Chapter 4

A Data-Driven Methodology for Motivating a Set of Relations

4.1 Introduction

In the previous chapter, it was argued that linguistic devices (in particular, cue phrases) can be taken as evidence for relations, provided these are thought of as constructs which people actually use when creating and interpreting text. This chapter describes how a set of relations can be determined and justified in the light of this argument. The methodology is incremental—it consists of a series of relatively simple linguistic tests, which can be performed quite systematically, with a minimum of inter-analyst disagreement.

To begin with, in Section 4.2, a suitable method is sought for defining 'cue phrases', without relying on terminology from existing theories of discourse. Cue phrases must be characterised independently in order to avoid circularity: since they are to be used to motivate relation definitions, no reference can be made to their role in signalling relations. Instead, a linguistic **test for cue phrases** is proposed, which makes use of readers' intuitions about the coherence of certain constructed mini-discourses.

Section 4.3 describes how a corpus of cue phrases is gathered using this test, and provides some preliminary discussion of its size, and the variation within it. In Section 4.4, a second linguistic test is presented, for classifying cue phrases into groups of synonyms and hyponyms: the test basically determines whether one cue phrase is **substitutable** for another. The results of the test are presented in the form of **substitutability diagrams**, which are explained and illustrated in Section 4.5. Using the test, a **taxonomy** of cue phrases is constructed: this taxonomy is described in Section 4.6.

4.2 Firming Up the Notion of 'Cue Phrase': A Test for Relational Phrases

The first task is to formulate a precise definition for the class of phrases which is under investigation. Some definitions of 'cue phrases' exist already, but these are often internal to the theory of discourse being proposed. For instance, Cohen (1984) defines 'clue words' as 'special words or phrases directly indicating the structure of the argument to the hearer'; Hirschberg and Litman (1993) define cue phrases as 'words and phrases that directly signal the structure of a discourse'. With such definitions, in order to decide what counts as a cue phrase, we already need to know what 'the structure of a discourse' is. In order to avoid circularity, the constructs used in the discourse theory must be justified using some other criterion.

As an alternative to this approach, cue phrases, if given an independent definition at the outset, can be used in motivating the constructs used in the discourse theory. This is the approach adopted here, and sanctioned by the arguments in the previous chapter.

In an attempt to come up with a precise yet theory-neutral definition of cue phrases, a linguistic test is proposed which picks out a certain set of phrases as they occur in natural discourse. The test is given in Figure 4.1 below. It is designed to pick out all sentence and clause connectives, but to stay away from methods of realising relations within a single clause.

In order to avoid any terminological confusion, we can refer to the class of phrases which pass this test as the class of **relational phrases**—although since this is quite a mouthful, the term 'cue phrase' will continue to be used, with this new technical meaning.

The central idea behind the test is that cue phrases have a function which extends beyond a single clause. They link clauses and sentences together to create larger units of text; therefore they cannot be made sense of when associated with one clause in isolation. Thus the clause

(4.4) *Because* Bill owed John money

is impossible to understand without prior linguistic context, but can be understood when the cue phrase is removed:

(4.5) Bill owed John money.

In order to make the test work, any anaphoric or cataphoric expressions in the clause to be isolated must be replaced by their referents; otherwise it would be impossible to interpret out of context regardless of whether or not it contained a cue phrase. However, propositional anaphora *within* the candidate phrase should not be substituted: thus complex constructions like *because of this* or *for this reason* will also be identified as cue phrases. There are good grounds for opting to allow anaphoric expressions within cue phrases: as Halliday and Hasan (1976) point out, many *bona fide* cue phrases

1. Isolate the phrase and its **host clause**. The host clause is the clause with which the phrase is immediately associated syntactically; for instance, if the passage of text to be examined is

   (4.1) ...John and Bill were squabbling: John was angry *because* Bill owed him money. That was how it all started ...

   then the isolated phrase and clause would be

   (4.2) *because* Bill owed him money.

2. Substitute any anaphoric or cataphoric terms in the resulting text with their antecedents, and include any elided items. For the above clause, this would result in

   (4.3) *because* Bill owed John money.

   Propositional anaphora *within the candidate phrase itself* should not be subsituted, however. Thus if the candidate phrase is *because of this*, the propositional anaphor *this* should remain.

3. If the candidate phrase is indeed a relational phrase, the resulting text should appear **incomplete**. An incomplete text is one where one or more extra clauses are needed in order for a coherent message to be framed. The phrase *because Bill owed John money* is incomplete in this sense: it requires at least one other clause in order to make a self-contained discourse. Even the fact that it could appear by itself on a scrap of paper (say as an answer to a question) does not make it complete; the question is essential context if it is to be understood.

   Note that it is only additional clausal material which is to be removed in the test. Any additional contextual information necessary for the comprehension of the clause (for instance, knowledge of the referents of definite referring expressions like *John* and *Bill*) can be assumed to be present.

4. Any phrases which refer directly to the text in which they are situated (such as *in the next section*, *as already mentioned*) are to be excluded from the class of relational phrases. Such phrases pass the test—but only because their referents have been expressly removed through the operation of the test itself.

5. Phrases which pass the test only because they include comparatives (for instance *more worryingly*, *most surprisingly*) are also to be excluded from the class of relational phrases. Stripped of the comparatives, such phrases do not pass the test. Comparatives like *more* and *most* introduce a very wide range of adverbials, bringing the compositional resources of the language quite strongly into play. Since we are more interested in stock words and phrases that have evolved to meet specific needs, phrases involving comparatives will not be considered as relational phrases.

6. Sometimes, more than one cue phrase can be found in the isolated clause (eg *and so*, *yet because*). In such cases, both phrases should pass the test when considered individually in the same context. In other words, the host clause should appear incomplete with either phrase.

Figure 4.1: Test for Relational Phrases

derive etymologically from phrases involving anaphora (consider words like *therefore* and *thereby*); so it seems reasonable also to allow as cue phrases expressions with an explicit anaphoric component.

The test is designed to give a reasonably objective way to pick out a set of phrases to act as the object for further study. It calls on analysts' intuitions, but there is no need for 'coherence relations' to be explicit in their minds: judgements are purely about whether or not given clauses make sense when isolated from their context. There are of course many cases where the test is hard to apply. One problem in the application of the test is to decide whether the required context for a mini-discourse is linguistic or non-linguistic. For instance, consider this discourse:

(4.6) *But* you can't just leave us here!

