

# Using Linguistic Phenomena to Motivate a Set of Coherence Relations

Alistair Knott

Department of Artificial Intelligence, University of Edinburgh

80 South Bridge, Edinburgh EH1 1HN, Scotland

Email: [A.Knott@ed.ac.uk](mailto:A.Knott@ed.ac.uk)

Robert Dale

Human Communication Research Centre, University of Edinburgh

2 Buccleuch Place, Edinburgh EH8 9LW, Scotland

Email: [R.Dale@ed.ac.uk](mailto:R.Dale@ed.ac.uk)

May 14, 1993

Running head: Motivating Coherence Relations

The authors would like to thank Karen Sparck Jones, and several anonymous reviewers, for comments on an earlier draft of this paper. The first author would also like to thank S.E.R.C. for Studentship Number 91308508.

## Abstract

The notion that a text is coherent in virtue of the ‘relations’ which hold between the elements of that text has become fairly common currency, both in the study of discourse coherence and in the field of text generation. The set of relations proposed in Rhetorical Structure Theory (Mann and Thompson [16]) has had particular influence in both of these fields. But the widespread adoption of ‘relational’ terminology belies a certain amount of confusion about the relational constructs themselves: no two theorists use exactly the same set of relations; and often there seems no motivation for introducing a new relation beyond considerations of descriptive adequacy or engineering expedience.

To alleviate this confusion, it is useful to think of relations not just as constructs with descriptive or operational utility, but as constructs with psychological reality, modelling real cognitive processes in readers and writers.

This conception of coherence relations suggests a methodology for delineating a set of relations to work with. Evidence that a relation is actually *used* by speakers of a language can be obtained by looking at the language itself—in particular by looking at the range of cue phrases the language provides for signalling relations. It is to be expected that simple methods will have evolved for signalling the relations we find most useful.

This paper presents a bottom-up methodology for determining a set of relations on the basis of the cue phrases which can be used to mark them in text. This methodology has the advantage of starting from concrete linguistic data, rather than from controversial assumptions about notions like ‘intention’ and ‘semantics’.

# 1 Introduction

The notion that a coherent text is more than simply a sequence of sentences is widely accepted; however, a number of different approaches have been taken to explaining what it is that provides textual coherence.

One dominant idea in recent text linguistics has been that there are implicit **relations** between the sentences of a text, so that (for example) the content of one sentence might provide elaboration, circumstances, or explanation for the content of another. These relations are held to bind a text together, and thus contribute to its overall coherence. They are sometimes made explicit by the speaker’s use of particular **cue phrases**, linguistic elements like *more precisely* or *for example*; but in other cases the hearer is left to infer the particular relation that holds between two sentences on some other basis. The idea of implicit relations in discourse was first proposed in the context of cross-linguistic studies (Ballard, Conrad and Longacre [1], Beekman and Callow [2], Grimes [5]); it has also been developed within the framework of systemic grammar (Martin [18], Martin [17]), and more recently by computational linguists (Hobbs [8], Grosz and Sidner [6], Mann and Thompson [16]) and psycholinguists (eg Sanders *et al* [22]). Following Hobbs and Sanders, we can refer to all such theories as theories of **coherence relations**: this term makes explicit the phenomenon which relations have been proposed to account for.

The present paper is written from the perspective of computational linguistics, and in particular from the perspective of ‘natural language generation’. In this field, a great deal of attention has recently been focused on coherence relations—especially on the **rhetorical relations** proposed in Mann and Thompson’s [16] **rhetorical structure theory** (RST). There has been considerable debate about the role of these relations in generation systems, and about which particular relations systems should use: much of this debate can be seen as having more general relevance for theories of discourse structure.

We begin in Section 2 with a brief description of RST’s conception of relations, and an attempt to explain why it has such appeal in the domain of text generation. In Section 3 we outline some problems which result from the widespread adoption of the relational concept; these arise mainly from the lack of a ‘standardised’ set of relations, and from the lack of consensus about how to work one out. Section 4 surveys some current suggestions about how a standard set of relations could be motivated. One suggestion is to make the set of relations *mirror the set of cue phrases* in a language. But while this is a reasonably clear criterion to use, it is not clear *why* relations thus justified should be especially appropriate for use

in a theory of discourse coherence. Another suggestion is to think of relations as modelling *psychological constructs* used by people when they process text. This idea gives relations a clearer role in a theory of discourse coherence; but difficulties arise in the empirical question of deciding ‘which relations people actually use’.

In Section 5, we propose that the two suggestions above can be combined, in such a way as to minimise both of their drawbacks: we argue that the cue phrases in a language are likely to *reflect* the set of relations actually used by those who speak it. In Section 6, we present a new methodology for motivating a set of relations based on this idea: the methodology consists of gathering and classifying a corpus of cue phrases.

## 2 Rhetorical Relations

According to RST, a written text is coherent (partly) in virtue of the relations which hold between its sub-parts or **spans**. The same relations can link spans as small as clauses or as large as whole paragraphs or chapters. RST gives a catalogue of twenty-three relations; these allegedly suffice to analyse ‘the vast majority’ of English prose texts. A sample RST analysis is given in Figure 1.<sup>1</sup>

(Insert Figure 1 around here.)

A number of features of RST relations make them particularly appropriate for use in text generation systems. Firstly, they are defined principally in terms of the likely *effects* of the two related spans on a reader; for instance, whether understanding one span increases the reader’s readiness to accept the other, or whether understanding both spans allows the reader to recognise a particular semantic relation as holding between them. The input to the text generation process is often expressed in similar terminology, as a goal to ‘achieve certain effects on a reader’. RST relations, which describe the effects that can be achieved by juxtaposing spans of text, seem very suitable constructs for processing input in this form.

A second attraction of RST is that it provides a uniform, *compositional* account of text. Relations can apply between spans as small as clauses, or between larger spans such as paragraphs or groups of paragraphs: presumably, the assumption is that intentions to produce small pieces of text are qualitatively similar to intentions to produce larger pieces. The fact

---

<sup>1</sup>The relations illustrated here link spans asymmetrically: in each case a **nucleus** span (which has an arrow pointing to it) is related to a **satellite** span (from which the arrow originates). The claim in RST is that satellite material is ‘less important’, and that it serves a subsidiary function with respect to the nucleus.

that the same relations can apply at different levels of hierarchy means they can be used in recursive planning mechanisms, where high-level goals are progressively broken up into simpler subgoals.

Given these features, it is clear why rhetorical relations have proved popular in text generation. Hovy [11] calls them ‘the basic building elements of coherent text’. The particular role that relations play in the generation process varies from system to system: in the original RST-based systems (Hovy [9], Moore and Paris [19]), they are incorporated into STRIPS-style planning operators; more recently (Vander Linden *et al* [26], Rösner and Stede [21]) they have been treated mainly as lower level guides to the syntactic realisation of sentences.

### 3 The Proliferation of Relations

Despite the attractions of the *concept* of rhetorical relations, there is little agreement between researchers about which *individual* relations to make use of. A text planner needs to specify a set of relations to work with; however, there are significant differences between the sets used by the various systems. Although the work inspired by RST takes the initial set of relations as proposed by RST as a starting point, each individual researcher uses relations not found in the original theory. These are sometimes specialisations of existing RST relations (for example, Rösner and Stede’s CAUSAL-CONDITION, which is a more specific version of the original RST CONDITION relation); sometimes generalisations over more than one relation (for example, Vander Linden *et al*’s RESULT, which is an amalgam of the RST relations VOLITIONAL-RESULT and NON-VOLITIONAL-RESULT); and sometimes completely new relations (for example, Rösner and Stede’s UNTIL, which is customised for instructional texts, where a given action must be performed until a certain condition is met). In other cases, relations seem to be defined orthodinally to those in RST: for instance, Hovy *et al*’s ANALOGY [12] covers some of the ground covered by the RST relations RESTATEMENT, EVALUATION and ELABORATION; but needless to say, these latter relations are not *always* analysable using Hovy’s ANALOGY.

Recently, Hovy [10] collected a large set of relations from other researchers, and organised them into a taxonomy: the idea was to produce a ‘standardised and covering set of relations’. The taxonomy has subsequently been reworked by Maier and Hovy [14] and Hovy *et al* [12]. The surprising feature of these classifications is the number of relations they involve: the latter study uses a systemic network with over 80 relations at the leaves and over 170 relations in the network as a whole. This is clearly a substantial departure from RST’s original set of relations.

### 3.1 Reasons for the Proliferation

This proliferation of relations is partly to be expected: its origin can be traced to a number of features of RST. For one thing, Mann and Thompson are themselves quite flexible about the set of relations in their theory; they are more concerned with establishing the rhetorical relation in general as a useful tool for text analysis.

