5 The study of discourse characteristics

5.1 Studying discourse characteristics

Discourse analyses focus on language characteristics that extend across clause boundaries. As a result, discourse characteristics are more difficult to identify and analyze than lower-level lexical or grammatical features. However, such analyses are important for both descriptive and applied linguistics. In fact, it turns out that the use of many lexical and grammatical features can only be fully understood through analysis of their functions in larger discourse contexts.

Most discourse studies identify salient discourse structures and exemplify those structures with illustrative text excerpts – such as identifying turn-taking structures in conversation or tracking the "themes" in a written text. However, it has proven difficult to apply these techniques to texts in a way that allows for generalizable results. Thus, although nearly all discourse studies are based on analysis of actual texts, they are not typically corpus-based investigations: most studies do not use quantitative methods to describe the extent to which different discourse structures are used, and relatively few of these studies aim to produce generalizable findings that hold across texts.

As a result of the lack of generalizable findings, we still know surprisingly little about the discourse similarities or differences across texts and registers. However, as the present chapter shows, corpus-based analyses can make a significant contribution in this area. Such techniques can be applied to a large body of texts to accurately describe the discourse characteristics of selected

registers, as well as the extent to which any individual text conforms to the expected discourse patterns of its register.

There are several reasons why discourse studies have generally not been corpus-based in the past. First, many discourse features cannot be identified automatically. The analysis of such features is often labor-intensive, requiring detailed consideration of language features in their larger textual contexts. For example, one goal of discourse analysis is to classify the kinds of information in a text, usually focusing on noun phrases as the primary carriers of referential information. Such analyses attempt to determine which pieces of information are already "known" by the reader/listener, versus those noun phrases that present "new" information. It is impossible to make distinctions of this type automatically; the discourse analyst must consider the previous textual context, and in some cases analysts even consider the background knowledge that readers/listeners use to understand a text. As a result, it requires a large commitment of time and energy to analyze extended texts in this way.

Furthermore, commercially available corpus analysis tools are not very helpful for investigations of discourse-level features. Standard concordancing packages are designed to produce a listing of specified target words with their immediate sentential contexts. Because such tools are not designed for complex grammatical or semantic analysis, they are also not suitable for discourse analyses. For example, concordancing packages provide no means for identifying all the nouns in a text, let alone classifying those nouns as known versus new referents.

Difficulties such as these might be taken to suggest that the corpus-based approach is not useful for discourse studies, and that it is simply not feasible to attempt broader studies with generalizable results. However, there are two major ways in which a corpus-based approach can be used to investigate discourse features. First, it is possible to develop and use interactive computer programs (similar to a spellchecker) to analyze discourse characteristics. Such programs can identify certain discourse characteristics more reliably and faster than humans can, while at the same time providing a means for the researcher to make judgements about areas that cannot be analyzed automatically. Second, it is possible to use automatic analyses to track the use of surface grammatical features over the course of a text. These analyses actually map the development of discourse patterns through texts; they can be used to compare texts, to find the typical patterns for a register, or to see how a particular text compares to the general pattern for the register.

These two types of analyses are exemplified in the present chapter through the following research questions:

1. How are references marked in different ways in different kinds of texts?

   This question is addressed in Section 5.2 with an investigation of the use of nouns and pronouns in four registers of English. Some specific questions relating to this issue are: What factors influence the choice between nouns and pronouns in a text? Which noun phrases present "given" (or "known") information, and which present "new" information? How are given and new referents distributed across texts?

2. How does the sequence of verbs within a text develop with respect to the marking of tense and voice?

   Some specific questions relating to this issue are: To what extent is there a prototypical sequence of verbs – or a "discourse map" of verbs – for all the texts in a register? To what extent do such discourse maps correspond to the underlying rhetorical divisions marked within a text?

   This second research question is illustrated in Section 5.3 with an investigation of the tense and voice of verbs in scientific research articles.

5.2 Reference types in spoken and written registers

Noun phrases are the major grammatical device used to refer to people, objects, or other entities in texts. However, texts from different registers often differ dramatically in the use of these "referring expressions." For example, consider the following two samples from news reportage and conversation, with all noun phrases italicized:

Text Sample 5.1: News reportage
Thortec International Inc. said *it* reached *agreements* with *an investor group* and *Wells Fargo Bank* under which *it* will receive *loans* and *an equity infusion* in return for *stock* that will reduce *the number of shares in public hands* by as much as 85 percent. *The engineering and consulting firm*, which has been plagued by *losses* for *five years*, said *the restructuring* is required to relieve *its debt burden* and *"acute shortage of cash."*

Text Sample 5.2: Conversation
A: Right, *I*'m ready. Have *you* locked *the back door*? [pause] *I* thought *we* were walking.
B: Well do *you* want to walk or do *you* want to go in *the car*?
A: Well *I* have to go to *the paper shop*.
B: Well *I*'ll drop *you* at *the paper shop* while *I* go round.
A: Oh *that*'s *a good idea*.