It is possible to imagine this discourse with no previous utterances at all. All the same, it needs to be interpreted as a reaction to a previously existing propositional attitude (in this case, perhaps an intention to leave), and so can arguably be interpreted as part of a relation between propositions.

It might be considered that the test is overly restrictive in some cases. For instance, as it stands, several phrases used to signal 'purpose'-type relations are excluded. Consider the following two cases:

(4.7) Bill escaped from prison *by* bribing a guard.

(4.8) Bob used the crowbar *to* lever open the window.

Bribing a guard is not a full sentence, and cannot stand alone; *Lever open the window* can be interpreted as a sentence, but only as an imperative, quite different from its original meaning. Other phrases used by Vander Linden (1992) to signal purposive relations are even more clearly ruled out, such as the preposition *for*:

(4.9) Bob used the crowbar *for* extra leverage.

Since the relation here is realised within a single clause, the candidate phrase's host clause contains both parts of the relation, and can *only* be interpreted when the phrase is present!

It would be useful to have a way of expanding the test to allow for additional phrases such as these. But the decision here has been to keep the test reasonably simple, rather than to extend it until it covers exactly the range of phrases we think should be included. Hopefully, when the present corpus of phrases has been analysed, a more principled method for identifying cue phrases can be found.[1]

[1] Note that as it stands, the test works less well in other languages, where connectives often exert a grammatical influence on their host clauses. (For instance, German subordinators can alter the position of the verb; conditional phrases in many languages require a clause in the subjunctive.) It may be that the standardisations of clauses required to overcome these problems are also sufficient to expand the test's English coverage in some of the ways required.

4.3 Gathering a Corpus of Cue Phrases

Using the test, a corpus of cue phrases has been gathered. In order to limit the scope of the investigation, the source texts were all from the same genre of discourse: 'academic' writing, such as can be found in journals or academic books. It is likely that different registers of text require slightly different sets of cue phrases: for instance, phrases like *just then*, *whereupon* and *sure enough* occur in narrative discourse but are unlikely to appear in academic articles. At the same time, the texts analysed in this study occasionally switch into different genres; for instance, a narrative genre is often adopted to describe a case study. (For this reason, the corpus contains some phrases which might not seem typical of 'academic' writing.)

All corpus analysis was carried out by the author. 226 pages of text were analysed altogether, from twelve different authors. This yielded a corpus of around 200 phrases. There was found to be extensive use of a core of phrases across all the authors: for instance, *and*, *since*, *if*, and *but* were used by all twelve; *on the other hand*, *however*, and *also* were used by eleven; and *then*, *for example*, *because*, *when*, and *although* were used by ten.[2]

Following the study, the corpus was somewhat enlarged, again by the author, as new phrases not encountered in the original analysis were discovered. Some of these phrases might well have been found if a larger amount of text had been searched. Others are not typically found in 'academic' discourse, but have been included because they provide interesting contrasts for subsequent discussion. For each new addition, contexts have been found in which new candidate phrases pass the cue phrase test. The enlarged corpus, containing some 350 phrases, is given in Appendix A. While there are doubtless other phrases still to be included, it is the largest corpus of connective phrases that I am aware of in the literature.

Sections 4.3.1 and 4.3.2 report the results of two preliminary analyses carried out on the corpus.

4.3.1 The Syntactic Diversity of cue Phrases

Cue phrases fall into five syntactic classes (as defined by Quirk *et al* (1972)):

- **Coordinators:** these always appear in between the clauses they link; the clauses can be in separate sentences or in the same sentence. If in the same sentence, no punctuation is required in addition to the coordinator; and if combined in a sequence with other cue phrases, coordinators always appear leftmost in the sequence. For example:

  (4.10) a. An object may move *but* it remains the same object...

  b. A general rule is needed to prevent comparative constructions. *Or* some rule is needed that will say: 'if a word cannot ...

[2] Of course, it might be objected that *if ... [then]* is really the cue phrase, rather than simply *if*; the test for cue phrases does not capture this as it stands. The issue of the inter-dependence of cue phrases is raised in Section 4.4.

- **Subordinators:** these introduce subordinate clauses in complex sentences. The subordinate clause can be on the left or the right of the main clause, but the subordinator is always on the left of the subordinate clause. For example:

(4.11) a. *Although* it is common sense that labels are related, this is a difficult idea to explicate.
b. One further illocution should be considered *before* we discuss some variants.

- **Conjunct adverbs:** these modify whole clauses, and can appear at different points within them, although there is often a default position for particular phrases. There are also syntactic constraints on exactly which positions conjunct adverbs can occupy: at the beginning of a clause, between subject and verb, between any auxiliary verbs, between auxiliary verb and main verb, after a copula if there is one, before a sentential complement if there is one. For example:

(4.12) a. The parallel between permissibility and possibility has been exploited by many linguists. There are, *however*, two important distinctions between them ...
b. We will select only those hypotheses we deem relevant. *As a consequence,* our discussion differs from the usual views ...

- **Prepositional phrases:** these often contain propositional anaphora referring back to the previous clause. For example:

(4.13) a. It has a high degree of opacity. *In that respect* it resembles glass.
b. The plate extends as far as the Pacific coast. *At this point* it slopes down.

The distinction between prepositional phrases and conjunct adverbials is often hard to make. I have tended to include phrases in the latter category if they are best analysed as idiom chunks, and in the former category if they retain a fair degree of compositionality—see Section 4.3.2 for further details.

- **Phrases which take sentential complements:** these often introduce a particular intentional stance with respect to the content of the clause they introduce. For example:

(4.14) a. An act that is physically impossible cannot occur. (...) *It follows that* the language used ... is often straightforward.
b. *It may seem that* we are making too much of orientation; but characteristic orientation is not an idiosyncrasy.

4.3.2 The Space of Cue Phrases

One finding that emerged from a study of the corpus was that the cue phrases fell into two groups. It was possible to envisage a compositional semantics for some phrases: for instance, the semantics of the phrases *many years later*, *a few years later*, and *twenty-five years later* can be seen to depend on the semantics of the constituent phrases *many years*, *a few years*, and *twenty-five years*; and these in turn depend on the semantics of the determiner phrases *many*, *a few*, and *twenty-five*.