Relation definitions have the status of applications of the theory rather than elements of the theory. One might want to change or replace the definitions...such changes are to be expected and do not cross the definitional boundaries of RST.

*Mann, Matthiessen and Thompson [15], p 48*

A more significant feature of RST is that its relations are proposed as ‘purely descriptive’ constructs. The motivation for the set of relations in RST is that it allows an analysis of ‘virtually every text’: you can represent nearly every text as a tree of relations, using the 23 relations RST provides. But this criterion serves to justify a great many different sets of relations, of which the RST set is only one. The space of texts which RST can describe can equally well be described by a larger set of relations obtained by subdividing existing relations, or a smaller set of relations obtained by amalgamating existing relations. To take an extreme example, one could use just two relations (for example, CAUSAL and NON-CAUSAL) to describe all the texts which RST can currently account for. Furthermore, it is possible to achieve descriptive adequacy with sets of relations which cut right across the RST set (such as Hovy’s ANALOGY, mentioned above). The point is that if relations are taken simply as useful descriptive constructs, there is no way of picking out *one particular* set of relations. (Indeed, we might question whether the notion of ‘descriptive adequacy’ makes much sense at all by itself. We really need to know, in addition, the *purpose* for which the description is claimed to be adequate.)

### 3.2 Problems with the Proliferation of Relations

It is clear that RST is very flexible about the set of relations to be used by analysts. Such flexibility is even held by some to be an advantage of the theory:

Generally a new domain has dictated modifications to the inventory of relations, but this very adaptability is one of its most useful features.

*Vander Linden et al [26], p 184*

At the same time, giving analysts such a free hand to choose relations creates some serious problems.

### 3.2.1 The Unfalsifiability of the ‘Relational Hypothesis’

Flexibility about individual relation definitions tends to undermine the empirical content of the relation construct in general. The hypothesis that ‘virtually any text can be analysed by representing its rhetorical relations’ becomes much less strong if relations can be created whenever they are needed: it is hard to think what evidence could be found which could disprove it. Rhetorical relations at this point hardly seem to be saying any more than speech act theory; that we must take intentions into account when representing text. The extra claim in RST—that text is coherent by virtue of the relations between its intentions—is virtually unfalsifiable without a method for specifying what is to count as a relation in the first place.

Even incoherent texts can be analysed according to the relations between the intentions in their spans. For instance, the text in (1) seems incoherent at first sight:

- (1) John broke his leg. I like plums.

Yet we could still define a relation which holds between the intentions underlying the spans in this text: perhaps we could call the relation **inform-accident-and-mention-fruit**. The relation might be multinuclear, with one nucleus intended to convey information about an accident, and the other intended to convey information about fruit. Clearly, we do not want to include these sorts of relations in any principled set of ‘rhetorical relations’.

To place the relational claim on a sounder empirical footing, we need to tighten the constraints on relationhood—for instance, by giving a fixed set of relations, or by giving rules for picking out a set of relations from some larger set of ‘potential relations’.

### 3.2.2 What are Relations Modelling?

The proliferation of different and incompatible sets of relations seems odd from a theoretical point of view: should we really have such a choice about which relations to use? Theoretical constructs are typically treated as corresponding to *real phenomena* underlying the data they describe, rather than as being purely synthetic: otherwise there seems little point in using the constructs at all. But in the present case, the emphasis on descriptive adequacy alone leaves no room for such realism: consequently, little attention is paid to the question of what relations actually model. This means it is possible that different researchers envisage different



roles for their relations (see the next section for examples), in which case it may make for confusion to use the label ‘rhetorical relation’ as a general term.

More seriously, there is a danger that researchers will look no further than descriptive adequacy in deciding on a set of relations to work with. This might happen if RST’s existing set were adopted, since it is just one of many descriptively adequate sets. It will also happen if RST’s set is altered simply to achieve descriptive adequacy in a different domain. In such cases, it is hard to speak of relations modelling anything at all; and hard to see why the construct is introduced in the first place.

Of course, the real situation is not quite as arbitrary as the preceding discussion would imply. There is still a reasonable consensus between researchers about which relations to use. But this very fact suggests that intuitions are at work which are not being acknowledged: while researchers rarely attempt justification for their choices beyond an adherence to the needs of descriptive adequacy, their choices are not as diverse as they might be given *only* this criterion.

## 4 Motivating a Set of Relations: Some Current Initiatives

Unless a more explicit approach is taken towards justifying a set of relations, the research programme based on RST may lose much of its momentum. What is needed is a way of giving the concept of a relation more substance: two current suggestions to this end are discussed below.

### 4.1 Linking Relations to Surface Linguistic Phenomena

One possible approach is to determine a set of relations by looking at the set of methods available for marking them in surface text. Ballard *et al* [1] and Longacre [13] follow this strategy:

It is our contention...that a surface taxonomy of form within a language determines a similar taxonomy of deep relations, and that the two taxonomies stand and fall together.

*Ballard et al [1], p 75*

Martin ([18], [17]) adopts a similar approach, presenting a classification of conjunctive relations based on the cohesive resources of a language. (Yet he is clearly interested in more than cohesion—he acknowledges that relations can be ‘explicit’ or ‘implicit’.)

**Cue phrases** are a particularly appealing source of evidence for determining a set of relations. The set of cue phrases is very large and diverse, and often reflects quite subtle distinctions between relations. Moreover, it can provide a useful indication of just *how* subtle a taxonomy should be: there is no need to make a subtle distinction in the taxonomy unless cue phrases exist which reflect it.

A number of computational linguists have made use of cue phrases in motivating a set of relations to work with. Hovy's current classification of relations (Hovy *et al* [12]) seems at least partially justified in this way. His relations are arranged into a hierarchy—a top-level distinction between IDEATIONAL, INTERPERSONAL and TEXTUAL relations is motivated on theoretical (Hallidayan) grounds—but lower down, it becomes hard to argue *a priori* for each of the hundreds of relations; instead they are distinguished by the different textual means by which they can be signalled. Textual realisation methods can also be used to arrange relations in the hierarchy: in an earlier paper (based on a slightly different taxonomy), Hovy suggests that cue words which can mark a given relation can also be used to mark any of its subordinates (Hovy [10]). For instance, *then* (which marks a SEQUENCE relation) can also be used to mark the subordinate TEMPORAL-SEQUENCE relation).

It is clearly convenient to use cue phrases to determine a set of relations in a text planner: once a structure of relations has been built, the step of converting it into a surface string of text is simplified. But there remain some questions about this approach. In particular, are there any good theoretical reasons why a taxonomy of relations *should* be built on the basis of cue phrases? This policy itself needs to be justified. It should be remembered that relations are intended to be elements in a theory of discourse coherence: so we have to ask, in what way are relations classified using cue phrases *particularly well suited* for explaining the phenomenon of discourse coherence? What insight can they give us which other sets of relations could not provide? Without an answer to this question, Hovy's relations must still be seen mainly as heuristic constructs for text generation; it is not clear what theoretical or explanatory significance they have. The argument applies equally to Longacre and Martin: what, if anything, makes the descriptions of text provided by their relations *particularly revealing*?

## 4.2 Relations as Psychological Constructs

An alternative approach to motivating a set of relations begins with a direct suggestion about what relations are: it takes them to be *psychological* constructs, which humans make use of

when creating and interpreting text. This approach can be seen as sharing Hovy’s conception of relations as constructs used in the task of text planning, but making an additional stipulation—that relations are to be those constructs which *people* use to create text, not just a set used by a text planning machine.

Associating relations with the *human* text planning mechanism endows them directly with explanatory significance. It is plausible that coherent texts have the structure they do *as a result of* the strategies we use for constructing them. Texts are human artefacts, constructed by humans for humans: an explanation of the way they are structured can be expected to make reference to the psychological machinery involved in processing them. The difficult question is deciding which set of relations we use as part of this machinery—if indeed we use relational constructs at all.

Hobbs [8] was perhaps the first to suggest an explicitly psychological account of relations. He refers to his relations as ‘text-building strategies’, employed by the writer to ease the reader’s comprehension process. The relations are motivated by a study of the type of inferences a reader needs to make in order to understand a discourse: thus they are seen as being tailored to the reader’s needs. It is suggested that relations might eventually be explained in terms of very general principles of ‘cognitive economy’—but what these principles are is not considered in any detail.

