What can a corpus tell us about lexis?

Rosamund Moon

1. Lexis and the lexicon

For corpus linguists, it is difficult to see how anyone can learn much about lexis without using a corpus, or could fail to learn something from each new corpus search. Lexis can be researched simply through exploring individual lexical items and their behaviour, or by using corpus data to examine the lexicon as a whole or to test lexical theory. This chapter gives an introductory overview of some key aspects of English lexis, all of which can be straightforwardly investigated with corpora. Data is drawn from the 450-million-word Bank of English corpus (BoE), created by COBUILD at the University of Birmingham (71 per cent British English, 21 per cent North American, 8 per cent Australian; 86 per cent written, 14 per cent transcribed spoken data).

The general lexicon

What do corpora tell us about the English lexicon? This kind of information is best obtained from a large reference corpus (ideally, at least fifty million words), since smaller and specialist corpora are likely to show skewings, with too few examples of rarer words. However, the obvious question – how many words are there in English? – is unanswerable, even with very large corpora. All corpora do is reveal which words are used in their constituent texts and how frequent they are.

A more useful question is how many words comprise the main vocabulary of a language, its central lexicon. Here word frequency lists provide a starting point, demonstrating that a comparatively small set of words accounts for a large proportion of text. Table 15.1 shows the cumulative proportions of BoE comprised by the N most frequent lemmas (base forms and inflections), and their approximate frequencies per million words of corpus text. Beyond this is a long tail of infrequent items, including hapaxes (words occurring once), of which many in BoE are names, numbers or errors.

Unsurprisingly, most very high frequency items are grammatical words: BoE’s top ten lemmas are the, be, of, and, a, in, to (infinitive particle), have, to (preposition) and it. It is usually the lexical words, though, which seem more interesting: BoE’s top 100 includes the noun lemmas year, time, person/people, day, man, way, verbs say, go, make, get, take,
Table 15.1 Distribution of lemmas in BoE

<table>
<thead>
<tr>
<th>Number of items</th>
<th>Percent age of BoE</th>
<th>Approximate rate of occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>top 10</td>
<td>23.5</td>
<td>8,000+ occurrences per million</td>
</tr>
<tr>
<td>top 100</td>
<td>44.9</td>
<td>1,000+ per million</td>
</tr>
<tr>
<td>top 1,000</td>
<td>68.5</td>
<td>100+ per million</td>
</tr>
<tr>
<td>top 10,000</td>
<td>88.9</td>
<td>5+ per million</td>
</tr>
<tr>
<td>top 25,000</td>
<td>92.5</td>
<td>1+ per million</td>
</tr>
<tr>
<td>top 100,000</td>
<td>94.5</td>
<td>1+ per 20 million</td>
</tr>
</tbody>
</table>

know, see, come, think, give and adjectives new, good. These are all fairly general words, many associated with semi-grammatical functions – for example, deictic uses of year(s), causative and phrasal uses of make:

rule rather than the exception. Last year 44 per cent of secretaries placed by a mere fling of fancy just a few years ago, but not any more.
government collapsed earlier this year. As an aside, the White House turn negative. <p> Over the past five years Central London has seen the UK’s Hadjout. <o> Beside her slept her two-year-old daughter, Belle, the child’s athletics. Few athletes, if any, have made a greater contribution to the details of all staff which will be made available to a range of companies make ILT membership worthless, and making entry so hard and time-consuming than curiosity. Her husband’s motives made no difference to the legal act which who are able to participate in trials make up only 5 per cent of the total.

This kind of data is replicated in other reference corpora, though exact rank orderings of words and rates of occurrence vary according to corpus composition and lemmatisation policies. The effect of composition is evident when we consider distribution: even high frequency words may have quite different frequencies in different types of text. See Hunston (2002: 3ff) for discussion of word frequency; also O’Keeffe et al. (2007: 31ff) with respect to applied linguistics. Leech et al. (2001) list frequencies in the British National Corpus; Biber et al. (1999) distributions of words within grammatical categories across discourse types.