Other phrases in the corpus, which we might refer to as simple phrases, are impossible to break down in this way. This might be because they are single words, or alternatively because they are idiom chunks, which are defined precisely as multi-word phrases whose semantics is not compositional. Examples of idiom chunks in the corpus include *on the other hand* (in contrast with the ungrammatical *off the other hand*), *after all* (in contrast with *before all*), and *given that* (in contrast with *taken that*).

A great many phrases seem to be *partly* compositional—for instance, the meanings of *on the one hand* and *on the other hand* can be thought to hinge on the meanings of *one* and *other*, but not on the meaning of *hand*: the phrase *on the other foot* is nonsense as a marker of a relation. There are many other phrases of a similar 'semi-compositional' status; but there seem to be no hard-and-fast rules for working out how such phrases are formed, and it is easiest at the outset simply to treat them as unanalysed atomic elements.

The existence of compositional cue phrases has an important consequence: it makes the class of cue phrases infinite in size. Phrases like *very very ... very many years later* are technically members of the class, even though in practice they will never occur. This means that in order to describe the class, it is necessary to lay down rules for how compositional cue phrases can be constructed. These rules will be syntactic in nature. For instance, the following two rules are helpful in expanding the charted space of cue phrases:

- There is a class of words which modify all subordinators and only subordinators; these words are *even*, *just*, *except*, *only* and *especially*. Thus we can construct cue phrases like *only where*, *except before*, and *just on the grounds that*. There are exceptions to this rule (for instance, **except in case*), but it still provides a useful generalisation.

- Temporal phrases can also be modified in a systematic way. The conjunct adverbs *earlier*, *afterwards* and *later*, as well as the phrases *before* and *after* (which can be conjunct adverbs or subordinators), can all be modified by any expression denoting a length of time; for instance *three days after*, *a minute earlier*, and *some time before*. The modifiers always precede the head phrases.

The general syntactic concepts of **head** and **modifier** can be used to analyse any phrase, regardless of its syntactic category. Compositional cue phrases can typically stand alone without modifiers—for instance, *later* and *after* by themselves are still cue phrases. In what follows, modifiers have been stripped wherever possible. To reiterate the point made in Section 4.2: we are not interested in phrases in which the full compositional power of the language is brought to bear; we are interested in the 'stock' words and phrases, which have evolved to meet specific communicative needs.

4.4 Organising the Corpus: A Test for Substitutability

Thus far, we have used a simple test for detecting cue phrases in text, and on the basis of this we have gathered a corpus of cue phrases. The phrases have been classified according to their syntactic properties, so that an idea can be obtained of the complete space of phrases. But since we are principally interested in cue phrases as signallers of discourse structuring strategies, a classification of phrases according to their function in discourse is our central objective.

In keeping with the data-driven methodology adopted thus far, the classification will be made by means of a simple linguistic test, rather than by making theoretical claims about the semantics or pragmatics of the phrases in the corpus. The test is to do with **substitutability**. Very broadly, if two phrases are inter-substitutable in a passage of discourse then they should be classified in the same category. If one phrase can always be substituted for another, but not *vice versa*, then the latter phrase should be classified in a category subordinate to that of the former phrase. In this way a taxonomy of synonyms and hyponyms can be constructed. It will also be interesting to represent those groups of phrases which can *never* be substituted for each other, and those which can *sometimes* be substituted for each other, in certain contexts.

The approach here is similar in many ways to that taken in the WordNet project (Beckwith *et al* (1990), Miller *et al* (1990)). WordNet is a lexical database organised on psycholinguistic principles: it comprises taxonomies of nouns, verbs and adjectives, which represent various different relationships between words, such as synonymy and antonymy. The classification of cue phrases makes use of different relationships, but the idea of a hierarchical taxonomy of words and phrases is the same.

The test for substitutability is given in detail in Figure 4.2. The main idea is that the tester considers a cue phrase in a context where it naturally occurs, and then considers which other phrases (s)he, *as a writer*, would be prepared to use in its place. This is a task which occurs quite regularly during the course of normal writing. The tester might imagine that the original phrase has been used recently in the preceding discourse, and needs to be changed for reasons of 'elegant variation'.

As it will be seen, the conditions for substitutability are slightly less constrained than those under which one phrase can simply replace another. To begin with, we are not interested in whether two phrases can take the same *grammatical* position in a clause; rather, we are interested in whether they have the same function in signalling discourse relations between the clause and other units. For instance, a conjunctive adverb like *nevertheless* might have the same function as a coordinator like *but*, but the latter can only appear at the beginning of a clause, so simple replacement will not always be possible. In view of this, candidate phrases can be substituted in the clause in a different position, from the original phrase, if necessary.

For another thing, when it comes to comparing the original text with its counterpart containing the candidate phrase, there are some factors which are not taken into account. Stylistic mismatches are ignored; *moreover* is thus a legitimate substitute for *and* in some contexts, even though the latter may be less formal. The following examples of substitutability are therefore legitimate:

Grease is the time, is the place, is the motion; $\left\{\begin{array}{l} \textit{and} \\ \checkmark\, \textit{moreover,} \\ \checkmark\, \textit{furthermore,} \end{array}\right\}$ Grease is the way you are feeling.

The size of the units of text being linked is also a factor to be disregarded. *Because* tends to connect smaller spans than *this is because*; but other than this, there is little difference between the two phrases. Finally, the amount of background knowledge possessed by the reader is treated as a variable in the test. The phrase *and* can be substituted for the phrase *because*, but only if we assume that the reader can *infer* that a causal relation is being expressed. See Figure 4.2 for further explanation of all these points.

Some Terms Based on the Notion of Substitutability

The test in Figure 4.2 identifies when one candidate cue phrase is **substitutable** for another phrase in a given context. If we generalise over all possible contexts, three different relationships between two cue phrases can be distinguished:

$always(x, y)$ $\Leftrightarrow$ in every context where y appears, x is substitutable for y

$sometimes(x, y)$ $\Leftrightarrow$ x is substitutable for y in some contexts where y appears, but not in all of them

$never(x, y)$ $\Leftrightarrow$ in every context where y appears, x is not substitutable for y.