Sanders, Spooren and Noordman [22] [23] also see relations as modelling psychological phenomena; but they go further than Hobbs, giving concrete suggestions about what these principles might be, and looking for experimental evidence for them. They seek to describe the space of relations using a small number of ‘cognitively basic’ concepts, which act as orthogonal parameters in defining individual relations. The primitive concepts are as follows:

- Basic operation: every relation is deemed to have either a **causal** or an **additive** component. **causal** relations are those where a ‘relevant’ causal connection exists between the spans; all other relations are **additive**.
- Source of coherence: every relation is coherent on **semantic** or **pragmatic** grounds. It is semantic if the spans are related in terms of their propositional content and pragmatic if they are related because of their illocutionary force.
- Order of segments: this distinction only applies to causal relations; they are deemed to have **basic** order if the antecedent is on the left, and **non-basic** order if it is on the right.

- Polarity: a relation is **positive** if its basic operation links the content of the two spans as they stand, and **negative** if it links the content of one of the spans to the negation of the content of the other span. Negative polarity relations typically involve either a violation of expectation, where the expectation derives from a causal basic relation; or a contrast, where the basic relation is additive.

These four parameters can combine to form twelve ‘complex’ relation types. For each type, Sanders *et al* provide one or more sample RST-like relations: for instance, CAUSE-CONSEQUENCE (basic operation = causal; source of coherence = semantic; order = basic; polarity = positive); or CLAIM-ARGUMENT (basic operation = causal; source of coherence = pragmatic; order = non-basic; polarity = positive). Each of the relations is associated with a ‘typical’ connective word used for marking it.

Support for this set of relations is provided by empirical experiments. One of these used discourse analysts as subjects: they were given definitions of all the relations, and asked to decide which relations were appropriate for a number of sample texts. In another experiment, ‘naive’ subjects were used: they were shown texts featuring implicit relations, and had to decide which cue phrase was most suitable. (Cue phrases were thus used as an experimental handle on the relations posited by the subjects.) Both experiments were designed to test how much agreement there is on how to use the relations: in both cases it was found that where there was disagreement over which relation to use, it tended to be over the value of a single parameter only.

These experiments provide good support for Sanders *et al*’s use of parameters to divide up the relation space. They also embody a sound methodology for investigating ‘psychologically real’ relations: empirical tests on analysts and naive subjects. But the actual parameters they choose present some problems.

Firstly, the experimental evidence for the ‘source of coherence’ parameter is less clearcut than for the other parameters: both analysts and naive subjects are apparently less attuned to it.

Secondly, the status of the ‘order of spans’ parameter as a ‘cognitively important’ primitive might be questioned. Sanders *et al* cite several studies which show the *effects* of presenting information to subjects in different orders; but such results do not permit the assumption in subjects of a specific mechanism attuned to span order. The *a priori* grounds for supposing such a mechanism seem weaker than in the case of the ‘basic relation’ parameter: causal relations are likely to play an important part in our knowledge representation systems, and in

the way we reason with them. On the other hand, ‘order of spans’ seems tailored particularly to the explanation of linguistic phenomena. Clearly analysts can use it, and writers are sensitive to it—but that doesn’t *imply* that it is itself a ‘cognitive primitive’.

Finally, it might be thought that Sanders *et al*’s methodology of empirical experiments is too blunt to work out all the subtleties in a classification of relations. As they acknowledge, their parameters only pick out *classes* of relations—some method has to be found for investigating the fine detail of the taxonomy in a principled way. The addition of new orthogonal parameters would soon boost the number of relations, but as we have seen, there are already potential problems with the evidence for some of the current parameters. Furthermore, there is no reason *a priori* to expect the entire space of relations we make use of to be neatly parameterised.

### 4.3 Summary

Two approaches to justifying a set of relations have been outlined; each has its advantages and disadvantages. One approach constructs a detailed taxonomy of relations using surface linguistic phenomena; but the explanatory significance of these relations is open to question. In the other approach, relations are given an explanatory role as psychological constructs; but the experimental methods for investigating these constructs are fairly blunt.

Our research, described in the following sections, attempts to combine elements from both these approaches: we suggest that evidence for ‘psychologically real’ relations can be sought in a study of linguistic phenomena.

## 5 A New Approach to the Motivation of Relations: Looking for Linguistic Evidence for Cognitive Strategies

If people actually *use* a particular set of relations when constructing and interpreting text, it is likely that the language they speak contains the resources to signal those particular relations explicitly. If people plan texts by building a structure of relations, and understand texts by working out this structure, then being able to mark relations explicitly in text will facilitate the communication process, by making it easier for a writer to indicate to a reader which relation is intended. The model we have in mind is something like that given in Figure 2: the communication of relations is seen as an *intermediate step* in the communication of a writer’s goals and ideas to the reader.

(Insert Figure 2 about here)

In such a model, the identification of relations is something which *really matters*: readers need to be able to do it. It is this which makes it likely that ways exist for identifying relations explicitly. Note that this argument does not go through if relations are thought of simply as convenient tools for discourse analysts: in this case, there is no way of arguing *why* they should be linked to a certain category of linguistic phenomena. On the other hand, the assumption of psychological reality, coupled with the assumption that language is somehow optimised for our communicative purposes, together constitute an *argument* why we can take linguistic marking techniques as evidence for a set of relations.

If the assumptions are accepted, a new methodology for determining a set of relations can be suggested. Studying the means available for marking relations in a given language should be able to tell us about the relations which people actually make use of. The methodology might be described in Hallidayan terms, as using the **cohesive** devices a language affords as evidence for a psychological theory of text **coherence**.

The appeal of this approach is not just to do with its emphasis on explanatory adequacy. It also has the advantage of starting from concrete linguistic data—the methods of signalling relations actually used in text. Typically, classification of relations has been done by defining them using concepts like ‘intention’ and ‘semantics’, whose representation is itself the subject of much debate. Using the present methodology, it is only after developing a taxonomy of linguistic data that it is necessary to characterise relations using these more abstract terms: a given linguistic strategy can be taken as evidence for the *existence* of some structuring mechanism which corresponds to it, even before we have a way of defining it in theoretical terms at all.

## 5.1 Cue Phrases: Linguistic Markers of Relations

Having argued that a set of relations can be motivated by appealing to certain linguistic strategies, it remains to identify a set of linguistic strategies to study. In this paper, we concentrate on **cue phrases**: these are perhaps the most obvious means of signalling relations in text. (Although there are others—for instance, tense and aspect can often be used to signal temporal relations.)

Many other researchers make reference to cue phrases under one name or another (Cohen [3]: ‘clue words’; Grosz and Sidner [6]: ‘cue phrases’; Schiffrin [24]: ‘discourse markers’). It is not certain that all of these researchers have the same conception of the role of cue phrases,

or even that they agree about what counts as a cue phrase—for the moment we can think of them pre-theoretically as ‘phrases whose function is to link spans of discourse together’. Examples of cue phrases are *because*, *meanwhile*, *but*, and *on the other hand*.

Beginning by looking at cue phrases is best seen as a working methodology. Cue phrases form a reasonably homogeneous group, which it is possible to isolate as the ‘object of investigation’. Furthermore, cue phrases are very versatile in the range of relations they can signal: for instance, *because*, *meanwhile*, and *on the other hand* all signal very different relations. Finally, the set of cue phrases is quite large, so a detailed taxonomy is a realistic target.

In addition, some facts about cue phrases seem to support our idea that the communication of relations is an important part of communication via a text.

- Cue phrases tend to be *simple* linguistic expressions. They are mostly single words or stock phrases that we might call **idiom chunks**; it is often possible to suggest that they have become simplified because they correspond to constructs which are in continual use when we process text. The evolution of a cue phrase is sometimes implicit in its etymology: for instance, it is possible to imagine how words like *anyway* or *because* have evolved from more complex formulas and phrases.
- There is substantial empirical evidence that cue phrases are useful for readers when they construct a mental representation of text. (See for example Segal *et al* [25], for an account of the role of the phrases *and*, *then*, *so*, *because* and *but*.)

## 6 A Data-driven Methodology for Determining a Set of Relations

In the previous section, it was argued that linguistic devices (in particular, cue phrases) can be taken as evidence for relations, provided these are conceived as constructs which people actually use when creating and interpreting text. This section reports on our initial investigations based on this argument. First, a suitable definition is sought for ‘cue phrases’; then, a corpus of cue phrases is obtained; finally, the corpus is analysed and classified. The resulting taxonomy seems promising as the foundation for a taxonomy of relations.

### 6.1 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 the category of cue phrases exist already, but these are

often expressed in theory-laden terms, the only independent characterisation being lists of examples: this is the case, for instance in Grosz and Sidner [6], and in Cohen [3]. Cohen's list of examples is more extensive than Grosz and Sidner's, but neither study claims to have a complete list.

In an attempt to come up with a precise yet theory-neutral definition, we have developed a test which picks out a certain group of phrases as they occur in natural discourse. In order to avoid any terminological confusion, we refer to the class of phrases which pass this test as the class of **relational phrases**. Given any candidate phrase situated in a passage of text, an analyst can decide whether it is a relational phrase by following the procedure outlined below.

