© Copyright Michael Stubbs 2001.

This file contains the contents pages, front matter and chapter 1 of *Words and Phrases: Corpus Studies of Lexical Semantics*, which was published by Blackwell in October 2001.

WORDS AND PHRASES: CORPUS STUDIES OF LEXICAL SEMANTICS

MICHAEL STUBBS

FRONT MATTER

Figures, Concordances and Tables
Acknowledgements [omitted here]
Data conventions and terminology
Notes on corpus data and software

INTRODUCTION

CHAPTER 1. WORDS IN USE: INTRODUCTORY EXAMPLES

1.1 Text and discourse: some distinctions
1.2 Language, action, knowledge and situation
1.3 Words and expectations
1.4 Language, logic and truth
1.5 Common-sense knowledge
1.6 Linguistic conventions
1.7 Possible and actual
    1.7.1 Example 1: the ambiguity of SURGERY
    1.7.2 Example 2: the (non-)ambiguity of BANK
    1.7.3 Example 3: the days of the week
    1.7.4 Example 4: lonely hearts ads
1.8 Summary and implications
1.9 Background and further reading
1.10 Topics for further study

CHAPTER 2. WORDS, PHRASES AND MEANINGS: BASIC CONCEPTS

2.1 Terminology
2.2 Words: word-forms and lemmas
    2.2.1 Example: the lemmas CONSUME and SEEK
2.3 Collocation
2.4 Words and units of meaning
2.5 Delexicalization
2.6 Denotation and connotation

2.7 Relational lexical semantics
    2.7.1 Semantic fields
    2.7.2 Synonyms, antonyms and hyponyms
2.8 Frequent and less frequent words
    2.8.1 Content and function words: lexical density
    2.8.2 Core vocabulary
2.9 Two examples
    2.9.1 Example 1: Bloomfield's analysis of SALT
    2.9.2 Example 2: CAUSE problems and CAUSE amusement
2.10 Summary and implications
2.11 Background and further reading
2.12 Topics for further study

CASE STUDIES

CHAPTER 3. WORDS IN PHRASES 1: CONCEPTS, DATA AND METHODS

3.1 Background
3.2 Communicative competence
3.3 Corpus methods: observing patterns
3.4 Terminology
3.5 Corpus, concordance, data-base
3.6 The Cobuild collocations data-base on CD-ROM
    3.6.1 The corpus
    3.6.2 The data-base
    3.6.3 Precision and recall
3.7 Data for semantics and pragmatics
3.8 Summary and implications
3.9 Appendix 1: measures of statistical significance
3.10 Appendix 2: further notes on the data-base
3.11 Background and further reading
3.12 Topics for further study

CHAPTER 4. WORDS IN PHRASES 2: A CASE STUDY OF THE PHRASEOLOGY OF ENGLISH

4.1 Frequency of phraseological units
4.2 Strength of attraction: word-forms, lemmas and lexical sets
4.3 Lexical profiles: comprehensive coverage of data
    4.3.1 Example 1: lexical profile for resemblance
    4.3.2 Example 2: lexical profile for reckless
    4.3.3 Example 3: lexical profile for backdrop
    4.3.4 Example 4: lexical profile for doses
4.4 A model of extended lexical units
    4.4.1 Example 5: lexical profile for UNDERGO
    4.4.2 Example 6: lexical profile for chopped
4.5 Summary and implications
4.6 Background and further reading
4.7 Topics for further study

CHAPTER 5. WORDS IN TEXTS 1: WORDS, PHRASES AND TEXT COHESION

- 5.1 Words and co-text
- 5.2 Routine and creativity
- 5.3 Variable phrases and textual cohesion
- 5.4 Antonyms and synonyms
- 5.5 Discourse prosodies
- 5.6 Lexical cohesion: textual examples
  - 5.6.1 Example 1: just large enough to see with the naked eye
  - 5.6.2 Example 2: causing untold damage
  - 5.6.3 Example 3: causing growing pains and undergoing a transition
  - 5.6.4 Example 4: undergoing rapid star formation
- 5.7 Collocations and coherence
- 5.8 Summary and implications
- 5.9 Background and further reading
- 5.10 Topics for further study

CHAPTER 6. WORDS IN TEXTS 2: A CASE STUDY OF A SHORT STORY

- 6.1 Public data and replicable experiments
- 6.2 Lexis and text structure
- 6.3 Analysis 1: frequency statistics (descending frequency order
  - 6.3.1 Frequency of function words: statistics
  - 6.3.2 Interpretation
  - 6.3.3 Frequency of content words: statistics
  - 6.3.4 Interpretation
- 6.4 Analysis 2: frequency statistics (keywords)
- 6.5 Analysis 3: frequency statistics (order of occurrence)
  - 6.5.1 Statistics
  - 6.5.2 Interpretation
- 6.6 Analysis 4: a vocabulary-management profile
  - 6.6.1 Types and tokens, vocabulary and text
  - 6.6.2 Youmans' method
  - 6.6.3 Eveline
- 6.7 A further note on replication
- 6.8 Limitations on the analysis
- 6.9 Summary and implications
- 6.10 Background and further reading
- 6.11 Topics for further study

CHAPTER 7. WORDS IN CULTURE 1: CASE STUDIES OF CULTURAL KEYWORDS

- 7.1 Data and citation conventions
- 7.2 Text and discourse
- 7.3 Case study 1: ETHNIC, RACIAL and TRIBAL
- 7.4 Case study 2: HERITAGE and CARE
  - 7.4.1 Keyword: HERITAGE
  - 7.4.2 Keyword: CARE

  - 7.4.3 Keyword: COMMUNITY
- 7.5 Case study 3: PROPER STANDARDS
  - 7.5.1 Keyword: STANDARD
  - 7.5.2 Keyword: PROPER
  - 7.5.3 Keyword: TRENDY
- 7.6 Case study 4: Little Red Riding Hood
- 7.7 Discursive formations
- 7.8 Summary and implications
- 7.9 Background and further reading
- 7.10 Topics for further study

CHAPTER 8. WORDS IN CULTURE 2: CASE STUDIES OF LOAN WORDS IN ENGLISH

- 8.1 Data
- 8.2 The etymological fallacy
- 8.3 Language change
- 8.4 Terminology
- 8.5 Words, politics and national stereotypes
- 8.6 Fields of knowledge and text types
- 8.7 A case study of German loan words in English
- 8.8 Frequency in the vocabulary versus frequency in texts
- 8.9 False friends: flak, blitz and angst
- 8.10 The OED and cultural keywords
- 8.11 A further note on vocabulary and text
- 8.12 Summary and implications
- 8.13 Background and further reading
- 8.14 Topics for further study

IMPLICATIONS

CHAPTER 9. WORDS, PHRASES AND CONNOTATIONS: ON LEXICO-GRAMMAR AND EVALUATIVE LANGUAGE

- 9.1 Connotations
- 9.2 Verbs, discourse prosodies and point of view
  - 9.2.1 Example 1: I was accosted in the street by a stranger
  - 9.2.2 Example 2: fears lurking just below the surface
  - 9.2.3 Example 3: LOITER and other verbs
  - 9.2.4 Inter-collocations: the example of STREET
- 9.3 A lexico-syntactic example: MAKE one's way somewhere
- 9.4 A note on syntax
- 9.5 A cognitive view
- 9.6 A syntactic example: BE-passives and GET-passives
- 9.7 Summary and implications
- 9.8 Background and further reading
- 9.9 Topics for further study

