Chapter 4

THE NEED FOR DISCOURSE THEME IN ANAPHORA RESOLUTION

> *The procedure is actually quite simple. First you arrange things into different groups depending on their makeup. Of course, one pile may be sufficient, depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo any particular endeavour. That is, it is better to do too few things at once than too many.*
>
> – *John D Bransford and Marcia K Johnson (1973)*[1]

In this chapter, we bring two more factors, which are interrelated, into play:

1 focus, and

2 discourse theme and discourse pragmatics.

In section 3.2.1 we introduced formally the concept of a focus set to model consciousness as a repository for antecedents, and we noted that the approaches described in section 3.1 do not explicitly use focus, but instead rely on a simple kind of history list to retain possible referents. In this and the following chapters we will consider in detail the problems entailed in focus:

1 Is an explicit focus really necessary?

2 What does focus look like? Is it just a set, or has it more structure than that?

3 How is focus maintained? What makes entities enter and leave focus?

We will also introduce the notion of discourse theme and ask ourselves:

1 Does an anaphor resolver need to use discourse theme?

2 How is theme related to focus?

3 How is theme determined?

[1] A paragraph said to have no theme, used in their experiments. Subjects found it very hard to comprehend or recall until it was given a theme by adding the heading *Washing Clothes*.

4.1. Discourse theme

To define the theme of a discourse, we appeal to the intuition as follows: The THEME or TOPIC of a discourse is the main entity or concept that the discourse is ABOUT – the subject central to the ideas expressed in the text, "the idea(s) at the forefront of the speaker's mind" (Allerton 1978:134). We use this intuitive definition because no more rigorously formal one is yet agreed on upon in linguistics.

A simple example: Is (4-1):

(4-1) The boy is riding the horse.

a statement about the boy or the horse? In this case, the answer seems to be clearly the former; *the boy* is the topic and *is riding the horse* is a comment about the topic.[2] As we shall see, however, the choice is not always as clear-cut as this. Much work has been done in attempting to capture precisely the concept of theme, and attempting to determine rules for deciding what the theme of a given text is. (See for example the papers in Li (1975).)

Let us begin by sorting out our terminology. To the confusion of all, different workers have used different nomenclatures, often describing the same concept with different words, or different concepts with the same words. I suspect that the failure of some people working in the field to realize that they and their colleagues were not talking the same language has hindered progress in this area. The following table summarizes terminology used:[3]

The boy	*is riding the horse*	**Used by**
topic	comment	Sgall et al (1973)
theme	rheme	Halliday (1967)
old	new	Chafe (1970)
given	new	Haviland and Clark (1974), Clark and Haviland (1977), and Allerton (1978)
logical subject	logical object[4]	Chomsky (1965)
focus	–	Sidner (1978a, 1978b)
psychological subject	psychological predicate	Hornby (1972)

[2]This is not the case in all contexts. If (4-1) were the answer to (i):

(i) Who is riding the horse?

then *the boy* would be the comment and *riding the horse* the topic.

[3]While the words in each column describe closely related concepts, it should not be inferred that they are precisely synonymous. In particular, Halliday (1967) and Allerton (1978) draw a distinction between theme and old, and between rheme and new (see section 4.1.1).

(See Allerton (1978) for a more detailed discussion of terminological confusion.)

In this thesis I will follow Allerton (1978) and use the words *theme* and *topic* interchangeably; but I will also need to make a distinction not yet commonly recognized explicitly in the nomenclature jungle: I will use *LOCAL THEME* or *LOCAL TOPIC* to refer to what a SENTENCE is about, and *GLOBAL THEME* or *GLOBAL TOPIC* to refer to what a DISCOURSE is about at a given point. These two concepts often coincide, but frequently don't. For example, in (4-2):

(4-2) Nadia's chinchilla is shaped like a pear with a brush for a tail. Its teeth are long, but not very sharp.

the local and global topics of the first sentence are both *Nadia's chinchilla*. In the second sentence the global theme is unchanged from the first sentence, while the local theme is now *Nadia's chinchilla's teeth*.

There are currently two major paradigms in investigating problems of discourse theme. The theoretical approach, initially centred in Europe, uses introspective linguistic analysis, and is typified by the work of Firbas (1964), Sgall, Hajičová and Benešová (1973), Halliday (1967), Chafe (1970, 1972, 1975) and many of the papers in Li (1975). The experimental approach uses the techniques of psycholinguistics, and is typified by the work of Hornby (1971, 1972) and Johnson-Laird (1968a, 1968b). First we will look at each paradigm in turn, and then at their applications in computational analysis of language.

