On inference theories and code theories: Corpus evidence for semantic schemas

MICHAEL STUBBS

Abstract

This article illustrates a method of studying the evaluative connotations of words and phrases, by studying their most frequent collocates in large corpora. The semantic schemas which can be identified in this way are clusters of preferred lexis and syntax, which often have conventional pragmatic connotations. Examples of individual schemas are given. A broader research project is then proposed to study the connotations of words and phrases in a significant area of social meaning: the many category terms in English for talking about groups of people in terms of the human life cycle (e.g., infant, baby, child, adolescent, teenager, youth, adult, senior citizen). Such analyses contribute to the long-standing debate between 'inference theories' and 'code theories' of language comprehension, and suggest that more is conventionally encoded in language structure than has often been suggested in recent work. In addition, they explore a central issue of corpus semantics: the relation between stability and variation in textual units.

Keywords: *schema; inference; encoding; connotation; corpus linguistics.*

This article illustrates a method of studying the evaluative connotations of words and phrases, and a way in which quantitative linguistic analysis of evaluative meanings can contribute to cultural studies. The article is in three main parts. First, I discuss the long-running question of how much meaning is encoded in a text versus how much we read into a text or infer from it. Second, I discuss some recent corpus-based work which provides evidence on how much pragmatic meaning is encoded in phraseological units. Third, I propose a research project which could use corpus methods to investigate a significant area of social meaning.

0165–4888/01/0021–0437 *Text* 21(3) (2001), pp. 437–465

Since I will be using corpus data to study words and phrases, I must start with a brief note on the sources of data and on my presentational conventions. The corpus data used in sections 2 and 3 amount to some 200 million running words. These are drawn from part of the Bank of English corpus held at Cobuild in Birmingham, from the British National Corpus, and from the *Times* and *Sunday Times* for 1995 on CD-ROM.[1] All examples of language use are attested in these sources, unless marked [I] for invented data. In presenting examples, upper case is used for lemmas, and italicized lower case for word-forms. For example, the lemma TAKE is realized by the word-forms *take*, *takes*, *taking*, *taken*, and *took*. Inverted commas are used for the meanings of linguistic expressions, and for quotes and technical terms.

1. Inference theories and code theories

Amongst the many dualisms which plague linguistics is the question of how much meaning is expressed in the text as opposed to how much is in the mind of the hearer or reader. How much do we get out of a text and how much do we read into it? How much is explicit and how much remains implicit? How much depends on linguistic knowledge and how much on encyclopedic knowledge? Traditionally, semantics has often been seen as a theory of those aspects of meaning which are expressed by sentences independent of context, whereas pragmatics is a theory of those aspects of meaning which are intended by speakers in context. However, it has also been questioned whether this distinction can be coherently maintained, and at the beginning of his textbook on pragmatics, Levinson (1983: 8) poses this as a serious problem for the field: 'aspects of linguistic structure sometimes directly encode (or otherwise interact with) features of the context'.[2]

Much work in recent semantics and pragmatics has therefore been a debate over the appropriate balance between 'code theories' and 'inference theories' of language comprehension.[3] I will illustrate a method of identifying the evaluative connotations of words and phrases, and argue that if attested examples of phraseological units are studied in large corpora, then this provides empirical evidence that pragmatic meanings are often conventionally encoded (in the text) rather than inferred (in the mind of the hearer/reader).

Clearly, there is a large inferential component in text comprehension, and it is inference theories which have recently been particularly influential. Particularly since the 1980s, but also in a wider tradition going back to the 1930s, a set of ideas about how we infer meaning from texts has had great intuitive appeal for scholars in psychology, sociology, artificial intelligence, literary theory, and linguistics.

A famous early version of this position was expressed very clearly in the 1920s and 1930s by Bartlett (1932) in his work on how people understand and remember narratives. Such work, within a tradition of gestalt psychology, showed that memory is an active process, which is influenced by what people are already familiar with, and expect to hear. After a period in which the idea was less prominent, it was increasingly used in an explosion of work from the 1970s onwards, in areas which had previously seemed often only tangentially related to each other. Much of the work was grouped under the label 'cognitive sciences', and it is this strand which is well known to linguists, especially through work on story grammars within artificial intelligence (e.g., by Abelson, Charniak, Minsky, Papert, and Schank). This work was particularly influential in showing just how much assumed knowledge is involved in story comprehension. Linguistic studies include work by van Dijk (1972), and by Hoey (1983) on problem-solution structures in stories.[4]

Such work often posits larger units of meaning, such as text macro-structures or story grammars. The hypothesis is thoroughly discussed in textbooks on the cognitive sciences (such as Johnson-Laird 1983), on discourse analysis (Brown and Yule 1983), and on literary theory (Cook 1994). The basic idea can be summarized as follows:

1. *Macro-schemas*. In order to understand language in context, we actively make sense of it, by using our knowledge of what is normal and what is to be expected. This knowledge is often represented in mental models or schemas. This view of comprehension therefore lays emphasis on two things which are logically related: the knowledge represented in such schemas, and the inferences which have to be carried out in context, in order to fill in what is not explicitly referred to (because it is assumed in the schemas).

Terms have proliferated. The term I will use, merely because it is probably the most frequent, is *schema* (plural *schemas* or *schemata*). But other terms, which often differ only in emphasis, include *frame*, *mental model, scenario* and *script*, and more general terms such as *data structure*, *knowledge structure* and *action stereotype*. (See Fillmore 1985: 223n for yet more terms.) Sometimes a single article uses different terms to make finer conceptual distinctions. However, all the variants share the notion that a schema is a structured mental representation of the typical features of a recurrent social event or situation. Such a schema contains taken-for-granted defaults: features which can be assumed to be present even if they have not been explicitly mentioned.

This definition implies both that a schema is something cognitive and individual, and also that it represents shared social knowledge.

This first model emphasizes the role of real world knowledge and inference, and correspondingly plays down the explicit coding of meanings in linguistic form. However, there is a second model, which—perhaps rather confusingly—also uses the term *schema*, but to refer to smaller units, roughly at the level of phrases. This work leads in a different direction, and proposes a different hypothesis, which I will summarize as follows:

2. *Phrasal schemas*. At this lower level of phraseology, pragmatic meanings are often conventionally encoded in lexicosyntactic form. The extent of nonlinguistic inferences has been exaggerated, and the extent of what is encoded has been correspondingly underestimated.

I will use the term *phrasal schemas* for this second case. However, the units are not fixed phrases, but abstract semantic schemas with variable lexis and syntax. For example, phrases such as *LIVE to a ripe old age* or *REACH a grand old age* not only denote that someone has lived to a relatively high age, say over eighty years. They also convey the cultural connotation that this is an achievement to be admired. (I discuss this example in detail later.)

The sets of work which are particularly relevant for model 2 are on case grammar, frame semantics, and construction grammar (by Fillmore, Kay, and others), on emergent grammar and emergent lexis (by Hopper, Bybee, and others), and on extended lexical units (by Sinclair, Moon, and others). Different scholars have arrived at comparable views, but terms and concepts are not standardized, and cross references within this work are at best unsystematic.

So, we have two areas of work, on larger and smaller units. They are closely related by their emphasis on the place of conceptual structures in language comprehension, but they differ in their views of how much meaning has to be inferred using general (nonlinguistic) reasoning, and how much meaning is conventionally encoded in linguistic form.