These three relationships exhaust the possible substitution relationships between x and y: for any pair of cue phrases, exactly one of them holds.

The definitions of *always*, *never* and *sometimes* form the basis of four composite relationships between cue phrases:

x and y are **synonymous** $\Leftrightarrow$ $always(x, y) \wedge always(y, x)$

x and y are **exclusive** $\Leftrightarrow$ $never(x, y)\ (\wedge never(y, x))$

x is a **hyponym** of y (y is a **hypernym** of x) $\Leftrightarrow$ $always(y, x) \wedge sometimes(x, y)$

x and y are **contingently substitutable** $\Leftrightarrow$ $sometimes(x, y) \wedge sometimes(y, x)$

Again, for any pair of cue phrases, exactly one of these relationships holds. The concepts of synonymity, hyponymity/hypernymity, exclusivity and contingent substitutability will be used in most of the discussion which follows.

Examples of Substitutability Relationships

In order to present examples of substitutability relationships, diagrams such as that in 4.20 will be used:

$$\text{First span of text.} \left\{\begin{array}{l} \textit{C1} \\ \checkmark\, \textit{C2} \\ \checkmark\, \textit{C3} \rightarrow \\ \checkmark\, \textit{(C4)} \\ \#\ \textit{C5} \end{array}\right\} \text{second span of text.} \tag{4.20}$$

1. Consider any cue phrase from the corpus in a text where it naturally occurs. Imagine you are a writer who has just produced this text, but needs to choose an alternative phrase (perhaps because you have just used the original phrase, and do not want to repeat it).
2. Remove the cue phrase from its host clause, and insert any other phrase from the corpus (the **candidate phrase**) into the same clause, at any appropriate position.
3. If need be, the punctuation of the new discourse can be altered to make it more suitable for the candidate phrase. For example, if the phrase *so* is being replaced by the phrase *this implies that*, it may be necessary to replace a comma with a full stop and create a new sentence.
4. If need be, the new discourse can be supplemented with additional or alternative cue phrases in other clauses. There are sometimes dependencies between the cue phrases in a text (for instance between *if* and *then*, or between *either* and *or*), so changing one phrase might require changes to others.
5. If it is possible to use the resulting discourse in place of the original discourse, then the candidate phrase is said to be **substitutable** to the original phrase in that context.

   The notion of 'being able to use one discourse in place of another' is expanded below.

   - It is not sufficient that the new discourse can be used to describe the same set of events in the world as those which the old one describes. For instance, the adverb *afterwards* and the subordinator *before* are truth-functionally equivalent in that they are both suitable for describing two events in temporal succession. But they are not always equally appropriate:

     (4.15) Bill was always interested in books. He could read *before* he could walk.

     (4.16) Bill was always interested in books. He could read; *afterwards* he could walk.

     In addition to describing the same eventualities, it must be ensured that the new discourse achieves the same goals as the old discourse achieved.
   - Some differences between the two discourses can nevertheless be overlooked—for one thing, *stylistic* discrepancies can be disregarded. The cue phrase *hence* can often be substituted for the cue phrase *so*, the only difference being in the 'formality' of the resulting discourse:

     (4.17) I'm just back from a holiday in France so there'll be no need to bring wine.

     (4.18) I have just returned from France; hence there will be no need to bring wine.

     Stylistic changes to the new text may thus be needed in order to accommodate the candidate phrase.
   - Different cue phrases are appropriate for linking portions of text of different sizes. For instance, *because* typically links clauses within a compound sentence; *this is because* typically links whole sentences. Such differences are to be overlooked in the test. In some cases, changing the punctuation is sufficient to accommodate the candidate phrase; but in others, it might be necessary to alter the length of the spans of text being linked, by substituting a précis or by adding additional relevant material.
   - A final factor to be disregarded is the amount of *background knowledge* the reader is assumed to possess. For instance, *and* can often be substituted for *yet*, but only if the reader will be able to infer the appropriate contrastive relation. Consider Example 4.19:

     (4.19) Mike was ravenous, yet he ordered watercress salad.

     Here, a substitution by *and* is only permissible if we can assume that the reader knows that petit fours are snacks, and hence that Mike's order comes as a surprise.

Figure 4.2: The Test for Substitutability

The items between the braces are all cue phrases. $C1$ is the original cue phrase; $C2$ is a phrase which in this context is substitutable for $C1$. $C3$ is substitutable for $C1$ in the given context, but it must first be moved to a different position in $C1$'s clause (in this case to the right). $C4$ is also substitutable for $C1$ in the given context, but it requires either a change to one of the other cue phrases in the text (due to a dependency between cue phrases), or a change to the size of the spans involved (due to the suitability of different cue phrases to spans of different sizes). All of these changes are permitted by the test for substitutability. Finally, $C5$ is not substitutable for $C1$ in the given context, even allowing for the changes the test allows.[3]

It should be noted that the text on either side of the braces can in principle be as long as is needed to make the original context clear. In practice, one or two clauses' worth of context will normally be given on each side. As was noted in Section 1.1.2, the idea of presenting the context of a text 'in its entirety' is problematic; however, it is hoped that the contexts provided in the examples which follow will be sufficient to give the reader a good idea of the texts.

A few examples of the test for substitutability can now be given. For instance, the phrases *to start with* and *to begin with* are intersubstitutable in all contexts, and hence termed synonymous: two examples of their intersubstitutability are given in texts 4.21 and 4.22.

$$\begin{array}{r}\text{Cyril set}\\ \text{about}\\ \text{preparing}\\ \text{breakfast.}\end{array} \left\{ \begin{array}{l} \textit{To start with,} \\ \checkmark\ \textit{To begin with,} \end{array} \right\} \begin{array}{l}\text{he put some porridge on to boil. Next, he}\\ \text{set out four bowls...}\end{array} \tag{4.21}$$

$$\begin{array}{r}\text{Sid's got no}\\ \text{hope of}\\ \text{winning the}\\ \text{race.}\end{array} \left\{ \begin{array}{l} \textit{To begin with,} \\ \checkmark\ \textit{To start with,} \end{array} \right\} \begin{array}{l}\text{he's out of training. For another thing he'll}\\ \text{be running against Otto Schultz, who hasn't}\\ \text{lost all season...}\end{array} \tag{4.22}$$