CHAPTER 10. DATA AND DUALISMS: ON CORPUS METHODS AND PLURALIST MODELS

- 10.1 Principles
- 10.2 Problems?
- 10.3 Dualisms and monisms
  - 10.3.1 Cartesian dualism
  - 10.3.2 Monism: version 1
  - 10.3.3 Monism: version 2
  - 10.3.4 The Saussurian paradox
- 10.4 Pluralist positions
- 10.5 Brute and institutional facts
- 10.6 Physical, psychological and social
- 10.7 Worlds 1, 2 and 3
- 10.8 A pluralist model
- 10.9 Performance data, corpora and routine behaviour
- 10.10 Summary and implications
- 10.11 Background and further reading

FIGURES, CONCORDANCES AND TABLES

FIGURES

- 3.1 Aspects of communicative competence. Based on discussion in Hymes 1972.
- 4.1 The prototypical uses of undergo.
- 6.1 Eveline span 35.
- 6.2 Eveline span 151.
- 10.1 Process and product, potential and actual. Based on discussion in Tuldava 1998: 13ff.

CONCORDANCES

- 2.1 Fifty random examples of CAUSE (verb).
- 4.1 Sample concordance lines for undergo.
- 8.1 Sample concordance lines for angst-ridden and teen(age) angst.
- 9.1 Fifty examples of GET-passives.
- 9.2 Fifty examples of BE-passives.

TABLES

- 4.1 Positional frequency table for undergo, span 3:3.
- 7.1 Positional frequency table for proper, span 3:3.

DATA CONVENTIONS AND TERMINOLOGY

1. An important feature of the book is that all data which are analysed in detail are attested in naturally occurring language use. Where necessary, the status of examples is indicated as follows:

- [A] attested, actual, authentic data: data which have occurred naturally in a real social context without the intervention of the analyst.
- [M] modified data: examples which are based on attested data, but which have been modified (e.g. abbreviated) to exclude features which are assumed to be irrelevant to the current analysis.
- [I] intuitive, introspective, invented data: data invented purely to illustrate a point in a linguistic argument.

Individual examples are not always marked in this way if their status is clear from the surrounding discussion.

2. Other conventions are as follows.

2.1. Single quotation marks ' ' for technical terms and for quotes from other authors.

2.2. Double quotation marks " " for meanings of linguistic expressions.

2.3. Italics for short forms cited within the text. e.g. The German sentence Sie soll sehr klug sein means "She is said to be very clever".

2.4. CAPITAL LETTERS for lemmas. Alternative terms for lemma include dictionary head-word and lexeme. A lemma is a set of morphological variants. Conventionally the base form of verbs and the singular of nouns are used to represent lemmas. (See chapter 2.)

e.g. do, does, doing, did and done are the forms of the lemma DO.

2.5. Asterisk * for ill-formed sequences, such as ungrammatical or semantically anomalous sentences.

e.g. *She must can come. *He is a vegetarian and eats meat.

2.6. A question mark ? before a form denotes a string of doubtful or marginal acceptability.

e.g. ?he mayn't come.

By definition, asterisked and questioned items cannot be attested, and intuitive judgements on ill-formedness should be treated with care. Corpus data sometimes reveal that forms which are thought not to occur, do occur and are used systematically.

2.7. Diamond brackets <...> enclose attested collocates of a node word. The position of the collocates relative to the node can also be given. See chapter 2.

node <N+1: ... list of collocates>

CAUSE <N+1: problems, trouble, damage>

That is, the lemma CAUSE is often immediately followed, one word to the right, by the word-forms problems, trouble and damage.

NOTES ON CORPUS DATA AND SOFTWARE

The first widely-used computer-readable corpora were set up in the 1960s and 1970s. The Brown Corpus is so named because it was prepared at Brown University in the USA by W. Nelson Francis. It consists of one million words of written American English, published in 1961, and sampled as text fragments of 2,000 words each. Such corpora may now seem rather small and dated, but they are carefully designed and can still provide useful samples of language in use. Many other corpora are now available. In preparing this book, I have used the following.

LOB (Lancaster-Oslo-Bergen) corpus

The LOB corpus was designed as the British equivalent of the Brown corpus: one million words of written British English, also published in 1961, and sampled as text fragments of 2,000 words each, from informative texts, such as newspapers, learned and scientific writing, and imaginative fiction. For a list of textual categories, see ICAME News, 5, 1981, p.4 (= International Computer Archive of Modern English, Bergen), and Biber (1988: 66ff).

FLOB and FROWN

These are the Freiburg versions of LOB and Brown, designed on the same lines as these earlier corpora, as samples of British and American English from thirty years later: material published in 1991. I am grateful to Christian Mair, University of Freiburg, for providing me with the sections of newspaper language.

London-Lund corpus

This corpus was constructed at University College London and the University of Lund. The corpus is about 435,000 words of spoken British English, and contains 5,000-word samples of the usage of adult, educated, professional people, including face-to-face and telephone conversations, lectures, discussions and radio commentaries. For further details, see Svartvik and Quirk (eds 1980), Svartvik et al (1982), and Biber (1988: 66ff).

Longman-Lancaster corpus

This corpus was constructed in collaboration between Longman publishers and the University of Lancaster. It consists of about 30 million words of published English, including fiction and non-fiction: samples from well known literary works and also works randomly sampled from books in print; and non-fiction texts from the natural and social sciences, world affairs, commerce and finance, the arts, leisure, and so on. For some purposes in this book, I have taken 2,000 word samples from 500 files, in order to construct a mixed sub-corpus of one-million words. Otherwise I have taken examples from an 8-million word sample of British English. Summers (1993) discusses the corpus design.

The Bank of English

Many individual examples in this book are drawn from the Bank of English corpus created by COBUILD at the University of Birmingham. COBUILD stands for Collins Birmingham University International Language Database. This corpus has been used in the design of major dictionaries and grammars (including Cobuild 1987, 1990, 1995a, Francis et al 1996, 1998). By the late 1990s, the corpus totalled some 330 million words, including fiction and non-fiction books, newspapers, and samples of spoken English. The corpus is available in different forms: primarily the Bank of English itself, and a 50-million word sub-corpus which is available over the internet as CobuildDirect. I have also used a data-base on CD-ROM (Cobuild 1995b), which was constructed from a 200-million word sub-corpus. (This is described in detail in chapter 3.) Renouf (1987) and Sinclair (1991: 13-26) describe the early corpus development, and Baker et al eds (1993) include several articles based on the corpus.

The British National Corpus

This is a 100-million word corpus of British English, 10 million words of spoken English, and 90 million words of written English. It has been used in the design of major dictionaries: OALD (1995) and LDOCE (1995). For further details, see Aston and Burnard (1998). Simple searches can be done free over the internet.