1.1. *The restaurant macroschema (and associated conventional phrases)*

One of the most frequently cited examples of model 1—itself a prototype of the main concept—is the restaurant schema (Schank and Abelson 1977: 40–46, 152–153, 222–225). This is an example of a larger schematic unit, which combines both an expected sequence of events and also expected phrases. Suppose someone says

(1) We found a rather good new restaurant in town last week. [I]

This will trigger a schema in hearers which allows them to take for granted several likely features of the event: that the restaurant contained tables and chairs, menus, waiters and other diners, and that there was a chronological sequence of events, with obligatory items such as ordering, eating, asking for the bill, and paying, and optional items such as ringing up beforehand to book a table, and asking for the wine list.

If we regard such schemas from a sociological point of view, we see that they help to maintain our sense of social reality, precisely because of what is left implicit. Sentences such as (1) imply a whole world, in which people go to restaurants to have an evening out, in which such places can be good or bad, and so on (Berger and Luckmann 1966: 172–173). If we regard schemas from a linguistic point of view, then they explain details of how we talk about such events (Schank and Abelson 1977: 40; Cook 1994: 13). For example, a first reference to a waiter can use a definite article (*The waiter was from France* [I]), even if no waiter has been previously mentioned. They also explain why some discourse seems coherent, even though there are no explicit linguistic markers of textual cohesion:

(2) A: We went to that new restaurant last night.
B: Was the food good? [I]

Such schemas are stable and widely recognized. Indeed, if it was possible to describe enough of them, then we would have a working definition of the culture, because we would have listed typical event and activity types, along with the taken-for-granted knowledge which people use to interpret norms and deviations from normal behavior. Other examples include going to school, going to the dentist, going on holiday, doing some gardening, doing some DIY, doing some shopping, and so on. Such schemas have more general and more specific forms (going on holiday versus going on a cheap last-minute package holiday to Spain). Similarly, the following (attested) utterances trigger more specific versions of the restaurant schema:

(3) a. a meal in a cheapish restaurant
b. he chose a Greek restaurant in Soho
c. an argument over the restaurant bill
d. he sallied forth to the restaurant-car in search of coffee.

Such schemas differ in their details across cultures, and expectations about going to a restaurant differ in Britain, Germany, and the USA (differences include whether you sit down or wait to be shown to a seat, how the menu is phrased, how tipping is done, and so on). The restaurant schema is well known to linguists, but a point which is perhaps less often made is that some of these cultural differences are observable in predictable and conventional phrases, which also differ across languages and cultures. Phrases used by waiters, and recognizable to any diner in Germany, include:

(4) a. *Kriegen Sie noch ein Bier?* 'Would you like another beer?'; but literally 'Are you getting another beer?', (to which I always want to reply, 'I hope so').
b. *Hat's geschmeckt?* 'Did you enjoy your meal?'; but literally 'Did it taste (good)?'.
c. *Zusammen?* 'Together?': i.e., 'Is one person paying the whole bill, or should I split it between the diners?'.

In this last case, not only is there no conventional phrase in English, but the German expression encodes an option (*zusammen oder getrennt?* 'together or separate?') which is standardly offered in German restaurants. Waiters in British restaurants don't like offering this: they provide a single bill and leave the diners to sort it out.

Mainstream linguistics has traditionally aimed at a clean division between linguistic and encyclopedic knowledge. However, some scholars have emphasized that work on schemas is at odds with this standard assumption. Nunberg and Zaenen (1992) discuss characteristic scenarios which allow the 'contextual filling-in of context', and argue that 'lexical content is integrated with knowledge representation in a broader sense'. Similarly, Hudson (1995: 35, 43) maintains that 'there are no standard arguments for the separateness of language' (from other social and cognitive activities), and argues that linguistic knowledge is 'part of a much more comprehensive knowledge structure'.

1.2. *Methods and theory*

These points about schemas have implications for both methods and theory. The concept of larger event structures plus associated conventional phrases suggests how systems of shared cultural meanings can be empirically investigated. Members of a culture interpret the world in (approximately) the same way. This cognitive claim cannot be directly studied, but the typical ways in which people talk about the world can be empirically observed, and recurrent ways of talking (*the restaurant car*, *a Greek restaurant*) provide one connection between language use and culture. In addition, two related dimensions of corpus semantics can help us out of a reliance on unobservable mental phenomena. Its methods are inherently quantitative and diachronic. Culturally significant phrases have a history of repetition (albeit with variation).

Therefore statistical methods can be used to estimate how frequently they occur, and how stable or variable they are.

In addition, there are implications for a theory of comprehension. First, communication would be impossible without the assumptions which are embodied in schemas. It is impossible to say everything: not only tedious, but impossible in principle (Garfinkel 1967). Second, such accounts explain how it is that more is communicated than is said. Meaning is underdetermined by linguistic form (Carston 1998), in the sense that hearers use their knowledge of schemas in order to make inferences about what speakers mean. What is said is merely a trigger: a linguistic fragment which allows hearers to infer a schema, which in turn provides default values which can lead to further inferences. So, it is not a mere convenience that we do not need to say explicitly what we mean, but a necessary characteristic of a functioning communication system (Levinson 2000: 27–30).

In summary, we are dealing with a series of theories which use concepts such as

i. schemas: mental representations of social norms
ii. the default values of these schemas
iii. the conventional ways we have of saying things
iv. the relations between linguistic and encyclopedic knowledge
v. the difference between what is said and what is meant
vi. the underdetermination of meaning by what is said
vii. culture as a system of (approximately) shared meanings.

2. Phrasal schemas

I now turn to the empirical part of the article, and illustrate how data from large corpora can be used to investigate the evaluative meanings of words and phrases. At the level of phraseology, it turns out that the balance between inference and decoding tips more in favor of decoding.

2.1. *Encoding pragmatic meaning*

The argument so far, with some caveats, has emphasized the inferences which hearers make in order to fill in taken-for-granted social knowledge. It might seem plausible that (core?) semantic meanings are encoded in linguistic form, whereas (noncore?) pragmatic meanings have to be inferred. However, as Levinson (1983: 8–9) points out, this distinction is impossible to maintain, since there are clear cases where pragmatic meaning is encoded in morphosyntax. In pairs of words such as *rabbit* and *bunny*, or *dog* and *doggie*, the first member of the pair is pragmatically neutral, but the second is pragmatically marked, in the sense of conveying information about speaker–hearer relations. Words such as *bunny* are restricted in speaker or addressee (e.g., used by or to children), and convey the speaker's evaluation of the referent (e.g., 'cuteness') or the context (e.g., informality). This distinction is encoded in the morphology quite independently of particular occasions of use. As I have shown in detail elsewhere (Stubbs 1996: 206–208), there is a large set of such words and word pairs with diminutive endings (-*y* or -*ie* in the spelling), where the diminutive conveys meanings such as childish, cute, feminine, informal and/or vague, and sometimes insulting:

(5) aunt, auntie; comfortable, comfy; nightgown, nightie; pup, puppy; stomach, tummy; Charles, Charlie; Jennifer, Jenny; etc.

Corpus studies have started to show that there are very many cases where pragmatic meanings are conventionally associated with a linguistic form. Here we come to the phrasal schema hypothesis. Theorists of phraseology have pointed out that idioms are often used to encode a 'recurrent situation of particular social interest' (Nunberg et al. 1994: 493). This point is developed by Moon (1998) in her corpus study of fixed expressions and idioms in English. She proposes 'idiom schemas', which she argues are compatible with the type of unit proposed in frame semantics (1998: 165). One of her many examples is represented by phrases such as the following:

(6) a. FAN the flames of [something]
    b. FUEL the flames of [something]
    c. ADD fuel to the flames
    d. ADD fuel to the fire

There is no single fixed phrase, but rather variations on a more abstract unit, which has preferred lexical realizations, but considerable lexical variation. The idiom schema is used to refer to a negatively evaluated situation, usually sociopolitical (e.g., *FAN/FUEL the flames of racism/extremism/discontent*). It refers to stereotypical situations, encapsulates shared experiences, and encodes ideological constructs. Moon's analysis is corroborated by examples which I have collected in other corpora, where prepositional objects of *FAN/FUEL the flames of* included

(7) anger, bigotry, despotism, evangelism, intolerance, rampant nationalism, prejudice, resentment, popular revolt, scandal, suspicion, vengeance.

