This is a revised version of a paper presented to the conference on Keyness in Text at the Certosa di Pontignano, University of Siena, June 2007.

THREE CONCEPTS OF KEYWORDS

Michael Stubbs, FB2 Anglistik, Universität Trier, D-54286 Trier. stubbs@uni-trier.de

ABSTRACT

The term "keywords" is widely used to refer to words which are important in some way, either in individual texts or in a given culture. The general idea is perhaps clear enough, but there are two problems. First, there are several different concepts of "keywords". This paper discusses three loosely related uses of the term, which derive from quite different academic traditions, and which are therefore only marginally compatible. Second, there is a very large gap between individual words, texts and culture. The paper argues that this gap can only be bridged by a theory which relates words, phrases and texts to the social institutions which are characterized by texts and text-types.

INTRODUCTION

Keywords are words which are claimed to have a special status, either because they express important evaluative social meanings, or because they play a special role in a text or text-type. From a linguistic point of view, they contribute to the long "search for units of meaning" (Sinclair 1996). From a sociological point of view, they are part of "a vocabulary of culture and society" (Williams 1976/1983). In work on keywords, semantic and social analysis are inseparable. However, the term "keyword" is used in several different senses which are only loosely related, and there is a large gap between individual words and the social world. Words occur in phrases and speech acts, which in turn occur in speech events in different social institutions. So, this paper is about how words relate to the world.

It is also about the state of the art in corpus linguistics. Many corpus studies since the 1990s analyse the words and phrases used in many different text-types, but these studies do not yet amount to a unified body of social theory. Corpus linguists like nothing better than empirical findings supported by levels of statistical significance. But outside this narrow circle, people want to know how it all hangs together, and how all the empirical information contributes to solving the great intellectual puzzles of language in society. How should all this work be evaluated? How does empirical linguistics contribute to "wider issues", and how can it be used "as a foundation for a broad range of intellectual exploration"? (Sinclair 2007: 1). One attempt to explain how it does all hang together is made by Searle. In the opening sentence of his 1969 book on *Speech Acts*, he asks "how

words relate to the world" (Searle 1969: 3). And in his later 1995 book on *The Construction of Social Reality* (Searle 1995), he shows that social reality depends on language, since there is a logical relation between speech acts, the inter-subjective world of social facts, and the structure of social institutions.

But these two approaches to understanding words and the world are very different in style. Corpus linguistics provides a powerful model of communicative acts (Sinclair 1996, 1998, 2005), which is firmly based on empirical facts and statistics, but it is often weak on social theory. Speech act theory also provides a powerful model of communicative acts, and this in turn provides the basis for a powerful social theory (Searle 1995), but it is weak on empirical linguistic data (it often uses only invented examples).

The paper uses the following presentation conventions. Italics are used for *word-forms*. SMALL CAPS are used for LEMMAS. Single quotes are used for 'meanings'. Double quotes are used for "quotes from other authors". Unless otherwise stated, the examples are from the British National Corpus (BNC), a corpus of one hundred million running words of written and spoken English.

PART 1. THREE CONCEPTS OF KEYWORDS

I will discuss three very different senses of the term "keywords". Sense 1 derives from cultural studies. It is well known from Raymond Williams' work (Williams 1976/1983), though there is also much earlier French- and German-language work. Sense 2 derives from comparative quantitative corpus analysis, which identifies words which are statistically prominent in particular texts and text collections. It is best known from Mike Scott's WordSmith Tools (Scott 1998). Sense 3 derives from work on lexico-grammar. In a 1993 article, Gill Francis proposes an ambitious project to discover the phrasal units which express taken-for-granted cultural meanings, and therefore to "compile a grammar of the typical meanings that human communication encodes" (Francis 1993: 155). These three concepts are very different, and possibly not even compatible, but they do at least share the notion of discourse as recurrent and conventional ways of talking which circulate in the social world and which contribute to ways of thinking about the social world. As J. R. Firth (1957: 29) expresses it in one of his more cryptic utterances: "We are in the world and the world is in us". The general idea of "keywords" is fairly clear, although the metaphor is rather vague. Keywords are the tips of icebergs: pointers to complex lexical objects which represent the shared beliefs and values of a culture.

KEYWORDS SENSE 1 (WILLIAMS 1976/1982): WORDS AND CULTURE

Sense 1 is explicitly cultural. As Wierzbicka (1997: 156) phrases it, keywords are a "focal point around which entire cultural domains are organized" (for German, she gives examples such as *Heimat* and *Vaterland*). For English-language scholars, the most famous example of sense 1 is Raymond Williams' book (1976/1983) *Keywords: A Vocabulary of Culture and Society*. However, the concept goes back much further. Forty years before Williams, J. R. Firth (1935: 40, 51) talked of "sociologically

important words, which one might call focal or pivotal words" (e.g. *work, labour, trade, leisure*).

In addition, thirty years before Firth, there was the beginning of a long tradition of German-language work on the use of *Schlüsselwörter* (= keywords). Dictionaries of words which are important in social and intellectual history were produced from the early 1900s up to the the fall of the Berlin Wall. This tradition is variously called *Schlagwortforschung* (= 'research on catch phrases') and *Begriffsgeschichte* (= 'the history of concepts'). Early 20th century examples include a historical dictionary (Ladendorf 1906) and a dictionary of keywords at the time of the Reformation (Lepp 1908). Late 20th century examples include a dictionary of *Brisante Wörter* (= 'controversial words') *von Agitation bis Zeitgeist* (Strauß et al 1989), a dictionary of controversial concepts in public discourse (Stötzel & Wengeler 1995), and a dictionary of keywords from the *Wende*, the fall of the Berlin Wall (Herberg et al 1997). Other work includes an article on *politische Vexierwörter* (= 'ambiguous political words') (Teubert 1989). There is also a long tradition of French-language work. In the 1950s Georges Matoré (1953) discussed *mots clés* (= 'key words') and argued that lexicography is a sociological discipline. Also from the 1950s is an article by Émile Benveniste (1954) who discusses the word *civilisation*, which is used rather differently from the word *civilisation* in English. His work is based on still earlier work by Lucien Febvre from 1930. This tradition was continued by Michel Foucault, who had his favourite keywords (e.g. *labour, madness, prison*) (Hacking 1986: 27).