In the case of *to start with* and *for a start*, the relationship is not bidirectional: the former phrase is more general than the latter.

$$\begin{array}{r}\text{Sid's got no}\\ \text{hope of}\\ \text{winning the}\\ \text{race.}\end{array} \left\{ \begin{array}{l} \textit{To start with,} \\ \checkmark\ \textit{For a start,} \end{array} \right\} \begin{array}{l}\text{he's out of training. For another thing he'll}\\ \text{be running against Otto Schultz, who hasn't}\\ \text{lost all season...}\end{array} \tag{4.23}$$

$$\begin{array}{r}\text{Cyril set}\\ \text{about}\\ \text{preparing}\\ \text{breakfast.}\end{array} \left\{ \begin{array}{l} \textit{To start with,} \\ \#\ \textit{For a start,} \end{array} \right\} \begin{array}{l}\text{he put some porridge on to boil. Next, he}\\ \text{set out four bowls...}\end{array} \tag{4.24}$$

In texts such as 4.24, *for a start* does not seem an appropriate substitution—it gives the text an argumentative tone which is lacking in the original, which is purely narrative. From examples such as these, we can conclude that *to start with* is a hypernym of *for a start*.

[3] It should be borne in mind that the text may still be *grammatical* with $C5$; it may even still *make sense* with $C5$. The point is that just $C5$ cannot be used as a *replacement* for $C1$ in the context. The hash sign used to indicate non-substitutability must therefore be interpreted somewhat differently from hash signs as conventionally used in linguistic examples, which often denote 'ill-formed discourse'.

The phrases *lastly* and *moreover* are contingently substitutable. In some contexts they are both appropriate:

Sid's got no hope of winning the race. For one thing, he's out of training. For another thing, he's best at altitude, and he'll be running at sea level. $\left\{\begin{array}{l}\textit{Moreover,}\\ \checkmark\,\textit{Lastly,}\end{array}\right\}$ he's pitted against Otto Schultz, who hasn't lost all season. (4.25)

But in some contexts, *lastly* cannot be replaced by *moreover*:

Cyril set about preparing breakfast. To start with, he put some porridge on to boil. Next, he set out four bowls. $\left\{\begin{array}{l}\textit{Lastly}\\ \#\ \textit{Moreover}\end{array}\right\}$ he sliced some bread ready for toasting. (4.26)

And in other contexts, *moreover* cannot be replaced by *lastly*:

Sid's got no hope of winning the race. For one thing, he's out of training. $\left\{\begin{array}{l}\textit{Moreover}\\ \#\ \textit{Lastly}\end{array}\right\}$ he's best at altitude, and he'll be running at sea level. In addition, he's pitted against Otto Schultz, who hasn't lost all season. (4.27)

When applying the test for substitutability, a question arises as to how subtle we should be in distinguishing between cue phrases. It is often noted that 'true synonyms' are extremely rare: indeed, in some of the above examples where substitutability is claimed, one phrase might appear slightly more appropriate to some readers even though no particular reason suggests itself. Typically, a *rule* can be envisaged which relates various features of a text to the cue phrases which are most appropriate. But in a context where both phrases are acceptable, one being just marginally better than the other, generalisations are often hard to make: In such cases, we will err on the side of generality, and allow that substitutability is possible. If, subsequently, we are able to find a reliable rule, of course this decision can be reversed, and subtler distinctions made. But it should be borne in mind that we are principally concerned with making broad classifications within the set of cue phrases, rather than descending into the minutiae of 'descriptive linguistics'—and the test for substitutability is perfectly adequate for this task.

4.5 Substitutability Diagrams

In this thesis, a diagrammatic representation of substitutability relationships is used. The diagrammatic notation allows information about many pairs of cue phrases to be presented simultaneously, in a form which is relatively easy to understand.

The diagrams consist of **nodes** containing (possibly empty) sets of cue phrases, connected by a structure of directed **arcs**. Figure 4.3 shows the simplest structural relationships that can exist between two nodes A and B: $hypo(A, B)$, $excl(A, B)$ and $cs(A, B)$. Informally speaking:

- If $hypo(A, B)$, then phrases which are in A are hyponyms of phrases in B.

- If $excl(A, B)$, then phrases which are in A are exclusive with phrases in B.
- If $cs(A, B)$, then phrases which are in A are contingently substitutable with phrases in B (provided no other relationship between A and B is documented—see Section 4.5.1).
- Phrases which appear at the same node are synonymous.

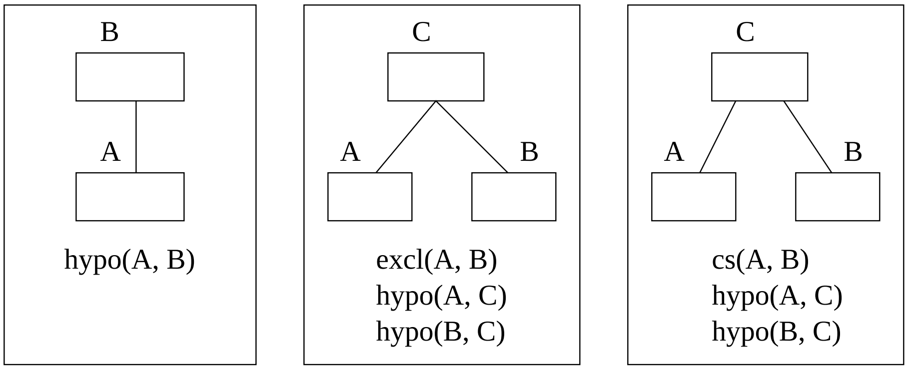

Figure 4.3: Three Possible Structural Relationships Between Nodes

Note that the $excl$ and cs relationships apply between 'sister' nodes; i.e. nodes which have a common mother node. The difference between $excl(A, B)$ and $cs(A, B)$ relates to whether or not their arcs meet on the mother node. This notation is chosen to allow the representation of hypernyms shared between exclusive or contingently substitutable phrases.