Corpus data also show that, on its own, FUEL as a verb often has disapproving connotations. Corpora do show a few positive collocates in object NPs (e.g., *hopes*, *imagination*, *revival*), but the most frequent collocates (especially in media texts) are semantically related words such as *speculations* and *rumors*, and FUEL is frequently followed by object nouns such as

(8) addictions, allegations, anger, antagonism, anxieties, argument, conflict, controversy, discontent, fears, guilt, hatred, problem, resentment, tension, trouble, violence.

With FUEL as a noun, *ADD* is the most frequent verb to the left, but *GIVE* also occurs:

(9) a. added fuel to controversy
    b. giving some kind of fuel to this violence.

As a verb, FAN has some similar uses, but it is more frequently literal:

(10) panic fanned by hysteria; fanned by a light breeze.

This example illustrates the following points. The meaning of ‘social disapproval’ is associated with the *FAN/FUEL the flames* schema by convention: it is not derivable from the words by processes of composition or inference (for example, as a conversational implicature). The meaning is context-free: it does not depend on its place in a textual sequence. That is, the meaning of ‘social disapproval’ is not defeasible. Nondefeasibility is normally taken to be a criterion for semantic, as opposed to pragmatic, meaning, but here we have a case of an attitudinal meaning which is not contextually variable.[5]

The schema is abstract and semantic, not a fixed phrase, since there is no single lexical item which is obligatory. From the point of view of culture, the unit encodes shared sociocultural values, and locates a concept within an ideology (Moon 1998: 161–163, 256–257). These are data structures for representing stereotyped situations. From the point of view of linguistic structure, the unit of analysis has a stable semantic content with a conventional pragmatic meaning: this provides a functional definition. There is preferred lexis and (as I will later show) often preferred syntax: this provides a formal definition. Such variable schemas are psychologically plausible, since we generally remember the semantic content (e.g., of a story) and not its textual form.

The idiom schema *FAN/FUEL the flames* is always metaphorical and always disapproving, but it fits into a more general pattern (*the flames of*) which is usually metaphorical, but only sometimes disapproving.

Without FAN or FUEL, the phrase *the flames of* does have literal uses, though even this phrase is usually used metaphorically, often with evaluative connotations, and often in romantic fiction:

(11) a. sat staring into the flames of the fire
b. the flames of desire/passion/romance.

The *FAN/FUEL the flames* schema therefore shows several features of what Kay and Fillmore (1999: 4–5, 19) identify as 'constructions'.

1. It directly encodes a pragmatic meaning (the speaker's disapproval of the escalation of an unpleasant social situation).
2. It is (almost?) always metaphorical. Its diachronic origin is easily visible in an extension of a literal meaning of 'fan the flames of a fire'. That is, the metaphor is not dead, but the schema is no longer used literally.
3. There is a 'smooth interaction' (Kay and Fillmore 1999: 7) between the idiom schema and other phrases: between the metaphorical and disapproving *FAN/FUEL the flames of (violence, etc.)*, the usually disapproving *FUEL (fears, violence, etc.)*, the usually metaphorical *the flames of (love, etc.)*, and the literal *fan the smouldering fire*, *the wind fanned the flames*.
4. Because of its lexical variability, the schema cannot be given a fixed phrase structure, but requires a more abstract representation (Kay and Fillmore 1999: 19).

2.2. *The (inadequacy of the) compositionality principle*

The varieties of formal semantics which derive from Frege's work are based on the compositionality principle. This is the view that the meaning of a sentence (specifically its truth value) can be derived from the meaning of its constituent parts plus their position in syntactic structure. It is idioms which provide the most obvious counterexamples to the principle, since they are, by definition, units whose meaning is not computable from their constituent parts (as with *spill the beans*, *hit the hay*). This is despite the fact that it may be possible to translate the constituent parts of such idioms: *spill* = 'betray', *the beans* = 'the secrets'. With *FAN the flames of*, it is possible to find literal equivalents: 'ENCOURAGE the virulence of'. However, the conventional socio-political connotation of speaker disapproval is not exhausted by such equivalences.

The following examples also show that the meaning of phrases may not be reducible to the meaning of their constituents. Consider two (invented) sentences:

(12) We went to a restaurant today. [I]
(13) We went to a museum today. [I]

We would normally interpret (12) as 'We had a meal in a restaurant' (with all the default interpretations which that involves), whereas we would interpret (13) as 'We wandered around looking at things', etc. But we do not interpret (12) as 'We paid entrance to a restaurant, where we wandered around looking at things, bought a couple of postcards and a book on Egyptian art'. It therefore seems that the units relevant to semantic interpretation are not individual words, but larger units such as *GO to a restaurant* and *GO to a museum*. Furthermore, in neither of these cases would *GO to* normally be interpreted literally, as simply 'Go to the location mentioned', but as 'Go into that location and do the kinds of things expected there'. Whereas, if I say

(14) We went to the station today (to pick up Susan). [I]

then this would normally be interpreted as 'Go to the building and wait there till she arrives'. These interpretations are, however, defeasible, so we can say

(15) We went to the station today, but just to buy a magazine at the newspaper kiosk. [I]

However, as Levinson (2000: 146–147) points out, there is sometimes a contrast between a noun with and without an article which triggers such interpretations. Thus

(16) GO to school/church/hospital/sea

means 'to go and do the associated stereotypical activities', in a recognized social role, such as pupil, member of a congregation, patient, sailor, and so on. These stereotypical interpretations are not defeasible, whereas *GO to the school* means 'to go the building for some other reason' (cf. Fillmore 1985: 236 on phrases such as *on land* and *at sea*).[6]

The delexicalization of verbs in longer constructions is evident on a wide range of common verbs. Very high frequency verbs such as TAKE and MAKE acquire most of their meaning in context from the co-occurring noun in phrases such as *to TAKE a look* (i.e., 'to look'). However, even a verb such as WRITE, which might appear to have a stable denotational meaning independent of context, is interpreted rather differently according to its object noun (Erman and Warren 2000: 54).

Compare the following:

(17) a. TAKE a look / a photo / a shower / a telephone call / the blame, etc.
b. MAKE the bed / a drink / a film / a friend / a mess / a note / money, etc.
c. WRITE a book / a cheque / a computer program / a piece of music / graffiti, etc.

Seuren (1998) collects together several other arguments that a strictly compositional calculus is unrealistic (see especially pages 384 and 401 for his conclusions).

2.3. *The phrasal hypothesis*

Within recent linguistics, there are complex strands of argument which revise the division of labor between syntax and lexis, by putting less emphasis on syntactic structure, and more on the controling role played by lexis. This idea has long been central to theories such as dependency grammar or valency grammar (Tesnière 1959). Within the generative tradition, there has been a shift towards a lexicalist position that 'sentence structure is to a large extent determined by lexical information' (Haegemann 1991: 25); systemic functional grammar (Halliday 1994) refuses to draw a line between lexis and grammar; and word grammar (Hudson 1995) uses word–word dependencies, and presents language as a network of knowledge, with no clear distinction between linguistic and encyclopedic knowledge.