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

(2)       ...John and Bill were squabbling. John fell *because* Bill pushed him: that was how it all started ...

then the isolated phrase and clause would be

(3)       *because* Bill pushed him.

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 give:

(4)       *because* Bill pushed John.

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.

(5)       Because Bill pushed John

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. If the candidate phrase is removed, the remaining clause should *no longer* be incomplete. In other words, it should form a coherent mini-discourse on its own. Thus the text

(6) Bill pushed John.

can stand by itself, without the need for extra clausal material.<sup>2</sup>

This test is designed to give a reasonably objective way to pick out a set of phrases to act as the object for further study, without relying too much on any particular theory of syntax. There are of course many cases where it is hard to apply the test. For instance, some phrases involve anaphora which is not easily substitutable: the phrase *for this reason* requires a substitution like *for the reason that . . .*, which is very ungainly. In such cases, it is preferable to leave the anaphoric expression in the phrase. Indeed, as Halliday and Hasan [7] point out, many *bona fide* relational phrases clearly derive from phrases involving anaphora (consider words like *therefore* and *thereby*); so this analysis is justified. Another 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:

(7) *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 this is arguably a kind of discourse context.

## 6.2 Gathering a Corpus of Relational Phrases

Using the test, we gathered a corpus of relational phrases. The source texts were all passages taken from academic articles and books. Note that different registers of text might contain different sets of relational phrases: for instance, phrases like *just then*, *whereupon* and *sure enough* occur in narrative discourse but are relatively rare in academic articles. We have concentrated for the moment on academic texts in order to limit the scope of the investigation.

---

<sup>2</sup>A final point: phrases which refer to places in the original text (such as *in the next section*, *as already mentioned*) will pass the test as it stands—but this is only because their referents have been expressly removed through the operation of the test itself. We will not consider such phrases here, though they are of interest in their own right.

226 pages of text were analysed, from twelve different authors. This yielded a corpus of around 200 phrases: see Appendix A for a list of these. 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.<sup>3</sup>

The following sections report the results of analyses carried out on the corpus.

### 6.2.1 The Syntactic Diversity of Relational Phrases

Relational phrases fall into four syntactic classes (as defined by Quirk *et al* [20]):

- **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:

- (8)
- An object may move, *but* it remains the same object. . .
  - A general rule is needed to prevent comparative constructions. *Or* some rule is needed that will say: ‘if a word cannot . . .

- **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:

- (9)
- Although* it is common sense that labels are related, this is a difficult idea to explicate.
  - 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,

---

<sup>3</sup>Of course, it might be objected that *if . . . [then]* is really the relational phrase, rather than simply *if*; our test for relational phrases does not capture this as it stands. We intend to return to this matter in subsequent work.

between auxiliary verb and main verb, after a copula if there is one, before a sentential complement if there is one. For example:

- (10) 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 . . .

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

- (11) 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.

### 6.2.2 The Space of Relational Phrases

One finding that emerged from a study of the corpus was that the relational 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 relational phrases has an important consequence: it makes the class of relational phrases infinite in size. Phrases like *a 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 relational 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 relational 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 relational 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 relational phrases can typically stand alone without modifiers—for instance, *later* and *after* by themselves are still relational phrases. In what follows, modifiers have been stripped wherever possible.

### 6.2.3 Classifying Relational Phrases According to Function

Thus far, we have used a simple test for detecting relational phrases in text, and on the basis of this we have gathered a corpus of relational 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 relational phrases as evidence for ‘psychologically real’ 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.

The test for substitutability must first be slightly elaborated, however. 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 straightforward substitution will not always be possible. In view of this, we should allow candidate phrases to be substituted in the clause in a different position, if necessary.

The substitutability test can be expressed thus:

1. Take any relational phrase from the corpus and place it in a context where it naturally occurs.
2. Remove the relational phrase from its host clause, and insert a different phrase from the corpus (the **candidate phrase**) into the same clause, at any appropriate position.
3. If the resulting discourse is consistent with the sense of the original discourse—*regardless of which context is chosen in step 1*—then the candidate phrase is **substitutable** for the original phrase.

Note that substitutability relations can be mutual. For instance, the synonyms *later* and *afterwards* can be substituted for each other:

(12) John went to bed. *Later/Afterwards* he remembered he had not put out the cat.

But sometimes the relationship is not bidirectional, if one phrase is more general than the other. For instance, *then* is substitutable for *later*, as in (13); but *then* can also mark presentational sequences, whereas *later* cannot, as can be seen in (14).

(13) John went to bed. *Then* he remembered he had not put out the cat.

(14) John is not right for the job. For a start he's not punctual—*then/\*afterwards* there are his manners to consider.

## 7 A Taxonomy of Relational Phrases

The substitutability test can be used to construct a hierarchical taxonomy of relational phrases. Ideally, such a taxonomy should have certain properties:

- All the relational phrases in a language should figure in it.
- All the substitutability relationships between phrases should be explicit in it.
- No phrase should appear in more than one category unless it has two separate meanings.

We have carried out a classification of the corpus of relational phrases according to these requirements. A portion of the taxonomy, dealing with phrases which signal position in a sequence, is given in Figure 3: the rest appears in Appendix B. (The titles given to the categories are purely to illustrate the kind of roles the cue phrases might have; no theoretical significance should be attached to them.)

⟨Insert Figure 3 around here.⟩

Some examples of the substitutability relationships in Figure 3 are given below.

- *Moreover* can be substituted for *above all*; note that the reverse is not always true.
- In turn, *next* can be substituted for *moreover*; and *next* is substitutable by *n-thly*, where *n* is some small number.
- *Above all* is also substitutable by *finally*.
- Note also which substitutions are never possible: for instance, *first of all* and *next* (though these phrases are both substitutable by *n-thly*); or *in the end* and *above all* (though both phrases are substitutable by *next*).

### 7.1 Discussion of the Taxonomy

The taxonomy is not yet perfect, and does not yet completely satisfy the criteria set out at the beginning of this section. But at least a reasonably clear method exists for querying and improving it: questions about the inclusion or placing of a given phrase will be decided on the evidence of concrete linguistic examples.

It is interesting to note that the relationship between cue phrases and the relations they signal is often many-to-many: this point has been made before, and it is neatly captured in

the taxonomy. A single phrase like *next* signals a number of different relations—including all those marked by the phrases in the categories below it. At the same time, since many categories contain more than one phrase, it is often the case that several phrases signal exactly the same relation or set of relations.

Because the taxonomy is a ‘pre-theoretical’ one, and makes no reference to theories of discourse structure, it can serve as a useful test-bed for such theories. Do they have a way of accounting for the pattern of phrases it exhibits? An answer to this question requires more research—but a few preliminary observations can be made here.

Firstly, it should be noted that the categories of cue phrases can often be mapped roughly back onto relations in RST’s original set. Relations like SEQUENCE, CONTRAST, CIRCUMSTANCE, CAUSE and RESULT all find natural correlates in the taxonomy.

At the same time, there are significant differences. These are sometimes to do with granularity: for instance, cue phrases marking a sequential relation do not form a single group, but are themselves classified into a number of subgroups. On the other hand, the distinction between VOLITIONAL-RESULT and NON-VOLITIONAL-RESULT finds no correlate in the taxonomy. Furthermore, some RST relations do not appear to have any relational phrases associated with them. EVALUATION and BACKGROUND, for instance, seem to fall into this class. And ELABORATION, while a prominent relation in many RST analyses, seems to have no single phrase to mark it—a particularly surprising finding.

The taxonomy also yields some interesting observations about Sanders *et al*’s parameterisation of the space of relations. For instance, the concept of ‘negative polarity’ is a useful one: relations involving a lack of similarity (marked by phrases like *whereas*) and those involving a lack of causality (marked by phrases such as *notwithstanding that*) can all be marked by the phrases *though* and *although*—what is more, these phrases are not sensitive to the order of the spans being related. (See Figure 7.) The class of relations marked by these latter phrases seems to correspond to all Sanders *et al*’s ‘negative polarity’ relations. (However, note that *but* also marks relations signalling a lack of causality or a lack of similarity; yet as Elhadad and McKeown [4] point out, and the taxonomy also shows, *but* and *although* are not intersubstitutable.)

There is also a certain amount of evidence (Figure 5) for Sanders *et al*’s **source of coherence** parameter (which, it will be recalled, takes a value of ‘semantic’ or ‘pragmatic’). As *a result* always indicates a semantic relation

(15) John felt sick. *As a result*, he stayed in bed.

