new words are overtly marked by quotation marks, 20 by introductory phrases, six by following phrases, and 65 explained by glosses; some are the same new items with multiple marking. So overt contextual help in our sample of journalism is offered for only about 2.5% of the total 4750 new compounds and derivations that we observed.

Nevertheless, the lack of overall anchoring surprised us. This lack is apparent even when words which are new to the computer but not to the human reader are ignored, and only genuine innovative uses are considered. It is true even though we have tried to maximally accept possible influences from the surrounding text; indeed, some of the 'aids' we list may seem to be scraping the barrel. But we do not believe this to be the case. Rather, we think that interpreting new words is a process which involves much more than linguistic abilities or even encyclopaedic knowledge. We view it as a holistic process which uses our linguistic abilities, engages our world knowledge, but also demands pragmatic inferences from the text-probably from a much larger portion of text than we actually examined2 and possibly from beyond the text as well in spoken interaction. If this is the case, it is surprising to us that linguistic abilities seem to be left with so much to do in interpreting these journalistic innovations.

Morphological analysis has not been discussed so far, since our objective has been to examine the established assumption that readers are provided with contextual help in interpreting novel words. Nevertheless, we have noticed that the morphological analysis of a word is a helpful guide to its 'semantic' meaning in many cases. Though it obviously cannot specify referential meaning, internal semantics often seems to be sufficient for an appropriate interpretation. Words which are interpretable by this means include: *milkophile, earth-lubbers, leafbuster, pottyloads, snoree-and* even *lampboard* (where this refers literally to a decoding device consisting of a board fitted with light-bulbs). Morphological analysis does have its limitations: it does not help with metaphor (see *catheads*, meaning 'a type of American biscuit' and, by extension, 'breasts') or polysemy (see *brimstones,* which could mean 'sulphurous stones/butterflies/viragos,' but in fact means 'whores'), but we find relatively few instances of these in our data.

We have not only found that morphological analysis can be sufficient to interpret a new word but that it is sometimes the only aid which allows the reader to interpret its meaning. Such a word is *post punk*, shown below, as it is contextualised in our data. The lexico-grammatical framework 'from X to Y' (Renouf, forthcoming) indicates that this word refers to an extreme on a scale, but the context offers no clue as to the precise meaning of the word itself:

> This scholarly anthology is both a cultural history and a literary odyssey. Ranging from fairytale whimsy to postpunk invective, from fables of oppression to those of liberation, it is full of unforeseen delights surprising us into reshaping our thoughts about familiar writers, about sexual politics and...

This all leads us to the realisation that morphological processing is an important facet of our linguistic competence: one which is used regularly to interpret incoming messages. While this is not a new idea, it has tended to be overlooked as far as word-formation is concerned, even if it has been accepted for inflexion.

There are problems with such a view, of course. One is that if we are correct, we cannot easily account for the poor performance evidenced by Gleitman and Gleitman's (1970) subjects. Their performance can be explained if it is the case that real speakers meeting real cases of word-formation rely heav ily upon the context to interpret the word, since they were not given contexts in those experiments. But we are suggesting that there is remarkably little help in most cases observed. There are two possible solutions to this apparent contradiction. One is to suggest that there is no real contradiction because Gleitman and Gleitman's subjects got some of the forms right, perhaps the proportion that people can cope with, but were overloaded by stimuli with no context, which is not natural (about half of all stimuli should provide reasonable help). The other is to suggest that we have not cast our net wide enough in the search for contextual clues. Given Gleitman and Gleitman's example of *house-bird glass*, we would have found help in such sentences as *She drank from the house-bird glass* or (with a different interpretation) *She caught sight of herself in the house-bird glass*, but *Give me the house-bird glass when Cheeky's finished with it* would not have recorded as helpful, even if *Cheeky* is known to refer to a budgie and provides a definite sin gular referent. Neither would we have recorded obvious contextual help in

I dropped it in the house-bird glass, although this establishes *house-bird glass* as a container of some kind. Thus, our figures are conservative, even with the number of factors we have considered. We feel this stresses the holistic nature of the interpretation process, which is grammatical and pragmatic as well as lexical and semantic. If all such clues are included, Aitchison and Lewis's 80% is probably a reasonable estimate.