Lemmatised frequencies are useful, but equally important are the relative frequencies of individual inflections and what these suggest about usage. While it is unsurprising that in BoE peas is twice as frequent as pea, it seems more surprising that fact is eight times as frequent as facts, though corpus data quickly shows why fact is so common:

cannot be made to grow elsewhere; in fact Devonshire gardens, full as they are probably fewer would even care. The fact is, we meet so many new people all any drug. The judge said: The mere fact that a customer picks up a bottle of a hiding. There was no denying the fact that players of the calibre of Brian and the insecurities, that are, in fact, the complex characteristics of

**Word formation**

General corpora also provide information about derivation and compounding, helping establish which potential words are actually institutionalised. In BoE, for example, principal derivatives from the high-frequency root colour/color include colourful, colourless, discolour, colourant, colourist, colourise, colouration, uncoloured...; also colourable, colourism, colouristic, recolour, etc. Formations such as colo(u)rous, miscolour are morphologically possible, but not found. Similarly with words formed with specific morphemes, for example the prefix hyper-: among the most frequent in BoE are hypertension, hyperactive, hyperinflation, hypermarket, hypertext, hyperventilate, hypersensitive. However, rare or hapax items often reveal more about creative processes of word formation, with hyphenated forms especially interesting: thus marginal items such as hyper-accurate, hyper-additive, hyper-animated, hyper-assertive, hyper-bubbling, hyper-blues...may suggest patterns or motivations for coinage. As for compounding, corpora show which formations recur and have specific meanings: for example, BoE has watercolour, colourway, colour-blind, colour-fast, colour-washed, colour-coded, hair colour (dye), adjectivals full-colour, two/four-colour (of printing), and so on.

**2. Phraseology and phrases**

It is much more interesting, of course, to look at words in their corpus contexts than in isolation. Here the interdependence of words becomes most obvious, showing how the phraseological patterning of words are critically important in relation to meaning as well as usage. A consistent finding in corpus studies has been the extent to which words occur as parts of phraseologies, whether collocational, structural, or both: for discussion, see Sinclair (1987; 2004: 24–48); Stubbs (2001: 80ff); O’Keeffe et al. (2007: 14–15, 59ff, 100ff). Phraseology is also covered in other chapters in this handbook (see chapters by Greaves and Warren and by Hunston), but it is too important with respect to lexis not to discuss here too.

**Collocation and patterning**

What can we learn about a word from looking at its collocates, the words with which it co-occurs? We might expect these to be least interesting in the case of items with specific meanings, for example a natural-kind entity like aphid:

> be found on the roots. Lettuce root aphid can overwinter on lettuce roots, by spraying under the leaves with the aphid-specific Rapid insecticide. \(< p > < c >\) are produced by this insect. This aphid pest attacks many plants in the livefood one can supply, since aphids, caterpillars and the like will The larvae of this tiny insect eats aphids, thus reducing the number that can insects like red spider, thrips and aphids. At Brisbane Botanic Gardens, Mt

(Here and elsewhere, concordances are randomised BoE samples.) Co-texts show something about the semantics of aphid: even without knowing what aphids are, we deduce that they are insects or insect-like, small, parasitic on plants, and preyed on by other insects. This is reinforced in listings of aphid’s significant collocates: control, spider, insects, species, mealybugs, plants, thrips, pests, caterpillars, larvae, attack ...
For many words, collocational and syntagmatic patterns are more noticeable, as even a few lines for the less concrete noun refuge show:

mythologised home and family as a refuge from a threatening world of change referred to the YMCA library as a refuge for many half-homeless wanderers Our consciences had been our last refuge. Their sanctity was destroyed. Rel < p > relief prior to reaching a place of refuge. In many cases religious persecution Park. < p > Losure: The river provides refuge for migrating ducks, geese and for him to leave Boston to seek refuge in England, Anne and he were forced of two policewomen who had sought refuge in a nearby doorway. Part of the car and he had regressed and taken refuge in amnesia as a defense against the more than 200 militiamen have taken refuge in a local monastery and sent the after more Cubans had taken refuge in the Spanish Embassy during the

Phraseological patterns include seek/take refuge, refuge in/from, last refuge, place of refuge; the first line contains a recurrent collocation with world, with world representing a difficult, frightening or uncongenial situation.