4.1.1. The linguistic approach

Chafe (1970:210-233, 1972) discusses the relationship between the topic of a sentence and the information in it which is not new. For example, in (4-1), it is assumed that the boy is already being talked about, and is therefore the topic, while the new information conveyed is what the boy is doing, riding the horse, and this is therefore the comment. Chafe describes given, or old, information as that already "in the air", used as a starting point for the addition of further information. Old information need not be explicitly spoken;[5] it may be something assumed to be known to both speaker and listener. For example, if I come up to you and say (4-3):

---

[4] *The horse* rather than *is riding the horse* is the logical object in Chomsky's nomenclature.

[5] A common literary device, for example, is to begin a novel with a sentence that presumes information, forcing the reader to immediately construct a mental frame containing this information, thereby plunging them straight into the story.

A similar phenomenon occurs when sentences are presented in a contextual vacuum, as are most of the example texts in this thesis. A series of experiments by Haviland and Clark (1974) showed that people take longer to comprehend sentences which presume ungiven information, implying that time is taken to create or invoke the mental frame required to understand the sentence.

(4-3) Hi! Did you hear that Ross was arrested on a morals charge?

it is assumed that we both know who Ross is. If I added the word *again*, it is also assumed we know about his previous arrest, and the new information that I am giving you is that it happened once more.

Halliday (1967) and Allerton (1978) refine the concept thus: given is what was being spoken about before, while theme is what is being spoken about now, these not necessarily being the same thing.

The concept of theme has been generalized somewhat by Chafe (1972) to that of FOREGROUNDING; if the topic is what is "in the air", then foregrounded items are those "on stage"; they are those "assumed to be in the hearer's consciousness" (Chafe 1972:50, 1974). When a lexical item occurs in a discourse, it automatically becomes foregrounded in future occurrences, says Chafe, until it retreats to the wings through lack of further mention. How long this retreat takes is unclear, and probably varies depending on other items taking the places, or "slots", of previous ones. Clearly, foregrounding is very similar to what we have been calling focusing.

In verbal discourse, a lexical item is signalled as being the theme or as being in the foreground by vocal tone, stress and gesture, as well as by textual devices. We see in (4-4) and (4-5) that the comment is stressed and the theme is not:

(4-4) What is Nadia doing?
Nadia is PRACTISING ACUPUNCTURE.
*NADIA is practising acupuncture.

(4-5) Who is practising acupuncture?
NADIA is practising acupuncture.
*Nadia is PRACTISING ACUPUNCTURE.

In written language the topic is usually indicated by syntactic, semantic and pragmatic cues, though italics or upper case may be used to simulate vocal stress.

We see, then, that the linguistic approach assumes that we have an intuitive idea of what topic is, and tries to formulate rules to formalize this idea. It has, however, yet to agree on any precise definition of theme, or produce any formal method for determining the theme of a sentence or discourse by computational analysis.

4.1.2. The psycholinguistic approach

To determine what subjects THOUGHT the theme of a sentence was, Hornby (1971, 1972) used the following experimental procedure: A number of pairs of pictures were drawn with each picture having three components, two objects and an action. The action was the same in each of the pair. A typical pair

showed *(a)* an Indian building a tepee and *(b)* an Eskimo building an igloo. For each pair, subjects were presented with sentences which described each picture with partial correctness. For the above pair, typical sentences were (4-6) and (4-7):

(4-6) The Indian is building the igloo.

(4-7) The one who is building the igloo is the Indian.

Subjects were asked to pick which picture each stimulus sentence "is about, even though it is not exactly correct" (1972:637). In the above example, most felt that (4-6) was nearest to *(a)* and (4-7) to *(b)*. The component that is the same in both picture and sentence (here, Indian and igloo respectively) is then assumed to be the psychological subject, or local theme.

Hornby found that the theme of a sentence is not necessarily either the syntactic subject or the first item mentioned, a result contrary to suggestions that word order determines theme (Halliday 1967) or that case relationships play a role independent of surface syntax (Fillmore 1968).

4.1.3. Lacunae abounding

Although much work has been done in the area of theme, there is little of substance to use. The linguistic approach has served to intuitively define for us the concepts of theme and foreground, but has given us no way to find them in a text, even though, as we will see, finding them is a necessity in NLU. Similarly, the psycholinguistic approach has so far shown us where not to look for rules about theme, but has not helped us find them.