Newspapers on CD-ROM

The CD-ROM versions of several newspapers can be a convenient source for some kinds of analysis. The text-types are obviously restricted, but perhaps less so than might appear at first sight, given the range of articles which appear in different sections of a major newspaper: not just news and political commentary, but also sports, and, especially in Sunday editions, cultural topics. As a subsidiary source of a few short examples, I have used The Times and The Sunday Times for 1995.

Institut für deutsche Sprache (Mannheim corpora)

The Institute for German Language in Mannheim holds large German language corpora of literary and non-literary texts. I have used these for the small comparative study reported in chapter 7.

There are many other corpora available and any attempt to give a comprehensive list would be quickly out of date. In any case, if one is attempting to make statements about general English, it is best to sample data from different, independent corpora, since all corpora have their biases. Given the data which it is convenient and more difficult to collect, corpora have often tended to over-represent mass media language, and to under-represent spoken language. Readers can consult corpus linguistics web-sites for information on what is currently available.

Similarly there are now many concordance programs and related suites of text processing software available, and a list of these would also be quickly out of date. Such programs allow texts to be searched for words, phrases and patterns, which can be displayed, in a convenient form, in contexts of varying size. In addition they allow frequency lists of various kinds to be prepared, and they may allow texts or corpora to be compared with each other in different ways.

If the main need is for simple word-frequency statistics and a concordance program, it often hardly matters which specific software is used. Some older software may be restricted in whether it recognizes non-English alphabets, or in the length of texts which can be analysed. As with computer software more generally, what is used is often a question of personal preference, availability and convenience. The large publicly available corpora, such as the Bank of English and the British National corpus, have their own powerful access software. I have used these, as well as both commercially available software (Longman MiniConcordancer, MicroConcord and WordSmith Tools: see Scott 1997a) and other batch software, written by my students in Trier. Again, consulting corpus linguistics web-sites will provide information on what is currently available.

As starting points for information in the world-wide web on corpora and software, use a search engine to look for 'Corpus Linguistics', 'ICAME' (= International Computer Archive of Modern English), 'Oxford Text Archive', 'Cobuild', and 'British National Corpus'. Links will take you to other relevant sites.

Indeed, for some simple investigation of phraseology, the world-wide web can itself be used as a vast text collection. Some search engines can find exact phrases in texts on the web, return the frequency of these phrases, and allow you to study how they are used in their original full contexts. Try, for example, searching for the phrases ripe old age, good old age and grand old age, and check which is the most frequent variant. Or find out how frequently these phrases occur in longer phrases such as LIVE to a ripe old age or REACH a grand old age.

CHAPTER 1

WORDS IN USE: INTRODUCTORY EXAMPLES

The topic of this book is words and phrases: how they are used, what they mean, and what evidence and methods can be used to study their meanings.

Here is an initial example of the kind of question the book deals with. The individual word *round* can mean "circular", and the individual word *table* can mean "a piece of furniture with a flat top, which people can sit at, so that they can eat, write, and so on". The phrase *round table* has one meaning which is simply due to the combination of these individual meanings: something which is both "round" and a "table". However, it is also used in longer phrases such as *round table talks*. This means that a group of people, with interests and expertise in some topic, are meeting as equals to discuss some problem. This meaning relies on additional cultural knowledge: we would not fully understand the phrase unless we also knew that it is often used of discussions between political groups who are trying to reach agreement after some conflict.

In other phrases, *round* and *table* mean quite different things (*a round number, a table wine, a timetable*). Most everyday words have different uses and different meanings. Indeed, in isolation, some words seem to have so many potential meanings that it is difficult to see how we understand running text at all. However, words do not usually occur in isolation, but in longer phrases, and in the following examples it is quite clear which meaning of round is relevant:

- they sat round the table; they ran round the table; they came round to my house; they came round to my way of thinking; a round dozen; a round of applause; a round of drinks; a round of golf; the doctor is on her round

So, our knowledge of a language is not only a knowledge of individual words, but of their predictable combinations, and of the cultural knowledge which these combinations often encapsulate:

- Knights of the Round Table; table manners; drink someone under the table; payments made under the table

Major questions throughout the book will therefore be: What do words mean? In particular, what do they mean when they are used in short phrases and in longer connected text? How do the meanings of words depend on their different uses?

Chapter 1 starts to provide answers to these questions, and to show the close relations between how words are used and what they mean, and chapter 2 discusses some concepts which are required for a systematic study of these relations. Many of these concepts (such as denotation and connotation, lexical field, and sense relations) are widely used within traditional semantics. Here I show how attested data collected in large corpora can be used to illustrate and develop these concepts. The book will probably be of most interest to students who have already done at least an introductory course in linguistics, although all such concepts are explained with detailed examples.

Chapters 3 to 8 then provide case studies of: (1) words in phrases: how words are used in predictable combinations, which often have characteristic evaluative meanings; (2) words in texts: how recurrent words and phrases contribute to the textual organization; and (3) words in culture: how some words, especially in frequent phrases, acquire layers of meaning, and become cultural 'keywords'. Chapters 9 and 10 discuss some implications for linguistic theory. Much of the book consists of case studies, which are sometimes quite complex and take up whole chapters, but I also suggest smaller studies and projects which students can carry out. Corpus data are increasingly available, either for use on individual desk-top computers or over the internet, and this means that students can carry out their own genuine descriptive studies.

For hundreds of years, dictionaries of English have recorded and defined the meanings of words, though they often differ considerably in which phrases they include. As evidence, dictionary makers have used both their own intuitions and also attested uses of words, often in the form of thousands of quotations from printed books. However, it is only since the mid-1980s that computer-assisted methods have been able to provide evidence about word meaning by searching across large text collections. So, I will also discuss questions of method and evidence: How do we know what words mean? What evidence do we have? Is this evidence observable and objective? How can large text collections (corpora) be used to study what words mean? A shorthand description of this approach is corpus semantics: using corpus evidence to study meaning (Teubert 1999a).

1.1. Text and discourse: some distinctions

Corpus semantics studies how words are used in text and discourse and uses observations of use as evidence of meaning.

The terms 'text' and 'discourse' are both used in different ways. I will use the terms to refer to naturally occurring, connected, spoken or written language, which has occurred in some real context, independently of the linguist. Usually, the terms mean stretches of language in use, such as conversations, lectures and stories: that is, units of language which are larger than single sentences. And text analysis is often seen as the study of how language is organized above the sentence or above the clause. It studies units such as (spoken) lectures or (written) short stories. For example, one could study the overall structure of a whole lecture, and note predictable differences in the language of different sections (introduction, main body of the argument, and conclusion), and how the lecturer marks divisions between sections, by saying things such as

- OK so let's look at the next main point now

However, there is a distinction between whole texts and long texts. Some whole texts may be very short, as in public notices, such as:

- Exit
- Private
- Wet paint
- Closed for lunch

It is therefore more accurate to say that text and discourse analysis study language in context: how words and phrases fit into both longer texts, and also social contexts of use. (Widdowson 1995.) This implies that language in use is an integral part of social action, and that we must study relations between language and culture. So, major themes in this book include: phrases and texts (the linguistic co-texts of language in use); socially recognized text-types (such as an advertisement or a short story); the social and cultural meanings conveyed by language; and the range of different functions which language serves (such as informing and evaluating).