Other work, which regards grammatical constructions as complexes of lexical, grammatical, semantic, and pragmatic information, also refuses to distinguish between linguistic and encyclopedic information. This work also insists on the importance of empirical attested data, on the functions of grammar in discourse, and on the place of routine, partly prefabricated multi-word units. It therefore suggests that the traditional units of word and clause are artifacts of invented data, and do not provide appropriate ways of analyzing attested language in use. For example, Hopper (1988) and Hopper and Traugott (1993) argue that grammar should be seen not as a static entity which pre-exists discourse, but as something which emerges from the 'enormously high proportion of repetitive or partly repetitive utterances', such as idioms, figures of speech, turns of phrase, sayings and clichés in spoken discourse (Hopper 1988: 120, 121).

In related work, Fillmore and his colleagues have investigated relations between underlying semantic and conceptual structures and their lexical and syntactic realizations. In early work on case grammar, Fillmore (1968, 1969) showed how the same deep semantic relations could be realized by different surface syntactic forms. In later work on frame semantics, Fillmore and Atkins (1994: 370) study the conceptual framework underlying the meaning of a word, and the linguistic realizations (lexical and syntactic) of the elements of this frame. And in work on construction grammar, Kay (1995: 172), Fillmore (1997), and Kay and Fillmore (1999) emphasize that pragmatic meanings are conventionally associated with specific morphosyntactic structures, rather than conveyed by conversational reasoning and inference. (See also Goldberg [1996] who makes useful comparisons between construction grammar and other varieties of cognitive linguistics.)

We currently lack an assessment of how far different models of such constructions are compatible or equivalent. I will now assume a simple and elegant model of the structure of extended lexical units, proposed by Sinclair (1996, 1998), who shows that such units depend on increasingly abstract relations of collocation, semantic preference, colligation, and discourse prosody.

2.3.1. *Collocation*

Collocates are word-forms and therefore directly observable in textual data. Their probability of occurrence can be stated.

2.3.2. *Semantic preference*

This is defined by a lexical set of frequently occurring collocates, which share some semantic feature such as 'change' (see section 3.4). An abstract set is not directly observable, but the preferred lexis can be listed (with probabilities of occurrence).

2.3.3. *Colligation*

This is a relation between lexis and a grammatical category, such as a preposition or modal verb. It is a more abstract category again, since it is the outcome of long sequences of analysis.

2.3.4. *Discourse prosody*

This is the speaker's evaluation of what is being talked about, and may well provide the point of the utterance. This is not directly observable, but recurrent collocates often provide replicable evidence of evaluative connotations. There may be no term easily available to label the prosody (see section 3.6). (Louw 1993 develops this idea in detail.)[7]

2.4. *Interpretation versus convention*

So, returning now to the main argument, a central question for pragmatics (e.g., Levinson 1983: 8–22; 2000 passim) is how much meaning is inferred on the basis of assumed general knowledge and general principles of rationality, and how much is conventionally encoded in lexical and grammatical form?

Since work by Grice (1975) and Sperber and Wilson (1995 [1986]) inference theories have been very influential. However, the fact that such theories are based almost exclusively on invented data causes problems. First, by definition, such work can investigate only what inferences speakers are potentially capable of, but not what they normally do. It investigates what is possible, not what is probable. Second, since the theory argues for the power of inferences, the suspicion is that the examples are created precisely to emphasize this power. This is done by inventing example sentences which lack the linguistic markers of pragmatic meaning which are frequent in attested language use.

The alternative argument, that pragmatic force is conventionalized, is indeed put forward by those (such as Fillmore, Kay, and colleagues) who work with attested data (and possibly also invented data in addition). The view from construction grammar and related approaches is expressed by Kay (1995: 172):

> Construction grammar places great emphasis on the fact that probably any of the kinds of information that have been called pragmatic by linguists may be conventionally associated with a particular linguistic form and therefore constitute part of a rule (construction) of a grammar.

Thus, we do not have to infer that a speaker who uses the schema *FAN/FUEL the flames of* disapproves of something: this is encoded in the construction. Similarly, a phrase such as *LIVE to a ripe old age* encodes approval (see later).

3. An illustrative project

So, the hypothesis is that phrasal schemas show that much more is conventional than is often assumed. Deciding on the appropriate balance between inference and code theories is an empirical question, which can be resolved only with a large amount of attested data. Conventionally encoded evaluative meanings are not observable in isolated instances, but only in repeated co-occurrences of lexis across large corpora (Channell 2000).

3.1. *Ways of talking about people*

I will now illustrate methods and concepts in more detail, by proposing a project to study a socially significant lexical field. The general problem for research in this area is how to relate language, cognition, and culture, and this task can be tackled empirically if it is seen as relating: recurrent phrasings (observable with computational help), phrasal schemas (not directly observable, but supported by corpus evidence of repeated co-occurrences), and cultural knowledge (as represented by shared schemas). With this aim in mind, it would not be sufficient to study individual phrasal schemas. Analyses would have to cover areas of meaning where the language has significant resources for categorizing situations and events. So, I will take a very few examples of the ways in which people are classified and talked about. Two concepts which are encoded in a large number of the words we use to talk about social life are 'groups of people' and 'the passing of time'. We constantly talk about people in terms of one or both of these dimensions.

The large number of approximate synonyms for 'groups of people' is not surprising, since the different ways in which people can be grouped is of inherent social interest. Thus, *group* is a neutral word, with many hyponyms, such as those in (18).

(18) band, bunch, crew, family, flock, gang, jury, rabble, team

Many such words are rough denotational synonyms, but differ sharply in connotation: for example, *gang* and *rabble* are clearly disapproving. There are also many terms for the structure and membership of such groups. Gangs have leaders and members; families have mothers, fathers and children; teams have captains and players; and so on. Many terms for 'large groups of people' include the following:

(19) army, crowd, horde, mob

Neutral *crowd* ('a large group of people') contrasts with pejorative *mob* (an 'unruly crowd', as in *angry mob*, *lynch mob*, *mob rule*), and with pejorative *horde* (a 'large, unruly, and possibly menacing group', as in *barbarian horde*, *screaming horde*).

Amongst many other ways of categorizing people are kinship terms, terms for professions (*butcher*, *baker*, and *candlestick maker*), and ordered sets of terms which signal degree of intimacy/distance, such as in (20).

(20) relatives, friends, acquaintances, neighbours, strangers

Krishnamurthy (1996) uses corpus data to analyze the approximate synonyms *ethnic*, *racial*, and *tribal*, which occur with different collocates, such as those in (21).

(21) clans, communities, minorities, nations, races, tribes

Partington (1998: 74–75) notes a range of words which have pejorative connotations, and which are used to refer to other people, but not to ourselves, such as in (22).

(22) cults, extremists, fanatics, fundamentalists, militants

(Compare Sacks [1992, vol. 1: 172, 399] on words such as *adolescent* and *teenager*, which are used of these groups only by adults.)

As well as words which directly denote groups (such as *army* and *family*), Persson (1995) points out that many words imply 'collective involvement', since they either denote activities involving groups of people or assume groups as a default. Examples include words denoting unrest, actions of many kinds, expression of feelings, attitudes, and values, as in, respectively:

(23) anarchy, riot; concert, demonstration; applause, cheer; fame, scandal

As Persson argues, the extent to which 'collective involvement' is encoded in the vocabulary goes unrecorded in dictionaries.