One clear difference between these two text samples concerns the form of the noun phrases. The sample from news reportage relies primarily on full noun phrases (*Thortec International Inc.*, *agreements*, *an investor group*, etc.), while the conversation sample uses pronouns (*I*, *you*, *we*, *that*) more commonly.

In addition to this difference, it is probably also obvious that these samples tend to use different types of reference. In particular, the conversation sample has a high percentage of "exophoric" – or text-external – reference, with the pronouns *I* and *you* referring directly to the speaker and addressee, rather than some previous entity in the text (see Section 5.2.1 for further explanation). The news sample does not include this kind of reference. Furthermore, because of the greater reliance on exophoric reference, more of the referents in the conversation sample are already known by both of the participants even when they are first mentioned (e.g., *I*, *you*, *the back door*, *the paper shop*), while more of the referents in the news sample are initially unfamiliar (e.g., *agreements*, *an investor group*).

In this section we illustrate how corpus-based analyses can be used to investigate the characteristics of referring expressions and to determine how registers differ in their use of reference. We present an analysis of the use of nouns and pronouns across four registers of English – examining their use to present given and new information and different kinds of reference. We use two spoken registers from the London–Lund Corpus and two written registers from the LOB Corpus: conversation and public speeches from the London–Lund, and news reportage and academic prose from the LOB.

The analytical procedures involved in this investigation are introduced step by step throughout this section. In Section 5.2.1, we provide an explanation of the informational characteristics that are included for analysis. In Section 5.2.2 we then describe the interactive computer techniques used to analyze those characteristics for all the nouns and pronouns found in the texts. Finally, Section 5.2.3 discusses the results of the analysis.

5.2.1 Characteristics of referring expressions

There are many characteristics of referring expressions that could be examined in order to better understand their use across texts and registers. For this sample analysis, we concentrate on four parameters:

- status of information: given versus new
- for given information, type of reference: anaphoric, exophoric, or inferrable
- for anaphoric reference, form of the expression: pronoun, synonym, or repetition
- for anaphoric reference, the distance between the anaphoric expression and its antecedent

Each of the noun phrases in a text can be classified according to the type of information that it presents: given or new. For example, if you look back at Sample 5.1 from a newspaper article, you can see that many of the noun phrases present new information, identifying a person or thing that has not previously been referred to in the text. Noun phrases of this type include: *Thortec International Inc., an investor group, Wells Fargo Bank, loans, an equity infusion, stock*. Other referring expressions present given information, representing an entity that has already been identified. In the first sentence of Sample 5.1, the pronoun *it* is used twice to mark a "given" referent, referring to the company *Thortec International Inc.*

Expressions presenting given information represent three kinds of reference relationships. Many of these expressions are "anaphoric." That is, they refer to a person or thing that has already been referred to in the text, i.e., the "antecedent." For example, the pronoun *it* in the first sentence of Sample 5.1 is anaphoric, referring to *Thortec International Inc.*, which is its antecedent.

However, other referents present given information because they refer to some person or thing in the external context. For example, in the conversation excerpt in Sample 5.2, the pronouns *I* and *you* refer directly to the speaker and addressee. *The back door, the car,* and *the paper shop* refer to physical objects present in the extended physical situation which are clearly understood by both participants. These are "exophoric" referents. Exophoric referents are given because their identity is known from the physical situation. In contrast, anaphoric referents are given because their identity is known from preceding textual references.

In addition to anaphoric and exophoric reference, there are some other expressions of given information which are more difficult to classify. For example, in Sample 5.1 the existence of *a restructuring*, referred to in the second sentence, is "inferrable" from the events described in the first sentence, but this noun does not have an anaphoric relation to any single preceding noun phrase nor does it refer to the external context. Similarly, the existence of a *debt burden* can be inferred from the fact that the company has been "plagued by losses," but again this does not represent an anaphoric relation. Thus, a category of "inferrable" is also necessary when classifying referents.[1]

Our third area of interest is to consider the different forms used to present anaphoric referents. Anaphoric referents are often expressed as a pronoun, as in the case of *it* used to refer to *Thortec International Inc.* in Sample 5.1. However, anaphoric referents can also be synonymous expressions, such as the use of *the engineering and consulting firm* in the second sentence of Sample 5.1 to refer to *Thortec International*. Finally, anaphoric referents can be a direct repetition of the original expression.