I believe that Hornby's experiments point us in the right direction: the theme of a sentence is a function of, inter alia, both its construction and the case relationships therein, and, if in a context, then of the topic of the previous sentence as well. It therefore remains to find this function. From this should follow rules for the foreground, which we can use in deciding when things no longer remain in focus. Despite the simplicity with which it can be stated, this goal is, of course, a major research problem. In the next chapter we will look at some recent approaches to it.

4.2. Why focus and theme are needed in anaphor resolution

Is a recency list really inadequate as a focus for anaphor resolution? Does discourse theme really play a role? In this section I will show that the answer to both these questions is "yes".

Taking an opposing view, Yorick Wilks (1975b) rejects the use of theme, except as a last resort, on the basis of the following examples:

(4-8) John left the window and drank the wine on the table. It was good.

(4-9) John left the window and drank the wine on the table. It was brown and round.

(These examples, together with (4-10), will be referred to below as the 'table' examples.) In (4-8), *it* clearly refers to the wine. In (4-9), things are not so clear; Wilks says that *it* must mean the table, and, uncoincidentally, the anaphor resolution component of his natural language system comes to the same conclusion, using the method of "preference semantics" (see section 3.1.7), whereby the table is chosen as the referent on the grounds that it is much more likely to be brown and round than the window or the wine. Since the wine (but not the table) is the theme here, Wilks concludes that we can therefore "reject all simple solutions based on [theme]"[6] (1975b:68).

The problem is that Wilks's interpretation of the sentence is wrong, or at best idiolectic. In my idiolect, (4-9) could only be describing the wine as brown and round (adjectives which make as much sense as many of the other terms often applied to wine).[7] Informants, speakers of American and Australian English, agreed. One described (4-9) as an absurdity, and when told that *it* meant the table replied that that possibility had not even occurred to them. When I included (4-9) in a conference presentation (Hirst 1977a), the audience laughed at it. Clearly, (4-9) is ill-formed.[8]

Example (4-9) is ill-formed because when *it* is encountered in the text, ***the table*** is no longer in focus; that is, it cannot be referred to anaphorically,

---

[6]The word in brackets was originally *focus*; where Wilks uses this term, he apparently means *discourse theme*, *topic*, or *focus of attention*. To avoid confusion with our sense of the word *focus*, I have amended this quotation.

[7]Compare Lehrer (1975), who showed that many oenological terms contain zero bits of information.

[8]This points out the danger, well known in linguistics but perhaps not in artificial intelligence, of losing one's intuition for even one's native language. (Spencer (1973) has shown that linguists have quite different intuitions regarding grammaticality and acceptability from non-linguists.) When generating sample sentences to demonstrate a point about the nature of language, it is surprisingly easy to come up with ill-formed or marginal sentences without being aware of the fact. (See also Carroll and Bever (1978), whose experiments suggest that linguistic intuition varies with context and mental state, including degree of self-awareness.) It is therefore advisable to at least test examples on informants (namely, long-suffering non-linguist friends) before using them. I have done this with important and/or contentious examples in this thesis, but nevertheless do not believe that I am necessarily innocent of generating ill-formed sentences myself. This is why I have, throughout this thesis, where possible, taken my examples from "real-world text", and given a complete citation of the source. Nevertheless, real-world text is sometimes suspect – people inadvertently write sentences they themselves would not accept, and some people are just plain illiterate – and in some instances I have marked real-world text used in this thesis as ill-formed when it grated my idiolect. (In section 7.3, I address the question of better alternatives for obtaining or testing linguistic data.)

A related problem is that of idiolects. Some examples in this thesis were acceptable to some but not all informants (all such examples are so noted). I concede that my difference here with Wilks may be merely idiolectic; however, his idiolect appears to be in a small minority (not that that proves anything).

4.2 Why focus and theme are needed in anaphor resolution

notwithstanding that only a period separates it from the *it*. (We will see in section 5.1.2 an explanation of why this happens.) Clearly, an anaphor resolver with nothing more than a history list ordered by recency would fail to find (4-9) ill-formed;[9] a similar language generator could erroneously produce it. Moreover, the recency-list approach would spuriously consider (4-10) ambiguous, though it isn't:

(4-10) John picked up the toy on the table. It was made of wood.

and then choose the wrong "possibility", namely the table being wooden, on grounds of greater recency and equal reasonableness.