Williams proposes a rather small set of around 120 words which are important in the culture: though quite how the culture might be defined, I am not sure. Four characteristics of his keywords are as follows. (1) First, Williams identifies words intuitively, on the basis of his extensive scholarship. He then uses the attested citations in the 12-volume *Oxford English Dictionary* as empirical evidence that his keywords have undergone historical shifts in meaning which have led to complex layers of meanings in contemporary English. They are "difficult words", as he puts it. (2) Second, only some of his keywords are in widespread use (e.g. *country, expert, family, genius*), whereas many are from an intellectual discourse and most native speakers of English would not have the slightest idea what they mean (e.g. *alienation, dialectic, hegemony, utilitarian*). But Williams has no explicit theory of the organization of the vocabulary (e.g. core versus specialized) or of text-types or discourse communities which could explain this distinction. (3) Third, Williams assumes that keywords do not just label, but help create, conceptual categories. He talks of "significant, indicative words in certain forms of thought" (Williams 1983: 15). Work on keywords necessarily implies a constructivist Whorfian perspective. (4) Fourth, Williams' particular interest is a marxist-socialist analysis of the social order. In his article on the discourse on the miners' strike in 1984-85, he discusses four "slippery" keywords / phrases: *management, economic, law and order* and *community* (Williams 1985). He discusses "the key issue in the whole modern organization of work": whether workers can control their own production or whether *management* simply means 'employer' (even in nationalized industries).

Example: The semantic field 'work and leisure'

A key semantic and cultural domain identified by both Firth and Williams is that of 'work versus leisure', and Williams (1976 / 1983) shows that related words (such as *work, career, job, labour, (un)employment*) have distinctly different uses and connotations. He shows that the word WORK has developed historically from general meanings of 'doing something', to more restricted meanings concerning the social relationship of paid employment.

One limitation in Williams' work is that he does not distinguish between different forms of a lemma, which often have very different collocates and uses. In a small corpus study (Stubbs 1996: 177-78), I showed that the word-forms *work*, *working* and *worker* occur in quite different compounds and fixed phrases:

- workaholic, workforce, workload, workplace, workroom, worksheet, workshop, workstation, worktop, workwear
- working class, working conditions, working mother
- aid worker, factory worker, office worker, social worker

Some of these constructions are highly productive: nouns immediately preceding *worker* include, amongst many others

- airport, bakery, bank, brewery, building, care, charity, coalface, community, construction, defence, farm, forestry, government, health, hospital, hotel, housing, kitchen, maintenance, morgue, sex, steel, welfare, youth

And some of these phrases are replacing other terms, for example

- bank clerk > bank worker, miner > coalface worker, agricultural labourer > farm worker

This is a clear example of how empirical linguistic data can contribute to a social analysis. As Teubert (2007: 59) points out: "without a word for *work* there would not be the social construct 'work'." In addition, this semantic field has undergone rapid historical change, is still strikingly productive, and this is an indicator of social change.

Summary of Williams' approach

Williams' work was within a cultural studies tradition, and was not intended to be a linguistic analysis, so you may think that the following points are unfair. However, from the point of view of linguistic analysis, Williams' list is not a good basis for a general theory of meaning. The list covers a very small set of words, many of which are very specialized. He has no way of providing comprehensive coverage of the vocabulary of a language, he has no theory of how the vocabulary of a language is organized, and he has no way of relating words to texts and text-types.

KEYWORDS SENSE 2 (SCOTT & TRIBBLE 2006): WORDS AND TEXTS

Sense 2 is statistical: keywords are words which are significantly more frequent in a sample of text than would be expected, given their frequency in a large general reference corpus. This concept is extensively discussed by Scott and Tribble (2006), who are very explicit about the merits and limitations of Scott's keywords software (part of the package WordSmith Tools: Scott 1998).

(1) First, "keyness is a textual matter" (p.65). Certain words characterise individual texts (such as *Romeo and Juliet*), as well as text-types and intellectual areas (such as medicine and natural science, p.29). (2) Second, the software turns texts into word-lists or lists of n-grams, and then compares the lists from different text collections. By filtering and sorting the lists, vast quantities of text are reduced to much simpler patterns (p.viii, 5, 40), which are invisible to the naked eye. (3) Third, content words directly indicate the propositional content of texts. However, although "keyness is a textual matter", since the texts have been ripped apart into lists of individual words and/or n-grams, the patterns ignore text segmentation. They are a feature of global textual cohesion, but not textual structure: unless the technique is adapted in various (minor) ways, which I will mention below.

I can illustrate some characteristics of this approach with two small case studies.

Example 1: Textual collocates, semantic fields and phraseology

The transcripts of the Hutton Inquiry are available in the world-wide-web [1]. This was an inquiry, chaired by Lord Hutton, into the death in 2003 of Dr David Kelly, a British government weapons inspector during the run-up to the war in Iraq. I compared the transcripts (around 930,000 running words) with the BNC as a reference corpus. As always, amongst the content words in the top 50 keywords were several proper names, plus words which clearly indicate some main topics of the transcripts. If you have forgotten the case, these keywords will remind you of major themes:

- Kelly, Dr, Hutton, Gilligan, Lord, Kelly's, BBC (and other proper names)
- dossier, intelligence, MOD, think, source, FAC, press, JIC, ISC, July, meeting, statement, evidence, committee, draft, September, letter, name, inquiry, document [2]

However, one gains a much better impression of both the content and of the formal nature of the discourse, via longer recurrent expressions. Here are some of the most frequent 4-grams, which all occur 125 times or more, in descending frequency [3].