4.5.1 Contingent Substitutability Relationships

The relationship of contingent substitutability is overridden by other relationships in diagrams where a conflict is present. For instance, in Figure 4.4, x and y are represented as exclusive (through the arcs that touch on the mother node), but also as contingently substitutable (through the arcs which do not touch). In cases where a conflict such as

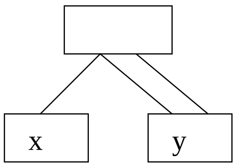

Figure 4.4: An Overridden Contingent Substitutability Relationship

this one is present, the contingent substitutability relationship is overridden. (In this example, of course, the additional arc representing contingent substitutability does no useful work, and it would be much clearer to leave it out; the diagram is just used as a simple illustration of how a contingent substitutability relationship can be overridden. But see the following section for cases where overridden contingent substitutability relationships do have a useful role.)

4.5.2 Complex Substitutability Diagrams

Complex substitutability diagrams involving many cue phrases can be created by combining the structures presented in Figure 4.3. These diagrams make use of inheritance: the phrases in a daughter node inherit the exclusivity and hyponymity relationships of the phrases in their mother node. Thus in Figure 4.5 (i), z is a hyponym of x, so by inheritance, z is a hyponym of w and exclusive with y.

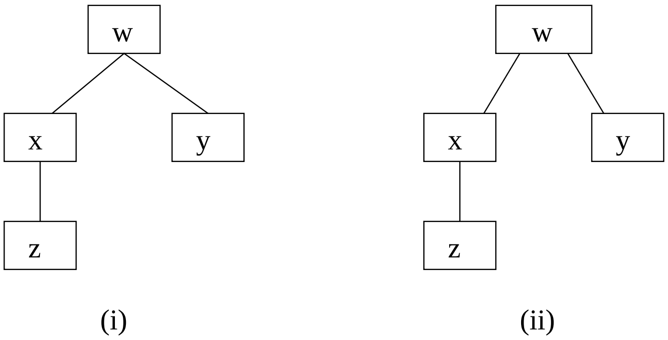

Figure 4.5: Two Examples of Inheritance

Contingent substitutability relationships are also inherited from phrases in a daughter node to phrases in its mother node. Thus, in 4.5 (ii), z and y are contingently substitutable. Note again, however, that these relationships can be overridden if they conflict with other relationships. Thus, in Figure 4.6, while x is contingently substitutable with y, z does not inherit this property because it is explicitly shown to be exclusive to y. The inherited contingent substitutability relationships between x and z, and between z and itself are also overridden.

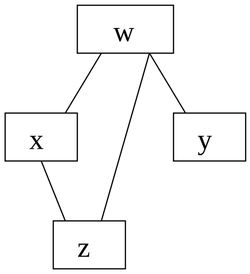

Figure 4.6: Overridden Inherited Contingent Substitutability Relationships

There are, however, cases where inheritance in a substitutability diagram causes genuine contradictions. For instance, Figure 4.7 is an illegal diagram: z is represented by inheritance as exclusive with itself; and exclusivity cannot be overridden.

Substitutability diagrams are intended to represent the relationship between *each pair* of phrases which appear in it—in other words, to provide all the substitutability information that it is possible to provide about the phrases involved. A final requirement to this end is to specify that diagrams must have a single top node. The diagram in

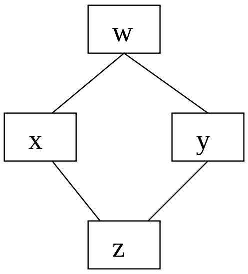

Figure 4.7: An Illegal Substitutability Diagram

Figure 4.8 is not permitted, because it does not document the relationship between x and y. Requiring that a single node dominates both x and y ensures that their

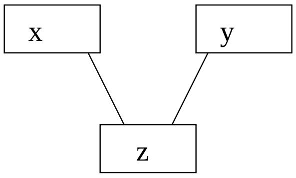

Figure 4.8: Another Illegal Substitutability Diagram

relationship will be represented.

4.5.3 Formalising the Semantics of Substitutability Diagrams

A set of rules for determining the relationships between phrases in a (legal) substitutability diagram is given in this section. The rules draw on the definitions of *always*, *never* and *sometimes*, which are re-iterated below:

$always(x, y) \Leftrightarrow$ in every context where y appears, x is substitutable for y

$sometimes(x, y) \Leftrightarrow$ x is substitutable for y in some contexts where y appears, but not in all of them

$never(x, y) \Leftrightarrow$ in every context where y appears, x is not substitutable for y.

The following rules are for deriving new substitutability relationships from existing ones:

$$\begin{aligned}
always(x, y) \wedge never(x, z) &\Rightarrow never(y, z) \\
always(x, y) \wedge always(y, z) &\Rightarrow always(x, z) \\
never(x, y) &\Rightarrow never(y, x) \\
sometimes(x, y) &\Rightarrow sometimes(y, x) \\
always(x, y) \wedge sometimes(x, z) \wedge \neg always(z, y) \wedge \neg never(y, z) &\Rightarrow sometimes(y, z)
\end{aligned}$$

The following rules are for deriving substitutability relationships from structures in a substitutability diagram. (They should be seen as replacements for the informal definitions given in Figure 4.3.)

$$\begin{array}{rcl} x \in A \wedge y \in A & \Rightarrow & always(x,y) \\ x \in A \wedge y \in B \wedge hypo(A,B) & \Rightarrow & always(y,x) \\ x \in A \wedge y \in B \wedge excl(A,B) & \Rightarrow & never(x,y) \\ x \in A \wedge y \in B \wedge cs(A,B) \wedge \neg always(x,y) \wedge \neg never(x,y) & \Rightarrow & sometimes(x,y) \\ x \in A \wedge \neg(y \in A) \wedge always(x,y) & \Rightarrow & sometimes(y,x) \end{array}$$

The intended definition of $sometimes(x,y)$ relies on a closed-world assumption about *always* and *never* relationships. In order to compute the complete set of relationships in a diagram, all the *always* relationships should first be computed, then all the *never* relationships, and finally the *sometimes* relationships.

4.5.4 Empty Nodes

Some nodes in a diagram do not contain any cue phrases at all. At the very top of the hierarchy, an empty category is necessitated by the formalism chosen for depicting substitutability relationships: if two phrases are exclusive or contingently substitutable, a common superordinate category must be shown whether or not they have a common hypernym. We can use the graph-theoretical category **top** (or $\top$) to fulfill this purpose (see Figure 4.9).