To show that the argument above does not rest solely on the idiolectic acceptability or not of (4-9), here is another example:

(4-11) If an incendiary bomb drops near you, don't lose your head. Put it in a bucket and cover it with sand.[10]

There are only two candidates for the first *it* here: ***an incendiary bomb*** and ***your head***. Semantics and world knowledge indicate the former, as its speaker presumably intended, yet the latter unambiguously "sounds like" the correct referent despite the nonsense resulting; and therein lies the jest. That ***your head*** is the referent despite the presence of a better choice means that the better choice violated other constraints which prevented it even being considered as a candidate in the resolution. These constraints are those of focus: ***an incendiary bomb*** was not properly in focus at the time of the first *it* and therefore was not available. However, ***your head*** appears to be the topic of the sentence despite the need to fracture the idiomatic expression, and is ipso facto the "dominant" item in focus.[11] When presented with (4-11), Wilks's preference semantics program would not, I think, see the humour, but would wrongly choose the bomb as the referent of *it*.

The above discussion demonstrates that focus is an integral part of language (or at least of English). Any anaphora resolution system should therefore take it into account; failure to do so will result in the wrong answers.

A second reason for maintaining a focus is that without it the number of possible referents grows with the length of the text. Clearly an NLU system

---

[9] An important point relevant here is the comprehension of ill-formed sentences: humans can do it in many cases, and it is desirable for computer natural language understanders to do so too. Baranofsky (1970), for example, gave heuristics for resolving the relative pronoun in sentences such as (i):

(i) *A man went to the fair who lost his mind.

Wilks might therefore defend his system as one which has the bonus advantage of understanding ill-formed sentences. But then he could not reject theme-based resolution on the basis of (4-9). In addition, we surely want such a system to try all possible well-formed interpretations first, and flag a sentence for which it is forced to make an assumption of ill-formedness.

[10] This text is of obscure origin, but is usually alleged to have come from a British air raid precautions leaflet during World War II.

[11] See section 5.1 for support for this assertion.

4.2 Why focus and theme are needed in anaphor resolution

attempting to read a scientific paper, for example, should not, on the fourth page, look back over all entities evoked by the entire preceding text for the most reasonable antecedent for an anaphor. But, as should be clear by now, a simple shift register, saving the last n possible antecedents or those from the last n sentences, is not enough.

We now agree that focus is necessary. The following examples demonstrate that discourse THEME plays a role in focus:

(4-12) Nadia hastily swallowed the licorice, and followed Ross to the bathroom. She stared in disbelief at the water coming out of the tap; it was black.

Wilks's preference semantics system will (as far as I can determine from his 1975b paper) choose *licorice* over *water* as the referent of *it*, because licorice is more likely than water to be black. The licorice should have been discarded from focus by the end of the first sentence of (4-12). It is out of focus because it is unrelated to the discourse topic or theme, the strange events in the bathroom, at the point the anaphor occurs.

Now consider this text, from *Wheels*,[12] in which the president of General Motors discusses with his wife charges brought against the motor industry by Vale, a Ralph Nader–like character:

(4-13) She continued, unperturbed, "Mr Vale quotes the Bible about air pollution."

"For Christ's sake! Where does the Bible say anything about that?"

"Not Christ's sake, dear. It's in the Old Testament."

His curiosity aroused, he growled. "Go ahead, read it. You intended to, anyway."

"From Jeremiah," Coralie said. "'And I brought you into a plentiful country, to eat the fruit thereof and the goodness thereof; but when ye entered ye defiled my land, and made mine heritage an abomination.'" She poured more coffee for them both. "I do think that's rather clever of him."

Vale is still available to Coralie in her conversation as an antecedent for "him" after eight intervening sentences of the conversation, and her anaphor is quite comprehensible to us in the written report of the conversation, despite ten intervening sentences which contain two other possible referents – the president of General Motors and Jeremiah. This is possible because Mr Vale and his quotation is the topic of the whole conversation. It may be objected that there is no possible confusion – Vale is the only referent for *him* that makes sense; in particular, Coralie would not refer to her husband in the third person when addressing him. But as we saw with (4-9) and (4-11), "making sense" is not enough. In any case, it is non-trivial to exclude the interpretation in which *him* means Jeremiah, and Coralie is commenting on something like the clever

[12]Hailey, Arthur. *Wheels*. New York, 1971, page 2. Quoted by Hobbs (1977).

use of language in the quotation. It is also apparent that the reference is to Mr Vale as a concept in consciousness rather than the words *Mr Vale*, which are almost certainly forgotten by the reader by the time the reference occurs.