- the Ministry of Defence; the Foreign Affairs Committee; weapons of mass destruction; the 45 minutes claim; the Joint Intelligence Committee; Intelligence and Security Committee
- I do not think; can I take you; thank you very much; I do not know; I think it is; I think it was; I am not sure; in relation to the

In order to capture characteristics of the discourse, we need this phraseology. The most frequent 5-gram (*can I take you to*) is part of the polite, formal, cautious, public usage of Lord Hutton himself. It occurs in utterances such as:

- can I take you to MoD/1/44 page 19. It is a memo ...

Example 2: Textual collocates, semantic fields and point of view

Friedrich (2007) assembled a 30-million-word corpus of newspaper articles, published between 1996 and 2006, in British periodicals such as *The Spectator*, quality papers such as *The Times*, and tabloid papers such as the *Sun.* He selected all articles which contained the word-forms *islam, muslim/s* and *middle east*. Using the BNC as a reference corpus, keywords in the top 50 in all three groups of newspapers included:

- *all sub-groups*: Iraq / Iraqi, war, Israel, Saddam, Palestinian, terrorism/t

Some individual papers were clearly distinguished by keywords in the top 50. For example, within their sub-groups, keywords in either *The Business* or *The Morning Star*, but not both, were:

- *Business*: oil, global, prices, investors, companies, economy, opec, markets, growth
- *Morning Star*: killed, troops, attacks, military, occupation, forces, soldiers

The technique can provide empirical evidence of how the same topic can be represented from different points of view. It is not surprising that the business periodical sees the topic from an abstract financial point of view, and that the tabloid sees the topic from a more concrete personal point of view of the soldiers who are involved, but introspection could not accurately predict the lexis used to create these representations.

Usually collocates are extracted from an immediate span of a few words to left or right of a node word, but a keywords analysis extends the span from phrases to texts. We are dealing here with what Mason and Platt (2006) call "textual collocates". This concept of keywords recalls classic structuralist semantics. The software can automatically extract sets of words, which fall into intuitively identifiable semantic fields, and provide an indication of how homogeneous the vocabulary is across a text. The meaning of a word derives, not directly from the relation between words and their denotation in the world, but from internal relations to other words. Again, it is clear that work on keywords implies a constructivist position.

Summary of Scott's approach

A major merit of Scott's approach to keywords is that it provides an empirical discovery method, based on frequency and distribution. Lists of keywords and phrases obviously differ in different text collections. Semantically related keywords are a good indicator of propositional content. Salient words and n-grams are identified, but not more abstract phraseology. Keywords are a mechanism of global textual cohesion, but for information about text structure, we have to adapt the technique. For example, Starcke (2007)

identifies keywords in a Jane Austen novel, then looks at their uneven distribution in the text, and uses this as evidence of text segmentation.

KEYWORDS SENSE 3 (FRANCIS 1993): PHRASES AND SCHEMAS

Although she does not use the term "keywords", Francis (1993: 155) proposes a third potentially more radical and ambitious approach to culturally significant units of meaning. She aims to identify what people regularly talk about: their conventional ways of expressing their shared values, such as "how difficult or easy life is made for us, how predictable things are, and how well we understand what is going on" (p.141). One example of the conventional phraseology which speakers use to express their lack of understanding is illustrated in the following concordance lines:

```
                cos I haven't the faintest idea who they are.
                  you haven't the faintest idea how to spell the word
                     I hadn't the faintest idea of his intention
                  but they had no faintest notion of man's perverse habit
                  If you have the faintest notion you might like it
               does not have the slightest idea that the size, shape and ...
                 we have not the slightest notion which parts are effective
                  I have only the foggiest idea of what is entailed.
      none of those would have the foggiest notion about it
      some of you may not have the foggiest what a fanzine is
"Don't know, can't say, haven't the foggiest," said Joe.
             I've no idea! Not the foggiest!
```

The meaning is expressed, not by individual words, but by a variable lexico-grammatical pattern, in which no single word-form is essential. The lemma HAVE is followed by one of a small number of superlative adjectives (*faintest, foggiest, slightest*), plus a word for 'idea', usually *idea* or *notion*, or one of a few other nouns (*she hadn't the faintest doubt, no more than the faintest inkling*). With *foggiest*, even the noun is optional. The whole unit is either negative or hypothetical or contains a word such as *only*.

Phrasal units and cultural schemas

There are other phrasal units which express the familiar experience of 'incomprehension' or 'exasperation'. For example, the 'CANNOT FOR THE LIFE OF ME' construction expresses irritation at not being able to understand something, often other people's unreasonable behaviour [4]. (The construction can have other pronouns and other psychological verbs.)

```
   I can't for the life of me understand what it is you see in it
   I can't for the life of me see what motive any of them can have
   I can't for the life of me see what that's got to do with you
  I cannot for the life of me see why children have to take so long
  I cannot for the life of me see why they're so resistant to it
I couldn't for the life of me see what the old git was moaning about
```

Similarly, the 'WHAT'S X DOING Y?' construction encodes 'unexpectedness' or 'incongruence' (Kay and Fillmore 1999: 4).

```
what's that  doing in there? get it out!
what's he    doing here at this time of night?
what's he    doing phoning this time of day?
what's she   doing with a young man like that?
what's Ellie doing all dressed up in Mother's clothes?
```

A major challenge for corpus lexicography is to discover such abstract extended phrasal units with their variants, and many other analyses have now followed the same approach, working bottom-up in order to discover constructions which express strong evaluative meanings: little schemas of cultural knowledge. These meanings are not evident to introspection, and are therefore not discussed in traditional speech act theory. They have to be discovered by empirical corpus analysis. Sinclair (1998) analyses the 'WOULDN'T BUDGE' construction, which signals frustration at trying to get something or someone to move, and failing. Channell (2000: 47-50) analyses the 'PAR FOR THE COURSE' construction, which signals that things have gone wrong, yet again, in just the way that you would have predicted. Stubbs (2007) analyses the 'NOT THE END OF THE WORLD' construction, which signals reassurance and sympathy for someone who has suffered some disappointment. Here are some attested examples:

- I tried to persuade him out of it but he wouldn't budge
- tough working conditions are often par for the course in catering
- it's disappointing but it's not the end of the world

Cameron and Deignan (2006) discuss the emergence of the phrase *emotional baggage*, which has stabilized recently as a preferred way of expressing a culturally shared schema. The phrase expresses "a negative view of past emotions and memories" (2006: 679). The following examples are from the BNC.