Under special circumstances, the number of terms for categorizing human beings may spiral almost out of control. Danet (1980) discusses a trial in the USA in which a doctor was accused of manslaughter following a late abortion which he had carried out. Examples from the 40 different terms used for a human being at a particular stage of life are listed in (23).

(24) baby, child, embryo, fetus, infant, neonate, offspring

Many other phrases for talking about people in terms of stage of life include:

(25) a. age group, age bracket, age of consent, come of age
b. in my younger days; in his/her day; in his/her heyday; (cut down) in his/her prime; thirty something; over the hill; burnt out; past it; ripe old age; twilight years.

Combining the two dimensions of meaning, 'groups of people' and 'the passing of time', there are many terms for talking about groups of people in terms of the human life cycle, from *babies* to *teenagers* to *senior citizens*. This area of everyday experience is classified in detail by the vocabulary, and English has many category terms, everyday and technical, which can be listed in ordered sets. A few examples include:

(26) a. infant, baby, child, adolescent, teenager, youth, adult
b. childhood, schooldays, youth, adulthood, old age, dotage
c. young, underage, youthful, middle-aged, elderly, old, senile

Individual words can easily be extracted from a thesaurus, but there is an important difference between words in the vocabulary and words in texts: between what is potentially available and what is actually chosen. A systematic study of this large area of classification would require a project well beyond the scope of this article. It would take us into theories of discourse (discursive formations) and the view that category systems organize the consciousness of a social group. Clearly, a stone is a stone is a stone: stones exist in the material world and are not created by discourse. However, the precise location of the distinctions between pebbles, stones, rocks, and boulders is created by a classification system (Hall 1997: 221). Or, to take a socially more significant example: people exist, but the categories to which we assign them (children, teenagers, adults) do not exist in a form unchanging for all time, and indeed have changed profoundly over the past century or so. The notion that children are distinct from adults, and should be dressed differently, treated differently, and so on, is a relatively recent historical notion (Ariès 1965). Groups of people are represented in different ways at different times in line with different social and cultural interests, and some categories have been produced by the language game, as the creation of different interest groups (*grey panthers*, *a new man*), of fashion and music industries (*pre-teens*, *teens*), and so on. (See Francis et al. (1998: 16) for other terms, from *in-laws* to *chattering classes*.)

All such categories allow us to locate other people in the social world, and sustain our sense of its order. In a famous paper, Sacks (1972) discusses how such words can contribute to textual coherence, since they signal implicit knowledge of social structure which is shared by members of a culture. This knowledge includes speakers' knowledge of 'membership categorization devices': these are lexical systems, such as family (*baby*, *mommy*, etc.) and stage of life (*baby*, *child*, *adolescent*, *adult*, etc.), plus their cultural associations (*babies cry*, *cry-baby*). Sacks sets out to demonstrate 'the fine power of a culture' which does not 'merely fill brains in roughly the same way', but 'fills them so that they are alike in fine detail' (1972: 332).[8]

I will restrict myself now to a few examples which illustrate how methods of corpus semantics can be used in such a project.

3.2. *Example 1:* Underage*,* youths*, and* teen

In simple cases, the evaluative connotations of individual words are shown in their most frequent collocates. For example, *underage* is almost always used to refer to activities which are illegal or socially disapproved of, most frequently in phrases such as in (27), and occasionally in references to underage applications to join some organization (such as the army).

(27) underage sex; sex with underage girls; underage drinking; underage smoking; underage driving

The word *youths*, despite its apparent denotation, is used only of males. This is explicit in phrases such as *teenage girls and gangs of youths*. In addition, the word has strongly disapproving connotations, and is used, especially in the British press, in contexts of violence and crime. As it is put in Cobuild (1995), 'Journalists often refer to young men as *youths*, especially when they are reporting that the young men have caused trouble'. Significant collocates and attested phrases include the following.[9] (Note those denoting 'groups'.)

(28) a. gang(s), group(s), hundred
b. black, white, unemployed
c. accused, armed, arrested, attack(ed), bombs, charged, chased, escaped, fight, hurled, police, threw/throwing
d. mobs of youths; a dozen marauding youths; gangs of youths hanging around; rioting youths have smashed shop windows; two youths were arrested; unemployed youths turn to crime; youths tempted into crime.

Teenage and *teenager(s)* are often used in disapproving ways. The word *teen* is used almost exclusively with reference to what is seen as low-grade pop music, television, and film. The most significant collocates and the often ironic and patronising phrases include those in (29) and (30).

(29) angst, idol, magazine(s), pop (star), pregnancy, sensation, terror, throb (*as in* heart-throb).
(30) teen angst years; teen rebel angst; teen cult; teen heart-throbs; teen idols; dire teen movie.

These collocations also illustrate how words form networks of mutual prediction. If we start with *teen*, then one of its most significant collocates is *angst*. Conversely, if we start with *angst*, then we find it frequently collocates with *teen(age)* (as in *teenage angst*, *angst-ridden teenager*). *Angst* is frequent in British journalism, where it is almost always used in highly disapproving and ironic ways (Stubbs 1998). Empirical evidence of this type of intercollocation is not systematically described in any work I am familiar with. (See also the following discussion on CAREER and LAUNCH.)

3.3. *Example 2: LIVE to a ripe old age*

Fillmore (1997) notes briefly that the phrase *ripe old age* is preceded by a definite or indefinite article, and that it frequently collocates with the lemma LIVE, but also with other verbs such as ATTAIN. He also poses the question: where does the unit end? I studied all 60 occurrences of *ripe old age* in the corpora listed: 200 million words of British and American English. There are clear central patterns, but, as always, considerable variation.

To the left of the phrase is almost always a definite or indefinite article. If there is a definite article, then to the right, the phrase is always followed by *of* plus a number and optionally *years* (e.g., *to the ripe old age of 70 years*). To the left of the article is usually a preposition, though this depends on the preceding verb: *LIVE to a ripe old age* but *REACH a ripe old age*. Other semantically related verbs include ATTAIN, SURVIVE TO, GO ON TO. Further to the left, are often verbs such as ASPIRE, HOPE, INTEND, STRIVE, SURVIVE, WANT, or other words expressing a possibility, as in (31).

(31) a. if you expect to live to a ripe old age
b. stand a better chance of living to a ripe old age

The phrase *ripe old age* is, however, not entirely fixed. The following variants are much less frequent, but all attested: *at ripe old ages*; *ripe age*; *good (old) age*; *grand (old) age*. In a very large text collection, frequencies were as follows: *ripe old age* circa 7,950, *good old age* circa 1,600, and *grand old age* circa 550. Of all examples of *ripe old age*, around eighteen percent occurred in the longer expressions *LIVE to* or *REACH a ripe old age*.[10]

The phrase is sometimes used ironically (*at the ripe old age of 29*), but the connotation of the whole unit is usually positive. Reaching a *ripe old age* is a good thing to do. The speaker admires the achievement of avoiding dangers and risks en route, to which there are frequent references in co-occurring vocabulary such as: *death*, *maximum life-span*, *perils of infancy*, *survive*. The admiration (and sometimes slight envy?) is explicit in examples such as (32).

(32) a. it is a major triumph of the 20th century that many more people survive to a ripe old age
b. he survived the perils of infancy to live to the ripe old age of 74.

