1 Preliminaries for model building

1.1 Introduction

Talking about language and meaning should, surely, be easy: the fact that we use language to pass on information to each other, to describe what we see around us, to reflect on our thoughts about ourselves, each other and the future, to confide in others about those thoughts and anxieties must mean that the concept of meaning for language is the heart and soul of what languages are about. Yet, as soon as we start probing what concept of meaning we should articulate, the phenomenon threatens to slip away under our fingers in a morass of open-endedness. So, for example, we can use words to mean the opposite of what the words themselves seem to mean, as in the first part of speaker B's reply to speaker A:

(1.1) A. How are things?
B. Everything's perfect. My computer's seized up for the second time in three days.

We can also use words loosely, but nevertheless successfully, as when one might say:

(1.2) I am shaking with fear.

And we can use words to convey something really rather different from what the words normally mean, as when one of us utters *She's an angel* to refer to a sister. Metaphorical use of words and phrases fades into ambiguity, along a cline of intermediate cases, as in (1.3), where neither *spend* nor *driven* relate to more familiar concrete interpretations associated with money or cars:

(1.3) I spend my life driven by terror.

Then, less conventionally, language can be used to convey meaning quite indirectly, as when, instead of greeting people as one joins them for lunch, one signally fails to greet them in any conventionalised way, starting immediately with:

(1.4) Don't ask me to construct a research proposal ever again. It's been a nightmare.

with a clear underlying message that the speaker is over-stressed, flustered and in need of calming down. (Notice in passing the use of *nightmare* to describe an event with no implication whatever that one had slept through it.) Indirect use of

language is common and pervasive; witness also the effectiveness of B's reply to A in the following exchange clearly implying that she doesn't like him:

(1.5) A: Are you inviting me to your party?
B: I'm only inviting people I like.

Through all this Pandora's box of data, certain aspects of language and its construal nevertheless stand out. First, despite all the problems of characterising the various uses to which language can be put, in using a language we have a clear capacity to combine words together to yield an interpretation for a sentence as a whole:

(1.6) No man I ever met kissed me when we were first introduced.

In processing (1.6), the individual words *no*, *man*, *I*, *ever*, *met*, *kissed*, *me*, *when*, *we*, *were*, *first*, *introduced* are parsed in turn, each adding to the information that has been established up to that point in the parse, progressively building up the meaning of the whole from those individual parts. Even without a time-linear parsing perspective, it is clear that the individual words combine with neighbouring words in a systematic way to determine some composite whole. This is known as the *principle of compositionality*, which takes it to be a universal property of natural languages that the meanings of complex expressions are constructed from the meanings of the words they contain and the way those words are put together by the syntax of a language. In other words, the meanings we ascribe to strings of words are not random, but determined, at least to a large part, by building blocks of meanings given by the words and modes of combination, including word order and grammatical processes such as passivisation, question formation etc.

Second, running somewhat counter to this idea, on almost every occasion of use of a sentence, its construal may depend on the immediate context in which the sentence is uttered. For example, in answers to questions, the answer in some sense completes the structure which the question, as its context, provides:

(1.7) A: What shall I give Eliot?
B: A teddy bear.

The string uttered may be just one word referring to some activity going on in the discourse situation, as when a parent shouts to a child reaching up to a saucepan full of boiling water:

(1.8) Don't.

or when a parent of a teenager looking at the waves beside their son holding his surfboard in his hand less dictatorially says:

(1.9) I wouldn't, if I were you.

This is part of a much more general phenomenon of *context-dependence*, which is in part conventionalised within a language. Some words have as their intrinsic content the signalling of the need to find a semantic value from the surrounding

context of utterance. These are pronouns and other so-called anaphoric expressions such as the determiner *the*, and VP (verb phrase) pro-predicate forms, like *do, do so* or *did, did so* below:

(1.10) John came in. He was sick.

(1.11) John came in. The poor dear was sick.

(1.12) John saw Mary and so did Sue.

These anaphoric expressions may even act as place-holders for getting their value from some subsequent part of the utterance:

(1.13) It is possible that I am wrong.

(1.14) She's an angel, my sister.

(1.15) If you want me to do so, I will come with you.