The fourth area that we examine in this illustrative analysis is the distance between the referring expression and its antecedent. For example, in the news reportage example above, the pronoun *it* occurs relatively close to the antecedent *Thortec International Inc.* The fuller synonymous expression *The engineering and consulting firm* has a greater distance from the original reference to this company.

Taken together, these four parameters can reveal many patterns in the use of reference in different registers. As you can see, however, there is a great deal of information to keep track of here, and analysis of even a couple thousand words of text could be very time-consuming. The next section explains how corpus-based analytical techniques can assist in studying the characteristics of referring expressions.

5.2.2 Interactive analysis techniques: coding the characteristics of referring expressions

In this section we describe the use of an interactive analysis program to code characteristics of referring expressions in texts. For the illustrative analysis here, a relatively small sample of the texts from the London–Lund and LOB corpora were coded: the first 200 words in forty texts (five texts from conversation, nine texts from public speeches, ten texts from news reportage, and sixteen texts from academic prose).[2]

Our investigation focused on the four informational characteristics reviewed in the last section, in addition to the form of the expression and the register of the text. Thus, the interactive program was designed to analyze and record up to six characteristics for each noun phrase:

1. register of the text
2. nominal form: pronoun versus full noun
3. information status: given versus new
4. if given, type of reference: anaphoric, exophoric, or inferrable
5. if anaphoric and a full noun, type of expression: synonym versus noun repetition (pronouns were already identified in step 2)
6. if anaphoric, the distance between the target referring expression and its antecedent

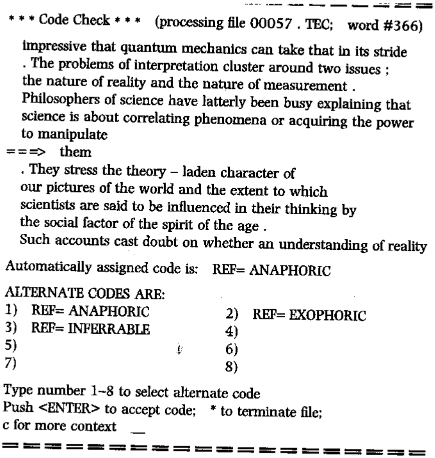
An interactive text analysis program is similar to a spellchecker in a word processor. In fact, to get a quick idea of the advantages

of using an interactive text analysis program, remember what it is like to proofread a paper without a spellchecker. Not only are spellcheckers many times faster, they are also many times more accurate because it is easy to miss typos when proofreading by hand. The same benefits hold for the use of interactive text analysis tools.

The first step for the present analysis was to grammatically tag all texts (as described in Chapter 3 and Methodology Boxes 4 and 5). The interactive program then cycles through each tagged text, stopping when it reaches a noun or pronoun (as identified by the grammatical tagger), and prompting the user to select the correct codes for that noun phrase. As with the spellchecker, there are several advantages to using this interactive text analysis program: first, the program identifies noun phrases automatically (so the analyst does not have to read through the text trying to spot noun phrases); second, the program provides an initial analysis of the informational characteristics of the noun phrase – when the initial analysis is correct, the user simply accepts that code; finally, the interactive program provides a list of other possible correct analyses to choose from, so that the user needs only to select the number corresponding to the correct analysis (if the initial analysis is not correct).

Figure 5.1 gives an example of a typical screen from the interactive program, showing how the codes can be accepted or edited. The referring expression currently being coded is presented in context and identified with an arrow (in Figure 5.1, the pronoun *them* is being coded). Underneath the text excerpt, the automatically assigned code is listed (in this example, "ANAPHORIC"), and then the alternative codes are listed. To choose one of these alternatives, the user simply types the number of the choice, or – as would be the case for this example – the user can push "ENTER" to accept the automatically assigned code.