Ripe old age is clearly an encoding idiom (Makkai 1972; Fillmore et al. 1988), but arguably also a decoding idiom. It could, in principle, mean ‘over mature, senile, past it’, but it does not; this is simply not its conventional connotation. This is partly explicable from the largely positive collocates of *ripe* itself, including:

(33) flavour, fresh, fruit, fruity, full, juicy

Also, when an adjective immediately precedes *old age*, this is most frequently (in 40 to 50 percent of cases) *ripe*, *grand*, *great*, or *good*. Further, these combinations fit into a broader pattern of phrases such as approving *venerable old age* versus disapproving *callow youth*. This again illustrates the 'smooth interaction' between the schema and other phrases (Kay and Fillmore 1999: 7).

This is a further example of a linguistically encoded cultural concept: a way of talking about the human expectation of three-score years and ten. It is therefore not surprising that such phrases turn up in newspaper headlines. Two examples are *A Grand Old Age* and *Refining the Idea of Ripe Old Age*.

Fillmore's question is certainly not fully answered: what are the boundaries of the unit? There is a frequent core element, *ripe old age*, though the only obligatory word is *age*. There is also a phrasal schema which consists of preferred collocates (often the verb LIVE), colligations (often a preceding verb plus preposition plus determiner), semantic preferences (often with words which express the achievement involved), and a positive discourse prosody (expressing the speaker's admiration). However, there are very few invariable phrases in English (Sinclair 1996: 83), and such schemas have a flexibility which allows phrases to fit into the surrounding co-text.

In summary, we have a phrasal schema, characterized by typical lexis and syntax, which is used to talk about people's lives. The recurrent use of an evaluative schema, instantiated in preferred but variable lexis, stabilizes the underlying phraseology and, arguably, the underlying concept. The schema is a fragment of a description of routine language use: a tiny fragment of the social world, concerning how people characteristically talk about groups of people and passing time.

3.4. *Example 3:* Ripe for change

The importance for semantic description of looking at phraseology, and not just at individual words, becomes even clearer if we contrast the schemas for *ripe old age* and *ripe for*. The word *ripe* has collocates such as *pick* and *fruit*. However, there are very few cases indeed where the phrase *ripe for* is used literally to refer to ripe fruit or other plants: almost all occurrences are abstract and metaphorical. The prototypical example is perhaps that in (34) (attested).

(34) the time was ripe for major change.

The most significant collocates (as measured by *t*-scores) of *ripe for* include those in (35).

(35) climate, conditions, situation, time

The noun or noun phrase following *ripe for* often concerns ‘change’ and often has negative connotations:

(36) a. change(s), development, expansion, growth, overhaul, reform, renewal, renovation, restoration, shift, transformation
b. assault, attack, insurrection, mockery, outbreak of cholera, revolution, show down.

So *ripe for* is frequently (in up to 50 percent of cases) part of a longer phrase. Often the implication is that things have got so bad that change is necessary:

(37) the time/conditions etc. BE/LOOK/SEEM ripe for ‘change’.
(38) a. is [the company] ripe for rationalization or a candidate for closure?
b. given the spate of recent disasters … the time is ripe for a major initiative in the field of emergency planning and hazard studies.

There are some recurrent subpatterns: *ripe for the picking*, or occasionally *plucking*. Variants with both *ripe* and *right* occur.

These examples show the need to define phrasal units in terms of prototypes, with central and frequent realizations, but open-ended diversity of lexical detail. Corpus studies reveal very clearly the repetition which is characteristic of language in use, but they show also that this repetition is probabilistic, not deterministic. Gumperz and Hymes (1972: 304) therefore see culture not as ‘replication of uniformity’ but as ‘organization of diversity’.

A possibly rather obvious, but crucial, point about the statistical nature of the evidence is that it comprises many occurrences of phrases which are independent of each other, in so far as they are in different texts, produced by many different speakers. These occurrences cannot influence each other directly. Nevertheless, a major finding of corpus semantics is the pervasiveness of intertextual patterns, which are observable in many texts in the culture. To that extent, the phrases are not independent of each other. Corpus semantics studies typical usage, but this leaves us with the unsolved puzzle of how individual competence relates to recurrent behavior in the discourse community.

3.5. *Example 4:* Signs of age, signs of ageing

Other words and phrases, as we have seen, occur with both preferred lexis and syntax. The phrase *signs of age* is usually used with reference to objects, not humans. It occurs in characteristic collocations and colligations: frequently preceded by a verb in an *-ing* form, often SHOW and/or START or BEGIN. A typical example is (39).

(39) beginning to show the first signs of age

Variants, such as *the signs of old age*, *signs of approaching age*, and *signs of ag(e)ing* are usually used of people, the latter with reference to ageing skin. Frequent preceding verbs include the semantically related SHOW or CAUSE, ACCENTUATE or HASTEN, and DELAY or COMBAT. Many uses have clearly negative connotations:

(40) a. causing such classic signs of ageing as constant tiredness
b. the freshness of youth or ... the signs of ageing.

These phrases fit into a wider pattern. The phrase SHOW *signs of* has a very negative discourse prosody, which is typically shown in following words, such as in (41).

(41) cancer, disease, exhaustion, faltering, fatigue, illness, jet lag, rot, substance abuse, violence, vulnerability.

The phrase *signs of* is also usually negative. It is followed by words such as those in (42).

(42) acrimony, a break-in, collapse, damage, defective vision, shyness, strain, trouble.

Even the word *signs* itself is predominantly negative. Its most significant collocates include the following:

(43) of, SHOW, danger, disease, distress, illness, increasing, ominous, stress, warning.

However, *signs* does have positive collocates, such as *encouraging*, *hopeful*, *positive*, and examples such as the following occur:

(44) a. the economy is showing signs of improvement
b. the housing market is showing signs of recovery
c. showing signs of an Islamic renaissance
d. those seeds that show definite signs of life.

As I have noted, we need a clearer concept of such related specific and general patterns. *SHOW signs of ageing* is a specific example of a more general schema, *SHOW signs of [something bad]*. This is in turn a more specific example of the generally, though not exclusively, negative discourse prosody of *signs of*. We therefore have a further example of the way in which the particular (idiomatic) and the general 'are knit together seamlessly' across such constructions (Kay and Fillmore 1999).

3.6. *Example 5:* CAREER

The status of CAREER as a cultural keyword is discussed by Williams (1976 [1983]). It is used to talk about progress in a successful life: characteristic uses refer not only to a *career on the stage* but also to *a stage in someone's career*. It has a very positive discourse prosody: it is used to refer to high-prestige occupations, often in public life, and people talk about careers in terms of a structured sequence, with beginnings, developments, and ends. This is shown by some of its most significant collocates, and characteristic examples:

(45) a. BEGIN, START, END, development, stage
b. after, during, early, long, throughout, years
c. distinguished, glittering, international, managerial, political, professional, promising, successful
d. at the peak of his career
e. finish my career on a high note
f. crown a distinguished career in radio and television
g. the best years of my career.

The data show another example of inter-collocation. A significant collocate of *career* is LAUNCH. In turn, other significant collocates of LAUNCH include the following:

(46) a. major, new
b. appeal, bid, campaign, initiative, inquiry, investigation
c. attack(s), offensive.

The concrete sense of 'LAUNCH a boat or missile' is much less frequent than the abstract use, as in *plan to launch a major new campaign*. LAUNCH connotes the start of something new, large, and important, and CAREER connotes a structured and successful professional life. A semantic feature is shared across the collocation CAREER–LAUNCH but we do not yet have the terms to label such semantic features or discourse prosodies. If it was possible to attach such features accurately to lexical items, we could predict that items with similar features would co-occur, and contribute to textual coherence.

4. Conclusions and implications

Corpus linguistics has documented a layer of organization between lexis and syntax, which is not recognized in much previous linguistic description. Substantial findings about this lexicosyntactic organization are presented in Francis, Hunston, and Manning (1996, 1998), and methods of analysis are described in detail by Hunston and Francis (2000).