Here is another example of reference to discourse topic:

(4-14) *Dear Ann:* No lectures on morality, please, I'm not asking you whether or not I should continue to sleep with this man. I have already decided that he is better than nothing. Now to the problem:

The guy's toenails are like razor blades. I get up some mornings and feel like I've been stabbed. I have mentioned this to him a few times, but he does nothing about it. I need help. – CLAWED-A-PLENTY

Answer: Buy King Kong a pair of toenail scissors. Be extra generous and offer to trim them for him. If he refuses, insist that he sleep with his socks on – or move to another bed.[13]

Them is the toenails in question, the topic of the second and third paragraphs, but not the actual text *the guy's toenails*, which is too far back to be recalled word for word. Nor is *them* a strained anaphor into *toenail scissors*, as the reference is ill-formed if the first two sentences of the answer are taken out of context. (In passing, we also notice in (4-14) the epithet *King Kong*, which requires a large amount of world knowledge and inference to recognize and comprehend.)

Lastly, consider this text:

(4-15) The winning species would have a greater amount of competitive ability than the loser as far as that resource axis of the *n*-dimensional niche is concerned (e.g. it would be more adapted to using that resource in that particular habitat).[14]

Not only is *the winning species* the local theme and the antecedent of *it*, but it is the only item in focus. None of the more recent NPs – *a greater amount*, *a greater amount of competitive ability*, *competitive ability*, *the loser*, *that resource axis*, *the* n-*dimensional niche*, *that resource axis of the* n-*dimensional niche* – can be referred to by this *it* regardless of the text that follows it. That is, there is NO text which could replace the text after *it* in (4-15) and make a well-formed sentence in which *it* refers to one of the more recent NPs.[15]

---

[13]From: Landers, Ann. [Advice column]. *The Vancouver sun*, 11 August 1978, page B5.

[14]From: Mares, M A. Observation of Argentine desert rodent ecology, with emphasis on water relations of *eligmodontia typus*. in: I Prakash and P K Ghosh (editors). *Rodents in desert environments* (= Monographiae biologicae *28*). The Hague: Dr W Junk b v Publishers, 1975.

[15]For support for this type of assertion, see section 5.6.

4.2 Why focus and theme are needed in anaphor resolution

4.3. Can focusing be tamed?

Implicit in the preceding discussion is the assumption that given any point in a text there is a set of focus sets associated with that point. It should be clear from our exposition so far that this is indeed the case. What is not so clear is how we can know the contents of these focus sets. For example, if the point is a pronoun, P, we are interested in knowing the contents of the nominal focus set F_n, which consists of all those concepts that P could refer to for some following text. More formally, F_n is a function of P and the preceding text t defined by:

(4-16) $Fn(t,P) = \{n \mid n$ is a noun phrase contained in t, or a concept evoked by t, and there exists t' such that tPt' is well-formed English text in which P refers to $n.\}$

At any given time, the nominal focus set F_n contains zero or more entities – foregrounded items – which are possible referents for anaphors. When a pronominally referent anaphor needs resolving, one of several cases can occur:

1 There is exactly one noun phrase in F_n which fits the basic syntactic and selectional constraints (see Chapter 6); it is chosen as the referent.

2 There are no suitable members of F_n; then either the alleged anaphor is really a cataphor or exophor, or the sentence is ill-formed.

3 There is more than one suitable member of F_n; then either *(a)* we need to choose one of these possibilities, or *(b)* the sentence is ambiguous.

Case 3*(a)* is the one of most interest here. Many apparent ambiguities can be resolved by knowing what the topic is. We have already seen one example of this:

(4-17) Ross asked Daryel to hold his books for a minute.

This is unambiguous in most idiolects because the topic indicates that *his* means *Ross's*. In general, the present topic is the default referent, and this is why we would like to be able to determine the topic of a sentence.

The definition of F_n above is clearly not of much use computationally, as it begs the question: it assumes the anaphor resolution capability of which it is itself a part. Therefore, if we intend to make use of focusing, we will need other, easier, rules to determine the contents of the focus sets. It is likely that such rules exist – humans, after all, have no problems – but finding them may be difficult. However, we have no choice but to search.

Let's summarize: In this chapter, I have tried to show that focus and theme are necessary in anaphora resolution, and that they are closely related. In the next chapter, we will look at the nature of this relationship and at some attempts to discover rules for focus.