The Routledge Handbook of Semantics

Nick Riemer

Corpus semantics

Publication details
Michael Stubbs
Published online on: 17 Aug 2015

How to cite: Michael Stubbs. 17 Aug 2015, Corpus semantics from: The Routledge Handbook of Semantics Routledge
Accessed on: 26 Oct 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
The term “corpus semantics” is a shorthand way of referring to a combination of data, methods and theory: empirical observational evidence from large computer-readable corpora is used to formulate hypotheses about meaning. The data-intensive methods depend on the ability to store and search large text collections for lexical and grammatical patterns. One of the most significant findings has been the pervasiveness, in all kinds of texts, of predictable collocations, and a major theoretical puzzle has been how best to model multi-word units of meaning.

Computer technology was used in the 1960s and 1970s to do preliminary work on word frequency and on collocations (Kučera and Francis 1967; Sinclair et al. 2004 [1970]; Allén 1975). However, it was only by the mid-1980s that computers became powerful enough to make realistic contributions to the construction of the first corpus-based dictionary (Cobuild 1987), and only from around the mid-1990s that the widespread availability of large corpora and user-friendly software made the data and methods available to scholars with no expertise in computing.

1 Data, definitions and initial examples

All unattributed examples of attested language use below are from the BNC (British National Corpus). This consists of 100 million words of contemporary English from over 4000 text samples, and can be accessed at http://www.natcorp.ox.ac.uk and via various user-interfaces (Hoffmann et al. 2008; Fletcher 2003). A clear advantage of empirical methods is that readers can check my findings in this corpus. Even better, of course, would be to check them in other independent corpora.

Double quotation marks are used for terms and quotations. Single quotation marks are used for meanings. Word-forms are in lower-case italic. LEMMAS and WORD FAMILIES are in upper-case italic. The term “lemma” is traditionally used for a class of a basic word-form plus its inflected forms which are all the same part of speech, e.g. the verb drink, drinks, drank, drunk, drinking. A “word family” (Bauer and Nation 1993) is a word plus both its inflected and derived forms, e.g. drink, drinker. But the form drink can be a verb or noun, and drunk can be a verb, adjective or noun. Empirical studies of corpus data have shown that different forms of a lemma often pattern in different ways from each other, and that members of a word family often pattern in a similar way.

Diamond brackets enclose collocates of a word-form or lemma, e.g. DRINK <beer, coffee, tea, wine etc.>. Square brackets, as in [A B], make a claim that this is a linguistic unit, not
Corpus semantics

merely a linear sequence of words. A slash, as in [A/B], indicates that the linear sequence is variable. For example, the following occurrences are evidence of a unit [HEAVY/DRINK] which is realized not only by variable forms of lemmas (as traditionally defined), but also by verbs and nouns (drank, drinker), and by adjectives and adverbs (heavy, heavily), in variable linear sequences and variable spans (in the sense of a window of orthographic words to the left and right).

she thought he drank too heavily
he was drinking heavily
he was a heavy drinker and smoker
both were heavy red-wine drinkers

This example confirms the view that multi-word units are “the biggest single problem that the lexicon presents for language analysis” (Hanks 2013: 61) and therefore constitute some of the most intractable problems in semantics: What are the basic units of meaning in a language? How can they be reliably and objectively identified and consistently and formally described? How are different units related? For example, are heavy drinker, heavy drug-user, heavy gambler and heavy smoker variants of a more general unit, in which heavy denotes ‘a lot’ and connotes the speaker’s disapproval of socially undesirable behaviour?

The combinations *heavily drunk and *heavy eater (in the sense of ‘glutton’) both sound odd to me, but both occur in many texts by native speakers in the Web. No two speakers use their language in exactly the same way, and a comparison of corpus-based dictionaries, prepared by experienced lexicographers, shows frequent disagreement about word meanings, especially connotations (Stubbs 2007). Since there is no sharp boundary between normal and unusual, the problem is to achieve the correct level of generalization about usage (Hanks 2013: 411), and a constant background question is whether a corpus can ever, strictly speaking, provide semantic data, since intuition is always required to interpret the data.

Kilgarriff (1997: 137) provides a good summary of the balance required. Semantic analysis can never be entirely objective, but corpora allow us to study language “with a degree of objectivity [. . .] where before we could only speculate”.