These cultural meanings are constantly changing, and new text-types are constantly appearing. For example, one new type of public language is the written and spoken messages which it is impossible to escape in many towns. Written messages have long been common: everything from road signs, to messages on gravestones, and advertising, outside shops, on posters or neon signs. Such language usually has no identifiable author, and is not addressed to anyone in particular. From time to time, old text-types appear in new places: for example, in the 1980s poems appeared in London underground trains, in the slots previously reserved for advertising. (This idea has now spread to buses in Stockholm.) To some extent you can avoid reading such written messages, refuse to accept leaflets pushed into your hand as you walk through town, and throw away the unsolicited mail that arrives at your house, but it is more difficult to escape the wide range of spoken messages in public places. Their early ancestors were produced by the town crier, and announcements in railway stations and airports have long been common, but many stores, as well as trains themselves, now have a constant stream of messages to staff, advertising jingles and music, which customers can escape by going to a different shop, but which the staff cannot escape at all. (Glück & Sauer 1990: 131-32.)

1.2. Language, action, knowledge and situation

Substantial arguments have been put forward that language, social action and knowledge are inseparable (chapter 1.9 gives references). The formulation which has had most influence on linguistics was put forward by J. L. Austin. As he put it (Austin 1962), people do things with words. Some actions can be performed only verbally (for example, apologizing or complaining), whilst others can be performed either verbally or non-verbally (for example, threatening). Austin pointed out that social relations - for example, being appointed to a job or getting married - can be created by people saying the proper words in their proper place. In addition, studies of how language is used in natural social settings, show that communication is impossible without shared knowledge and assumptions between speakers and hearers, and show that communicative competence and cultural competence are inseparable. So, a study of how words are used can reveal relations between language and culture: not only relations between language and the world, but also between language and speakers with their beliefs, expectations and evaluations. A major finding of corpus semantics is that words and phrases convey evaluations more frequently than is recorded in many dictionaries.

Austin's argument means also that language and situation are inseparable. In some games and rituals, the relation may be deterministic: actual words and phrases may be laid down as part of the proceedings (for example, in religious ceremonies). Most everyday uses of language are much more flexible, although we are rarely, if ever, entirely free in what it is appropriate to say. Words occur in expected sequences in phrases and texts.

1.3. Words and expectations

Occasionally, a single word or phrase may be enough to identify the text-type: if you hear the word *furlong*, you are probably listening to a horse-racing commentary, and if you hear the phrase *warm front*, it is probably a weather forecast. Any choice of words creates a mini-world or universe of discourse, and makes it likely that other words will be co-selected in the same context. So, much of this book concerns such expectations and the mechanisms of co-selection.

Sometimes, individual words can trigger assumptions and frames of reference, and words can acquire implications if they are repeatedly co-selected with other words. For example, in a large collection of texts, I found that the word GOSSIP frequently occurred in phrases with very negative connotations such as:

- baseless gossip; gossip-mongering; idle gossip; juicy gossip; name-calling and malicious gossip; scandal and gossip; sleazy gossip; titillating gossip; her affairs became common gossip

Even if everyone does it, and even if it can have positive functions of maintaining group solidarity, it is evident from such phrases that gossip is often talked of as an activity to be disapproved of.

Does the word GOSSIP imply to you a woman speaker, or can men also gossip? To what extent do such words for speech acts carry (in this case sexist) implications about speakers? In the text collection, I found that the word also occurred frequently in phrases such as

- the mothers stood gossiping in the alleys
- the women gossiped and the men smoked
- a gossiping old woman

Men certainly also gossip (though they may call it something different, such as *male bonding*!), but if the word is habitually used in such phrases, then this is likely to contribute to a stereotype that gossiping is something which mainly women do. As Cameron (1997: 455) puts it: 'both sexes engage in gossip, [...] but its cultural meaning [...] is undeniably "feminine".'

There are many terms in everyday English for different kinds of language behaviour, and by studying how these terms are used, it is possible to study the logical relations between them, and whether they have positive or negative connotations. TALK is a general word. CHAT is a sub-category of TALK: friendly talk. GOSSIP is a different sub-category: talk in which secrets are revealed and/or details of other people's lives are discussed, with the implication that the topics are trivial. CHATTER is rapid talk. PRATTLE is foolish talk. BABBLE is incoherent talk. CHAT and GOSSIP need more than one person, but babies can BABBLE on their own. GOSSIP, CHATTER and PRATTLE are disapproving terms. PRATTLE is definitely insulting. (See chapter 2.7.2 on superordinate terms and hyponyms.)

Evidence of such meanings comes from the typical phrases in which the words occur. Attested phrases containing the words above include:

- a friendly chat; they sat around chatting amiably; chat show
- chatterbox; chattering classes; scatter-brained chatter; he chattered without stopping; chattering all day long; a constant stream of chatter; he chattered away about nothing; the chatter of voices; his teeth chattered; the chatter of monkeys; chattering wheels
- he prattled away incessantly; she prattled on
- voices babbled; babies babble; he babbled on; he babbled away; the babbling river; the stream babbling and gurgling

1.4. Language, logic and truth

The relations between language, action, knowledge and situation imply that meaning is not restricted to matters of information and logic. We are not dealing only with whether statements are true or false, and everyday conversation often contains utterances which are logically tautological or contradictory. For example, I was on my way with a colleague to his place of work at a weekend, when he discovered that he had forgotten the key to an office he shared with others. He decided not to go all the way back home for the key, and remarked:

- either it's open or it isn't [A]

([A] = attested data: see Data Conventions and Terminology).

From a purely logical point of view, this utterance conveys no information whatsoever: an office must be either open or locked. But he was implying something like: "Well, it's not all that important; if I can get in, then fine, but if not, it doesn't really matter; I'm not going all the way back home." In everyday uses of language, there is more to meaning than logic, and what is ill-formed from a logical point of view, may be quite normal in conversation. Different factors interact to determine the appropriateness of utterances: not only their logical structure and truth value, but also their rhetorical functions. We need to know what speech act is being performed in what speech event.

Truth conditions involve more than a correspondence between a sentence and the state of the world. In fact, sentences generally do not have a truth value at all. It makes no sense to ask whether the sentence *She came yesterday* is true or false. It depends who she is, when yesterday was, and where she came to. Only some sentences which express propositions about general states of affairs, such as Ice floats on water, will be true on each occasion of use. So, we have to distinguish between sentences (linguistic forms) and utterances (the use of sentences on specific occasions), or between what are sometimes called eternal sentences and occasion sentences (Seuren 1998: 317). Using substantial corpus data, Channell (1994: 115, 119) also shows that many utterances contain vague language, such as

- this impossible task of handling umpteen jobs with very little in the way of training
- there was no kind of social contact, there was no coffee room or anything

It is not possible to say exactly how many the speaker meant by *umpteen*, or exactly what the speaker meant by *no coffee room or anything*, but hearers generally have no problem in supplying a reasonable interpretation, such as "a place where staff could meet informally during work-breaks". Such utterances cannot be judged simply true or false, but show the kinds of inferences on which language in use often depends.

Furthermore, the concept of truth is applicable only to a narrow range of sentence-types. Only statements can be true or false, but not questions, requests or orders, expletives, promises, counter-factuals, such as

- If you weren't a policeman, I wouldn't have let you in [A]

and other utterances which express probabilities, beliefs or intentions. Truth and falsity are also problematic with respect to evaluative utterances. If someone says *That's super!*, then this may tell us something about the speaker, but little about the world.