(16) \* John stayed in bed. *As a result*, he felt sick.

while *it follows that* always indicates a pragmatic relation:

(17) John's been in bed all day. *It follows that* he feels sick.

(18) \* John feels sick. *It follows that* he's been in bed all day.

However, for *negative polarity* relations, there are no phrases of this kind to make the semantic/pragmatic distinction explicit. This may suggest that readers and writers are not normally so concerned with the distinction in this case, and it might explain why evidence for this parameter was particularly scanty for negative polarity relations in Sanders *et al's* experiments.

A final point: the test for cue phrases picks out phrases like *incidentally*, *anyway*, and *by the way*: these are often thought of not as signallers of relations, but of 'pushes' or 'pops' on a focus stack. They are classified separately in the taxonomy (Figure 9); it has not yet been decided how to treat them.

## 8 Conclusion

This paper has presented a new methodology for motivating a set of discourse relations. Relations are treated as psychological constructs, operative in the processes of reading and writing; and evidence for these constructs is sought in a study of the linguistic means available or marking them explicitly in text. The methodology is based around two simple tests: one is for picking out what we have termed **relational phrases** in discourse, the other is for deciding on the substitutability relationships between the phrases. On the basis of these tests, a taxonomy of relational phrases can be built up for a given language, which can be used as the basis for a classification of discourse relations.

Treating relations as psychological constructs gives them an explanatory role in a theory of discourse structure: they form part of an account of why discourse is structured the way it is. This avoids the problems of arbitrariness which beset many current classifications of relations. At the same time, linking relations to relational phrases allows a fine-grained taxonomy to be constructed—one which can be justified systematically at every level by appealing to concrete linguistic evidence.

Two tasks remain outstanding for future work. Firstly, it is important to obtain evidence that the tests for relational phrases and for substitutability can be relied on to produce



similar results when used by different analysts: at present, they have only been applied by the two authors. Secondly, the taxonomy as it stands is of linguistic data rather than of relations proper: no relation definitions have yet been framed. It remains to work out a way of defining the categories in the taxonomy. Whether this can be done in terms of any of the existing theories of discourse structure is a question which we intend to address in future work.

## References

- [1] D Ballard, R Conrad, and R Longacre. The deep and surface grammar of interclausal relations. *Foundations of Language*, 4:70–118, 1971.
- [2] J Beekman and J Callow. *Translating the Word of God*. Zondervan Publishing House, Grand Rapids, MI, 1974.
- [3] R Cohen. A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics*, pages 251–258, 1984.
- [4] M Elhadad and K R McKeown. Generating connectives. In *COLING-90*, pages 97–101, 1990.
- [5] J Grimes. *The Thread of Discourse*. The Hague: Mouton, 1975.
- [6] B J Grosz and C L Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, pages 175–203, 1986.
- [7] M Halliday and R Hasan. *Cohesion in English*. Longman, 1976.
- [8] J A Hobbs. On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information, Stanford University, 1985.
- [9] E H Hovy. Planning coherent multisentential text. In *Proceedings of the 26th ACL Conference*, pages 163–169, Buffalo, 1988.
- [10] E H Hovy. Parsimonious and profligate approaches to the question of discourse structure relations. In *Proceedings of the 5th International Workshop on Natural Language Generation*, Pittsburgh, 1990.

- [11] E H Hovy. Unresolved issues in paragraph planning. In R Dale, C Mellish, and M Zock, editors, *Current Research in Natural Language Generation*, pages 17–45. Academic Press, 1990.
- [12] E H Hovy, J Lavid, E Maier, V Mittal, and C Paris. Employing knowledge resources in a new text planner architecture. In *Proceedings of the Sixth International Workshop on Natural Language Generation*, pages 57–72, Trento, Italy, 1992.
- [13] R Longacre. *The Grammar of Discourse: Notional and Surface Structures*. Plenum Press, New York, 1983.
- [14] E Maier and E H Hovy. A metafunctionally motivated taxonomy for discourse structure relations. In *Proceedings of the 3rd European Workshop on Natural Language Generation*, pages 38–45, Judenstein, Austria, 1989.
- [15] W C Mann, C M Matthiessen, and S A Thompson. Rhetorical structure theory and text analysis. Technical Report RR/89/242, Information Sciences Institute, Marina del Rey, CA, 1989.
- [16] W C Mann and S A Thompson. Rhetorical structure theory: A theory of text organisation. Technical Report RR/87/190, Information Sciences Institute, Marina del Rey, CA, 1987.
- [17] J Martin. *English Text: System and Structure*. Benjamin, Amsterdam, 1992.
- [18] J R Martin. Conjunction: The logic of English text. In J S Petöfi and E Sözer, editors, *Micro and Macroconnectivity of Texts*, pages 1–72. Helmut Buske Verlag, Hamburg, 1983.
- [19] J D Moore and C L Paris. Planning text for advisory dialogues. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 203–211, 1989.
- [20] R Quirk, S Greenbaum, G Leech, and J Svartik. *A grammar of contemporary English*. Longmans, London, 1972.
- [21] D Rösner and M Stede. Customising RST for the automatic production of technical manuals. In *Aspects of Automated Natural Language Generation*, pages 199–214. Springer-Verlag, 1992.

- [22] T J M Sanders, W P M Spooren, and L G M Noordman. Towards a taxonomy of coherence relations. *Discourse Processes*, 15:1–35, 1992.
- [23] T J M Sanders, W P M Spooren, and L G M Noordman. Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics (to appear)*, 1993.
- [24] D Schiffrin. *Discourse Markers*. Cambridge University Press, 1987.
- [25] E M Segal, J F Duchan, and P J Scott. The role of interclausal connectives in narrative structuring: Evidence from adults' interpretation of simple stories. *Discourse Processes*, 14:27–54, 1991.
- [26] K Vander Linden, S Cumming, and J Martin. Using system networks to build rhetorical structures. In *Aspects of Automated Natural Language Generation*, pages 183–198. Springer-Verlag, 1992.

## A The corpus of relational phrases

Phrase	Category
above all	Conj-adverb
actually	Conj-adverb
after	Subordinator
after all	Conj-adverb
afterwards	Conj-adverb
all in all	Conj-adverb
also	Conj-adverb
although	Subordinator
and	Coordinator
anyway	Conj-adverb
as a consequence	Conj-adverb
as a result	Conj-adverb
as soon as	Subordinator
at any rate	Conj-adverb
at first blush	Conj-adverb
at first view	Conj-adverb
at the outset	Conj-adverb
because	Subordinator
before	Conj-adverb
by comparison	Conj-adverb
by the same token	Conj-adverb
certainly	Conj-adverb
consequently	Conj-adverb
correspondingly	Conj-adverb
despite the fact that	Subordinator
either	Conj-adverb
equally	Conj-adverb
even	Conj-adverb
even then	Conj-adverb
every time	Subordinator
except insofar as	Subordinator
first	Conj-adverb
firstly	Conj-adverb
for a start	Conj-adverb
for instance	Conj-adverb
for the simple reason	Subordinator
further	Conj-adverb

Phrase	Category
accordingly	Conj-adverb
admittedly	Conj-adverb
after	Conj-adverb
after that	Conj-adverb
again	Conj-adverb
all the same	Conj-adverb
alternatively	Conj-adverb
always assuming that	Subordinator
and/or	Coordinator
as	Subordinator
as a corollary	Conj-adverb
as long as	Subordinator
as well	Conj-adverb
at first	Conj-adverb
at first sight	Conj-adverb
at the moment when	Subordinator
at the same time	Conj-adverb
before	Subordinator
but	Coordinator
by contrast	Conj-adverb
by the way	Conj-adverb
clearly	Conj-adverb
conversely	Conj-adverb
despite that	Conj-adverb
earlier	Conj-adverb
else	Coordinator
essentially, then	Conj-adverb
even so	Conj-adverb
eventually	Conj-adverb
except	Conj-adverb
finally	Conj-adverb
first of all	Conj-adverb
for	Subordinator
for example	Conj-adverb
for one thing	Conj-adverb
for this reason	Conj-adverb
furthermore	Conj-adverb

Phrase	Category
given that	Subordinator
however	Conj-adverb
if ever	Subordinator
if only	Subordinator
in a different vein	Conj-adverb
in addition	Conj-adverb
in case	Subordinator
in contrast	Conj-adverb
initially	Conj-adverb
in particular	Conj-adverb
in spite of that	Conj-adverb
in that case	Conj-adverb
in the case of X	Conj-adverb
in the first place	Conj-adverb
in this way	Conj-adverb
inasmuch as	Subordinator
indeed	Conj-adverb
it follows that	Phr.w/scomp
it might seem that	Phr.w/scomp
last	Conj-adverb
later	Conj-adverb
likewise	Conj-adverb
merely	Conj-adverb
more Xly	Conj-adverb
most Xly	Conj-adverb
much sooner	Conj-adverb
neither is it the case	Phr.w/scomp
next	Conj-adverb
nonetheless	Conj-adverb
not because	Conj-adverb
not that	Conj-adverb
notwithstanding that	Subordinator
now	Subordinator
obviously	Conj-adverb
on condition that	Subordinator
on one side	Conj-adverb
on the contrary	Conj-adverb
on the one hand	Conj-adverb

Phrase	Category
hence	Conj-adverb
if	Subordinator
if not	Conj-adverb
if so	Conj-adverb
in actual fact	Conj-adverb
in any case	Conj-adverb
in conclusion	Conj-adverb
in fact	Conj-adverb
in other words	Conj-adverb
in short	Conj-adverb
in sum	Conj-adverb
in the beginning	Conj-adverb
in the end	Conj-adverb
in the meantime	Conj-adverb
in turn	Conj-adverb
incidentally	Conj-adverb
instead	Conj-adverb
it might appear that	Phr.w/scomp
just as	Subordinator
lastly	Conj-adverb
let us assume	Phr.w/scomp
meanwhile	Conj-adverb
merely because	Subordinator
moreover	Conj-adverb
much later	Conj-adverb
naturally	Conj-adverb
nevertheless	Conj-adverb
no doubt	Conj-adverb
not	Conj-adverb
not only	Conj-adverb
notably	Conj-adverb
notwithstanding that,	Conj-adverb
now that	Subordinator
of course	Conj-adverb
on one hand	Conj-adverb
on the assumption that	Subordinator
on the grounds that	Subordinator
on the one side	Conj-adverb

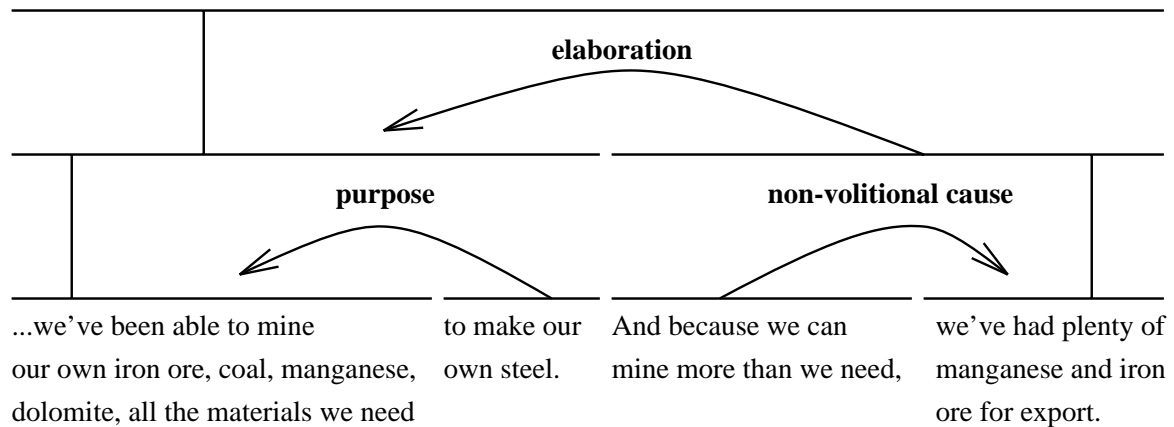
Phrase	Category
on the other hand	Conj-adverb
once	Subordinator
once more	Conj-adverb
or else	Coordinator
overall	Conj-adverb
presumably because	Subordinator
provided that	Subordinator
put another way	Conj-adverb
reciprocally	Conj-adverb
second	Conj-adverb
similarly	Conj-adverb
simultaneously	Conj-adverb
so	Subordinator
specifically	Conj-adverb
subsequently	Conj-adverb
summarising	Conj-adverb
suppose	Phr.w/scomp
supposing that	Subordinator
surely	Conj-adverb
that is to say	Conj-adverb
the more often	Subordinator
then again	Conj-adverb
thereby	Conj-adverb
third	Conj-adverb
this time	Conj-adverb
thus	Conj-adverb
to begin with	Conj-adverb
to start with	Conj-adverb
to summarise	Conj-adverb
to the degree that	Subordinator
too	Conj-adverb
ultimately	Conj-adverb
unless	Subordinator
we might say	Phr.w/scomp
when	Subordinator
where	Subordinator
wherein	Subordinator
while	Subordinator

Phrase	Category
on the other side	Conj-adverb
once again	Conj-adverb
or	Coordinator
otherwise	Conj-adverb
plainly	Conj-adverb
previously	Conj-adverb
providing that	Subordinator
rather	Conj-adverb
regardless of that	Conj-adverb
secondly	Conj-adverb
simply because	Subordinator
since	Subordinator
so that	Subordinator
still	Conj-adverb
such that	Subordinator
summing up	Conj-adverb
suppose that	Phr.w/scomp
sure enough	Conj-adverb
that is	Conj-adverb
the fact is that	Phr.w/scomp
then	Coordinator
thereafter	Conj-adverb
therefore	Conj-adverb
thirdly	Conj-adverb
though	Subordinator
to be sure	Conj-adverb
to conclude	Conj-adverb
to sum up	Conj-adverb
to take an example	Conj-adverb
to the extent that	Subordinator
true	Conj-adverb
undoubtedly	Conj-adverb
until	Subordinator
what is more	Conj-adverb
whenever	Subordinator
whereas	Conj-adverb
wherever	Subordinator
yet	Coordinator

## **B A preliminary taxonomy of the corpus of relational phrases according to function**

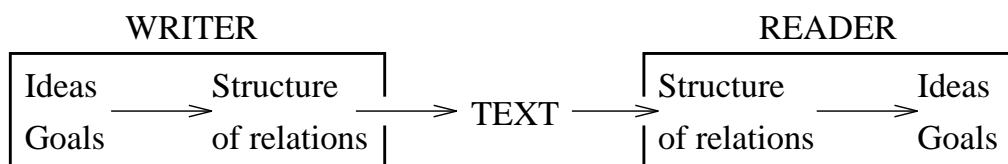
⟨Insert Figures 4, 5, 6, 7, 8 and 9 here.⟩