- She couldn't help but sound cynical. "... It won't help, you know. You'll still carry that emotional baggage with you wherever you go, wondering what you said or maybe didn't say that frightened him off."
- No, he isn't the cause of my nightmares. My past is. It's the emotional baggage I'm hauling around that's causing all the trouble.

This use extends the tendency, over hundreds of years, for *baggage* to be used metaphorically in critical and pejorative ways:

- she was afraid the old baggage was going to start asking awkward questions (BNC)
- corpus linguistics offers a fresh start without the baggage that has accumulated over the years (Sinclair 2004a: 185).

Cameron and Deignan (2006: 678) take an explicitly constructivist view and regard this as a case of "an expression becoming fixed and a concept becoming delineated". This conventional way of expressing a critical opinion provides a label which does not refer to an objective thing in the external physical world, but to a subjective perception of the social world. The world is full of entities which exist only because speakers think they exist.

Work on speech acts is best known to linguists in Searle's (1969) version. However, the kinds of acts discussed in Searle's puritan and analytic world are invented, and often rather trivial: speakers ask each other to pass the salt, open the window and take out the garbage. By working bottom-up, a major discovery of corpus study has been extended lexical items which express much more complex and subtle evaluative speech acts. Unfortunately we are all familiar with the experience of irritation or incredulity when things go wrong, and we have conventional ways of expressing our feelings, but semantic categories such as 'incomprehension', 'exasperation', 'indignation' 'frustration' or 'world-weary despair' do not generally appear in descriptions of English.

Example: phrasal units around verbs of perception

Two aspects of this work require development. First, no-one quite knows how to talk about such linguistic and conceptual units, and various metaphors are used. Units are said to crystallize, emerge, reify and stabilize. Second, a list of isolated examples does not amount to a theory. We have the problem again of how to provide comprehensive coverage of the vocabulary of the language. An intermediate step would be to study a well-defined semantic set. For example, the five verbs of sensation and perception, FEEL, HEAR, SEE, SMELL, TASTE, have distinct grammatical and phraseological characteristics. They are not normally used in continuous *-ing* forms, and they are often preceded by a modal (*can* or *could*) which adds little or nothing to the meaning of the main verb (F. Palmer 1974: 117).

In addition, their literal meanings are "marginal to the way in which they are used" (Sinclair 2004b: 278). FEEL is frequently used in conventional expressions of disagreement (*I don't feel like it*) and other non-literal expressions (*get a feel for it*). HEAR is frequently used in conventional expressions of politeness in closing a letter (*I look forward to hearing from you*), or regret (*I'm sorry to hear*), or of rejecting an offer of help (*I wouldn't hear of it*), or to mean 'have recently learned' (*I hear he's got a new job*). SEE is most frequently used in the meaning 'understand' (*if you see what I mean*), and in conventional expressions of tentative agreement (*I don't see why not*) and surprise-cum-complaint (*I've never seen anything like it*). SMELL has many non-literal uses (e.g. *to smell a rat, smelling of roses*). TASTE is certainly used literally, but often in one text-type: recipes. Otherwise, it frequently means 'the ability or lack of ability to judge what is appropriate' (*good taste, bad taste, in poor taste, in the worst possible taste*), or in the extended sense of 'experience' (*a taste of things to come, his first taste of freedom*), or in idioms (*a taste of their own medicine*).

Several uses of FEEL provide further conventional ways of expressing uncertainty or tentativeness or surprise and/or lack of understanding of something:

```
But I can't help feeling that there must be more to it than that
But I can't help feeling the situation has intensified somewhat
but I can't help feeling the future looks black
But I can't help feeling that this is strictly a US product
But I can't help feeling we've missed all the really vital ones
yet I can't help feeling just a bit sorry for the Italians
```

```
          I got the feeling something was going on. It was too quiet.
          a strange feeling ... that something was going to happen
     an overwhelming feeling that we must do something to stop it
         There was a feeling in his bones that something was ...
I suppose it was the feeling that we were doing something secret
```

That is, we have a set of high frequency words, whose primary use does not correspond to their literal meaning. Native speakers cannot generate comprehensive lists of such uses from introspection, but they immediately recognize them, in concordance lines, as preferred and conventional ways of expressing evaluative pragmatic meanings. So, we need a more systematic method of discovering these longer expressions, their forms and their non-compositional meanings. (Stubbs 2007 provides a case study).

Summary of Francis' approach

In summary: This third approach is corpus-driven [5]. It holds the promise of being able to discover speech acts and cultural schemas which are entirely missed by the introspective data used in speech act theory. The examples from Francis and others illustrate one of the major theoretical proposals to come out of corpus studies: what Sinclair (1996, 1998, 2005) has called extended lexical units. In contrast to Williams' examples, which often come from intellectual discourse, Francis' examples are from everyday sociolinguistic acts. These examples now have to be approached from both ends: What are the non-compositional meanings of expressions which contain frequent words (such as FEEL)? And what are the conventional ways of conveying frequent speech acts (such as the expression of incomprehension or uncertainty)?