1.5. Common-sense knowledge

One of the major problems in studying language in use is to disentangle linguistic knowledge from background cultural assumptions. What is said can be a long way from what is meant, and a very large amount of language in use is indirect or vague in different ways. For example, I heard the following exchange between a husband and wife:

He: When will dinner be ready?
She: It's a very small chicken

The second utterance is not a direct answer to the question. In order to see its relevance, He had to assume that She was being co-operative, and make a series of inferences along the lines of: we're having chicken for dinner, it's a small chicken, I know how long chickens take to cook, I know when the chicken went into the oven, you can't predict exactly how long these things take, but dinner will be ready soon. An important distinction (Widdowson 1978) is between cohesion and coherence. There is a cohesive link between the two utterances, in so far as dinner and chicken are in the same semantic field, but much still depends on inferring the point of the utterance from common sense knowledge.

Language in use sets up expectations, and whenever two utterances occur in sequence, hearers will attempt to relate them: to use the first as a frame for the second. In a useful article on the inferences we perform on language, Brown (1994: 17) gives this example from local radio:

- The Suffolk doctor whose wife has been reported missing stayed firmly in his house today. Police have been digging in the garden. [A]

The relations between these two sentences seem obvious and natural, though they are not stated explicitly. Why is the doctor not named? Who reported his wife missing? What is the relation between his wife going missing, his staying indoors, and police digging in his garden? Why do we assume they were digging in his garden?: that is, the garden around his house? These examples pose problems for semantics, and no general method has been found for automatically identifying the referent of definite noun phrases such as the garden. This frequently requires information which is not explicit in the co-text, and in this case the information depends on inferences about suspicious circumstances, crimes, and common police procedures. The example also illustrates a further key distinction between meaning and reference. We know what the phrase *the garden* means, but we do not necessarily know what an utterance of the phrase refers to. (See chapter 2.6 on denotation and reference.)

Such examples depend on schematic knowledge: sets of taken-for-granted knowledge about how the world works. We all have such schematic knowledge about schools, so that when a school is mentioned in conversation, we automatically assume it to be populated with teachers and pupils, without need for them to be explicitly mentioned. Suppose I say:

- I used to teach in a school. The pupils were horrible. [I]

I do not expect the definite noun phrase to be questioned with *What pupils*?

People certainly say unexpected things in jokes and ironic remarks, but this is itself an indication that there are expectations to be broken. We often recognize the existence of norms only when they are broken. Each utterance sets up a frame with built-in default expectations; but these default values can be overridden. However, it might be that discourse has no clear-cut mandatory rules, but rather depends on maxims of co-operativeness or guiding principles (Grice 1975). One reason why discourse structure is likely to be less deterministic than phonological or syntactic structure is that discourse is the joint construction of at least two speakers. It is difficult to see how A could place absolute constraints on what B says.

It is plausible that languages are tightly patterned at the lower levels of phonology, morphology and syntax, and that discourse is more loosely constructed. Nevertheless, menus, stories and telephone conversations have beginnings, middles and ends, and that is already a structural claim. Jakobson (1971: 242-43) puts it like this:

> [I]n the organization of linguistic units there is an ascending scale of freedom. In the combination of distinctive features into phonemes, the freedom of the individual is zero [...] Freedom to combine phonemes into words is [...] limited to the marginal situation of word coinage. In forming sentences with words the speaker is less constrained. And finally, in the combination of sentences into utterances, [...] the freedom of any individual speaker to create novel contexts increases substantially, although [...] the numerous stereotyped utterances are not to be overlooked.

Jakobson here mentions explicitly 'numerous stereotyped utterances', but perhaps did not appreciate their extent. In chapters 3 and 4, I will show that the freedom to combine words in text is much more restricted than often realised. I will also show that the lexical organization of text is distinctly different from the linguistic organization which has usually been described at lower levels.

1.6. Linguistic conventions

We often rely on social knowledge in order to make inferences which are not expressed in the textual cohesion, and we must distinguish what can be inferred from the language itself, and what must be inferred from real world knowledge. For example, a sentence such as *My sister is sick* has as one of its presuppositions "I have a sister". As Horn (1996, citing earlier work by Grice and others) points out, if this presupposition is not part of hearers' real world knowledge, the communication does not break down simply because a felicity condition for such a sentence has been broken. The hearer infers the presupposition, and is more likely to say *Oh dear!* than *What sister?* The presupposition signals that the information is non-controversial rather than common knowledge.

Even in the case of utterances which are very indirect indeed, there may be a balance between inferences based on the particular situation of utterance and inferences based on predictable linguistic patterns. I heard the following utterance from a surgeon to a patient in a hospital:

- Right! A little tiny hole and a fishing expedition, is that it?

The intended meaning of this utterance would be completely irretrievable without knowledge of the specific situation of utterance. I assume that the surgeon intended to convey the information that "I am going to operate on you and remove your appendix", and in addition to convey reassurance by implying "but don't worry, I do this kind of thing every day, it's routine, I know what I'm doing, and I can even joke about it". How can I make such inferences? Parts of the original utterance can, as it were, be translated: *a hole and a fishing expedition* meant "a surgical incision and the removal of the appendix". There is obviously some relation in meaning between a fishing expedition and a surgical operation, but we would not expect this to be recorded in dictionaries. In individual utterances, it may be that idiosyncratic meanings will always depend partly on specific context-bound interpretations and will never be fully explicable.

However, other parts of the surgeon's utterance have conventional meanings. The combination *little tiny* has connotations of the language used to children. Evidence for these connotations comes in turn from the frequent use of little in attested phrases such as

- beautiful little; charming little; cute little; lovely little; nice little

The connotations of little become even clearer if *little* is contrasted with *small*, which often occurs in rather formal phrases such as

- comparatively small, exceedingly small, relatively small

Typical phrases are pretty *little girl*, but *comparatively small quantity*. (See chapter 7.6.)

So, the surgeon's utterance is multi-functional. As well as indirectly expressing propositional meaning, he simultaneously expressed interpersonal meaning: the social relations between doctor and patient, and authority, reassurance, and joking. And although part of the meaning depends heavily on social context, part is also conveyed by linguistic convention. How connotations can be identified more formally is a major topic in chapters 7, 8 and 9.

1.7. Possible and actual

A brief summary of the argument so far is the slogan 'meaning is use'. Words do not have fixed meanings which are recorded, once and for all, in dictionaries. They acquire, or change, meaning according to the social and linguistic contexts in which they are used. Understanding language in use depends on a balance between inference and convention. Here are more detailed examples which use textual data to show that our communicative competence relies on knowledge of what is expected or typical.

1.7.1. Example 1: the ambiguity of SURGERY

In isolation, many individual words are ambiguous or indeterminate in meaning, but this hardly ever troubles us in practice, because the phrases in which they occur are not ambiguous. For example, *surgery* can mean

[1] a medical procedure involving cutting a patient's body open
[2] the branch of medicine concerned with these things
[3] the room or house where a doctor works
[4] the time of day when a doctor sees patients.