The interactive text analysis program relies on the computer to perform those parts of the analysis that can be done automatically, while retaining human decision-making for those difficult analyses that involve meaning distinctions. Specifically, to code the six characteristics listed above, the program operates as follows:



```
*** Code Check ***   (processing file 00057 . TEC;   word #366)
   impressive that quantum mechanics can take that in its stride
   . The problems of interpretation cluster around two issues ;
   the nature of reality and the nature of measurement .
   Philosophers of science have latterly been busy explaining that
   science is about correlating phenomena or acquiring the power
   to manipulate
===>   them
   . They stress the theory – laden character of
   our pictures of the world and the extent to which
   scientists are said to be influenced in their thinking by
   the social factor of the spirit of the age .
   Such accounts cast doubt on whether an understanding of reality

Automatically assigned code is:   REF= ANAPHORIC

ALTERNATE CODES ARE:
1)  REF= ANAPHORIC              2)  REF= EXOPHORIC
3)  REF= INFERRABLE             4)
5)                              6)
7)                              8)

Type number 1–8 to select alternate code
Push <ENTER> to accept code;   * to terminate file;
c for more context   _
```

Figure 5.1 Sample screen from interactive program to code referring expressions

Characteristic 1. Register of the text

The register is recorded at the beginning of each text, and does not require any further analysis by the interactive program.

Characteristic 2. Nominal form: pronoun versus full noun

The program automatically records the grammatical category of the noun phrase (noun versus pronoun) using the information coded into the texts by the grammatical tagger.

Characteristic 3. Information status: given versus new

Pronouns are automatically coded as given information. For each noun, the program automatically checks whether there is an earlier occurrence of the same noun in the text. If there is, the repeated noun is automatically coded as given information. All other full nouns are pre-coded as new information. These nouns are then checked interactively to determine whether they actually represent given information.

Characteristic 4. If given information, type of reference: anaphoric, exophoric, or inferrable

The pronouns *I* and *you* are automatically coded as marking exophoric reference (i.e., referring directly to the speaker/writer or listener/reader). Third person pronouns are automatically labelled anaphoric but checked interactively to identify exophoric and inferrable occurrences.

Nouns with given informational status are automatically labelled anaphoric but checked interactively to identify exophoric and inferrable occurrences.

Characteristic 5: If anaphoric and a full noun, status as synonymous versus noun repetition

If nouns have been coded as anaphoric and an earlier occurrence of the same noun was found in the text, the referring expression is automatically identified as a noun repetition. Other anaphoric nouns are coded as synonymous.

Characteristic 6: Distance between the target referring expression and its antecedent

For this analysis, the antecedent of all anaphoric nouns and pronouns must be identified. For repeated nouns, the antecedent is automatically pre-coded as the earlier occurrence of the same noun; these antecedents are checked interactively to determine if there is a closer synonymous expression. For all other nouns and pronouns, the user of the interactive program must type in the antecedent.

The distance between the target referring expression and its antecedent can then be computed automatically. The program simply counts the number of intervening noun phrases occurring in between each referring expression and its antecedent.

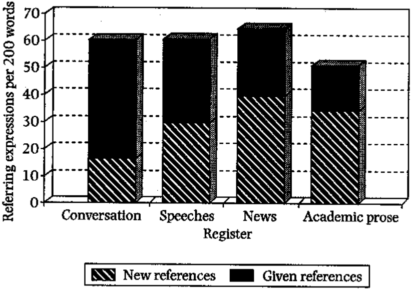


Figure 5.2 Frequency of given versus new referring expressions

As each noun phrase is analyzed, the codes are recorded in the text, with lines such as

<<<Ref = anaphoric

and

<<<Status = given

Another computer program is then used to analyze the coded texts and create a file listing the informational characteristics of each noun phrase. Finally, a statistical analysis package is used to compile frequency counts showing the interaction among these characteristics.

5.2.3 Patterns in the use of referring expressions in spoken and written registers

When the counts for the characteristics of all the noun phrases are compiled, the findings reveal a great deal about the use of different kinds of referring expressions in different registers. For example, Figure 5.2 plots the distribution of given and new referring expressions across the four registers. First, even considering only the overall frequencies of referring expressions, Figure 5.2 reveals some surprising differences. News reportage has the largest number of referring expressions of the four registers, while academic prose has the lowest. Interestingly, conversation and public speeches both have relatively frequent referring expressions, despite their general characterization as verbal rather than nominal.[3]

It is perhaps even more interesting that these four registers have striking differences in their reliance on given versus new references. At one extreme, over 70 percent of all referring expressions in conversation present given information. At the opposite extreme, over 65 percent of all referring expressions in academic prose present new information. News reportage shows the most frequent occurrence of new references, although they make up a slightly lower proportion than in academic prose.

The striking patterns of reference in conversation – both the surprisingly large number of total referring expressions and the extreme reliance on given references – can be better understood by considering the different types of given referring expressions. Figure 5.3 breaks down the occurrences of the given references to show the frequencies of exophoric pronouns, anaphoric pronouns, and anaphoric nouns.[4]

As you can see from Figure 5.3, exophoric pronouns account for over half of all given references in conversation. These pronouns are almost all used for reference to the speaker (*I*) or the hearer (*you*), although there are also exophoric references to third persons and objects present in the situational context (e.g., *she*, *he*, or *it*). Text Sample 5.3 illustrates the extremely common exophoric references typical of conversation. References to the speaker and hearer are italicized, while third-person exophoric references are marked in bold with square brackets (exophoric adverbial nouns are marked as well).

Text Sample 5.3: Conversation

A: What are *you* doing **[this afternoon]**?
B: *I*'m going **[home]**. *I*'ve got to teach about half past one. Can *you* pick **[your own trousers]** up?
A: no *I* don't think it'll be likely. *I*'ve got **[this meeting at three thirty]** and . . .

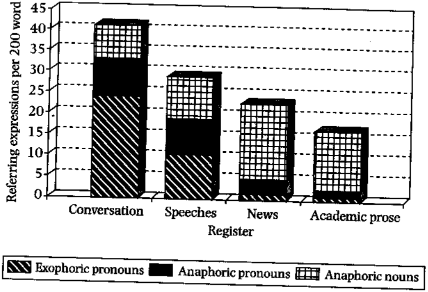


Figure 5.3 Frequency of exophoric and anaphoric referring expressions

> B: well what are *we* doing **[this weekend]**? **[Tomorrow]** *I*'ve got **[dancing class]** in **[the morning]**
> A: well *I*'ve nothing down anyway at all.
> B: well **[they]**'re open **[tomorrow afternoon]** up till three o'clock. *you* remember **[last time]** *we* went, and **[they]** were closed.

In contrast, Figure 5.3 shows that this reliance on exophoric reference is not found in either written register; in fact, apart from an occasional reference to the author (*I*), exophoric references are almost completely absent from these written registers. Instead, the written registers have a greater reliance on anaphoric references. Further, Figure 5.3 shows that the preferred nominal forms used for anaphoric reference are different in the written registers from the spoken registers: conversation and speeches make about equal use of anaphoric pronouns and full nouns, while the written registers use full nouns most of the time.

Text Sample 5.4 illustrates the patterns typical of written exposition, including the reliance on full noun phrases and the high proportion of expressions marking new information. The only exophoric reference in this passage refers to the authors (*us*; also note the use of the possessive pronoun *our*). In the text sample below, full noun phrases marking new information are italicized; anaphoric references are given in bold face marked by brackets; and exophoric references are given in bold face **CAPS**:

> *Text Sample 5.4: Academic prose*
> *The NPL Russian–English automatic dictionary* is organized on *a stem-paradigm basis* wherein there is for *most nouns* and *adjectives a single entry* for all *their inflected forms* and for *most verbs* only *one or two entries*. **[This]** is in contrast to *the full-form type of dictionary organisation* wherein *each inflected form* of *every word* has *a separate entry*. The decision to organise **[our dictionary]** on **[this basis]** was made so as to be able to accommodate **[it]** on *the magnetic tape store* available to **US** on *the ACE digital computer* of *our laboratory* . . .

There are two types of anaphoric reference in the passage. The first type refers back to a specific object introduced earlier: for example, the *Russian–English automatic dictionary* is later referred to as *our dictionary* and *it*. The second type of anaphoric reference refers back to a concept introduced in a preceding chunk of discourse; for example, the pronoun *this* and the noun phrase *this basis* both refer to the organization of the Russian–English dictionary. In addition, noun phrases like *the decision* identify inferrable information – readers are expected to figure out that the researchers had made some kind of decision concerning the organization of the dictionary.

As noted above, the computer program also coded the distance between anaphoric referring expressions and their antecedents in a text, with distance defined as the number of intervening noun phrases. For example, in Sample 5.1 above, the company *Thortec International Inc.* is referred to four times – by the company name, twice by the pronoun *it*, and once by a synonymous phrase *the engineering and consulting firm*. The passage is repeated here for convenience with these referring expressions italicized:

> *Thortec International Inc.* said *it* reached [agreements] with [an investor group] and [Wells Fargo Bank] under which *it* will receive [loans] and [an equity infusion] in [return] for [stock] that will reduce [the number] of [shares] in [public hands] by as much as 85 percent. *The engineering and consulting firm* . . .