**Figure 1:** Example of a Rhetorical Structure Tree (from Mann and Thompson [16])

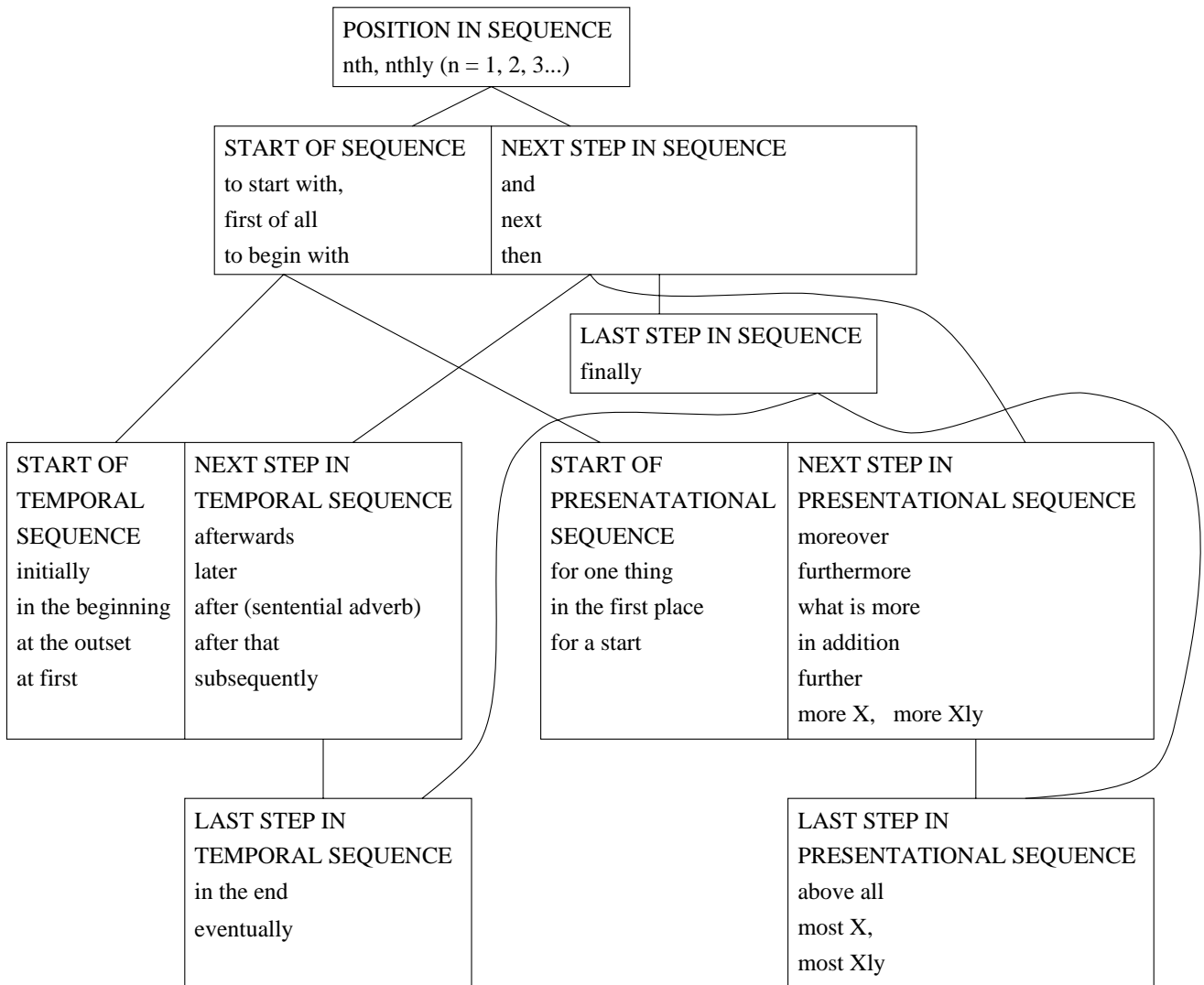




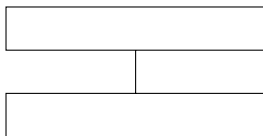
**Figure 2:** A Model of ‘Communication Via a Text’



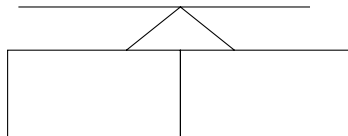
**Figure 3: Sequences**



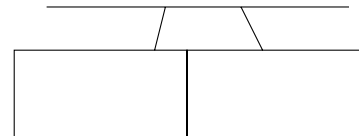
**KEY**



Phrases in daughter category can always be substituted by phrases in mother category

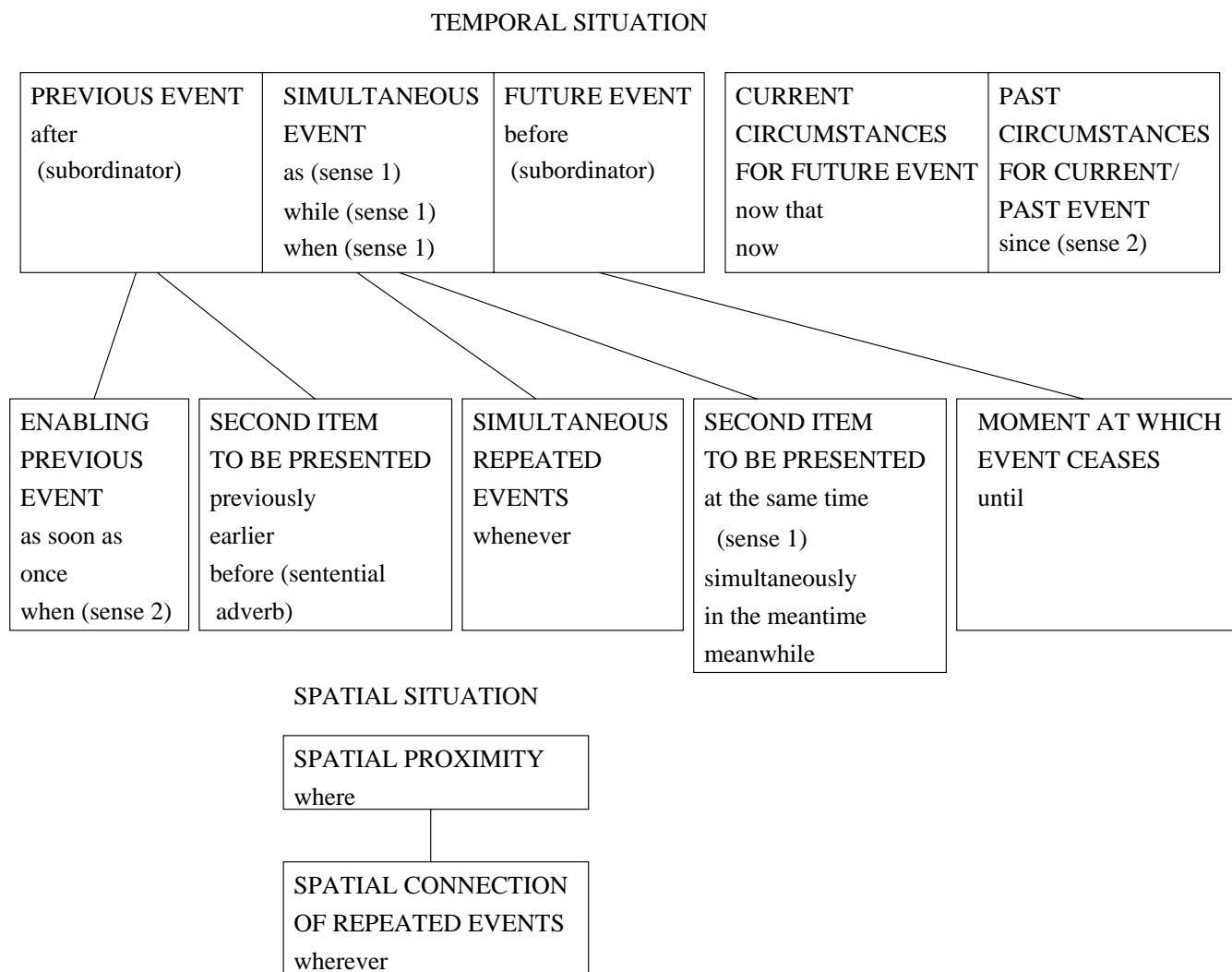


Exclusive categories: phrases in one category will never be substitutable for phrases in the other

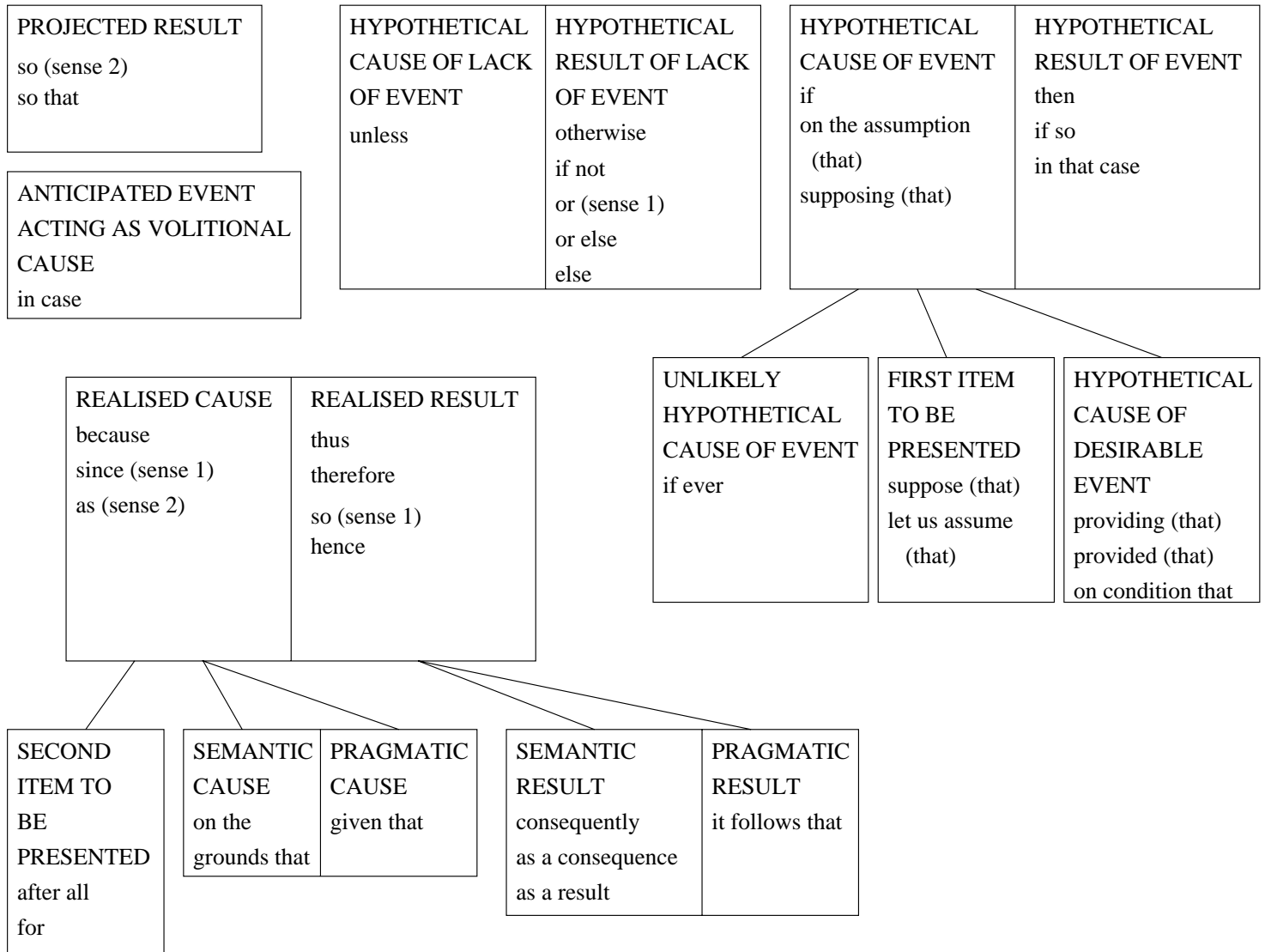


Phrases in one category may be substitutable for phrases in the other in certain contexts, but not always