However, the hypothesis of phrasal schemas is at an early stage, and I do not wish to gloss over substantial unsolved descriptive problems. These include how to identify the boundaries of such units in texts; how to define them as units in the vocabulary (as prototypes with central cores plus optional features?); how statistical facts can best be summarized; how relations of mutual prediction (e.g., TEEN–ANGST and LAUNCH–CAREER) can be discovered and stated; how relations between phrasal schemas of more or less generality (e.g., *signs*, *signs of [something bad]*, *signs of age*, etc.) can best be stated; and how connotations or discourse prosodies can best be labeled (e.g., is a feature such as 'admiration for an achievement' too specific or is it shared by other lexical units?).

We currently have detailed statements of only a few phrasal schemas, in formats which are roughly compatible, but certainly not identical. Some of the most detailed analyses in the literature are provided by Fillmore, Kay, and O'Connor (1988), Louw (1993), Fillmore and Atkins (1994), Stubbs (1995), Sinclair (1996, 1998), Kay and Fillmore (1999), and Channell (2000). The possibility of describing such lexicosyntactic units across the whole language is shown clearly by Francis and colleagues (1996, 1998). They also show how analysis of lexico-grammar can identify semantic and pragmatic features which are shared by sets of lexical items, but given the impressive scale of this work, the delicacy is often rather coarse: the work is still based on individual words as the units of description (it is words which are listed in the index), and many social implications remain to be followed up.

Relations between lexical units are usually conceived of as relations between words (lemmas) in the vocabulary (language system). However, corpus data show different types of relation, which are not mere performance, since they are shared by many speakers, but nor do they fit into standard views of competence.

Broad hypotheses to be tested include the following:

1. Our comprehension of discourse depends on both decoding (a linguistic process) and inference (a more general, not exclusively linguistic, process). Little is known about how much each of these processes contributes, but corpus studies show that the contribution

of conventional encoded meanings is larger than is claimed by the (recently dominant) inference theories.

2. Corpus linguistics cannot directly study comprehension (which is unobservable), but only co-occurrence patterns across large text collections (which are observable with computational help). This reveals a level of organization between lexis and syntax, which is not equivalent to or reducible to the levels that linguistics traditionally studies. This level of recurrent phrasal schemas can be inferred from frequently co-occurring lexis and syntax, and the schemas can be made explicit by using the models of extended lexicosyntactic units which have been proposed independently within theories of lexico-grammar, construction grammar and related approaches.
3. The techniques of analysis which I have illustrated show how the analysis of cultural keywords can be extended from individual words to socially significant networks of related phrasal schemas. This can show how people routinely talk about and evaluate significant areas of their social world (Hunston and Thompson 2000), and therefore help to make the link between different levels of social reality (Carter and Sealey 2000), between the predictable behavior of the discourse community and individual cognition.

Corpus linguistics studies not what is possible, but what is probable. It is inherently social: its data comprise attested communicative acts. It is also inherently diachronic: the semantic units which can be discovered by computational techniques have occurred countless times before. Their meaning can be empirically investigated in the history of their occurrences (Teubert 1999a, 1999b). One of the central puzzles for linguistics is the balance between individual creativity (made possible by lexical variety) and stability (which can be estimated with statistical methods). Corpus semantics can therefore make a small empirical contribution to studying the age-old problem of freedom versus constraint in human behavior.

Notes

[*] For discussion of topics in this paper, I am most grateful to Andrea Gerbig and Gabi Keck. For detailed and constructive critical comments on a previous draft, I am most grateful to two anonymous referees.

[1] I have used the fifty-million-word CobuildDirect corpus, available at http://titania.cobuild.collins.co.uk/form.html, and the hundred-million-word British National Corpus, available at http://thetis.bl.uk/lookup.html.

[2] In more recent work, available only after this article was first submitted for publication, Levinson (2000) provides further detailed discussion of the difficulty of correctly deciding the division of labor between semantics and pragmatics.

3. This dichotomy is placed at the beginning of the standard work on relevance theory (Sperber and Wilson 1995: 2–3), which then argues for the importance of inferential mechanisms in language comprehension. Relevance theory is comprehensively criticized by Levinson (2000).
4. A second set of work, less influential in linguistics, has been within phenomenologically influenced sociology, e.g., Garfinkel (1967), who acknowledges the influence of earlier work by Schutz, initially published in German (as *Der Sinnhafte Aufbau der Sozialen Welt*) and available in English from the 1950s and 1960s (Schutz 1962). In turn, these ideas have roots in Husserl's work on the natural attitude of mind and what we usually accept 'as a matter of course'. A systematic review of such work would go far beyond the scope of this article, and would have to discuss the 'cultural turn' and the 'cognitive turn' in much work in the social sciences (the title of Cicourel's 1973 book is *Cognitive Sociology*).
5. There are major implications here for text analysis, which I can only note but not discuss in detail. If speaker attitudes are signaled in text in such ways, are they deniable? Would it be possible to establish systematically, via text analysis, that a text is expressing particular attitudes towards a topic? If yes, what are the implications for critical discourse analysis, (not to mention libel law)?
6. Levinson (2000) argues that generalized conversational implicatures play a much more important role in comprehension than is usually recognized, i.e., he argues for pragmatic inferencing (with default interpretations) not semantic decoding. Nevertheless, he admits that cases such as *GO to school* are a 'miscellaneous set' and that the implicatures here are 'conventionalized' (2000: 146–147).
7. In the Firthian tradition, a prosody is a feature spread over two or more segments, and the best known application of the concept is probably in phonology. The concept of semantic prosody is developed by Sinclair (1991: 70–75, though the term is not used there) and by Louw (1993). Partington (1998: 66–67) and Channell (2000) also have good discussion and examples.
8. It would be worth studying the links between Sacks's work and corpus semantics. As Levinson (1983: 287, 295) points out, conversational analysis uses inductive methods—to search for recurring patterns across attested data—and argues that introspection is unreliable in identifying such patterns. But conversational analysis pays correspondingly little attention to the wider social context of utterance. These comments could be applied equally well to corpus linguistics (which hardly existed when Levinson's 1983 book was published). As far as I know, these parallels have not been studied.
9. Measures of significant attraction between node and collocates have been fairly thoroughly discussed in the literature. When I use the term 'significant' here, I mean significance as measured by a *t*-score: see Church et al. (1991), Clear (1993), Stubbs (1995).
10. The text collection comprised the world wide web documents, accessed in November 1999 by the search engine at http://www.alltheweb.com. It is impossible to know how many running words this might represent: the search engine claimed to access 200 million documents, not all of which are in English.

References

Ariès, P. (1965). *Centuries of Childhood*. NY: Random House.
Bartlett, F. C. (1932). *Remembering*. Cambridge: Cambridge University Press.

Berger, P. and Luckmann, T. (1967 [1966]). *The Social Construction of Reality*. Harmondsworth: Penguin. [Originally published 1966.]

Brown, G. and Yule, G. (1983). *Discourse Analysis*. Cambridge: Cambridge University Press.

Carston, R. (1998). *The Semantics/Pragmatics Distinction: The View from Relevance Theory*. (UCL Working Papers in Linguistics, 10.) London: University College.

Carter, A. and Sealey, R. (2000). Language, structure and agency: What can realist social theory offer to sociolinguistics? *Journal of Sociolinguistics* 4 (1): 3–20.

Cicourel, A. V. (1973). *Cognitive Sociology*. Harmondsworth: Penguin.