However, in different phrases, the ambiguity disappears. For example, senses [1] to [4] are conveyed unambiguously by these attested examples respectively:

[1] plastic surgery; he had to undergo surgery; patients who need surgery
[2] progress in surgery has made heart transplants possible
[3] she had her surgery in Cemetery Road; he had to be rushed to the surgery
[4] she was taking evening surgery; his surgery ends at eleven.

It is not difficult to find words in the immediately surrounding text which discriminate between these different senses with a high degree of probability. For example, sense [1] is signalled by co-occurring verbs such as *carry out, need, respond to* or *undergo*, or adjectives such as *cosmetic, extensive, major* or *successful*. (The phrase *undergo surgery* is analysed in chapter 4.4.1.) It is possible to invent examples where the verb *remove* could occur in sense [3]: *we had to remove it through his surgery door* [I]. But the phrase *surgery to remove* signals sense [1]: *surgery to remove two wisdom teeth*.

In other words, cases of apparent multiple ambiguity at word level are usually illusory: they dissolve in context. Combinations of words in phrases are therefore a good candidate for the basic semantic unit of language in use. Instead of regarding the meaning as being carried by the individual word, we could see things as follows. The word *surgery* conveys a rather general meaning: "something to do with medicine". It is the phrase which conveys the precise meaning. The following formulation might be a slight exaggeration, but it makes a useful point: it is not the words which tell you the meaning of the phrase, but the phrase which tells you the meaning of the individual words in it. I will return to this point later (chapter 9), since it questions the principle of compositionality, that is, the assumption that the meaning of larger units (such as phrases) is equal to the sum of the meanings of smaller units (such as words).

Translators obviously have to take such cases into account. It does not make sense to ask for a translation of *surgery* into German. The word would have four different translations for the four senses I have identified: [1] *Operation* or *operativer Eingriff*, [2] *Chirurgie*, [3] *Praxis* or *Sprechzimmer*, and [4] *Sprechstunde*. In turn, these individual German words would have different translations into English, according to context: *Praxis* can mean "doctor's practice", but also "practice" in the sense of "practice versus theory".

1.7.2. Example 2: the (non-)ambiguity of BANK

Here is a similar example of a semantic problem which is posed in many introductions to linguistics. In isolation, the word BANK is ambiguous, and dictionaries distinguish two main senses. Sense 1 is the "place where you keep money", either the institution thought of as the abstract organization, or as a particular building. Sense 2 is a little more difficult to define precisely, since there is a range of related meanings. It means an "area of sloping, raised ground" (grassy bank), often the raised ground around a stretch of water (river bank) or under shallow water (sand bank), or something of the same general shape (bank of fog, bank of switches). Let us call these the "money"-BANK and the "ground"-BANK senses. It is certainly possible to invent sentences, and to imagine circumstances, where the word is still ambiguous:

- the supermarket is opposite the bank [I]

However, even such sentences are most unlikely to be ambiguous, in practice, in a larger context. Depending on what has been said previously, this could mean "opposite a bank of daffodils". However, a hearer is most likely to assume a parallel construction and to assume that the supermarket and the bank are both buildings.

So, in isolation the word is ambiguous, but this statement depends on a very artificial assumption, since the word never occurs in isolation. It either occurs in a physical context, for example, on a sign above a building (and probably also in a phrase such as *Bank of Scotland*), or it occurs in co-text, with other words around it. I studied all occurrences of *bank* (n = 82) and *banks* (n = 28) in their linguistic contexts in a corpus of one million words of written English (LOB: see Notes on Corpus Data and Software). In the vast majority of cases, any potential ambiguity was ruled out due to words within a short span to left or right. Many occurrences were in fixed phrases which signalled unambiguously the "money" or "ground" sense:

- bank account, bank balance, bank robbery, piggy bank
- canal bank, river bank
- the South Bank (= "an area along the Thames in London"), the Left Bank (in Paris), Dogger Bank, Rockall Bank, Icelandic Banks (= "fishing areas in the Atlantic")

In addition, the word usually co-occurred, within a few words to left or right, with other words which clearly signalled one or other semantic field:

- cashier, deposit, financial, money, overdraft, pay, steal
- cave, cod, fish, float, headland, sailing, sea, water

So, the two senses occurred in complementary distribution, either in one lexical context or the other, not both. Even in short phrases, only very few cases remained ambiguous, such as

- the Worthing bank murder case

(Worthing is a town in the south of England). However, even here, everyday expectations probably tip the interpretation towards the sense which goes along with *bank raid*. Often, the sense was over-determined, and occurred both in a fixed phrase and alongside several disambiguating words in lexical strings such as

- money - deposits - Bank of England - paid - instalment
- shallows - sea - cod - Icelandic Banks - haddock

This simple case illustrates several principles which will be central to the whole book. (1) It is impossible to observe the meaning of a word: meaning is an invisible (arguably mental) phenomenon. However, it is quite possible to observe evidence from which meanings can be reliably inferred. A major type of evidence of the meaning of a word is the other words round about it, especially repeated patterns of co-occurrence. (2) The meaning of a word is not independent of the environment, including the co-text, in which it occurs. In fact, it is rather misleading to talk of a word occurring in an environment. A word predicts that other related words will occur round about it, and the co-text predicts the word, or one very like it. (3) Invented and decontextualized examples may exaggerate difficulties of interpretation. A theory of semantics should deal primarily with normal cases: what does typically occur, not what might occur under strange circumstances. (4) Findings such as those above, from one small corpus, are predictions which can be checked on other corpora. Readers can check my findings about BANK on data from other corpora. (These principles were first discussed thoroughly with reference to corpus study in Sinclair ed (1987) and Sinclair (1991). See also chapter 6 on replicability.)

So, one of the main topics of this book is how large corpora can be searched for observable patterns which provide evidence of what words mean.

1.7.3. Example 3: the days of the week

The BANK example illustrates the difference between what speakers can say and what they usually do say, but we have to deal with both: it would be misleading to base a description only on frequency of actual occurrence. Frequency becomes interesting when it can be interpreted as typicality, and speakers' communicative competence includes tacit knowledge of behavioural norms.

In corpora of 150 million words, I found that the words for different days of the week differed considerably in frequency. Rounded to the nearest 50, occurrences were:

Sunday	17,350
Saturday	14,600
Friday	10,650
Monday	9,500
Wednesday	8,150
Thursday	6,900
Tuesday	6,750

It would be absurd to base a description of English merely on frequencies, and to argue that *Sunday* is over twice as common as *Tuesday*, and should therefore be twice as prominent in our description. However, the category days-of-the-week is culturally structured, and there are cultural reasons why people talk most often about the weekend, less often about the beginning and end of the working week, and less often again about the days in the middle of the week.

The seven words also tend to occur in different phrases, such as

- Friday night; Saturday night; Sunday afternoon; Monday morning; that Monday morning feeling; Monday morning blues

Of course, it is formally possible (i.e. grammatical) to say *Sunday night*, but *Saturday night* is more frequent, and this is a fact with cultural significance. Words have a tendency to co-occur with certain other words, and culturally and communicatively competent native speakers of English are aware of such probabilities and of the cultural frames which they trigger.

The words for the days of the week are not the names of something which exists independently in the external world. Suppose you are shipwrecked and washed ashore on a desert island, where you lie in a coma for some time. You wake up, but do not know how long you have been unconscious. There is no way to observe what day of the week it is, and no way to find out. The week is a cultural reality, whose conventions are maintained by talk (and other social activities). This does not mean that the days of the week are not real: they are real, and they have a real effect on our behaviour. However it means that they are mental and social constructs which are maintained by language and its use. They do not refer directly to the external world, but only indirectly, via cognitive representations, to a reality which they have helped to create.

1.7.4. Example 4: lonely hearts ads

The next example illustrates that many things are possible, but that what actually occurs is often very predictable. Lonely hearts ads are an example of a text-type which is highly conventionalized and restricted in its forms and meanings. A page of ads can be read in any order, but readers can predict the semantic structure of any individual ad, the speech acts it expresses, and much of its vocabulary and grammar: each ad is a standard solution to a standard problem. Here are two attested examples:

- MUCH-TRAVELLED engineer/manager/lecturer, now retired into writing and sociobiological research, seeks female friend/lover, similarly fit and active, to share and exchange ideas, interests and ambitions. Box ... etc.
- ATTRACTIVE PROFESSIONAL, degree educated woman, 43, divorced, one child, Bristol area, would like to meet similar man, 45-55 for caring relationship. Box ... etc.

It would be possible to write ads in different forms, but in practice their form is very restricted, due largely to constraints on space (and the amount of money one is willing to spend). Some of the main patterns are as follows. The propositional content varies very little. There is an obligatory proposition: "X is looking for Y", with a small amount of variation then possible: "with a view to friendship, marriage, sex, etc". And there is a request: "please get in touch", traditionally via a box number, and more recently via recorded telephone messages. The most frequent syntactic structure is also simple:

- NP1 seeks NP2 for X

This corresponds to the content: "person 1 seeks person 2 for friendship, etc". The most frequent verbs are *seeks* and *would like to meet* (sometimes abbreviated to *wltm*).

The NPs (noun phrases) have a head noun denoting a person, which is always marked as male or female, though this may require some inferences. Optional, but frequent (especially for the sender), is their profession. There is a description of the addressor, and a description of the hoped for addressee (often in less detail: after all, the writer knows what s/he is like, but being too specific about the addressee might cut out too many potential responses). The NPs are often long and complex, mainly due to the occurrence of relative clauses, and (very frequently) strings of adjectives which usually denote personality and appearance. Other frequent linguistic features include: elliptical sentence structure, and lexical abbreviations, which may not always be interpretable to readers unfamiliar with the genre (e.g. *gsoh* = good sense of humour; *ns* = non-smoker: *tlc* = tender loving care). These statements certainly do not account for all examples which occur, but for the most frequent patterns. Deviations from the basic schema (such as using humour or self-deprecation) can be interpreted only with reference to the prototype.

By looking at large corpora, it is also possible to state the most frequent vocabulary which occurs. In a corpus of 200 million words (Cobuild 1995b), the word seeks occurred 7,847 times. It does not occur only in lonely hearts ads, but it often does, as can be seen from the ten words which most frequently co-occur with it (within four words to left and right):

- female 1113, black 972, male 785, attractive 619, similar 568, guy 499, lady 493, man 425, caring 401, professional 389

In turn, the word *caring* occurred 4,814 times. Its ten most frequently co-occurring words (within four words to left and right) were

- seeks 401, loving 353, honest 336, sincere 194, make 159, very 155, more 149, people 149, children 128, kind 128

This starts to show the typical phrasings used in such texts. For example, the following phrases all include both seeks and caring:

- seeks a sincere, caring single lady
- caring, Christian-minded, romantic, seeks attractive, reliable female
- Black male, 31, seeks caring lady
- various interests, own flat and car, seeks caring, ambitious lady
- kind, honest, reliable, boyish, 35, seeks caring, genuine female, for lasting relationship
- male, 35, quiet, honest, caring, seeks down-to-earth female, 25-34, for lasting friendship

In this example I have deliberately taken a text-type which is much more restricted in its forms than much language use. However, all language use is restricted to some extent, and a main topic of this book will be to show just how strong the co-occurrence relations between words often are.

1.8. Summary and implications

In this chapter I have introduced some of the topics which arise in the study of language in use. The meaning of words depends on how they are combined into phrases, and on how they are used in social situations. It follows that their meaning depends on both linguistic conventions and also on inferences from real-world knowledge. These linguistic and social expectations mean that, although we are in principle free to say whatever we want, in practice what we say is constrained in many ways. The main evidence for these constraints comes from observations of what is frequently said, and this can be observed, with computational help, in large text collections.

Two shorthand ways of referring to the approach I take in this book are (1) 'meaning is use' and (2) 'corpus semantics'.

> (1) 'Meaning is use' is convenient phrase, but is merely a shorthand way of referring to a complex set of ideas. The meaning of words and phrases differs according to their use in different linguistic and social contexts.
>
> (2) 'Corpus semantics' refers to an approach to studying language in which observational data from large text collections are used as the main evidence for the uses and meanings of words and phrases.

A corpus is a large sample of how people have used language. Meanings are invisible and cannot be observed directly, but if we put (1) and (2) together, then we have empirical observational methods which can be used in semantics, since words acquire meanings from their frequent co-occurrence with other words. I have also introduced the following supporting concepts:

> (3) Expectations. Our interpretation of what other people say or write depends partly on our expectations of what is likely to occur. Our communicative competence involves knowledge (often unconscious) of what is probable, frequent and typical.
>
> (4) Real-world inferences. Sometimes our interpretations depend on non-linguistic knowledge: that is, our background, encyclopedic knowledge of the everyday world (such as why policemen might be digging in a garden). Meanings are not always explicit, but implicit. Speakers can mean more than they say.
>
> (5) Linguistic conventions. However, our (unconscious) knowledge of what is probable also involves expectations of language patterns. Our knowledge of a language involves not only knowing individual words, but knowing very large numbers of phrases (such as *river bank, sand bank, bank clerk, piggy bank*), and also knowing what words are likely to co-occur in a cohesive text (*bank, water, fish,* or *bank, money, robbery*).
>
> (6) Text-types. Different text-types have different patterns of expectation. For example, most lonely hearts ads use a restricted set of verbs which have typical subjects and objects. The semantic pattern is very simple, and much of the vocabulary and grammar is predictable, but there is scope for considerable lexical variation.

So, the central programme of corpus linguistics is to develop a theory of meaning (Teubert 1999a, b). When people hear or read a text, they are usually interested in its meaning, not in its wording or grammar, and they generally remember its content, not how the content was phrased. Yet, as Pawley (2000) puts it, recent linguistic theories have often not recognized that anything is being said at all. In following chapters, I will discuss more detailed examples of the predictable co-occurrence of words and other linguistic patterns, and methods for studying patterns of co-occurrence and of probability.

1.9. Background and further reading

An important line of thought on language in use has its origins in problems with concepts of truth and falsity, which were realised in philosophy from the late 1800s onwards. Levinson (1983) and Seuren (1998: 377, 384) discuss the formal semantic background to such work.