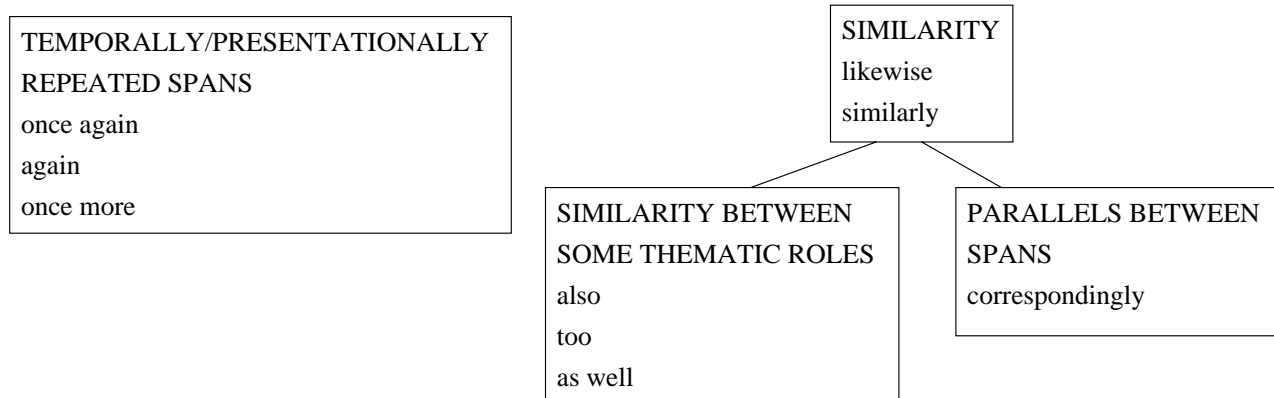
**Figure 4:** Situating an Event



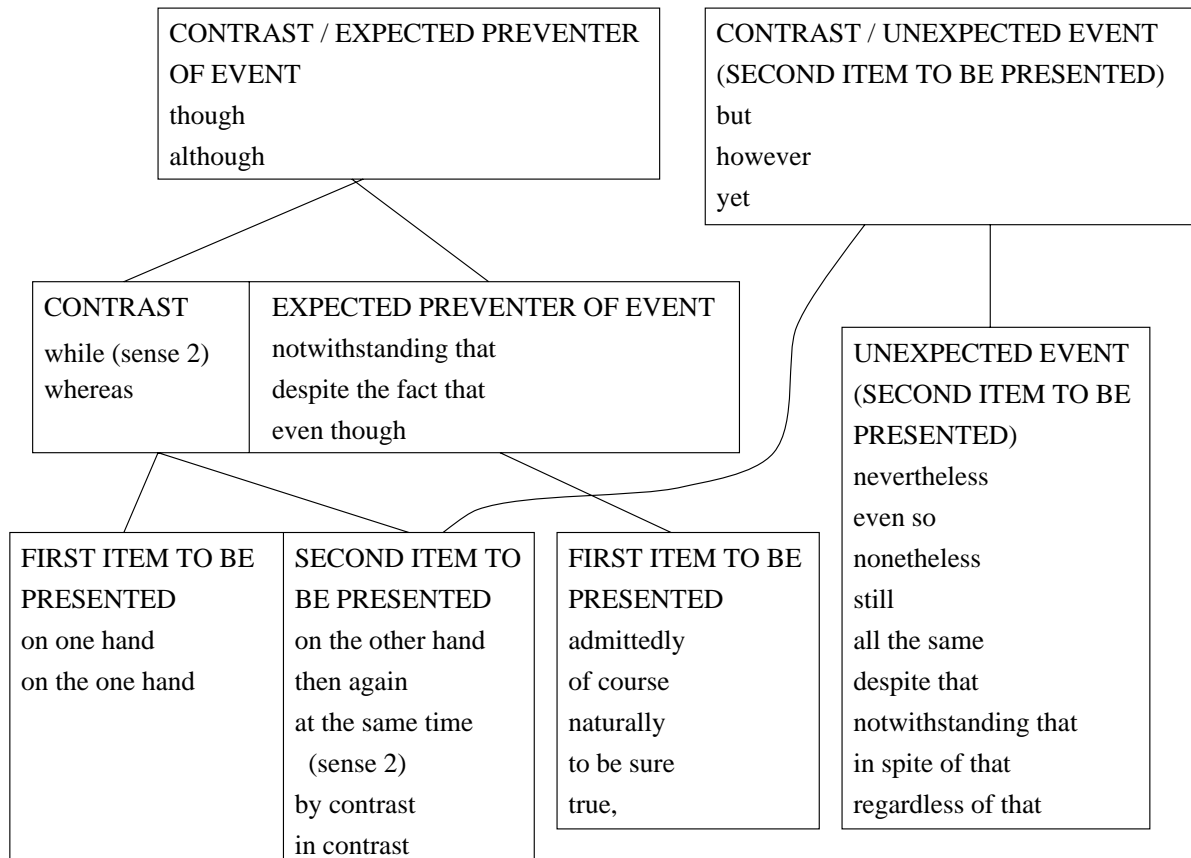
**Figure 5:** Causal/Purpose Relations



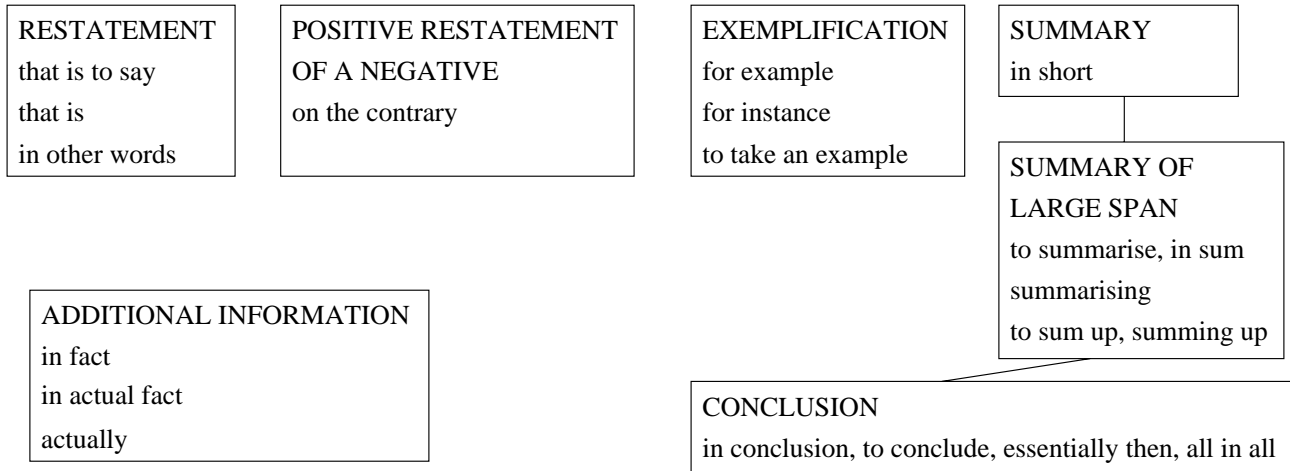
**Figure 6:** Similarity Relations



**Figure 7:** Contrast/Violated Expectation/Choice



**Figure 8: Clarifying Statements**



## Figure 9: Interruptions

DIGRESSION  
incidentally  
by the way

RETURN TO PREVIOUS POINT  
in any case  
in any event  
anyway