Corpus data for verbs demonstrates recurrent or mandatory grammatical structures, typical subject and object realisations, and so on. Verbs of motion are typically followed by adverbials or prepositional phrases of direction or manner, even when metaphorical:

against the blood-red sky; a man walking along the curving road; lights order. They left the others. As they walked, at first in silence, Serena tried remaining and isn’t considering walking away before or after that deal protect identities. Why would a mother walk out and leave her children in the in the world. The first person to walk to both the North Pole and the figure with an aristocratic demeanour, walked with the help of a cane. His

With other verbs, grammatical structure and kinds of collocate seem more restricted. For example, comply is typically preceded by something indicating coercion, necessity, willingness, etc. (incentive, must, force, fail(ure), hesitate), and/or is followed by with, itself typically followed by a noun phrase indicating a constraint (agreement, decision, law, obligation ...). The specific meaning of comply is inseparable from such a pattern:

which is a powerful incentive to comply. Ms Coward might have mentioned the recited. President Milosevic must comply in full with the agreements he made assault weapons and ammunition. To comply, owners must either remove them as external structures, to be complied with or rebelled against but not surprising that Franco hesitated to comply with this condition until dwindling 1998 be increased by $246 million to comply with the decision by the New Jersey that would force signatories to comply with nuclear safety standards. The that de Klerk has also failed to comply with UN resolutions demanding the or changed their name. Failure to comply with the new law is punishable by a bout, unless and until he complies with his obligation to fight

Fixed expressions and idioms

In many cases, recurrent collocates, especially when syntagmatically fixed, represent some kind of multi-word lexical item. In the following, of course, on course (to/for) and stay the course have holistic meanings:

> today,' Woods said. This golf course is not easy. Birdies are hard to
> him to drop out of his medical course < /h > < b > By JOJO MOYES < /b > < p >
> swamped by that silly role. And of course often she had to play it while I
> who would have guessed? But, of course, it makes sense.' Then I saw the
> of between 80 and 85%. That does, of course, leave a fairly sizeable margin of
> < p > He doesn't stand a chance of course and neither did Ronald Reagan,
> in some future century? And of course, if we wish to explore our
> < /h > < p > Steffi Graf is still on course to retain her title at the women's
> fixed and invariant during the course of therapy. I cannot even predict
> varieties a year, very few stay the course, despite the fact they are good,'

In fact, BoE evidence for course is dominated by phrasal uses, particularly of course. Similarly, much of the corpus evidence for high-frequency verbs such as take and get consists of occurrences in phrasal verbs and other fixed expressions, while an important feature of data for nouns such as hand, head and heart is the way in which they are used metonymically and metaphorically, including in idioms.

It is difficult to generalise about multi-word items since the range is wide and behaviour varies greatly. Some individual phrasal verbs and compound nouns are themselves very frequent, as are fixed expressions like of course, as well (as), for example, in particular, take place, take part, sort of... . all with important grammatical or discourse functions. However, corpus studies of idioms, proverbs and other such items suggest that most are infrequent, tending to occur mainly in journalism and fiction. A number of claims have been made in the literature about the essential ambiguity of idioms, or their frozenness and syntactic defectiveness; these claims can be correlated with corpus data. Do literal counterparts to idioms actually occur? (only rarely). Are passivisable idioms actually used in the passive? (in BoE bury the hatchet is passive in about 5 per cent of instances, spill the beans in fewer than 1 per cent). While many idioms are indeed lexically frozen, many others are unstable. Thus corpus searches need to be flexible and recursive to find all idiom occurrences, including creative exploitations and interruptions: for example, keep something under one's hat and ... under one's cap/urban/bonnet; or upset the proverbial/political/establishment applecart. Proverbs too are often shortened or varied: BoE has evidence for canonical what's sauce for the goose is sauce for the gander, truncated what's sauce for the goose, variants what is good for the goose (is good for the gander), and manipulations (liberal goose/conservative gander, etc.). The discourse contributions of such expressions, their evaluations and cohesiveness, become apparent from their corpus contexts:

I seem to recall that it was Ashdown who ... almost upset the whole Alliance apple-cart. He did so by overturning the leadership on a major question of defence policy. Sauce for the Czech goose should not mean arsenic for the Slovak gander.

3. Meaning

Context and meaning

By forcing us to consider words in context, corpus concordances make us aware of how far the meanings of words are derived from context – even raising the question of whether words have independent meanings at all. This interdependence of meaning and context is clearest in the case of semantically depleted words, such as much-discussed *take*:

```
court. < subh > Boasts < /subh > It would take a huge investment of police time and
that Whitbread crews are willing to take boats into treacherous waters and
so that I know that I’ve the time to take erm antibiotics and i+ erm not wait
< p > It would seem that Nine is taking its rival very seriously indeed.
in our work, what will it mean that I take off certain Jewish holy days but
send 2300 more men to the Balkans to take part in a future peacekeeping force.
< p > mcilroy said: David practises taking penalty kicks for half-an-hour
the early 1960s an economic miracle took place in West Germany under the ‘
would foment a revolution. Well, it took us a long time, but we finally
which was lucky. After two hours they took us to our hotel and made us collect
```