7. Conclusion

At the outset, we assumed that it was possible to deduce part of the meaning of an innovative word from the meaning of the known context. We have found this to be true for new compounds and derivations with the aid of the kinds of anchoring which we have identified. However, we have discovered that the degree of support provided by the surrounding context is generally low: the individual kinds of support are diverse and sometimes indirect, and the interpretative process diffuse linguistically over many types of support. The degree and type of contextual anchoring can be expected to vary across word formation classes, texts, and domains, but in whatever form it is present, we have confirmed that the process of interpretation is a holistic one.

We have found that almost all our new compounds and derivations are semantically, if not referentially, interpretable by means of their internal components and that morphological processing is not just a vital back-up procedure to contextual analysis but that it is probably the single most efficient starting-point for deducing the meaning of journalistic neologisms (see Bauer, forthcoming, for a summary of psycho linguistic evidence to endorse this finding).

Notes

1. Proper names are not the focus of this paper, but as brand names, they bring a real-world dimension to the text, and as collocates of the new word, they help the reader who recognises them to construe its specific meaning and reference. They are the most precise semantic specification that can be provided (multi-referential and, thus, ambiguous items such as *John Smith* notwithstanding). Below, the brand name *Tetley* explains *teamen,* which is unknown and morphologically difficult to construe:

> A former professional wrestler, he was the voice behind the *Telley* teameD *ads* and the teacher in Kes.

Familiarity with the television advertisement is presupposed, and this explains the word *teamen,* which refers to its cartoon characters.

2. We noted at the outset that our method allowed us to consider hapaxes within a 30-word span of context. In some cases, this has had an effect on our results. We cited above, for example, a context for the hapax *ladettes,* concluding that it offered little interpretative help. It transpires that the wider context is an article about laddishness, thickly sown with derivations of *lad.* In such a lexical domain, the form *ladettes* is easily interpretable. On the other hand, another hapax, *smoulderers,* cited earlier, is marooned in the wider context of an article with only *smouldering* for company. It is likely that the lexical field will be richer if the new hapax is closely associated with or names the central topic. We have no way, short of studying every article, of establishing this or of knowing the proportion of larger contexts which are helpful for this reason.

References

Aitchison, Jean and Diana Lewis. 1995. "How to handle wimps: incorporating new lexical items as an adult." *Folia Linguistica* 29: 7-20.

Baayen, R. Harald and Anneke Neijt. 1997. "Productivity in Context: a case study of a Dutch suffix." *Linguistics* 35: 565-587. Bauer, Laurie. Forthcoming. *Morphological Productivity.* Cambridge: CUP. Collier, Alex. 1993. "Issues of large-scale collocational analysis." *English Language Corpora: Design, Analysis and Exploitation: Papers from the 13th 1CAME Conference, Nijmegen* 1992, eds. Aarts, Jan, Pieter de Haan, and Nelleke Oostdijk, 289-298, Amsterdam: Rodopi.

Collier, Alex and Mike Pacey. 1997. "A Large-Scale Corpus System for Identifying Thesaural Relations," *Corpus-based Studies in English-Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME* 17), ed. Ljung, Magnus, 87-100. Amsterdam: Rodopi.

Gleitman, Lila R. and Henry Gleitman. 1970. *Phrase and Paraphrase: some innovative uses of language.* New York: Norton. Kastovsky, Dieter. 1986. "Productivity in Word Formation." *Linguistics* 24: 585-600. Nagel, Rainer. 1997. "Zur Behandlung textueller Wortbildungsvorgiinge vor einem prozeduralen Hintergrund." *Zeitschrift flir Anglistik and Amerikanistik* 45: 1-19.

Pacey, Mike, Alex Collier, and Antoinette Renouf. 1998. "Refining the Automatic Identification of Conceptual Relations in Large-scale Corpora." *Proceedings of the Sixth Workshop on Very Large Corpora, ACUCOLlNG, Montreal,* 15-16 *August 1998.* 76-84.

Plag, Ingo. 1996. "Selectional restrictions in English suffixation revisited: a reply to Fabb *(1988)." Linguistics* 34: 769-798.

Renouf, Antoinette. 1993. "A Word in Time: first findings from dynamic corpus investigation." *English Language Corpora: Design, Analysis and Exploitation,* eds. Aarts, Jan, Pieter de Haan, and Nelleke Oostdijk, 279-288. Amsterdam: Rodopi.