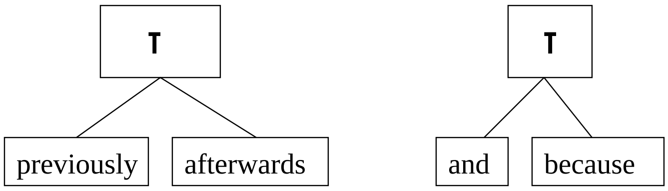

Figure 4.9: Two Uses of the Empty 'Top' Category

Empty nodes can also appear lower down in the taxonomy. These are not essential—a diagram can always be redrawn without them—but they often make diagrams easier to read. Imagine we have three phrases, X, Y and Z. X and Y are exclusive, so are X and Z; but Y and Z are contingently substitutable. Figure 4.10 shows two alternative ways of representing all these relationships: in many cases, the method involving the empty category is neatest.

4.6 The Taxonomy of Cue Phrases

The central task is now to incorporate as many of the cue phrases as possible into a single substitutability diagram. Ideally, the aim is to document the substitutability relationship between each pair of phrases in the corpus. Of course, this is a huge number; assuming there are N phrases in the corpus, the total number of relationships

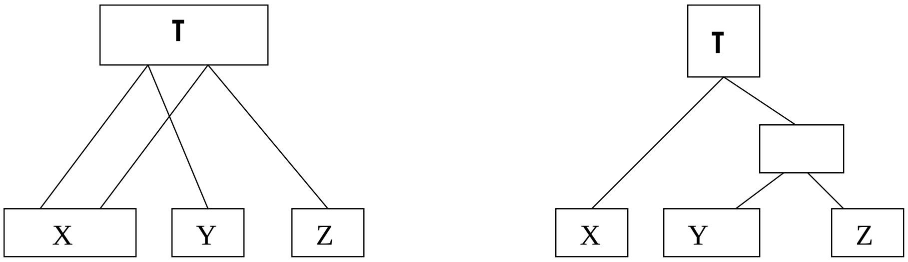

Figure 4.10: Empty Categories Lower Down in the Taxonomy

documented will be $(N * (N - 1))/2$. For $N \approx 330$, as in the present case, the number of relationships will be over 50,000. Clearly it makes sense to begin by looking for all the relationships within a *subset* of the complete corpus.

Currently, around 150 phrases have been incorporated into a single substitutability diagram, documenting some 11,000 relationships, which will be referred to as the **taxonomy** of cue phrases. The complete diagram is given in Appendix B, along with details of its organisation, and copious examples to motivate it. The following sections present an extract from the taxonomy, and summarise some of its most important characteristics.

4.6.1 Construction of the Taxonomy

As with the corpus of cue phrases, the taxonomy of cue phrases was constructed entirely by the author. It would have been preferable to construct the taxonomy on the basis of the judgements of a sizeable group of people (ideally, people without any theoretical experience of discourse analysis, given that the substitutability test is designed to recreate a task that forms a part of ordinary writing). However, the amount of data needed in order to build a taxonomy of any reasonable size from scratch makes such an experiment quite infeasible, bearing in mind the huge number of relationships that must be documented. Instead, the decision was taken to build a taxonomy reflecting the author's own intuitions, which could then be used and tested more systematically in subsequent experiments on groups of naive readers and writers. Such experiments have yet to be carried out; however, they would be very valuable as a follow-up to the present study.

The amount of data required to build the taxonomy also dictated that most of the examples used to motivate substitutability relationships were hand-crafted. It would have been preferable to search for appropriate examples in a corpus, but again, this would have been prohibitively time-consuming. A corpus-based study would certainly shed useful light on the taxonomy as currently constructed; but again remains to be pursued in follow-up work.

4.6.2 An Extract from the Taxonomy

A small portion of the taxonomy, dealing with some of the phrases which signal position in a sequence, is given in Figure 4.11. Pre-theoretical titles have been assigned to some

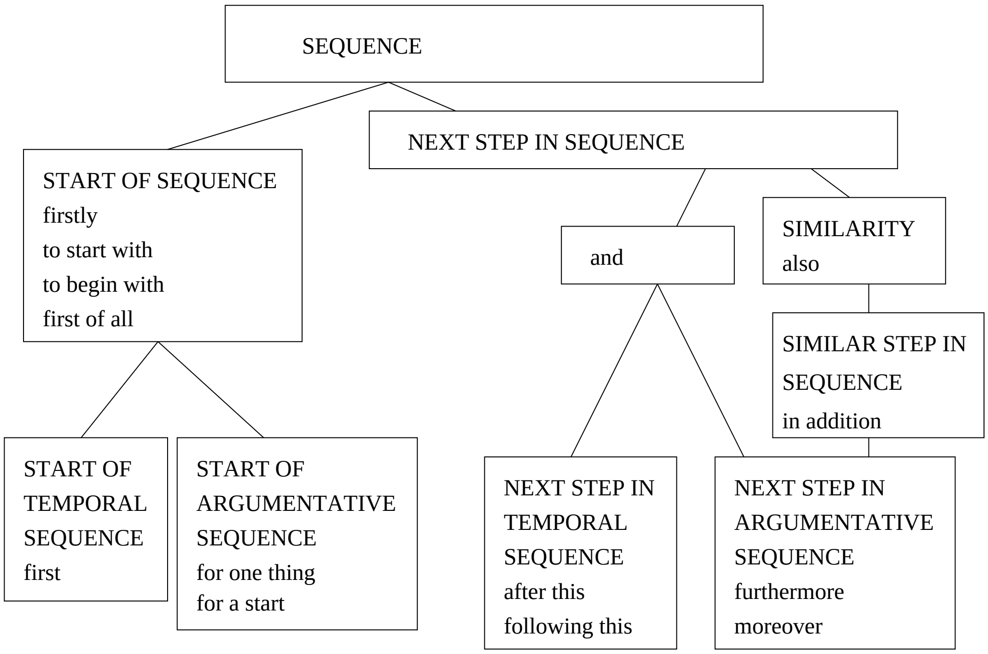

Figure 4.11: A Portion of the Taxonomy of Cue Phrases

of the categories; but this is just to give some idea of what functions the phrases might have: the taxonomy still just represents substitutability information.

Some examples of the substitutability relationships in Figure 4.11 are given below.

$$\begin{array}{r} \text{Bob set about cleaning the} \\ \text{house.} \end{array} \left\{ \begin{array}{l} \textit{To start with,} \\ \checkmark\, \textit{To begin with,} \\ \checkmark\, \textit{First,} \\ \#\ \textit{For one thing,} \\ \#\ \textit{Furthermore,} \\ \#\ \textit{And,} \end{array} \right\} \begin{array}{l} \text{he swept the floors; next he} \\ \text{washed them; and lastly, he} \\ \text{tidied the cupboards.} \end{array} \tag{4.28}$$

$$\begin{array}{r} \text{Bob set about cleaning the} \\ \text{house. To start with, he} \\ \text{swept the floors and washed} \\ \text{them;} \end{array} \left\{ \begin{array}{l} \textit{in addition,} \\ \checkmark\, \textit{after this,} \\ \checkmark\, \textit{following this,} \\ \checkmark\, \textit{(and)} \\ \checkmark\, \textit{also} \rightarrow \\ \#\ \textit{For one thing,} \\ \#\ \textit{Furthermore,} \end{array} \right\} \text{he tidied the cupboards.} \tag{4.29}$$

$$\begin{array}{r} \text{Television is bad for us. It} \\ \text{kills creativity;} \end{array} \left\{ \begin{array}{l} \textit{and} \\ \checkmark\, \textit{furthermore,} \\ \checkmark\, \textit{also,} \\ \checkmark\, \textit{moreover} \\ \#\ \textit{after this} \\ \#\ \textit{for one thing} \end{array} \right\} \begin{array}{l} \text{it promotes an unhealthy} \\ \text{kind of `crowd mentality'.} \end{array} \tag{4.30}$$

$$\text{Television is bad for us.} \left\{ \begin{array}{l} \textit{For one thing,} \\ \checkmark\ \textit{To begin with,} \\ \checkmark\ \textit{Firstly,} \\ \#\ \textit{Moreover} \\ \#\ \textit{And} \\ \#\ \textit{First} \end{array} \right\} \begin{array}{l} \text{it kills creativity; in} \\ \text{addition it promotes an} \\ \text{unhealthy kind of `crowd} \\ \text{mentality'.} \end{array} \tag{4.31}$$

$$\begin{array}{r} \text{Jim jumped off the cliff, so} \\ \text{Bill} \end{array} \left\{ \begin{array}{l} \textit{also} \\ \#\ \textit{and} \\ \#\ \textit{in addition} \\ \#\ \textit{furthermore} \\ \#\ \textit{for one thing} \end{array} \right\} \text{jumped off.} \tag{4.32}$$

4.6.3 Some General Remarks about the Taxonomy

Two of the taxonomy's most significant characteristics should be mentioned straight away, as they have an important bearing on its organisation, and on the theoretical interpretation it will subsequently be given.

For one thing, a degree of hierarchy is found throughout the taxonomy. Chains of 2 or 3 hyponymic nodes are fairly common. The 'most general' cue phrase is *and*, which has over 30 separate hyponyms. In other words, the degree of generality of cue phrases in the taxonomy is an interesting variable to study.

Another important finding is that the taxonomy does not divide neatly into large exclusive subgroups of phrases. For any candidate grouping, many phrases can be found which fit into more than one group. (*And* and *then*, for instance, have many other uses aside from signalling position in a sequence.) In fact, most of the variation between cue phrases is represented at a relatively low level, in the microstructure of the taxonomy. This is interesting, because it already suggests that a classification scheme based on the taxonomy is unlikely to identify any one dimension of variation amongst relations as 'dominant'—an assumption which is characteristic of many existing classifications of relations.

4.6.4 The Global Organisation of the Taxonomy

The task of representing *all* the substitutability relationships between *all* the phrases in the corpus is an extremely complex one. One of the main difficulties is the fact just alluded to, that the phrases in the corpus do not separate neatly into exclusive categories. This lack of modularity makes it difficult to work with a subset of the phrases in isolation.

To solve this problem, phrases are organised at a high level into a number of *non-exclusive* categories: SEQUENCE PHRASES, CAUSE PHRASES, RESULT PHRASES, RESTATEMENT PHRASES, TEMPORAL PHRASES, NEGATIVE POLARITY PHRASES, ADDITIONAL INFORMATION PHRASES, HYPOTHETICAL PHRASES, SIMILARITY PHRASES, and DIGRESSION PHRASES. These categories have no theoretical significance at all, and should just be thought of as providing an expedient way for spreading the taxonomy over several pages. Two types of cue phrase are then identified: **exclusive phrases**, which belong just to one category; and **multicategory phrases**, which belong to two or more categories. This distinction introduces a certain amount of modularity into the taxonomy, and greatly reduces its complexity. To begin with, there is a diagram showing the exclusive relationship between the 'exclusive phrases' in every category. Then a substitutability diagram is given for each separate category, showing the relationships between its exclusive phrases, and between its exclusive phrases and all the multicategory phrases. Finally, there is a substitutability diagram for all the multicategory phrases. In this way it is ensured that the relationship between each phrase and each other phrase is represented.

The reader is referred to Appendix B for a closer look at the taxonomy. It is not yet perfect, of course: there are still cue phrases in the corpus which have not been incorporated; and it is still not very hard to find counterexamples to some of the relationships it documents. But at least a reasonably clear method exists for querying and improving it: questions about the placing of a given phrase will be decided on the evidence of concrete linguistic data.

4.7 Summary

This chapter has described the incremental construction of a taxonomy of cue phrases. Initially, a **test for cue phrases** was employed to gather a corpus of cue phrases from naturally occurring texts. Then this corpus was organised into a taxonomy, using a **test for substitutability**. Using these two tests, it should be possible for several people to arrive at very similar taxonomies of cue phrases. And where there are discrepancies, the reliance in both tests on concrete linguistic examples should provide a convenient way for alternative analyses to be discussed.

The next chapter shows how the taxonomy of cue phrases can be used to motivate an isomorphic taxonomy of coherence relations.