2 Historical perspectives

The layperson’s notion of basic units of meaning – at least for speakers of European languages – has long been the “word”. Dictionaries list words, and in several Germanic languages they are simply called “word-books” (e.g. German Wörterbuch). This looks like a case of speakers being misled about the general nature of language by the nature of their own writing systems. If we ignore complications such as hyphenated words, the writing systems of languages such as English, German and Russian signal word boundaries by spaces. Other languages do things differently: for example, in written Chinese each character represents a single syllable morpheme, but word boundaries are not marked.

In addition, dictionaries (for English, German, Russian, etc.) generally present language units (word-types) out of context in an alphabetic sequence that is irrelevant to both their semantic relations and their use in communication. This contributes to the odd but widespread notion that each word has a “literal meaning”; that is, a basic inherent meaning when it is not being used to communicate anything (Toolan 2008: 156).

A very different view, that the meaning of words depends on their verbal environment, was already clear, from the Middle Ages onwards, to the scholars who constructed
concordances. Ordering concordance lines alphabetically by given search words facilitates finding instances, but what is listed is not word-types, but word-tokens along with their co-text. Cruden (1737) defines the function of his famous biblical concordance thus:

A Concordance is a Dictionary, or an Index [. . .], wherein all the words used [. . .] are ranged alphabetically, and the various places where they occur are referred to, to assist us in [. . .] comparing the several significations of the same word.

It would be more helpful to distinguish between a concordance and a dictionary. A concordance sets out the data in a convenient form which allows patterns of co-occurrence to be seen. But this is only the evidence that must be interpreted by the dictionary-maker. A slogan often quoted in this connection, and usually attributed to much later work by Wittgenstein (1958, §43) and Austin (1962), is “meaning is use”. This is a convenient slogan, but a more precise formulation would be that patterns of use, especially patterns of recurrent collocation, provide empirical observational evidence of meaning.

The first scholar to use technology for constructing concordances was Roberto Busa. In 1948, in collaboration with IBM, he began converting the complete works of Thomas Aquinas into digital form, initially on 11 million punch cards. Twenty million lines of text were published as the Index Thomisticus, nowadays available in an online version with a powerful search engine. For many, his work in the 1940s defines the starting point of computer-assisted text analysis. Busa (2007) provides his own account of this work.

Ideas are often proposed independently by different researchers at approximately the same time. The idea that the distribution of a word – specifically its co-occurrence frequencies with typical collocates – can disambiguate word senses was put forward in various places in the 1950s and 1960s. In a paper on machine translation, Weaver (1955) used the term “statistical semantics” to mean that probabilistic patterns of language usage in large text collections can provide evidence of what words mean. This is similar to the well-known statement by Firth (1957a: 11) that “you shall know a word by the company it keeps” and that part of the meaning of a word is its “habitual collocation” with other words. Detailed examples were provided by McIntosh (1961), still several years before corpus resources became generally available. He predicts that nouns that follow the adjective molten will be similar in meaning, though it would be “laborious” to list them. With modern methods, of course, it is not at all laborious, and findings from the BNC confirm his predictions exactly. The top noun collocates, immediately to the right of molten, with a frequency of 5 or more, include superordinate metal along with several hyponyms:

metal 31, rock 26, lava 15, iron 14, lead 10, glass 9, material 9, gold 8

McIntosh further argues that we learn the meanings of words from our experience of them in different contexts, and this can also be illustrated from corpus data. The following occurrences of molten provide partial paraphrases of the word in the immediate co-text: something that is molten is “red hot”, can “flow”, can “cool” from its molten state, and so on.

– seem to glisten, as if they were red hot, molten
– placed in moulds for molten lead to flow around them
– the basalt cooled from its original molten state
– as the powder passes through the flame, particles can become molten
Finally, he points out that the co-text influences the meaning of a word: for example, *pig* in the phrase *molten pig iron*. McIntosh (1961: 330–31) thus sets out major principles that have been thoroughly documented by corpus study: “the lexical item and the word are not co-extensive” and the collocates of a word go “a long way towards constituting [its] meaning”.

Sinclair (1998: 8) makes more explicit the significance of the *pig iron* example via the descriptive category he calls “reversal”: “the precise meaning of a word is determined more by the verbal environment of a word [than by] the parameters of a lexical entry”. That is, most words are ambiguous in isolation, but their potential meanings are restricted – and often determined – by their phraseology. Sinclair’s example is that *white* in *white wine* denotes a colour specific to wine, often pale green, and not what *white* might seem to mean out of context (when it is not being used to communicate anything).