But such signals, which direct the hearer to context to establish their interpretation, are, apparently, not necessary. We may deliberately leave out portions of sentences, knowing that our hearer will be able to recover the intended interpretation from the surrounding context. This is the phenomenon called *ellipsis*:

(1.16) John has finished his homework, but Sue hasn't.

(1.17) I am seeing someone today, but I don't know who.

(1.18) John will be interviewing the President, Harry the Vice-President.

(1.19) I persuaded Tom to visit Mary in hospital, and Sue did Harry.

(1.20) If you want me to, I will come with you.

These various ellipsis phenomena have been analysed as heterogeneous, not subject to a single form of explanation; but what underlies them all is the fact that the context, in some sense, provides the meaning for the elliptical expression.

However, the reliance of meaning on context can go further even than this, with speakers and hearers switching roles midway through an utterance:

(1.21) A: What shall I give
B: Eliot? A teddy bear.
A: Or a dinosaur?

What A says in (1.21) can be taken as a context which provides enough information for B to take over as though he had been the speaker – he doesn't have to start from the beginning and say everything silently before providing the continuation. Equally, A is able to switch into being a hearer as though she had been listening throughout that utterance. Just like B, she doesn't have to parse everything from scratch again. To the contrary, she just picks up as hearer from where she leaves off as speaker: the context is sufficient for her to parse from that point. As this example shows, this switching can happen successively. This is not just

a random performance error or sloppiness. It is something we can all do fluently, and from very early on in child language acquisition. Universally children love the kind of game where you sing to them:

(1.22) A: Old MacDonald had a farm. And on that farm he had a
B (child): Dog.

So it appears that, though we can use words to successively build up a composite whole, this process has also to be sensitive to how the context contributes to such a process.

The concept of compositionality is made more problematic by the third aspect of language: the variation in how much meaning a word may have, and, accordingly, how essential words are to the point being made. Some are critical, as is each word in (1.6) and the only word in (1.8), others barely make any difference. The first word in (1.13) seems purely a prop required by English word order; and yet others, like the *that* in (1.23) and *there* in (1.25), make no difference at all as can be seen by the paraphrases in (1.24) and (1.26) which omit *that* and *there* and still seem to mean the same thing:

(1.23) No man that I ever met kissed me when we were first introduced.

(1.24) No man I ever met kissed me when we were first introduced.

(1.25) There is something I must tell you.

(1.26) I must tell you something.

So we have to articulate the precise nature of structural and meaning properties of natural language, in order to determine the precise role that words in a language play in this process of establishing interpretations for sentences. With (1.25)–(1.26), we stumble on a different puzzle. Though there is a difference in the order of the words, and clearly some structural relation between what is expressed by *there is* and the remainder, the resulting meaning is the same. But this might suggest, perhaps, that the structural properties of sentences have to be seen as something different from just the provision of a basis for interpretation, as there can be strings with different structure and yet the same interpretation.

1.2 Explaining semantics: starting from words?

In making a first stab at the problem of compositionality, one might assume that one should first look at word meanings, and then define a process of combining those meanings together. So, let us suppose, one should be able to turn to a dictionary and take definitions from there as a starting point. It might come as a surprise to someone coming to the study of semantics for the first time, but any such move turns out to be a complete failure. Despite long and very rich traditions of dictionary-making, there are really very few words for which we can successfully provide definitions at all. There are verbs of causation such

as *kill*, *blacken*, *paint*, there are kinship terms such as *bachelor*, *mother*. But the list stops almost at that point. But it's worse than that; for, even within this list, such verbs have their interpretation very largely determined by context. As Jerry Fodor vividly spells out in detail (Fodor 1998), the concepts of painting the Sistine Chapel, painting one's sitting-room wall, painting one's signature on the painting and painting one's face red do not all involve 'causing some surface to be covered with paint' as a dictionary definition might lead us to expect: paint factories that explode and totally cover some road with paint have not thereby painted the road. To get even remotely close to a reasonable definition one has to shift into a definition such as 'cover a surface with paint having the primary intention so to do'. Now this might be closer to what, upon reflection across a reasonably broad range of usages, an analyst might think had to be specified as the meaning of the word. But is this what the child has to learn in order to use the word *paint*? When the child says in tears, *Mummy*, *you've painted my dinosaur*, have they not used the word correctly unless they have some complex intention-attribution on the part of the mother in mind? And it won't do to say the word is simply ambiguous according as these different concepts are invoked, as otherwise, by that criterion, every word of the language will be ambiguous. And, though indeed we might conclude that there is much more ambiguity than might be considered at first sight, we certainly should not trivialise this as applying to every single word in the language.