Channell, J. (2000). Corpus-based analysis of evaluative lexis. In *Evaluation in Text: Authorial Stance and the Construction of Discourse,* S. Hunston and G. Thompson (eds.), 38–55. Oxford: Oxford University Press.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In *Lexical Acquisition*, U. Zernik (ed.), 115–164. Englewood Cliff, NJ: Erlbaum.

Clear, J. (1993). From Firth principles: Computational tools for the study of collocation. In *Text and Technology*, Baker, M., Francis, G., and Tognini-Bonelli, E. (eds.), 271–292. Amsterdam: Benjamins.

Cobuild (1995). *Collins Cobuild English Dictionary*. London: HarperCollins.

Cook, G. (1994). *Discourse and Literature*. Oxford: Oxford University Press.

Danet, B. (1980). 'Baby' or 'fetus': Language and the construction of reality in a manslaughter trial. *Semiotica* 32-1/2: 187–219.

Erman, B. and Warren, B. (2000). The idiom principle and the open choice principle. *Text* 20 (1): 29–62.

Fillmore, C. J. (1968). The case for case. In *Universals in Linguistic Theory*, E. Bach and R. Harms (eds.), 1–88. NY: Holt, Rinehart and Winston.

—(1969). Toward a modern theory of case. In *Modern Studies in English*, D. A. Reibel and S. A. Shane (eds.), 361–375. NJ: Prentice-Hall.

—(1985). Frames and the semantics of understanding. *Quaderni di Semantica* VI (2): 222–254.

—(1997). Lectures on construction grammar. Available at http://www.icsi.berkeley.edu/˜kay/bcg/lec02.html (accessed 11 May 1999).

Fillmore, C. J. and Atkins, B. T. S. (1994). Starting where the dictionaries stop: The challenge of corpus lexicography. In *Computational Approaches to the Lexicon*, B. T. S. Atkins and A. Zampolli (eds.), 349–393. Oxford: Oxford University Press.

Fillmore, C. J., Kay, P., and O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64 (3): 501–538.

Francis, G., Hunston, S., and Manning, E. (1996). *Grammar Patterns 1: Verbs*. (Collins Cobuild.) London: HarperCollins.

—(1998). *Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

Garfinkel, H. (1967). *Studies in Ethnomethodology*. NJ: Prentice Hall.

Goldberg, A. (1996). Jackendoff and construction-based grammar. *Cognitive Linguistics* 7–1: 3–19.

Grice, H. P. (1975). Logic and conversation. In *Speech Acts: Syntax and Semantics*, vol. 3, P. Cole and J. L. Morgan (eds.), 41–58. New York: Academic Press.

Gumperz, J. J. and Hymes, D. H. (eds.) (1972) *Directions in Sociolinguistics: The Ethnography of Communication*. NY: Holt, Rinehart and Winston.

Haegemann, L. (1991). *Introduction to Government and Binding Theory*. Oxford: Blackwell.

Hall, S. (1997). The centrality of culture. In *Media and Cultural Regulation*, K. Thompson (ed.), 207–238. London: Sage.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar*. 2nd edition. London: Arnold.

Hoey, M. (1983). *On the Surface of Discourse*. London: Allen and Unwin.

Hopper, P. (1988). Emergent grammar and the a priori grammar postulate. In *Linguistics in Context*, D. Tannen (ed.), 117–134. NY: Ablex.

Hopper, P. J. and Traugott, E. J. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.

Hudson, R. A. (1995). Identifying the linguistic foundations for lexical research and dictionary design. In *Automating the Lexicon*, D. E. Walker, A. Zampolli, and N. Calzolari (eds.), 21–51. Oxford: Oxford University Press.

Hunston, S. and Francis, G. (2000). *Pattern Grammar*. Amsterdam: Benjamins.

Hunston, S. and Thompson, G. (eds.) (2000). *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Kay, P. (1995). Construction grammar. In *Handbook of Pragmatics*, J.-O. Östmann and J. Blommaert (eds.), 171–177. Amsterdam: Benjamins.

Kay, P. and Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: The *What's X Doing Y?* construction. *Language* 75 (1): 1–33.

Krishnamurthy, R. (1996). Ethnic, racial and tribal: The language of racism? In *Texts and Practices: Readings in Critical Discourse Analysis*, C. R. Caldas-Coulthard and M. Coulthard (eds.), 129–149. London: Routledge.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

—(2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In *Text and Technology*, M. Baker, G. Francis, and E. Tognini-Bonelli (eds.), 157–176. Amsterdam: Benjamins.

Makkai, A. (1972). *Idiom Structure in English*. The Hague: Mouton.

Moon, R. (1998). *Fixed Expressions and Idioms in English*. Oxford: Clarendon.

Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language* 71 (3): 491–538.

Nunberg, G. and Zaenen, A. (1992). Systematic polysemy in lexicology and lexicography. In *Proceedings of Euralex II* (Tampere), H. Tommola et al. (eds.). Available at http://www.parc.xerox.com/istl/members/nunberg/Euralex.html (accessed 21 May 1999).

Partington, A. (1998). *Patterns and Meanings*. Amsterdam: Benjamins.

Persson, G. (1995). On collective involvement in English. In *Studies in Anglistics*, G. Melchers and B. Warren (eds.), 125–136. Stockholm: Almqvist and Wiksell.

Sacks, H. (1972). On the analysability of stories by children. In *Directions in Sociolinguistics*, J. J. Gumperz and D. Hymes (eds.), 329–345. NY: Holt, Rinehart and Winston.

—(1992). *Lectures on Conversation*. 2 volumes. Edited by G. Jefferson. Oxford: Blackwell.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Erlbaum.

Schutz, A. J. (1962). *Collected Papers 1: The Problem of Social Reality*. The Hague: Nijhoff.

Seuren, P. A. M. (1998). *Western Linguistics*: *An Historical Introduction*. Oxford: Blackwell.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

—(1996). The search for units of meaning. *Textus* IX: 75–106.

—(1998). The lexical item. In *Contrastive Lexical Semantics*, E. Weigand (ed.), 1–24. Amsterdam: Benjamins.

Sperber, D. and Wilson, D. (1995 [1986]). *Relevance: Communication and Cognition*. 2nd edition. Oxford: Blackwell. [1st edition published 1986.]

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2 (1): 23–55.

—(1996). *Text and Corpus Analysis*. Oxford: Blackwell.

—(1998). German loanwords and cultural stereotypes. *English Today* 14 (1): 19–26.

Tesnière, L. (1959). *Élements de syntaxe structurale*. Paris: Klincksieck.

Teubert, W. (1999a). Corpus linguistics: A partisan view. Available at http://solaris3.ids-mannheim.de/ijcl/teubert_cl.html (accessed 24 Nov. 1999).

—(1999b). korpuslinguistik und lexikographie. *Deutsche Sprache* 4: 292–313.

van Dijk, T. A. (1972). *Some Aspects of Text Grammars*. The Hague: Mouton.

Williams, R. (1976 [1983]). *Keywords*. London: Fontana.

Michael Stubbs is Professor of English Linguistics at the University of Trier, Germany. He was previously Professor of English, Institute of Education, University of London, 1985–1990; and Lecturer in Linguistics, University of Nottingham, 1974–1985; and is Senior Honorary Research Fellow, University of Birmingham. He has published widely on educational linguistics and discourse analysis. His most recent book is *Text and Corpus Analysis* (Blackwell, 1996). His next book, entitled *Words and Phrases: Corpus Studies of Lexical Semantics*, is due to appear in 2001.