PART 2: THE DUALISM OF AGENCY AND STRUCTURE

So far, I have discussed three approaches to keywords, with their strengths and weaknesses. The three approaches share the general idea that certain words and phrases convey meanings which are socially important, but they come out of very different traditions (cultural studies, quantitative corpus analysis and lexico-grammar), they are only loosely conceptually related, and perhaps only marginally compatible. Williams relates individual keywords and cultural concepts, Scott extracts sets of keywords from texts, and Francis shows how conventionally phrased speech acts express widely shared everyday values. What is missing is an explicit model of the relation between phraseology, speech acts, texts and text-types, and social institutions.

Institutions are abstract structures, which change historically, but which typically exist over long periods of time. They nevertheless depend on their constituent speech events, which in turn depend on speech acts which last for only seconds. The problem is to relate things of very different scales in time and space, and the main ontological nightmare involves the dualism of agency and structure. Trying to work out the ontology of social institutions has provided "problems in the foundations of the social sciences" (Searle 1995: xii) for around 150 years. Searle (1995: 3-4) admits that he can hardly bear the metaphysics involved in ordering a beer in a French café (well, that's his problem, but you can see his theoretical point). You sit down at a table, and utter a conventional French sentence, which expresses a predictable speech act; a waiter brings a beer, you drink it, utter another conventional speech act or two, pay the bill and leave.

It hardly occurs to you that you must know about speech acts (such as question and request), and also about money, property, ownership (the waiter doesn't own the beer but he sells it to you), and so on. These aspects of social reality are not physical facts, but seem just as robust and objective (Searle 1995: xi and passim, Collin 1997, Miller 2007).

Although we tend to think of institutions such as Parisian cafés, churches, universities and all the rest as places, they are better regarded as placeholders for patterns of activities (Searle 1995: 57). This is what gives us the opening we need to relate language use and social institutions. Social institutions and speech events are the same thing looked at from different points of view [6]. But since Searle does not use any empirical language data, he does not discuss texts and text-types. Corpus studies can fill this gap, by providing empirical data on the lexical patterns which make up phrases, texts and text-types. And text-types make up the activities which are referred to in a shorthand way as universities, churches and all the rest. Corpus studies can document how such an emergent model works.

TEXTS IN SOCIETY: SOME VERY BANAL OBSERVATIONS

The following remarks are hardly original, but they are, I hope, are obvious. That is, I hope you agree that they are simply obviously true and even banal, because I will just state them, then take them for granted, and use them to introduce some points which are certainly not obvious, in so far as they have provided problems in the foundations of the social sciences for the last hundred and fifty years or so, and which remain unsolved. I will then propose that corpus methods can provide a new way of looking at these old puzzles.

There is an inherent and logical relation between social institutions, the professionals who work in them and their clients, and the language which is used there. The rough idea can be illustrated as follows.

Scientists	write research papers	for their peers	in specialist journals.
Professors	give lectures	to students	in universities.
Preachers	give sermons	to congregations	in churches.
Doctors	give consultations	to patients	in doctors' surgeries.
Employers	ask questions	of potential employees	in job interviews.
MPs	give speeches	to other MPs	in parliament.
Journalists	write editorials	for readers	in newpapers.
Judges	pass sentences	on the accused	in courtrooms.

The social roles are interdependent: a preacher cannot give a sermon unless there is (at least) a (potential) congregation; a doctor can only give a consultation to a patient; there are no judges without defendants. And just to make this stunningly banal point even more obvious, you don't find other combinations of speakers, speech events and social institutions. You don't find scientists giving sermons to patients in job interviews. We have what seem like natural rules which determine which combinations are possible or not. These rules are not regulative: it is not that scientists go around giving sermons and

other people tell them not to. They are constitutive rules which define how society works.

It would be a major undertaking to make such a list anything like comprehensive. However, the list could not be continued indefinitely, because the social world is structured around such speech events. Indeed, if we could list and describe all such speech events, we would have defined the culture. A Hallidayan formulation would be that social reality is an edifice of meanings; the network of meaning potential is what we call the culture. There is a range of social institutions, which are staffed by professionals. Part of their professional communicative competence requires them to engage in particular speech events, with their peers and with their clients. These speech events can be described in terms of their conventionalized speech acts, which are realized by words, phrases, discourse structure and so on.

The model seems most obviously applicable to public and professional spheres, but similar statements can also be made about private spheres: members of the family, friends, acquaintances and neigbours chat, argue and gossip with each other, and express their frustration and incomprehension (so, of course, do doctors and professors). One main difference to the public sphere is that anyone can be a neighbour, but not anyone can be a doctor. Another difference is that professionals may be authorized and/or required to carry out certain speech events (e.g. give lectures every Monday morning). It may be advisable to wish your neighbours "good day" when you see them, but you do not have to, and you do not require special authorization to do so.

When I put these points to a group of students recently, one student objected that the model doesn't work with, for example, novelists. But it seems to me that the student had in mind a rather dubious notion of the free and creative artist, who works outside social conventions. Novelists write novels for readers who have bought their work, and this is only possible within a set of conventions and institutions. The whole institution of literature, in the sense of fiction, is historically quite recent: for example, *Robinson Crusoe* is often discussed as the first novel. And literature relies on other institutions, including publishers, not to mention university courses which define the conventional canon.

A more detailed discussion would have to distinguish between at least three senses of the concept of social institution. For example, with respect to medicine, it can refer to a specific place (e.g. hospital, clinic, surgery, etc), or to a whole social system (e.g. the National Health Service) or to a body of knowledge or an academic subject. (This third sense is the one assumed by Scott & Tribble 2006: 82-3.) The same applies to other institutions such as the law and theology, in which people receive a professional training. We therefore find educated groups who have the communicative and textual competence to act in these areas, including command of a technical vocabulary and its phraseology and speech acts. It is the established members of an occupational group or discourse community who create and maintain the genres (Swales 1990). Distinctions are also necessary between different professionals and lay persons (e.g. junior doctor, consultant, nurse, surgeon, patient, etc), and between different speech events (e.g. consultation, diagnosis, case history, prescription, etc). We might also have to identify

specific texts (e.g. the Hippocratic Oath; and an oath is, of course, a speech act). But in principle, we could construct a list such as the following:

In religious institutions, including (Christian) churches (Catholic, Church of England, etc), priests give sermons and baptize and marry people, and together with members of the congregation, they sing hymns, say prayers, and so on, often using specific texts (such as the Bible, the Book of Common Prayer, the marriage ceremony, etc), to which explicit reference is often made in conventional phraseology (*Our reading today is taken from* ...). A church service (speech event) consists of a sequence of hymns, sermon, prayer, announcements (and other speech acts).