Trying to abstract an essential meaning seems pointless: each use needs a different explanation of meaning or paraphrase. There are some fixed expressions (*take part, take place, take something seriously*); others are simply recurrent usages, and a proper sample would provide many more. What is the commonality of meaning, the meaning of ‘take-ness’? Another case is the noun *heed*:

```
their own vendettas with little heed to a common tactical plan. Air cover
and male chauvinism to give much heed to claims that the equal rights
all-American halls of fame to pay any heed to what the Olympic Museum has done
also starting to run. Anna paid no heed but kept on going, in seconds
the only contributor to pay serious heed to the four directions, so crucial
Edwina Currie should have taken heed. It was Mr Winterton who said, the
with millions of women taking heed of such advice. Reading the book I
card but after he failed to take heed of the warning, they were forced to
Businesses might need to take heed of pressure groups if the latter
pursuing its rational ends without heed of the wider consequences. Darwin
```

We might feel that *heed* means ‘attention, notice’, but corpus evidence shows that it normally occurs in a restricted range of structures, including the curiously synonymous *pay/give/take heed*. If its usage is so restricted, how versatile can its meaning really be? Can meaning be separated from phraseology? We can pursue this point with the adjective *heinous*, which a dictionary might gloss as ‘odious, wicked’:

```
apparently sensible people consider heinous. < p > A concept of obscenity
university student. This was a most heinous crime, savage and brutal in
by the rope were guilty of such heinous crimes. Many innocents had already
blacks. < p > While offering their own heinous schemes for raising Americans
```

202
Certain crimes, it said, were so heinous that the normal immunities granted to their behaviour is criminal. Heinous. Unjust. Unfair. Inequitable. Her

Heinous indicates unacceptability, but part of its meaning is supplied through collocation with crime and related words. While take and heinous represent very different kinds of word in terms of meaning, a corpus perspective suggests that both reflect the way in which meaning is a construction of context, not intrinsic to the word. It might be objected that there are plenty of other words in the lexicon which are semantically full, have specific denotations and are not collocationally restricted: words like new, walk, year or aphid, peas, hypertension. Any number of words could be substituted in the following concordances, changing meanings, but still maintaining syntactic and collocational well-formedness:

The larvae of this tiny insect eats aphids, thus reducing the number that can medication they had received for hypertension. The fish oil supplement used by publishers are convinced that the new book is guaranteed to be a best-seller and spices, then add the tomatoes and peas. Simmer until cooked – about 15-20 against the blood-red sky; a man walking along the curving road; lights turn negative. Over the past five years Central London has seen the UK’s

What can corpora tell us about the meanings of such words? One response is that corpora show us typical contexts in which we encounter a word – even if we first acquire them through ostension (walking, peas, aphids) – and that part of our knowledge of a word relates to its associations with other words: cf. Firth’s comment ‘One of the meanings of night is its collocability with dark, and of dark, of course, collocation with night’ (1957: 196).

Polysemy

Another aspect of word meaning to explore is polysemy: how many different senses or uses words have, and how these are distinguished in context. Collocates discriminate by indicating topic and semantic field, with different senses often associated with different phraseological patterns and structures. The noun race is relatively straightforward:

The reigning champ has won 2,641 races – three behind Arthur Stephenson. years to promote racial equality. Are race relations getting worse? Have your

Mr. ERROT: Every religion, every race, every profession, every age group colored by the signifiers of gender, race, class, and culture that obtain and nuclear war in which the entire human race was wiped out except for the the title to be decided in the last race was probably the best thing as it the personalities in the presidential race: the current president, Joaquin with victory in the Five Valleys road race at Port Talbot, South Wales, on overall lead in the Paris-Nice stage race yesterday, despite an attack on his be $41 better off by not going to the races each Saturday and going to the

The primary sense distinction between ‘competitive activity’ and ‘ethnic grouping’ is clear, reinforced by words such as champ, won, title, victory on the one hand, relations, religion, gender, human on the other. Since race derives from two distinct roots, it is homonymic
rather than polysemous; a more synchronic view, consistent with corpus approaches, is to see this in terms of a major semantic discontinuity and lack of collocational overlap.