A second important line of thought, which relates language and social action, has sources in twentieth century linguistics, anthropology and philosophy. Malinowski (1923) talked of language as a 'mode of action' or 'behaviour', and related ideas of meaning as use were proposed by Firth (1957) and Wittgenstein (1953). The two classic books on speech act theory are Austin's discussion of *How to Do Things with Words*, based on lectures given in 1955 (published as Austin 1962), and Searle's discussion of *Speech Acts* (Searle 1969). Searle (ed. 1971) contains important papers by Austin, Searle and Strawson. Cole and Morgan (eds. 1975) edited a collection on Speech Acts, which contains important papers on indirect speech acts: Grice (1975), published there for the first time, but already widely circulated in manuscript form, Gordon and Lakoff (1975, originally published 1971), and Searle (1975). Searle (1976) is another important updating of the theory.

In this work, from the 1960s and 1970s, the term 'speech acts' became standard, although it is slightly unfortunate as a term, since such acts can be performed in both speech and writing. There are differences between the acts which are performed in spoken and written forms, and indeed some acts can only be performed in writing (such as a signature, a last will and testament). 'Language acts' would have been more accurate, but is scarcely used as a term.

There are useful textbook accounts available in many places. Again, Levinson (1983) and Seuren (1998) explain the impact of such ideas within linguistics, and give a wider account of their history. Seuren discusses ideas about language and logic from the Greeks and Romans onwards, and the problems recognized in truth-conditional semantics in the late 1800s and early 1900s, and gives an account - by someone actively involved - of attempts within Chomskyan linguistics to integrate semantics into syntactic theory. He argues however, that the Firthian approach to studying language in context 'proved largely sterile' (p.170), and Halliday is not mentioned at all in his account. From my approach in this book (and in Stubbs 1996), it is clear that I disagree with this judgement.

Lyons (1977) and many articles in linguistic encyclopedias and reference books, such as Sadock (1988), provide more general discussion of different approaches to the study of language in use. See also Saville-Troike (1989), Mey (1993), Schiffrin (1994), and Van Dijk ed. (1997a, b). By the 1990s, these broad strands of work on language in use had led to constructivist theories of social organization (Searle 1995), and radical reinterpretations of Grice's theories (Levinson 2000).

Firthian work has also been radically developed by corpus methods: Sinclair (ed. 1987) and Sinclair (1991) were the first books to demonstrate the methods; Hunston and Francis (2000) give a detailed account; and Partington (1998) is a good introductory textbook. It is neo-Firthian work which is most immediately relevant to the methods I discuss in this book.

1.10. Topics for further study

(1) Collect your own data on the actual occurrence of words for related speech acts in phrases and texts:

- GOSSIP, NAG, CARP, COMPLAIN, WHINGE
- PROMISE, OATH, VOW, PLEDGE, GUARANTEE

Study the words they repeatedly co-occur with, and use this evidence to provide a description of their meanings, including whether they express the speaker's attitude to the language behaviour: approving, neutral or disapproving.

(2) In chapter 1.7.3, I discussed words for days of the week and some phrases in which they occur. Analyse other sets of words which form well-defined semantic sets: for example, months of the year, numbers, or girls' names and boys' names. Some sets are small and finite (months); others are larger and open-ended (professions); others have clear central members, but also other members about which there might be dispute (colours). Investigate why members of such sets differ in frequency, and consider what relations this shows between what is possible in the language system, what is frequent in language use, and how choices express cultural meanings. See Firth (1957: 12) for a classic proposal for such studies.

(3) In chapter 1.7.4, I described some central features of lonely hearts ads. Study further examples, and propose a more formal description of their vocabulary and grammar. Useful references are: Mills (1995: 167-69); Yule (1996: 250-51), who gives American examples; Sandig and Selting (1997), who give German examples (translated into English); and Nair (1992), who gives examples of Indian matrimonial advertisements, which are different, but also highly conventionalized.

(4) Study other text-types which are highly restricted in form (vocabulary and grammar) and function, such as other kinds of classified ads, weather forecasts (newspapers or television; perhaps contrasting general purpose or specialized shipping forecasts), horoscopes, and menus. In these cases too, the possible or expected vocabulary and grammar can be specified in detail.

Notes on Corpus Data and Software. Introduction. . 1. Words in Use: Introductory Examples. Text and Discourse: Some Distinctions. Language, Action, Knowledge and Situation. Words and Expectations. Language, Logic and Truth. Common--Sense Knowledge. Linguistic Conventions. Possible and Actual. Example 1. The Ambiguity of Surgery. Example 2. The (Non--)ambiguity of Bank. Example 3. The Days of the Week. Example 4. Lonely Heart Ads.Â Some features of the site may not work correctly. Corpus ID: 61031773. Words and Phrases: Corpus Studies of Lexical Semantics. @inproceedings{Stubbs2001WordsAP, title={Words and Phrases: Corpus Studies of Lexical Semantics}, author={M. Stubbs}, year={2001} }. M. Stubbs. Published 2001. Recent papers in Phraseology, Lexical Semantics , Cultural Studies. Papers. People.Â The English-Portuguese phraseologies were extracted from corpora of TV series subtitles, such as Bones, Game of Thrones and Star Trek, 30 being selected for this paper discussion. The methodology adopted was Corpus Linguistics (CL) and WordSmith Tools 6.0 was the lexical analysis program used.Â Afterwards, word lists were made; we chose terms and verbs that allowed the selection and analysis of phraseologies from TV series by using the Concord tool and most of the analyzed terms were registered on VoTec (FROMM, 2007), an online platform for terminological and terminographical research open to public access. This book fills a gap in studies of meaning by providing detailed case studies of attested corpus data on the meanings of words and phrases. Categories: Linguistics\\Linguistics.Â Library of Congress Cataloging-in-Publication Data Stubbs, Michael, 1947Words and phrases : corpus studies of lexical semantics I Michael Stubbs. p. em. Includes bibliographical references and index. ISBN 0-631-20832-1 (hbk.: aJk. paper)-ISBN 0-631 - 20833-X (pbk. : alk. paper) 1. Semantics-Data processing. I. Title. P325.5.D38 S78 2001 401 ' .43'0285-dc21 2001018462 A catalogue record for this title is available from the British Library. TITLE: A Study of Semantic Clash in Skit-Humor within the Framework of Co-Selection. AUTHORS: Lele Sun. KEYWORDS: Semantic Clash, Co-Selection, Skit-Humor. JOURNAL NAME: Open Journal of Social Sciences, Vol.8 No.5, May 13, 2020.Â Verbal humor accounts for the majority of the punchlines in the skits. It is believed that semantic clash is the primary factor for verbal humor which may elicit laughter. Chinese verbal humor in the skit-humor could not be transferred into English because the controversial attitude is lost in the process of conversion. This, in turn, indicates that the semantic clash may contribute essentially to the punchlines in Chinese skit-humor. Lexical semantics (also known as lexicosemantics), is a subfield of linguistic semantics. The units of analysis in lexical semantics are lexical units which include not only words but also sub-words or sub-units such as affixes and even compound words and phrases. Lexical units include the catalogue of words in a language, the lexicon. Lexical semantics looks at how the meaning of the lexical units correlates with the structure of the language or syntax. This is referred to as syntax-semantic interface.