Unless we are content to invoke lexical ambiguity for a word each time its interpretation in some use is at all distinct from that of previous uses, this flexibility of use suggests that there is something else going on between words themselves and the actions/events/objects in the world which they describe, a topic which we shall return to in Chapter 8. In the mean time, even in the vanishingly few cases which can be given some superficially appropriate definition, it seems that we have to invoke different *concepts* for what it is that the word *paint* can be used to express; and these are arguably indefinitely rich and variable, complex and highly context-dependent. So, in all cases, the idea that there might be a unique correlation between words and some meaning that they express on the basis of which composite phrasal/sentential meanings can be explained turns out to be a non-starter.

This difficulty was recognised early in the systematic study of language, and some argued in consequence that the energies of the linguist should be directed to capturing the various *sense relations* which a word enters into, as the basis for capturing a more restricted sense of word meaning. This would be at least a step forward, since one would be expressing sense-relations between words, hence at least indicating the web of meanings into which a word can be seen to fit. In fact, this is what regular dictionaries do in practice: they define the meaning of a word by giving some other expression(s) of the language to which it might be said to correspond. We can indeed identify a number of sense relations that hold between words (and phrases); and a large body of work has been put into this enterprise. In particular, such work flourishes in computer science language-directed research,

where it plays an important role in developing more intelligent search strategies than just blind pattern-matching.

It is generally assumed that there are at least three basic types of sense relations:

(1.27) *Synonymy*: sameness of sense (pullover/sweater).

(1.28) *Hyponymy*: sense inclusion (cat/animal, house/dwelling).

(1.29) *Antonymy*: oppositeness in sense (cold/hot, dead/alive, big/small).

From these basic relations we can derive a web of connections among words in a language that permit a wide range of inferences over the sentences that contain them. So from (1.30) we can infer (1.31) (among many others):

(1.30) Joan's pullover is yellow.

(1.31) Joan's sweater is yellow.

from (1.32) we can infer (1.33):

(1.32) I do not like animals.

(1.33) I do not like cats.

and from (1.34) we can infer (1.35):

(1.34) This water is cold.

(1.35) This water is not hot.

There are many extensions to these basic relations that we will not go into here, including complex and non-traditional approaches to lexical relations that try to derive the intuitive inference from (1.36) to (1.37):

(1.36) John wants a hamburger.

(1.37) John wants to eat a hamburger.

There are very interesting challenges here as to how to distinguish what each word contributes; and giving classificatory lists of what is synonymous with what, what is an antonym of what, etc. may seem like a first step in meaning analysis – part of the gathering of data that is an essential prerequisite of theory construction. Certainly, discussion of such issues is always incorporated in basic linguistic semantics textbooks, but it quickly becomes apparent that these are little more than a distraction from the task of defining a general characterisation of what the meaning of a word consists in. Far from providing any such explanation, they simply presuppose that this question has been answered, and the classifications of uses of these words merely constitute a basis for gathering together those words that have the same or related meaning. All these lists are doing is indicating relations between words, not providing explanations of why these relations hold and how. At most, then, they set out the problems to

be explained; but merely looking through them, however assiduously, never in and of itself leads to the explanation that has to be constructed. That has to come from some external reflection of what it is that brings words together into these various classes.

Indeed, the discipline of collecting up appropriate databases of semantic relations provides a good illustration of an *inductive* approach to meaning. *Inductivism* is a term for the methodology which presumes that classifying data, facts under some description, is a necessary step in establishing theoretical explanations of phenomena, and, if done properly, can constitute a base from which theoretical explanations emerge. Moreover, as the argument would have it, the bonus of the inductivist methodology is that researchers are not imposing their own world-view or preconceptions about the data on the data themselves, for these are analysed prior to any such theory construction. However, this view of theory construction, and more particularly of linguistic theorising, is doomed to failure. Theories come from having an idea and then formulating a theory sufficiently precisely around it so that we can evaluate whether that idea is fruitful. It never comes just from making lists of data, as there are just too many data to know what to look for. As the philosopher Karl Popper notably pointed out to students, the instruction 'Observe!' is impossible to conform to, even in informal situations (Popper 1965), let alone when in search of a theory. One needs to have a hypothesis about what it is one is hoping to find, as driven by some background theory. Only then can the search be sufficiently focused to yield fruitful results, either to confirm one's current theoretical hypothesis or, through negative results, to lead one to modify one's theory and, that way, to gradually improve it. We need to know what it is that we are looking for. Observation alone, so Popper claimed, will never yield theoretical results. This was, at the time of Popper writing in the mid-1930s, a controversial stance, when an inductivist methodology of solemnly collecting supposed facts held sway. But now, in a modified form, this is a standard enough view of scientific practice. With inductivism never in principle able to lead to conclusions, but merely to confirmations of hypotheses, we need to state our theories about some phenomenon, in this case linguistic meaning, in terms that are sufficiently precise: in particular, they must either be falsifiable or at least sufficiently precise so that they can lead to other falsifiable hypotheses, each to be tested in their due turn.

For the particular challenge we face in linguistics, what we need in explaining meaning is some basis for formulating a model from which to start to explore a **formal** account of the basis of meaning for natural language; and then, having constructed such models, we evaluate them by assessing their ability to withstand constant attempts at refutation. This is of course just standard methodology for science as applied to semantic theorising. But it is pressing nonetheless. For if we want an explicit characterisation of the nature of language, and more particularly of interpretation of the lexical and phrasal expressions within that, we cannot fail to take up the challenge of constructing formal models to reflect the insights about language that we want to express.

1.2.1 Constructing a semantic theory

So what do we do to construct a theory of semantics? First, we have to set out the criteria that we expect minimally to be met by any part-way reasonable explication of interpretation for natural language expressions. And then we turn to putative models of language to see how well they can meet the target of satisfying those criteria. We have already touched on such minimal criteria. In the first place is the problem of the compositionality of meaning: the meaning of sentences and the phrases that make them up are dependent on the contribution made by the words they contain and the way such sentences are constructed – word order, voice alternations, and so on. An adequate semantic theory must provide an account of how the meanings assigned to words are put together in a systematic way by the syntactic constructions of a language to yield interpretations. And this process, whatever it is, must allow for recursive complexity in order to account for the multiple-embedding properties of natural languages.

Howsoever we characterise this relationship between a sentence as a form and its interpretation(s), there must be appropriate characterisation of the syntax–semantics relation, for there is, as we've seen, a systematic relation between the way the words are structured into units and the way in which they themselves contribute to the whole process of interpretation, however small a slice of meaning any individual word provides. Prediction of semantic relations such as *synonymy*, *hyponymy*, *entailment*, etc., must also be expected to be included in this list of criteria by which a putative semantic theory might be judged. Again, whatever the basis upon which interpretations of expressions are constructed, both simple and complex, there are systematic relations between expressions in virtue of such interpretations; and these a semantic theory should surely be able to characterise, much like a syntactic theory is expected to characterise which strings of a language constitute wellformed sentences.

Now a test of whether we are getting the right semantics for sentences is whether this specification will yield appropriate relations between sentences, said to hold in virtue of their meaning. As we saw above, certain inferential relations hold between sentences simply by virtue of the lexical relations that hold between the words they contain. But there are relations that hold between sentences by virtue, if you like, of their structure and of the grammatical expressions they contain. As with *homonymy*, *synonymy* and *antonymy*, we might thus recognise three primary relations that hold between sentences ($\models$ is the notational symbol for 'derivability in virtue of semantic content'):

(1.38)
a. *Entailment*: a sentence S_1 entails ($\models$) sentence S_2 if and only if the propositional content of S_1 includes that of S_2.
   King's College is on the Strand and is very noisy. $\models$
   King's College is very noisy.
   King's College is on the Strand. $\models$
   There is a building on the Strand.
b. *Paraphrase*: a sentence S_1 paraphrases sentence S_2 if and only if the propositional content of S_1 is identical to that of S_2 (mutual entailment).

Mary fed the cat. ⊨
The cat was fed by Mary. ⊨
It was the cat that Mary fed.

c. *Contradiction*: a sentence S_1 contradicts sentence S_2 if and only if the propositional content of S_1 necessarily excludes that of S_2 (S_1 entails the negation of S_2).
Mary likes dogs, but hates cats.
Mary does not like dogs.

Being able to predict these relations, presumed to hold among sentences, is one of the primary driving factors behind theoretical approaches to semantics, capturing entailments in particular. As we shall see as the book progresses, there are a number of different ways of going about this task with differing levels of success and with different implications for the nature of natural-language semantics and the way human beings understand what is said.

1.3 Breaking out of the language circle

In the search for a genuine basis for explaining what the intrinsic content of expressions of language consists in, and how they induce entailment relations, there are two alternative approaches that have been put forward, both serious contenders for success: one is a *representationalist* view that involves assuming representations of content as part of the explanation, the other involves only a mapping from words onto so-called *denotations*, that is, what the words can be used to make assertions about.

1.3.1 The language-of-thought hypothesis

On the first, psychologically based view, we use language to express concepts, and it is the concepts that we have constructed from words with which we reason about the world around us, not the words themselves. The words do no more than provide procedures to enable us to construct such concepts; and it is these which are systematically combined to form complex composite concepts, propositions, with which we reason. On this view, language is just one type of *input system* on a par with vision and other vehicles for retrieving the information that the world around us provides. With language and vision alike, the stimuli which these input systems process enable us to construct concepts with which we reason about the world around us. A systematic property of such input systems is, however, that the information which the particular stimulus intrinsically carries systematically under-determines the interpretations imposed upon it – indeed an input system must have this property to be economical and flexible. So the input stimuli we manipulate depend on context for the way in which they are interpreted. In the language case, too, it is the concepts that we use words to construct which denote the objects we use our words to refer to, not the words themselves. So, on this view, all cognition – vision, language-processing, hearing, processing smells – involves analysing input stimuli from which we

construct concepts that we take to be the content of the input information. This may seem a far-fetched view of language, and worse of vision, but in fact such a perspective is becoming a mainstream view of vision. As Francis Crick put it:

> It is difficult for many people to accept that what they see is a symbolic interpretation of the world – it all seems so like the 'the real thing'. But in fact we have no direct knowledge of objects in the world. (Crick 1994: 33)

On this view, there is ample room for incorporating theories of context. The concepts that we construct from words may naturally be said to be determined by context in one of two ways: either in interaction with those concepts that have just been constructed out of words uttered just previously, or from information independently constructed from other input, such as vision. The *underspecification* of language and its dependence on context is then seen to be a systematic part of what it means to be a sub-system of a cognitive system, clearly an advantage as an explanation of the psychological basis of semantic interpretation. On this view, input stimuli constrain but do not fully determine their interpretation, and this underspecification interacts with information provided by the immediate cognitive context to determine the concepts that we take to be the content of what it is that we see, hear, understand, and so on. This is a mind-internal process, hence computational and, in this sense, syntactic, a mapping from one form of mind-internal representation to another. Essential to this form of explanation is an internal system of conceptual representations – the so-called *language of thought*.

The instigator of this language-of-thought view, Jerry Fodor, puts it thus:

> It's entirely natural to run a computational story about the attitudes [beliefs, intentions and other kinds of thought] together with a translation story about language comprehension; and there's no reason to doubt, so far at least, that the sort of translation that is required is an exhaustively syntactic operation. . . Syntax is about what's in your head, but semantics is about how your head is connected to the world. Syntax is part of the story about the mental representations of sentences, but semantics isn't. (Fodor 1989: 419)

From a general viewpoint, as a programmatic statement, this perspective might seem surprisingly common sense, if you like it at all, despite the more abstract view of word meaning that it imposes. For the words, on this view, serve to provide constraints on the concepts that, in context, they are taken to express. It has to be said, however, that for a long time this perspective has mainly been articulated in the form of programmatic statements with little or no attempt to give formal substance to them. And, remember, we have committed ourselves to the working assumption that providing formal models that substantiate such programmatic statements is an essential prerequisite for any serious contender for an account of the nature of interpretation of a linguistic signal; and