In a legal institution, the courtroom, witnesses and defendants swear to tell the truth, they are cross-examined by barristers, and the judge gives a summing up. It is at this level of speaker roles, speech events and speech acts that people understand such social events. Here, for example, is a short extract from the UK Government site on the Criminal Justice System in England and Wales, which is designed to explain to potential jurors their responsibility and what happens in court [7]. Almost the whole extract consists of explaining the relation between the participants, the speech events and the speech acts, which I have underlined.

> "Once all 12 jurors are in the jury box ... the court clerk will call out each name and each member of the jury will be sworn in. They must either take an oath on a holy book of their choice or they must affirm. This is similar to swearing in, but without the holy book. ... All criminal trials follow similar procedures. A defendant or number of defendants will have been accused of a crime. The prosecution advocate opens the case by explaining the accusations and setting out the facts they will seek to prove during the trial. Witnesses for the prosecution will be called. They take an oath, or affirm, to tell the truth and are then questioned and cross-examined. Next, witnesses for the defence may be called. If they are, they too will take the oath, or affirm, and be questioned and cross-examined. ... When all the evidence has been given to the court, the prosecution and defence advocates may make their closing speeches. They will talk directly to the jury as they argue their cases. The judge will explain the law and summarise the facts of the case. Then he or she will clarify the duties of the jury before they go to the jury room to consider their verdict."

Much of the phraseology in churches and courtrooms is largely fixed, for traditional and/or legal reasons. This is less so in many other institutions, but the same points hold in general. Media institutions (e.g. radio, television, newspapers) have presenters, news readers, journalists and so on, who take part in and/or write news broadcasts, documentaries, editorials and so on. Scientific and technical institutions (e.g. universities, research laboratories) have scientists and researchers who write lab reports, publish articles and so on.

There are few studies which use empirical corpus data to show the micro-macro relations across this range between the linguistic features of text-types and social institutions. However, Atkinson (1999) provides an exemplary case study of the development of a scientific journal and its research articles (text-type) within the Royal

Society of London (social institution). He used a sample of texts from seven 50-year intervals between the late 1600s and the late 1900s, in order to study the development of an area of institutionalized knowledge along with its changing norms of language use. The discourse community of The Royal Society is its scientists, who were initially gentlemen amateurs, and only later the professional scientists which we know today. Their journal was *The Philosophical Transactions of the Royal Society of London*. The content of the journal evolved from the form of polite letters, often narratives of observations, which relied on the trust and authority of their gentlemen authors, into experimental reports, with a highly conventionalized structure of theory-methods-discussion, with explicit sub-headings, which is so familiar in modern scientific articles. As Atkinson (1999: xvi) puts it, "the evolution of these forms of meaning" from person-centred to object-centred discourse, was "an integral part of the changing form of scientific life".

Atkinson studies things both top-down and bottom-up (p.56). He studies how a particular text-type developed over time within the institution: this gives the macro perspective. And he studies the rhetorical features of individual texts: this gives the micro perspective. In addition, he uses both qualitative and quantitative methods: traditional rhetorical analysis of the discourse organization of the articles, and also multi-dimensional analysis of the significant co-occurrence of conventionalized linguistic features. [8]

PUZZLES OF SOCIAL THEORY

We could not operate in the social world if we did not take such things for granted. A social theory of language has to deal with the following relations:

(1) Objective behaviour and subjective meaning. For example, sitting down and drinking a beer versus entering a café, ordering the beer and paying for it. For most of the time, human beings cannot be wrong about what they are doing. If you go into a café and issue a speech act, such as ordering a beer, you must understand what you are doing, you must be doing it intentionally, and if you do not do it in conventional ways you risk causing, at least, confusion.

(2) Scales of time and place. We need a model which explains the relations between large macro structures which exist over long periods of time (social institutions) and small micro events which last a few moments (speech acts and their conventional idiomatic expressions).

(3) Structure and agency. The model must also explain how structure and agency interact. They seem to be different kinds of things: one is not reducible to the other, but both are real.

(4) Cause and effect. You can become a doctor only if the required social institutions exist. These institutions create expectations, but it is the agency of the participants, including their speech acts (e.g. giving advice) and speech events (e.g. a medical consultation) which creates and maintains the institutions. The conventions and the institutions are independent of any individual person's activities, but dependant on the

cumulative behaviour of the speech community. Corpus methods are good at describing recurrent behaviour across groups of speakers.

The overall model combines structure, knowledge and agency. The social institutions provide the structure. The speakers have the knowledge (communicative competence). The speech events and speech acts are the intentional behaviour of the agents. The conventional phraseology is part of the linguistic system. The evidence for the speech acts and their phraseology is the recurrent textual traces which can be studied in large corpora.

The social institutions are abstract structures, which depend on agency. A university is something which happens. It exists because teachers and students engage in particular kinds of language behaviour which creates conventional social relations between them. In addition, part of being a university, a church, a court of law and so on is being thought to be one, and these social institutions have had their statuses assigned to them by constitutive rules (Searle 1995: 34, 50).

The professionals in such institutions are people with the communicative competence to utter the appropriate speech acts in the conventional way in the required speech events. In addition, they are social groups of a particular kind (Sealey & Carter 2004: 111). You can only belong to a category such as doctor or priest intentionally, and you have to be authorized to have such a status [9]. Having such a status confers rights, obligations and powers, including the power to issue certain speech acts: professors can fail students, priests can marry people, employers can employ you.