See also has multiple senses, with a primary distinction between vision and comprehension. In context, these are distinguished through the nature of what is seen—physical objects (bodies, trees, a horse, text) or entertainment, and ideas (contrast, one's status, situations):

the rest of the family. < M01 > Yes. I see. And where did you settle in Britain had lunch at Grill & C. then went to see Billy Liar with Albert Finney being authority in her behaviors (that is, saw herself as having no real' power).

12 The contrast can be clearly seen in the worlds of employers and class gauging local opinion. I don't see it as a national forum,' he says. because it is now clear for all to see that he leads two Conservative described what he saw. ANATOL LIEVEN: I saw the bodies of four Lithuanian She glanced out of the window and saw long, thin trees standing in lines supernatural powers—such as seeing through solid objects—by onto the same chip as the microarrays (see Water gets weird', New Scientist,

Adjective senses can be distinguished through types of noun collocate, the entity being described. For colourful, we might group together physical objects and places (cloth, gardens); people and their lives, sometimes with an implication of notoriety; and then more abstract things (lively descriptions, mixtures of culture):

The centre of Harare is at its most colourful at this time of year: the avenues made of transparent plastic with colourful balls inside. The base has how to manage a football team. The colourful businessman, a close pal of Serb back legs chained and covered in colourful cloth full of pattern and finery, just fade away. < p > Mr Norris' colourful description of what a Clarke although the settings are colourful enough. The mix of images and his address by noting Fairfax's colourful history, saying this had included show of Ericas make the garden so colourful. I have therefore arranged that the imagination wandering. < p > The colourful mix of cultures adds a further speculation and finance. The colourful yet elusive baron, aged 72 this

One analyst might say that these senses are substantially the same, a core idea with contextualised metaphorical applications. Another might find further nuances of meaning. There are arguments for and against fine-grainedness of sense distinction: what corpus data contributes is a way of having the argument in the first place. With take, an extreme example of a polysemous word, we have already seen that its meanings are tied up with phraseology. It would be impossible to investigate this robustly without a corpus. See Sinclair (2004: 24–48, 131–48) for discussion of meaning, ambiguity and phraseology; see Hunston (2002: 38–66) for further examples and analysis.

Metaphor, connotation and ideology

Figurative senses such as see 'understand' and colourful 'lively' are institutionalised: corpora also contain large quantities of more creatively figurative language, but it is not always easy to locate. Systematic studies of metaphor or metonymy may start by searching for a
specific item, such as a set of verbs of vision/cognition, or a metaphor-rich word like heart. The following sample omits literal uses, and includes fixed expressions (break one's heart, change of heart, take heart) and more generalised allusion to the heart as source of emotions:

To see a little kid just breaks your heart. <p> I tell them how brave their welcomed Mr Trimble's change of heart', as did Sinn Fein and the
in the group, though Ireland can take heart from their efforts in frustrating
But it was a sight to gladden the hearts of supporters of the English bid.
the highest standard comes from the heart, not a textbook. 7 Innovate or die.
states in his introduction # at the heart of this book's analysis is the
fit as he was when he captured the hearts of all Australians along with his
form of resistance - in the mind, the heart. Outwardly, she would try to comply
Bobb and her record label is a real heart warmer. Her first single, Dreams,
blood was even more overpowering. My heart was beating so fast it felt like

The last line is arguably literal, but such uses typically also imply psychological states (one's hands shake, one's stomach churns ...).

Connotational meaning seems to depend on intuition and suggestion, but are there traces in corpus data? If we follow Firthian principles of meaning by collocation, then collocates may well provide answers. For example, a word such as claret denotes a kind of red wine while connoting a certain lifestyle:

Bosses said they now sell more claret and champagne than cheap plonk. The
is renowned predominantly for claret and Sauternes, it also produces
Sainsbury's. Classic medium-ranking claret from the finest recent Bordeaux
in port and Stilton after duck and claret is not the wisest behaviour for a
waiting for me alongside a glass of claret. Sorry,' I said to Jack. I'll now
advising Mr Chairman on what claret to buy in for his cellar. Claret?

Its collocates include bottle, glass, drink; general adjectives such as good, old, classic; wine-related items burgundy, bordeaux, champagne, port, cnu, decanter; vintage, fruity, fine, full-bodied; foods duck, stilton. As a set, they seem to point to that intuited claret lifestyle. See Stubbs (1996: 85–6, 106ff) for discussion of connotation and further examples.

The recurrent lexicogrammatical patterning of words such as refuge, comply, storm also helps us explore connotation in relation to semantic prosodies (Louw 1993; Stubbs 2001: 198ff and passim; Hunston 2002: 141ff, whose term is 'discourse prosody'). We can start by identifying canonical semantic structures:

(take) refuge from + UNDESIRABLE SITUATION + in + SAFE PLACE
someone (+ MODAL) + complies with + OFFICIAL INSTRUCTION

The interesting cases are those which deviate from these patterns, belittling, criticising or creating irony by subverting the canonical evaluation or downtoning a semantic feature:

presented by some contributors as a refuge from academic life where there is
and he had regressed and taken refuge in amnesia as a defense against the
to see the outside world, forced to comply with the whims of researchers. Yet
Following on from connotation, prosodies and collocation is the use of corpus data to explore ideologically significant items. We might look, for example, at gender issues through he/she (cf. discussion of gendered pairings in Section 4) or ethnicity through race; or else look at ‘keywords’ in Raymond Williams’ terms (1983). One of Williams’ words is bourgeois; corpus examples point to something of what it now suggests:

novels is a caricature. Not all bourgeois consumers of print read novels; between them did much to define bourgeois manhood in the nineteenth increasing prosperity, and bourgeois satisfaction, if not complacency, capitalist economic order and of bourgeois society. Although their recipes descended, and definitely not nice, bourgeois, unsoulful Ealing), and images of of this study reflects the ideas of bourgeois Western culture, in which the

This is reinforced in its collocates (society, culture, family/families, life, values; revolution, ideology, democracy, liberalism, hegemony, class), which in turn suggest discursive associations to be examined. Stubbs (1996: 157–95; 2001: 145–93) has extended discussion of culturally loaded items and keywords; Krishnamurthy (1996) looks at racial, tribal and ethnic; Baker (2006) provides many further case studies; see also O’Halloran (this volume).

4. Sets and synonyms

Lexical sets

Words fall into lexical sets, fitting into semantic fields, though these are not always easily identified from corpus data. However, where a corpus/subcorpus is limited to a particular field, especially if technical, we can learn something about the lexis with which that field is discussed. For example, in the BoE subcorpus of business-focused texts, almost 40 per cent of its top 100 words are lexical and topic-specific (firm(s), shares, securities, investment, company, market, business, income, value, sales …). Similarly, collocate listings for words may reveal co-hyponyms: those for aphid include a range of insects and other small creatures, often sharing a semantic feature ‘parasitic’ (ants, blackfly, greenfly, caterpillars, lacewings, meatybugs, moths, slugs, snails, spiders, thrips, weevils …). Another approach to exploring sets is through phraseological frames, looking at what kinds of word occur within a particular slot. For example, the noun slot in the phraseology (QUANTIFIER) + NOUN + ago is realised by not just the obvious set of time words (years, months, weeks, days, seasons …) but also items considered as periods of time (generations, games, moons, vintages, albums …). Cf. the way that comply with (see above) is typically followed by collocates which indicate a constraint: another kind of semantic set.

Synonyms

It is often said that English has no perfect synonyms: that is, words which can be used interchangeably in any context. Corpora make it possible to test this by examining collocation, phraseological structure, genre, variety and frequency. Asylum seems to occupy a similar semantic space, ‘place of safety’, to refuge (roughly twice as frequent: see Section 2, ‘Collocation and patterning’, for concordances):
of course, extended to every case. Asylum was never carried to the point of
black economy by illegally employing asylum seekers. There, not too
be reviewed sympathetically. If asylum is refused to this group, the white
However, although the numbers of asylum seekers coming to the UK had risen
caused by the growing number of asylum-seekers in Calais. Among the
the writer who sought political asylum in Paris earlier this week. Dr
A CHINESE doctor seeking political asylum in the United States said that he
last March after being refused asylum in Italy and France. Choice
and she has married since seeking asylum. In the meantime she is happy to
of visitor’s refusals under the Asylum and Immigration Appeals Act (1993)

But even a few lines show differences: seek* seems to collocate with both, but take* only
with refuge; asylum is more politised. Another pairing is colourless and drab:

Mulcahy and starring a rather colourless Alec Baldwin in the title role of
Then the cast seems a trifle colourless. And finally the script does lie
silicate. X-ray analysis of the colourless crystals reveals five oxygen
and Hermann-Otto Solms, a colourless economist. All these hopefuls
on the outside with something colourless? It can be exasperating. ‘As a
all bleached glare and a strangely colorless ocean. Air temperature is forty-
You’ve been very kind,’ but this colourless understatement made me despair of
to gas can’t it? And if the gas is colourless, you could say it disappears, but
scratch. The downtown I remember - drab, Calvinistic, with white men in dark
His eyes roamed anxiously around the drab conference room, and he forced his
for exotic ingredients during the drab days of rationing - you were lucky
housewives in the study led a pretty drab existence. On the average, we found
he walks past it along the rather drab green corridor. But after enjoying
drive home the message of a typically drab memo, chart or graph. Color can
cutting a swathe through the grey, drab ordinariness of contemporary music
a gash of brilliant colour in our drab surroundings, was our school. There,

Drab, roughly twice as common as colourless, here refers to appearances, experiences and
situations; colourless means ‘lacking pigment, colour’ or is applied metaphorically to
uninteresting performances and people. While drab/colourless (and dull/dreary/lacklustre) are
only partial synonyms, there are many more problematic pairs like wide/broad or begin/start
where corpora can similarly help disentangle their different usages. See Partington

Antonyms and opposites

Conventional antonyms can also be explored with corpus data, in particular to establish
whether they share ranges of reference and phraseological patterns. Are, for example,
morphological counterparts such as acceptable/unacceptable, important/unimportant and col-
ourful/colourless also counterparts in usage? What about happy with its two antonyms,
unhappy and sad? A third choice, not happy, may need to be considered too:

<p> Jones said: He probably wasn’t happy because he wasn’t in the squad and
make the home happy,’ yet she is not happy playing that role. <p> Arlie
getting on with my life. They weren’t happy that I hadn’t returned to school. I
crossin after he said he was not happy with the environmental impact
force. Many of these men are not happy working alongside the American

Of other types of oppositeness, one of the most obvious to investigate is gendered
pairings, such as man/woman, boy/girl, husband/wife, where collocation often shows up
gender stereotyping. For example, looking in the BoE books subcorpora at the pattern
ADJECTIVE + husband/wife, it appears that collocationally a husband is abusive, unfaithful,
wayward, hardworking, drunk, a wife good, perfect, battered, pregnant, beautiful... there are
interesting implications ideologically. For studies, see those by Pearce of man/woman

5. Lexis in spoken language

This final section looks briefly at lexis in the most distinctive of all discourse subtypes,
spoken interaction, here drawing on BoE’s twenty-million-word subcorpus of British
conversation and local radio broadcasts. One significant feature is its lexicon: smaller and
more homogenised, with fewer types, just under 47,000 non-hapaxes. Correspondingly,
its top 100 and 1,000 types comprise a larger proportion of the corpus, or to put it
another way, are used proportionately more. Lexical items in its top 100 include mean,
sort, thing, want, all more common than in written English; also discourse markers and
phatics yeah, yes, right, well, okay, er, ern, mm, oh. Many are strongly patterned phraseo-
logically, all have important pragmatic functions. See McCarthy and Carter (1997) for an
overview of spoken lexis; O’Keeffe et al. (2007: passim) for extended discussion.

Phraseology

The following, extracted from BoE, shows some of the features of spoken lexis:

AA ... and good shoes ... and ... and buy the things to me that are ern practical
they need ... th that ... that something that might need to be in use constantly
every day and therefore has to be you know a good thing. Erm and not to buy it
on silly frivol you know ... d+ ... don’t spend your money on silly frivolities and
BB Mm
AA the things that you don’t need. I mean it’s okay to do that providing that
you’ve already taken care of all the other needs.

Features include hesitations, repetitions and relexicalisations; the chunked nature of the
language is also clear. Among more lexical words, know and mean occur in phraseologies with
discourse functions; others have fairly general meanings (good, thing/things, need/needs) or
occur in common collocations or phrases (in use, every day, spend money, take care). Only
shoes, buy, practical, constantly, silly, frivalities have fuller meanings or seem more independent.
Predictably, if we compare data drawn from written and spoken subcorpora for a word
like know, we find phraseological differences. From British broadsheet journalism:

the story first.’ But she must have known, as a scandal-peddler herself, how
</date> It was the stare of a man who knew he was going to get his way. We’d
cover their tracks these days. I know his wife by sight, though not to
food because, though we think we know how to cook, we do not know how to
33 sites have been identified. Nobody knows how many people were killed in the
as previous Celtic players have been known to do, will now come ready salted.
moor. This painting will become known to the rest of the group as Mike’s
not the end of the world’, and we knew we really were in trouble. I have

and from the spoken subcorpus, where 94 per cent of occurrences are as the base form know:

models saying well we don’t really know < ZG1 > < M02 > Yeah. < ZF1 > < ZF0 >
know. They don’t talk actually you know < tc text = laughs > < tc text = pause >
< ZF1 > I do + erm I < ZF0 > I don’t know how you would whether you would say
They know the basics of a bike you know how to balance them < M02 > Yeah.
we were < ZF0 > we were chasing you know maybe fighting the Belgians or the
system installed < ZG1 > now < ZG0 > You know one of the things we were talking
say it’s a virus’ cos they don’t know what else it is. < F03 > Yeah. < F01 >
Yeah that’s what you were asking me I knew you had something to get back to er

Even in terms of collocates, there can be marked distinctions between speech and writing. If we take a structure such as ADVERB + happy to explore the ranges of sub-modifiers used, and compare realisations in BoE generally with those in the spoken subcorpus, there are interesting contrasts. Items such as very, quite, perfectly, really, reasonably are found in both subcorpora; in BoE overall, further significant collocates include blissfully, deliriously, ecstatically, fantastically, gloriously, idyllically, infectiously, insanely, irrationally, radiantly, serenely, supremely, unspeakably, wonderfully, etc. But the only such item to occur significantly in the spoken subcorpus is deliriously, suggesting a simpler vocabulary, characteristic of the spoken lexicon.

**Meaning and usage**

Perhaps the most important and complex aspect of lexis in spoken language is the way in which words and phraseologies fulfil pragmatic functions: meaning effectively has to be explored in terms of function, as in the case of know. While I know, I don’t know, you know are not semantically opaque, they contribute interactionally, indicating hesitation and uncertainty, appealing to shared knowledge or understanding, pre-empting contradiction. With thing, corpus data shows its functions as a proform or vagueness marker, or, in the formula the NOUN is, prefacing a reason, point or new information:

< M01 > something some development thing that will indicate they’ve er < ZG1 > and FX think October would be a good thing you know because FX said it sets d finished with it. But the only good thing about it was that erm it was < ZF1 > lectures say it is the most important thing in the whole lecture course I think and the customer and that sort of thing. < FOX > Right. < M02 > How you’re < ZF1 > the o+ < ZF0 > the only thing I would say < tc text = pause > that and the line-to-line movement of the thing. Erm there’s a much more er basic it’s a very good question. I mean the thing is if you were to kill er the

I+ on the er < F01 > Er thermostat thing? < F03 > Yeah. < F01 > Fan. < FOX > Mm. bit older and what’s been the worse thing about getting a little bit older? I
For words and phrases like these, corpora provide a means for exploring usage, co-textual patterning and positioning, and so on. Less directly, concordances of such words may suggest other items or phenomena to explore: patterns of repetition and relexicalisation, particular phraseologies, and other narrative devices, as in:

in those days < ZF1 > it < ZF0 > it was very very hard er for the family like < F01 > Mn. < HOI > They were really very very hot on quality. < F01 > Mn. Mn.
Yes, < F02 > But her he’s very nice very very nice and I like him very much.
words. < HOX > Thirty million a year is very very roughly a hundred thousand a
and suited well for what she does very very well which is pantomime. Erm
out in practice they saw it as a good thing and I’m sure that’s < ZF1 > will
I think it’s going to be a good thing for Birmingham. It’s going to be a
that. And I think that’s er a good thing. It sort of challenges them and
on the estate as well which is a good thing. Erm so t so < ZF1 > th < ZF0 > there
Is he a fool is he wise is it a good thing? And if you want to talk about

Discourse pragmatics can only be studied properly with full transcripts and intonational data, yet there is still much to learn about the spoken lexicon from corpora. It is also salutary to remember that assertions made about lexis in homogenised general corpora provide only part of the truth: there may be extraordinary dissimilarities within corpora, alongside the extraordinary similarities of lexical and phraseological patterning, and patterns of conventionalised usage.

Further reading


Stubbs, M. (2001) Words and Phrases: Corpus Studies of Lexical Semantics. Oxford: Blackwell. (This includes a stimulating array of corpus studies of words and phrases, along with discussions of the relevant linguistic issues.)

References


210