Computers do not, of course, understand meanings: all they can do is identify and count recurrent sequences of characters. This is, however, essential to linguistic analysis, and studying recurrent strings across large amounts of text (especially with the help of computer-generated concordance lines) has made visible new kinds of evidence of extended phrasal units. As Harris (1988: 19) puts it:

> [T]he very essence of language seems to depend on the possibility of regular recurrence of verbal items of various kinds. [...] Any general analysis of how language works is thus forced to tackle the notion of linguistic units.

This was explicitly recognized in early computer-assisted work by Allén et al. (1975: vol 3, xxxiii). In their investigation of a one-million-word Swedish newspaper corpus, the “first level of description” was “all collocations having at least two identical instances” in the corpus. Recurrence was “the methodological foundation of the investigation”. This characteristic of language was also widely recognized in the British Firthian tradition of functional linguistics which led directly to modern corpus linguistics. Halliday (1978: 4), Firth’s student, argues that “in real life, most sentences are not uttered for the first time” and that “a great deal of discourse is more or less routinized”. In the first computer-assisted study in this tradition, Sinclair – Firth’s student and Halliday’s colleague – studied the relation “between statistically defined units of lexis and postulated units of meaning” in a 135,000-word corpus of spoken texts (Sinclair et al. 2004 [1970]: 6).

Nevertheless, until the development of corpus methods, most linguists greatly underrated the role of recurrent phraseology in modelling language use. There have long been dictionaries of clichés, catch phrases and the like, which contain thousands of examples of idioms that are not semantically transparent (e.g. *axe to grind*, *bee’s knees*, *can of worms*, etc.). But the wider phraseological tendency of language in use, and the cultural functions of these phrasal units, was initially recognized by only a few scholars, such as Bolinger (1976) and Pawley and Syder (1983), who argue that native speakers know thousands of idiomatic units, which vary only slightly in form, and which are conventional labels for culturally recognized concepts.

Many ideas are described as new or ground-breaking, whereas they have long histories. It is often impossible to find the first mention of an important concept, and some of the basic ideas in this field were put forward over long periods of time, forgotten about and independently rediscovered, often much later. However, more important than mere mention is the development of a concept into a systematic research programme. It is this recognition – of the theoretical significance of phraseology – that has developed rapidly since the 1960s.
3 Current contributions and research

3.1 The discovery problem

Research on the phraseological tendency of language in use must be approached from two directions: software can automatically extract data from large corpora, but the linguist’s intuition is necessary to interpret these data. The following concepts have proved useful in providing initial evidence, as far as possible automatically, of phrasal units of meaning.

(1) “N-grams” are fixed sequences of orthographic word-forms. For example, the 4-grams at the end of and at the same time are frequent in the BNC, and are easy for software to identify, but they are only raw surface data. Another 4-gram is that there is a, but we probably would not want to claim linguistic status for it, even though it occurs frequently. The 4-gram [on the other hand] seems a much better candidate for a linguistic unit: it is also frequent, it is a fixed phrase, its meaning is non-compositional (i.e. not predictable from the meaning of the individual words), and it is listed separately in many dictionaries as a discourse marker. However, if we specify search parameters in advance, we may find only what we are looking for, since a recurrent n-gram may be only part of a longer more abstract unit. For example, in no uncertain terms is a 4-gram with rare internal variation, but it is usually preceded by the verb TELL, by one of its hyponyms (e.g. EXPLAIN, REPLY, URGE), or by a roughly synonymous word or phrase:

were told in no uncertain terms
informed them in no uncertain terms
put him right in no uncertain terms
were shown the door in no uncertain terms

N-gram data are useful, but have to be carefully interpreted.

(2) “Skip-grams” are n-grams such as A x B or A B x C, where x = one or more variable word-forms. (Cf. Fletcher (2003) on “phrase-frames” and Renouf and Sinclair (1991) on “collocational frameworks”.) For example, at the x end of is a 5-skip-gram, whose variants include at the far end of and at the top end of. But again, if we search for skip-grams of a fixed length, we will miss examples such as at the very cheap end of the market; at the as yet unfashionable end of the terrace. These seem – intuitively – to be variants of the same phrasal unit, and illustrate a very general problem. There is a difference between recurrent word strings and what we would intuitively regard as more abstract semantic units. Also, a x of is very frequent. Frequent realizations include: a number of; a series of; a great deal of; a wide range of; a large number of. This raises the interpretative problem of how to make valid semantic generalizations about the variable words. Many involve measurements, but not all (e.g. a result of; a matter of; a kind of).

(3) “Concgrams” (Cheng et al. 2006; Greaves 2009) are units that allow three different parameters of variation in the recurrent association of two or more words: their realization (different forms of a lemma or word family), their linear sequence and their span (e.g. A B, B A, A x B, A x y B, A B C, B C A, etc.). Cheng et al. (2009) study variants and their frequencies of the two-word concgram [PLAY/ROLE]. The BNC has over 3,800 occurrences of [PLAY/ROLE] in a span of 5, of which around 12 per cent are [PLAY/ROLE/important], for example:
will play a crucially important role
playing an active and important role
may play important and unanticipated roles
have an important role to play
one important role which typifications play
the important and unacknowledged role which women had to play

Other approximate synonyms of important occur (e.g. crucial, major, key, leading, vital), and it is relatively easy in this case to agree on ‘important’ as a superordinate semantic label. Variants such as play a minor role occur much less frequently, and are often signalled to be against expectation. However, in addition, [PLAY/PART/important] also occurs over 400 times, with many of the same adjectives. The occurrence of ROLE and PART in the same context is distributional evidence that they are related in meaning. But this leaves the problem of how best to summarize the main patterns without being swamped by details. We have to decide whether [PLAY/ROLE/important], [PLAY/PART/crucial] etc. are all variants of a more abstract pattern.

(4) “PoS-grams” (Fletcher 2003) are sequences of parts of speech. This concept allows us to identify frequent ADJ-NOUN pairs (e.g. long time), NOUN-NOUN pairs (e.g. world war), ADV-ADJ pairs (e.g. most important), etc. These patterns lie outside what are usually regarded as “collocations” in the sense of frequent lexical co-occurrences, but PoS-grams are realized by n-grams. For example, the top 4-PoS-gram in the BNC is:

preposition + determiner + singular noun + of

Its top realizations are: at the end of, as a result of, in the case of.

Clearly PoS-grams can be identified only in a grammatically tagged corpus where we can look only for sequences defined by the tag-set used. Sinclair (2004: 190–91) is therefore sceptical of such annotation and favours a “plain text policy” which avoids the potential circularity of tagging the corpus (possibly using pre-corpus assumptions) and then searching for the tags. This might well reveal recurrent sequences of old categories, but will not reveal any new categories.

N-grams and similar linear strings can be found rapidly and accurately with software, but – as illustrated above – they are often realizations of more abstract patterns which can be recognized only via the linguist’s intuitions.

(5) “Collocations” are defined in many ways and are only recognizable to software if operationalized. Firth (1957a) seems equivocal as to whether collocation is an observable linear sequence of orthographic words (“quite simply the mere word accompaniment”, p. 11), or an abstract psychological relation of order (“mutual expectancy”, p. 12). It seems most obvious to interpret his famous definition (Firth 1957b: 196) as referring to word families:

Meaning by collocation is an abstraction at the syntagmatic level […]. One of the meanings of night is its collocability with dark, and of dark, of course, collocation with night.

But he does not make explicit the possible variability in sequence, span and word-form, which – as I have shown above – can be documented with data on coneprams:
one dark and stormy night
on the darkest of nights
belonged to darkness and the night
the dull day was darkening into a cold night

night darkened the cup of the valley
the night-darkness of his hair and eyes
through the lonely night in pitch darkness
at nightfall just as the sky darkened

Nor does Firth recognize that different forms of a lemma often pattern differently: see Sinclair (1991: 53–64, 154) on YIELD, and Stubbs (2001: 28) on SEEK.

Second, collocations can be semantically transparent or not. Palmer’s (1933) famous definition seems ambiguous in this respect: “a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts”. Sinclair therefore argues that words do not merely co-occur, but that they are co-selected: a single choice of two or more apparently independent words can “constitute a single choice” (Sinclair 1991: 112), and therefore create a new meaning. This is the main evidence that they are components of the same lexical item. A widely cited example is naked eye (Sinclair 1996) which is hardly ever used in the sense of ‘unprotected eye’, but has a non-compositional meaning, and is used when talking of something which is ‘difficult to see without a telescope/microscope etc’, in uses such as:

– just about visible with the naked eye and easily seen with binoculars
– because they are so faint not a single one is visible to the naked eye

Corpus study has shown that the Fregean concept of compositionality applies much less widely than is often assumed. The most frequent noun-noun combinations in the BNC include the following, none of which are entirely semantically transparent: world war, interest rates, world cup, health service, trade union. For example, world cup means ‘an international football (soccer) championship that takes place every four years between men’s teams from different countries’.

(6) The term “collostruction” denotes a unit identified by measuring the attraction or repulsion of words to slots in grammatical constructions (Stefanowitsch and Gries 2003; Gries and Stefanowitsch 2004). This documents one of the essential claims of corpus analysis, that it is impossible to separate lexis and syntax, since different meanings of words are associated with distinctive formal patternings.

In a quantitative study of collocations (Stubbs 1995), I showed that the lemma CAUSE is used overwhelmingly to talk about undesirable states of affairs. Characteristic collocations, averaged across several independent corpora, included:

CAUSE <problems, death, trouble, damage, cancer, difficulties>

With data mainly from small corpora of only 3.3 million words, I used two simple statistical tests to check probabilities of co-occurrence of word-forms. Bambrook et al. (2013: 156–61) replicate my analysis on a corpus of 450 million words, and corroborate my basic findings. However, Klotz (1997) points out that purely lexical data miss a generalization about the relation between lexis, syntax and semantics. We can say: X caused problems. X caused
Corpus semantics

cancer. X caused him problems. But not: *X caused him cancer. That is, lexis interacts with grammar: words for illnesses and diseases do not occur in the ditransitive construction. This point is further refined by Stefanowitsch and Gries (2003) who use statistical methods to distinguish between three syntactic structures in which CAUSE occurs:

- transitive: \( X \) caused problems
- prepositional dative: \( X \) caused harm to \( Y \)
- ditransitive: \( X \) caused \( Y \) distress

They argue (p. 222) that

The transitive construction occurs exclusively, and the prepositional dative predominantly, with external states and events; in contrast, the ditransitive construction encodes predominantly internal (mental) states and experiences.

Hunston (2007: 251–53) agrees that CAUSE occurs with negatively evaluated situations when human self-interest is involved, but points out that the expression of speaker attitude interacts with text-type, and that there is often no negative evaluation in scientific and technical texts (e.g. The effect of the tides is like a brake, causing the spin of the earth to slow). Finally, Smith and Nordquist (2012) compare Early Modern and Present-day English and show that the negative semantic associations of CAUSE have developed over time.

This shows the power of empirical methods. An initial study is checked against a much larger independent corpus, missing details are filled in, partial counter-examples are discovered and explained, and more refined methods are developed.

3.2 Sinclair’s model

Underlying any theory of language are assumptions about social interaction, but much linguistic description ignores both topic (what speakers are talking about) and communicative function (why they are talking about it). Harris (2003: 166) refers to Chomsky’s “perverse rejection of the notion that languages are forms of communicational activity identified with particular communities”.

One of the best worked-out models of multi-word semantic units, which attempts to relate language form and communicative function, is proposed by Sinclair (1996, 1998). His model consists of an obligatory core, plus four parameters. Collocation and colligation define the form of the unit. Semantic preference defines its propositional relation to its co-text, and semantic prosody defines its communicative function (the reason for phrasing something in this way at this point in the text). In more detail:

1. Collocation is a relation between co-occurring word-forms. It can be identified on largely objective observable evidence of concrete word-tokens, although it is often useful to group word-forms into abstract lemmas and word families.
2. Colligation is a relation between words and co-occurring abstract grammatical classes (e.g. passive or negative) which it is often possible to identify automatically. However, grammar shades into semantics. For example, both grammatically
negative and positive clauses can express the same idea (he would not move versus he refused to move).

(3) Semantic preference is a relation between co-ordinated topical choices in text. It concerns propositional content, and can be identified by co-occurring lexis, but only if the analyst understands what the text is about (its topic) and has an intuitive knowledge of semantic fields.

(4) Semantic prosody expresses the speaker’s motivation or communicative purpose. It is similar to illocutionary force. It can be identified only if the analyst has an intuitive understanding of speakers’ evaluations and attitudes. Here semantics shades into pragmatics.

As we move from (1) to (4), we move from features that are objectively observable to features that require the analyst’s subjective interpretation: from (mere) surface sequence to underlying (theoretical) order.

For various reasons these relations were not previously recognized. The units do not refer to independent things or events in the external world. Their realizations are typically variable and discontinuous. Their constituents have different strengths of internal attraction. The semantic prosodies often have no traditional labels.

Sinclair’s model can be summarized even more simply as follows:

<table>
<thead>
<tr>
<th>Form</th>
<th>Content</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>collocation and colligation:</td>
<td>strings of words/grammar</td>
<td>FORM</td>
</tr>
<tr>
<td>semantic preference:</td>
<td>reference and predication</td>
<td>CONTENT</td>
</tr>
<tr>
<td>semantic prosody:</td>
<td>purpose, speaker intention</td>
<td>FUNCTION</td>
</tr>
</tbody>
</table>

This now looks very similar to the original definition by Searle (1969: 23–24) of a speech act, which has an internal structure of acts of different kinds:

<table>
<thead>
<tr>
<th>Form</th>
<th>Content</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>utterance act:</td>
<td>mere strings of words</td>
<td>FORM</td>
</tr>
<tr>
<td>propositional act(s):</td>
<td>reference and predication</td>
<td>CONTENT</td>
</tr>
<tr>
<td>illocutionary act(s):</td>
<td>purpose, speaker intention</td>
<td>FUNCTION</td>
</tr>
</tbody>
</table>

There are also similarities with models proposed in construction grammar, though Sinclair and Muraenen (2006: 31) regard the differences as substantial since construction grammar takes little account of the linearity of text.

Sinclair’s model can compactly describe empirical data. Here is an example (from Atkins and Rundell 2008: 376), which illustrates recurrent lexical patterns around the core CONSIGN... to ...

will be consigned to the dustbin of history
have been consigned to the rubbish heap
should be consigned to the scrap heap of history
was soon consigned to the wannabes pile
have been consigned to oblivion
will be consigned straight to hell
he was now consigned to outer darkness
has been consigned to the dustbin as a forgotten tradition
simply consigning him to a museum of dead ideas
would consign 100,000 workers to the dole queues
Corpus semantics

(1) Collocation. Frequent collocates are: dustbin, history, oblivion, rubbish, scrap.

(2) Colligation. The verb is usually passive. It is not always easy to distinguish sharply between propositional content and evaluative attitude (and perhaps especially here since the most frequent collocates have unpleasant denotations), or to find appropriate labels for attitudinal meanings. However:

(3) Semantic preference. The meaning is occasionally neutral: something is moved from one place to another (suitcases were consigned to his office for safe-keeping). But most often by far, the change is for the worse. Something no longer wanted is being disposed of, or someone is being marginalized or condemned to an unpleasant fate, usually permanently (Rundell 2002). The set of collocates is open-ended, but most denote ‘the past’ (e.g. history books, museum), or ‘rubbish’ (e.g. wannabes pile, scrap heap), or some very unpleasant place (e.g. hell, perdition, outer darkness).

(4) Semantic prosody. In addition, the speaker is often conveying a critical attitude towards this move and/or sympathy for who/whatever is to be disposed of and forgotten about. If it is reported that old people are “consigned to institutional care” the implication is that this is not their choice.

Out of context, some uses seem positive (Isabella was consigned to comfortable retirement). But the wider co-text reveals that Isabella is a queen who has been arrested, and the next sentence reports that Mortimer, arrested at the same time, was executed (BNC text F9L, history book). Any analytic method displays some things and hides others. A concordance is a visualization technique: it rips texts into fragments, and presents them in ways that show otherwise invisible patterns. This is its purpose: to show things from a different point of view. However, leaving concordance lines as isolated fragments divorces them from their communicative function in the original texts, and can imply that language consists of independent formal structures.

Through their linear syntagmatic organization, extended lexical units contribute to textual cohesion. They integrate lexical, syntactic and textual analysis (Sinclair and Mauranen 2006: 155). Therefore, if we combine quantitative and qualitative methods, and study how the lexical item [CONSIGN x to ‘an unpleasant fate’] is used in texts, we see that it often emphasizes a textual contrast between an earlier desirable state and a later undesirable state. In the following examples there are contrasts between achievements/oblivion, life/death, along with other markers of textual contrast (e.g. if... then...; in that case... rather than...).

If the achievements of the Thatcher years were not to be consigned to oblivion, then a tactical retreat was necessary. (BNC text AHN, broadsheet newspaper.)

In that case, the electorate might well prefer to be Red and live to fight another day rather than be consigned to a nuclear crematorium. (BNC text ABA, book on world affairs.)

Consigning something/someone to the dustbin/to oblivion does not have a purely discourse-external reference. It has reality as a social construct: the kind of event that has been talked about frequently in the past (all good things come to an end?). Francis (1993: 141, 155) emphasizes the cultural functions of such units, and thereby tackles the complex web of intentions, beliefs and social values which characterize language use:

[W]e can compile a grammar of the typical meanings that human communication encodes [ ... ] the ways in which we typically evaluate situations [ ... ] how difficult or easy life is made for us, how predictable things are, and how well we understand what is going on.
Michael Stubbs

Extended lexical units are ideal types: coherent gestalts with a formal structure and logical relations between the parts. Their function is textual (they make texts hang together), intertextual (they create discourse objects) and social (they encode cultural schemas).

More generally, Sinclair’s model operationalizes Morris’s (1938) classic definition of the relations between syntax, semantics and pragmatics. Collocation and colligation relate signs to other signs: this is the traditional definition of syntax. Semantic preference relates signs to the world: this is the traditional definition of semantics. Semantic prosody relates signs to speakers: this is the traditional definition of pragmatics. The model therefore meets a major criterion of the significance of an idea, by relating things that were previously often not clearly related, namely lexis, syntax, semantics and pragmatics.

The model is based on a large amount of data, but on close study of only a small number of cases, perhaps a few dozen. Some of the more widely cited case studies include Louw (1993), Stubbs (1995), Channell (2000), Partington (2004) and Hunston (2007). It is not yet clear how many such items there are – presumably there are thousands – how variable they can be, and how the model can be tested. It is easy to obtain confirming instances, but unclear what would constitute a counter-example, or even whether the concept of counter-example is fully applicable when the analysis is based on probabilities.

3.3 More on semantic prosodies

The feature of Sinclair’s model that immediately attracted attention was “semantic prosody”. The first use of the term in print was by Louw (1993), following the first example of a prosody around the phrasal verb SET in by Sinclair (1991: 74), who comments:

The most striking feature of this phrasal verb is the nature of subjects. In general, they refer to unpleasant states of affairs. [A few] refer to the weather; a few are neutral, such as reaction and trend. The main vocabulary is rot, decay, malaise, ill-will, decadence, impoverishment, infection, prejudice, vicious (circle), rigor mortis, numbness, bitterness, mannerism, anticlimax, anarchy, disillusion, disillusionment, slump. [...] The subjects of set in are also [...] largely abstractions: several are nominalizations of another part of speech.

The term “prosody” derives from Firth’s work on phonology, where he argued that a single phonetic feature can stretch across several units, as traditionally defined. For example, since the vowel in the word man occurs between two nasals, it will itself be nasalized. In the words, limp, lint and link, the final consonant clusters consist of two bilabials, two alveolars or two velars. English has no words such as *limt or *linp. It is therefore misleading to regard the consonant clusters as consisting of two phonemes: both are selected together as one unit. In analogy, “semantic prosody” refers to a single semantic feature instantiated across several words which are simultaneously co-selected.

It might seem more helpful to label this “pragmatic prosody” or “discourse prosody”, since it has communicative and text cohesive functions. However, Sinclair’s term “semantic prosody” emphasizes that the evidence is not inferred from non-linguistic knowledge about the social context, but is observed in the co-text. Other literature argues that the term is not clearly defined, and tries to distinguish more clearly between “semantic prosody” and “connotation” and/or between “semantic prosody” and “semantic preference” (Morley and Partington 2009). Hunston (2007) distinguishes between the implied attitudinal meaning of a word and Sinclair’s own use of “semantic prosody” to refer to the discourse function of a unit.
of meaning. Stewart (2010) provides a book-length attempt to define the term, and a comprehensive list of previous work, but discusses “semantic prosody” independently of the model of which it is only one parameter, and reaches no clear conclusions about a better definition.

Debates about terminology and definitions risk missing the essential point that the most important thing is empirical description, and that much previous work neglected the fact that co-occurrence phenomena, and the resulting evaluative prosodies of extended lexical units, are much more extensive than previously realized.

### 3.4 The vexed question of statistics

Raw frequencies alone cannot identify semantic units. For example, if a word is itself infrequent, then clearly the collocations in which it occurs cannot be frequent, and would be missed if raw frequency was our only search criterion. The lemma *QUESTION* occurs over 42,000 times in the BNC. The word *vexed* occurs only 165 times, but when it does occur, it has nearly a one-in-three chance of occurring in the 2-gram *vexed question(s)*. Less frequent 2-grams include the approximately synonymous *vexed problem* and *vexed topic*. That is, the attraction is asymmetrical: *vexed* predicts *QUESTION* much more strongly than *QUESTION* predicts *vexed*. However, for several reasons, I am more interested in what corpus data can tell us about the nature of extended units of meaning than with statistical methods.

First, many different statistics have been used to measure the strength of association between co-occurring word-forms (Clear 1993, Stubbs 1995, Evert 2008), and there is considerable disagreement about which statistical methods are appropriate. The log-likelihood test is widely regarded as more reliable than some other tests, since it takes into account the relative frequency of words: it ignores chance combinations of high-frequency words, and also combinations of rare words (Dunning 1993; Rayson 2012). However, like many other statistical measures of collocational strength it also fails to distinguish symmetrical and asymmetrical relations (Gries 2013).

Second, there is more general disagreement about whether such statistics are appropriate at all. At one end of the debate are scholars such as Gries (2010), who laments the lack of statistical sophistication amongst (many) corpus linguists. At the other end are those such as Sinclair (2008: 28), who argues that all the statistics used assume the possibility of randomness and, since this never holds for words in texts, statistics are largely irrelevant in the study of meaning. In words often attributed to Ernest Rutherford (1871–1937), “if your experiment needs statistics, you ought to have designed a better experiment”.

Third, whatever methods are used, the linguist still has to interpret the data generated. There is sometimes a suspicion that analysts try out different statistics, and then select the one which supports their intuitions (which is like shooting an arrow at the barn door and painting a bull’s eye around where the arrow lands).

### 4 Future directions

Corpus semantics is essentially diachronic and sociolinguistic (Teubert 2005), in so far as we analyze not what could hypothetically be said, but what frequently has been said in the past, by many different speakers in many different kinds of texts. The focus is on “normal, central, typical usage” (Hanks 2013: 7). The essential problem for analysis is variability. Many phrasal units have a well-defined core and canonical realizations, but fuzzy peripheral features and boundaries. In addition to the variables discussed above (frequency, linear sequence, span, word-form/lemma), further variables are only beginning to be studied and
pose problems of choosing the appropriate level of descriptive delicacy. Only a few representative studies are noted below.

Many years ago Firth (1935) showed that different units occur in different text-types. More recent studies (e.g. Gray and Biber 2013) compare n-grams in different written and spoken genres. Studies by Hoey (2005) and Mahlberg (2009) show that different phraseology tends to occur at different places in texts and thereby has text-organizing functions (which they call textual colligation). Change over time is documented in many corpus-based studies of grammaticalization (e.g. Lindquist and Mair 2004, Hoffmann 2005), which analyze the role of frequency in the emergence of phrasal units. The development of semantic prosodies over time is studied by Smith and Nordquist (2012). The extent to which semantic units and their prosodies might be universal or comparable across languages has only just begun to be studied, but see Lewandowska-Tomaszczyk (1996) and Xiao and McEnery (2006) on Polish and Chinese. The extent to which automatic recognition of phrasal units is possible is discussed by Sinclair et al. (1997) who have designed language-independent software.

The most extensive applications have been in lexicography. Since the late 1980s, the compilation and analysis of large corpora has had a major influence on methods used to compile English-language dictionaries and grammars, especially monolingual dictionaries for advanced learners. Many of the principles are set out by Sinclair (1987) and Hanks (2013). By 1995 other major British dictionary publishers (Chambers, Cambridge University Press, Longman, Macmillan and Oxford University Press) had followed Cobuild (1987) and were also using corpus data and methods.

Applications in other areas have been relatively modest. They include the following. Cotterill (2001) illustrates potential forensic applications by examining the semantic prosodies of words and phrases used to describe domestic violence in the O.J. Simpson trial. Dam-Jensen and Zethsen (2008) illustrate applications in the training of translators by showing how extended lexical units convey evaluations that are not accessible to introspection and which have not been part of traditional dictionary definitions.

5 Conclusions

As often happens in scientific areas, several principles that have turned out to be central to corpus semantics were proposed over a long period of time. Major ideas were proposed, then forgotten about, then rediscovered when circumstances changed – the invention of new technology was often crucial. Scholars from the Middle Ages onwards realized clearly the potential of language use for studying meaning. A series of other ideas then paved the way for rapid development from the 1980s onwards when new technology could help with visualizing patterns in large corpora. However, corpus evidence of meaning is indirect, and the linguist still needs traditional skills to see the wood for the trees in the vast amounts of data that technology makes available.

Further reading

Hanks, P. 2013 Lexical Analysis: Norms and Exploitations. Cambridge, Mass.: MIT Press. Dozens of examples from an experienced lexicographer of the implications of corpus analysis for both practice (dictionary making) and theory.

Corpus semantics

References


**Related topics**

Chapter 3, A history of semantics; Chapter 10, Semantics and pragmatics; Chapter 14, Sense relations; Chapter 15, Semantic shift; Chapter 24, Compositionality; Chapter 28, Interpretative semantics.