The text-types often have everyday names (e.g. church sermon, news broadcast), and these things are also intentional events. You cannot give a church sermon without intending to. But there is no systematic and comprehensive classification of such things. Various classifications of text-types have been proposed, but they seem only to work as general dimensions or ideal types, based around communicative functions such as informative or persuasive, or around structural types such as descriptive, argumentative or narrative. Most real texts are mixed.

Having proceded top-down, we now arrive back at texts, which are the only thing which corpus linguists can observe: the traces which are left by these activities. Technology has profoundly changed the traces which people leave and how these traces can be analysed.

I assume that Searle's view of these things is basically correct. He proposes that language plays a special role in institutional reality, and that the social world has a hierarchic structure which explains the relation between speech acts and social institutions. A certain kind of speech act is a promise. A certain kind of promise is a contract. A certain kind of contract is a marriage vow. Only certain kinds of people are authorized to perform marriage ceremonies, because they themselves have entered into other kinds of contracts and have been authorized by other people who are authorized to authorize them. Though, oddly enough, I'm not sure whether Searle says explicitly that different kinds of promises and contract formation are the most basic trait which pervades social behaviour (Wilson 1998: 189).

Although social institutions are complex, they are the result of simple rules which are applied recursively. This is Searle's (1969) proposal: speech acts depend on constitutive rules and social reality is constructed recursively with speech acts at the bottom level (Searle 1995). However, since Searle does not use any empirical language data, he does not discuss texts and text-types. Corpus studies can fill this gap by providing empirical data on the lexical patterns which make up phrases and texts. The predictable co-occurrences of patterns make up text-types, and text-types make up the activities which are referred to in a shorthand way as universities, churches and all the rest. Corpus studies have shown how such an emergent model can work, and corpus methods provide the possibility of describing complex systems by tracing causation across many levels (Wilson 1998: 207), from phraseology and speech acts to social institutions.

CONCLUDING COMMENTS

This article is in two distinct parts: part 1 was about keywords and argued bottom-up; part 2 was about social institutions and argued top-down. This is the problem with the concept of keywords: the large gap between individual words and the social world. The term is used in different senses, which are related in only a loose way. It may be a productive concept, but it cannot stand on its own. It assumes other concepts, such as cognitive schema, textual collocates and semantic fields, text and text-type. Concepts of text and text-type in turn imply the concept of social institutions.

So how do we get back to keywords? Quantitative corpus data provide evidence of semantic units (extended phrasal units) and thereby extend the empirical basis of speech act theory. This gives us the basis of a theory of language as social action, of the relations between language use and language system, and of the relations between phraseology, texts, text-types and social institutions. There is a series of questions in linguistics which are all logically the same question. How does something arise from nothing? How do extended phrasal units of meaning arise from recurrent collocations? How do social institutions arise from recurrent speech events? How do structures arise from agency? How does the macro arise from the micro? How do the properties of whole systems emerge? The answer is: by recursive application of constitutive rules. Social institutions and text-types imply each other: they are different ways of thinking about the same thing.

Speech act theory asks the right questions, but does not have the data or methods to answer them. It tries to do ordinary language philosophy without attested data on ordinary language. Corpus linguistics has the data and the methods, but has not yet co-ordinated studies in a way which can answer cognitive and social questions. It has not yet moved from description to explanation. If this line of argument can be worked out successfully, it will show how corpus data and methods can help to solve puzzles in the foundations of the social sciences.

ACKNOWLEDGEMENTS

For valuable comments on a much earlier version of this paper, I am grateful to Kieran O'Halloran and to lecture audiences in Italy in June 2007, at the Catholic University of Milan and at the Certosa di Pontignano, University of Siena.

NOTES

[1] The transcripts may be reproduced free of charge providing that Crown Copyright is acknowledged. They are available at http://www.the-hutton-inquiry.org.uk/index.htm (accessed July 2007). For help in constructing the corpus I am grateful to Simone Dausner.

[2] *MOD* is the Ministry of Defence, *JIC* is the Joint Intelligence Committee, *FAC* is the Foreign Affairs Committee, *ISC* is the Intelligence and Security Committee.

[3] The n-grams were extracted with Bill Fletcher's software kfNgrams, available at http://www.kwicfinder.com/kfNgram/kfNgramHelp.html (accessed July 2007).

[4] I am grateful to Katrin Ungeheuer for pointing out this construction to me.

[5] As far as I can determine, Francis (1993: 137, 139) is the first reference in print to a "corpus-driven" and "data-driven" approach to linguistic analysis. Francis distinguishes this approach from the use of a corpus merely to find examples for a theory which has been independently formulated.

[6] This way of thinking about things has been proposed by Halliday in his analogy of the relation between weather and climate (Halliday 1991).

[7] The website is at http://www.cjsonline.gov.uk/juror/the_trial/index.html (accessed 15 December 2007). A related web-page lists other participants and the speech acts they perform, including Clerk, Crown Prosecution Service Representative, Expert Witness, Probation Representative, Usher, Defence Representative, Dock Officer.

[8] Miller (2007) points out that the Alan Sokal case shows how formal features of a text-type may fool some of the people some of the time, but cannot fool all of the people all of the time. Sokal, a professor of physics, wrote an article which had all the stylistic features of an academic discussion of post-modernism, and submitted it to a peer-reviewed journal, where it was published (Sokal 1996). These aspects of the world (academic journal, peer-reviewing, etc) are institutional facts. However, as Sokal then admitted, the content of the article was gibberish and he had written it as a hoax. Style had triumphed over content, but only in the short term.

[9] This is quite different from other categories which sociologists use, such as "unemployed men over 50". It is also quite different from a category such as "patient" or "member of the congregation" (which anyone can be), and from a social category such as "adolescent" (which is not voluntary and has no necessarily associated conventions, even if we have a stereotype of how adolescents behave). It is because

these statuses have to be authorized, that people can pretend to be doctors or professors when they are not so authorized (Searle 1995: 48).

REFERENCES

Atkinson, D. 1999. *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. London: Lawrence Erlbaum.

Benveniste, É. 1954. Civilisation: contribution d'un mot. Reprinted in *Problèmes de linguistique générale*, 336-45. Paris: Gallimard, 1966.

Cameron, L. & Deignan, A. 2006. The emergence of metaphor in discourse. *Applied Linguistics*, 27, 4: 671-90.

Channell, J. 2000. Corpus-based analysis of evaluative lexis. In *Evaluation in Text*, S. Hunston & G. Thompson (eds), 38-55, Oxford: Oxford University Press.

Collin, F. 1997. *Social Reality*. London: Routledge.

Firth, J. R. 1935. The technique of semantics. *Transactions of the Philological Society*. 36-72.

Firth, J. R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, 1-32. Special Volume, Philological Society. Oxford: Blackwell.

Francis, G. 1993. A corpus-driven approach to grammar: principles, methods and examples. In *Text and Technology*, M. Baker, G. Francis, & E. Tognini-Bonelli (eds), 137-56. Amsterdam: Benjamins.

Friedrich, F. 2007. *A Corpus-Based Study of British Newspapers*. Unpublished Staatsexamensarbeit, Universität Trier.

Hacking, I. 1986. The archaeology of Foucault. In *Foucault: A Critical Reader,* D. C. Hoy (ed.), 27-40. Oxford: Blackwell.

Halliday, M. A. K. 1991. Corpus studies and probabilistic grammar. In *English Corpus Linguistics*, K. Aijmer & B. Altenberg (eds), 30-43. London: Longman.

Herberg, D., Steffens, D. & Tellenbach, E. 1997. *Schlüsselwörter der Wendezeit*. Berlin: De Gruyter.

Kay, P. & Fillmore, C. J. 1999. Grammatical constructions and linguistic generalizations: the what's x doing y? construction. *Language*, 75, 1: 1-33.

Ladendorf, O. 1906. *Historisches Schlagwörterbuch*. Berlin: K. J. Trübner.

Lepp, F. 1908. *Schlagwörter der Reformationszeit*. Leipzig.

Mason, O. & Platt, R. 2006. Embracing a new creed: lexical patterning and the encoding of ideology. *College Literature*, 33, 2: 155-70.

Matoré, G. 1953. *La méthode en lexicologie*. Paris: Didier.

Miller, S. 2007. Social institutions. *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/social-institutions. [Accessed 10 December 2007.]

Palmer, F. 1974. *The English Verb*. London: Longman.

Scott, M. 1998. *WordSmithTools*. Version 3. Oxford: Oxford University Press.

Scott, M. & Tribble, C. 2006. *Textual Patterns*. Amsterdam: Benjamins.

Sealey, A. & Carter, B. 2004. *Applied Linguistics and Social Science*. London: Continuum.

Searle, J. R. 1969. *Speech Acts*. Oxford: Oxford University Press.

Searle, J. R. 1995. *The Construction of Social Reality*. London: Allen Lane.

Sinclair, J. 1996. The search for units of meaning. *Textus*, 9, 1: 75-106. Also in Sinclair 2004a, 24-48.

Sinclair, J. 1998. The lexical item. In *Contrastive Lexical Semantics*. E. Weigand (ed), 1-24. Amsterdam: Benjamins. Also in Sinclair 2004a, 131-48.
Sinclair, J. 2004a. *Trust the Text*. London: Routledge.
Sinclair, J. 2004b. New evidence,new priorities, new attitudes. In *How to Use Corpora in Language Teaching*, J. Sinclair (ed.), 271-99. Amsterdam: Benjamins.
Sinclair, J. 2005. The phrase, the whole phrase and nothing but the phrase. Unpublished plenary lecture, *Phraseology 2005*, Louvain-la-Neuve, October 2005.
Sinclair, J. 2007. Introduction. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert *Text, Discourse and Corpora*, 1-5. London: Continuum.
Sokal, A. 1996. Transgressing the boundaries: toward a transformative hermeneutics of quantum gravity. *Social Text*, 46-47 spring/summer 1996: 217-52.
Starcke, B. 2007. *Korpusstilistik: Korpuslinguistische Analysen literarischer Werke am Beispiel Jane Austens*. Unpublished PhD thesis, Universität Trier.
Stötzel,G. & Wengeler, M. 1995. *Kontroverse Begriffe: Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik*. Berlin: de Gruyter.
Strauß, G., Hass-Zumkehr, U. & Harras, G. 1989. *Brisante Wörter von Agitation bis Zeitgeist*. Berlin: De Gruyter.
Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.
Stubbs, M. 2007. Quantitative data on multi-word sequences in English: the case of the word *world*. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert *Text, Discourse and Corpora*. London: Continuum. 163-89.
Swales, J. M. 1990. *Genre Analysis*. Cambridge: Cambridge University Press.
Teubert, W. 1989. Politische Vexierwörter. In *Politische Semantik*, J. Klein (ed.), 51-68. Opladen: Westdeutscher Verlag.
Teubert, W. 2007. *Parole*-linguistics and the diachronic dimension of the discourse. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert *Text, Discourse and Corpora*, 57-87. London: Continuum.
Wierzbicka, A. 1997. *Understanding Cultures through their Key-words*. Oxford: Oxford University Press.
Williams, R. 1976/1983. *Keywords. A Vocabulary of Culture and Society*. London: Fontana.
Williams, R. 1985. Mining the meaning: key words in the miners' strike. *New Socialist*, March 1985.
Wilson, E. O. 1998. *Consilience: The Unity of Knowledge*. London: Abacus: