Table 5.1 *Average distance measures for four registers*

	Average distance
Conversation	4.5
Public speeches	5.5
News reportage	11.0
Academic prose	9.0

The intervening referring expressions in this case are marked with square brackets. Thus, because there are no intervening expressions between *Thortec International Inc.* and the first *it*, this occurrence of the pronoun *it* has a distance of 0. The third reference to the company (again with a pronoun *it*) has a distance of 3, because there are three intervening referring expressions between the two occurrences of *it* (i.e., *agreements, an investor group,* and *Wells Fargo Bank*). Finally, the synonymous expression *the engineering and consulting firm* has a distance of 7 from the preceding reference to the company.

Based on such distance measurements for each anaphoric referring expression, it is possible to compute the average distance for each text and register. As Table 5.1 shows, the average distance scores are quite different across the four registers. In conversation and speeches, the average distance is considerably lower than in the written expository registers. This makes sense given the difference in the production and comprehension circumstances of written and spoken registers. Conversation and speeches must be comprehended on-line, and conversation must also be produced on-line. In those circumstances, co-references with short anaphoric distance are easier to understand. In contrast, writers and readers have as much time as they need to produce and comprehend written expository texts, and as a result, such texts have greater distances between co-referential expressions.

The frequent use of exophoric pronouns referring to the speaker or listener is also a major factor in the low average distance for conversation. For example, Sample 5.2 is repeated below with all exophoric references to Speaker A marked with italics and all exophoric references to Speaker B marked in bold **CAPS**. All other referring expressions are marked with square brackets.

Table 5.2 *Average distance measures for pronominal versus full noun anaphoric expressions*

	Average pronominal distance	Average full noun distance
Conversation	3.0	9.0
Public speeches	3.5	10.0
New reportage	3.0	13.5
Academic prose	2.5	10.0

> A: Right, *I'm* ready. Have **YOU** locked [the back door]?
> *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.

In this short excerpt, the many references to Speaker A have an average distance of less than 1, since many of these references occur in sequence with no intervening referring expressions. In contrast, references to Speaker B have a larger average distance of about 5 because of the many intervening referring expressions.

It turns out that the distance between two co-referential items is closely related to the form of the anaphoric referring expression: pronouns tend to occur much closer to their antecedent than repeated full nouns. Further, as Table 5.2 shows, this generalization holds across registers. That is, whether in conversation or expository writing, pronouns tend to occur relatively close to their antecedent, while full noun phrases tend to be used when the distance is greater. There is a natural explanation for this pattern: the greater the number of intervening referring expressions, the greater the chance for ambiguity and confusion over the intended reference of pronominal forms. Thus, full noun expressions are preferred for anaphoric reference over large distances, regardless of the register.

The pattern seen in Table 5.2 (i.e., pronouns associated with shorter referential distances than nouns) helps to explain the overall register differences in average distance. That is, because conversation relies heavily on pronominal reference – in particular repeated references to the speaker and hearer – the overall average distance in that register is small. In contrast, because the written expository registers have many referring expressions presenting new information, a large number of different expressions will typically intervene between any two co-referential expressions. As a result, the average distance between co-referential expressions is large in news reportage and academic prose, and anaphoric full nouns are preferred to anaphoric pronouns.

Results such as these show the importance of including multiple factors in studies of the referential discourse characteristics of texts. Obviously, this analysis could be extended in many ways. A larger number of texts and longer text samples are required to insure generalizable results. In addition, a number of other distinctions could be investigated, such as considering the way in which the different kinds of referring expressions are distributed across grammatical contexts (e.g., in main clauses versus dependent clauses).

This sample analysis has also illustrated the usefulness of interactive computer analyses, in combination with automatic techniques, for analyzing a discourse system. The following section shows another application of corpus-based techniques to discourse studies: actually tracking the progression of discourse features within a text.

5.3 Linguistic correlates of rhetorical structure: discourse maps of verb tense and voice

A major concern in discourse analysis is the way that meaning unfolds over the course of a text. From a discourse perspective, the meaning of a text cannot simply be derived from the meanings of the individual sentences in the text. Rather, there are systematic ways in which the grammatical resources of sentences work together at a discourse level. Choices among nominal forms in relation to the marking of given and new information, discussed in the last section, is one such discourse system.

The present section considers the marking of verb tense and voice, and the ways in which choices among the available tense/voice options can reflect larger rhetorical divisions within a text. That is, written texts are commonly divided into smaller sub-texts – chapters or sections – which are often marked overtly with titles. However, such divisions are not merely conventional ways of segmenting a text. Rather, they often reflect major shifts in communicative purpose within the course of a text.

It turns out that differences in communicative purpose also correspond to linguistic differences. However, linguistic features are more subtle indicators of purpose than titled section divisions, and as a result, it is more difficult for developing readers and writers to control such differences in use. Further, communicative purpose actually evolves continuously over the course of a text, rather than changing abruptly at marked section/chapter boundaries, making it even more important to understand the ways in which linguistic features reflect differences in purpose. In this section we illustrate this association by tracking the use of verb tense and voice over the course of a text.

Despite the central concern with structure larger than the sentence, it has been difficult in practice for discourse analysts to track the allocation of grammatical resources over the course of a text. One reason for this difficulty is simple bookkeeping: it is difficult to compile grammatical information from an entire text, and difficult to decide how to present such information once it is collected. Corpus-based analyses can help overcome these difficulties. Once the computer programs are developed for a given analysis, they can be applied to multiple texts of extended length with relatively little additional effort.

The following discussion illustrates the application of corpus-based methods to track the use of verb tense and voice across the major sections of research articles in experimental science. This is one of the few English registers that clearly distinguishes among internal purpose-shifts. That is, irrespective of discipline, experimental studies tend to follow a standard four-part organization: Introduction, Methods, Results, Discussion (I-M-R-D). Because each of these sections is overtly marked in the text and has distinct communicative functions, this register provides an ideal testing

Table 5.3 *Mean scores (per 1,000 words) of selected linguistic features across the I-M-R-D sections of English medical research articles (N = 19)*

	Section			
Linguistic feature	Introduction	Methods	Results	Discussion
present tense	47.9	21.1	35.9	60.6
F = 29.25; p < .001; r² = .549 (54.9%)				
past tense	20.7	48.5	40.3	13.0
F = 36.74; p < .001; r² = .605 (60.5%)				
agentless passives	18.4	39.9	16.9	16.3
F = 33.17; p < .001; r² = .580 (58.0%)				

site for discourse-analytic techniques. In the present section, we illustrate such techniques through analysis of a few texts; however, these same techniques could be applied to much larger corpora.

A simple first step is to compare frequency counts of linguistic features across the text-internal sections, interpreting any linguistic differences in terms of the primary communicative purposes of each section. To do this, we treat each article section as a separate text and compute counts for each of those texts; then we are able to compare the average or "mean" frequency counts for each type of section. This is a different kind of research design from those used previously in this book: it treats each text rather than each occurrence of a linguistic feature as the unit of analysis. (For further information about determining the unit of analysis in corpus-based studies see Methodology Box 8.)

Table 5.3 presents the mean scores (normalized per 1,000 words of text) for three tense and voice features across the I-M-R-D sections of nineteen medical articles published in 1985. The texts are taken from the ARCHER Corpus (see Chapter 8 and the appendix for descriptions of the ARCHER Corpus); all articles are from either the *New England Journal of Medicine* or the *Scottish Medical Journal*. The texts were grammatically tagged and the counts compiled with a program written for that purpose.

As Table 5.3 shows, present tense verbs occur most frequently in Discussion sections, and somewhat less frequently in Introductions. The use of present tense verbs in these sections reflects an emphasis on the current state of our knowledge and the present implications of research findings. In contrast, Methodology sections use considerably more past than present tense verbs, reflecting a focus on the reportage of past events and procedures. The Results sections in these articles have roughly equal frequencies of present and past tense verbs, although other studies based on science texts have found a much higher proportion of past tense verbs in Results sections (reporting the results of the study as past accomplishments rather than present findings). Finally, it should be noted that Methodology sections are also marked by their extremely frequent use of agentless passives, presenting events impersonally, with no acknowledged agent.

The following text samples, taken from an article published in the *Scottish Medical Journal*, illustrate these patterns in the use of verb tense and voice. Sample 5.5 comes from the Methodology section of the article and relies heavily on verbs in the past tense and passive voice to describe the procedures used in the experiment. Sample 5.6, in contrast, comes from the end of the Discussion section of the article. It relies almost exclusively on present tense verbs as it explains the implications for future studies.

> *Text Sample 5.5: Methodology section, medical research article* (past tense verbs are italicized, passive voice capitalized)
>
> Two hundred and fifty patients with rheumatoid arthritis who *required* change of NSAID *WERE* ENTERED into two consecutively run studies. In the first, 100 patients *WERE* randomly ALLOCATED to either feprazone (maximum dose 600 mg daily), or flurbiprofen (maximum dose 400 mg daily). In the second study, 150 patients *WERE* randomly ALLOCATED to ketoprofen (maximum dose 400 mg daily), benoxaprofen (maximum dose 600 mg daily), or fenbufen (maximum dose 900 mg daily). No additional information or advice *WAS* GIVEN other than that normally presented when non-steroidal anti-inflammatory drugs ARE CHANGED. . . .

> *Text Sample 5.6: Discussion section, medical research article* (present tense verbs are italicized)

> It *is* certainly apparent that further examples of similar remedies in this category *are* not needed. There would seem little point in marketing yet another "me too" drug unless acceptability by this method *allows* more than 50 per cent of patients to tolerate therapy for six months. Only then would the major exercise necessary to seek rare side effects *seem* justified.

The statistics reported in Table 5.3 show that all three grammatical features (present tense, past tense, and passive voice) have significant differences across the I-M-R-D sections, and these patterns are relatively strong (with r^2 values over 50 percent; see Methodology Box 9 for more on the reporting of statistics). These findings support earlier research showing that there are systematic linguistic differences across the sections of experimental research articles.

Frequency counts like these provide useful average characterizations of each section, and by considering such patterns across all four sections, it is possible to obtain an overview of the discourse organization of an article as a whole. However, average frequency counts cannot be used to analyze the way that a text actually develops. For example, it would be useful to know whether article sections are in fact coherent internally, or whether there are systematic patterns of variation within sections. Further, we also need an analytical approach that can analyze the discourse structure of non-experimental texts, which do not have overtly marked sections.

One approach for tracking the sequence of verbs over the course of a text is to use a graphic display, drawing a "map" of the progression of verbs through a text. We illustrate that approach here with the analysis of two medical research articles in ecology. (The texts are taken from the Corpus of Writing in the Disciplines described in Chapter 6; both texts come from the journal *Ecology*.)

A computer program was written to cycle through grammatically tagged texts and record the tense and voice of each verb phrase. For this illustrative analysis, the program made only a two-way distinction for tense: past versus non-past. Non-past tense includes both present tense and verb phrases with modal verbs (in which case the main verb is not marked for tense). A two-way distinction

was also made for voice: active versus passive. Non-finite clauses (i.e., infinitives and participial clauses) were excluded from the analysis.

The output of the program is set up in columns to show the movement between past and non-past tense, and between active and passive voice. When the program finds an occurrence of a verb phrase, it records the characteristics in the appropriate columns. The result is a visual display of the sequence of tense and voice as the text progresses.

Figure 5.4 displays the complete verb maps for the two articles in ecology. The sequence of verbs in each text moves from the top of the figure to the bottom, with each occurrence of a verb marked for tense and voice. Thus, in Text A the first verb is non-past and active (marked in the "NP" position for tense and the "A" position for voice); the second verb is non-past and passive (marked again in the "NP" position for tense, but the "PS" position for voice); and so on. The beginnings of the I-M-R-D sections are also marked on the figure.

As Figure 5.4 shows, the discourse progression of verbs in these two experimental articles is remarkably similar. Introduction sections in both articles are written mostly in the non-past tense (using present tense or modal verbs) and the active voice. In contrast, Methodology sections are written mostly in the past tense and passive voice. Results sections maintain the use of past tense but shift strongly to the use of active voice. Finally, Discussion sections maintain the predominant use of active voice but frequently switch back and forth among present tense, modal verbs, and past tense verbs.

These overall patterns correspond to our expectations from the frequency counts for each section. However, the maps can be additionally used to identify systematic departures from the expected patterns, leading to the identification of rhetorically salient shifts in the discourse. One such area of study relates to the transition zones between sections. That is, rather than shifting abruptly from one section to the next, these verb maps show that writers often begin a transition at the end of one section, or continue a transition into the beginning of the following section.

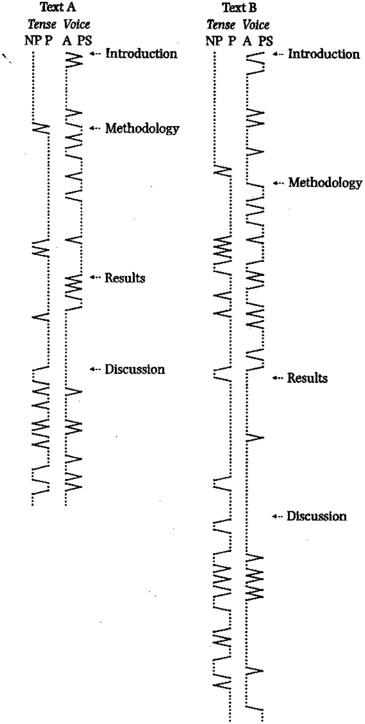

Figure 5.4 Verb maps of two research articles from ecology (NP = non-past; P = past; A = active; PS = passive)

For example, Figure 5.4 shows that the end of the Introduction section in Text B shifts into the past tense with active voice, a marked pattern for this section. In addition, this marked pattern continues for the first sentence of the Methodology section, before shifting to the more expected use of past tense with passive voice.

Text Sample 5.7 gives the full text corresponding to this sequence of verbs in Text B. The first two sentences in this extract present typical introductory material using present tense verbs and modal verbs to provide background information. However, the last paragraph of the Introduction section in this article shifts to provide a brief synopsis of the methodological steps followed in the study, and this shift in purpose corresponds to a shift to the past tense. Unlike typical Methodology sections, this synopsis is written as a personal account, with first-person pronouns as subjects and verbs in the active voice. Interestingly, this marked pattern is maintained for the first sentence of the Methods section, after which the exposition shifts to the more expected pattern providing methodological details in the past tense with passive voice.

> *Text Sample 5.7: Research article* (non-past tense is marked by bold face; past tense is marked by italics; passive voice verbs are given in CAPS)
>
> Introduction
>
> . . . .
>
> Because leaf area **IS LINEARLY RELATED** to sapwood cross-sectional area, if leaf area **is** static, sapwood volume **can only increase** if tree height **increases**. Because most height growth **occurs** early in stand development, sapwood volume and maintenance respiration **may not increase** greatly between intermediate and old-growth stands.
>
> In this study, we *tested* the hypothesis that higher maintenance respiration for woody tissues **would reduce** net primary production in older stands. We *estimated* maintenance respiration (R) for woody tissues from sapwood volume, sapwood temperature, and a correction for diurnal temperature amplitude. We then *compared* estimates of R with measured changes in above ground wood production for a chronosequence of subalpine lodgepole pine (pinus contorta ssp latifolia) stands in Colorado.

> Methods
> To reduce site differences, we *selected* three adjacent stands growing within 300 metres. The youngest stand (40 years old in 1986) *WAS ESTABLISHED* when a plot *HAD BEEN CLEARFELLED* for a regeneration study . . .

Two other marked transition zones in these texts occur between the Methodology and Results sections. In Text A, the use of passive voice verbs (interspersed with active voice) continues over the beginning part of the Results section. Text B shows a different kind of marked transition at this boundary, shifting to the present tense and active voice at the end of the Methodology section. Similar to the example discussed above, marked shifts such as these identify particularly interesting parts of a text, deserving more detailed discourse analysis to understand the associated shifts in communicative purpose.

Research of this kind could be extended in many ways. For example, it would be useful to investigate the patterns of modal verbs in these texts, as well as the use of perfect and progressive aspect verbs. For these characteristics, too, corpus-based techniques can act like a pointer, locating those detailed portions of a text that are particularly interesting from a discourse-analytic perspective. They do this by providing a comprehensive analysis of the text. That is, to identify those parts of a text that are marked rhetorically (as reflected in unexpected linguistic/discourse characteristics), it is necessary to first study complete texts, tracking the details of linguistic form as they develop over the course of a text. Corpus-based techniques are ideal both for providing comprehensive analyses of this type and for highlighting text chunks that depart from the expected norms.

5.4 Conclusion

This chapter has illustrated the usefulness of corpus-based techniques for investigating issues in discourse analysis. Specifically, it showed how analyzing the characteristics of referring expressions revealed interesting differences across spoken and written registers, and how constructing verb maps provided important

insight into the rhetorical structure of research articles. For both analyses, larger-scale studies with more texts would be necessary to insure generalizable results. However, even the relatively small analyses in this chapter showed how such investigations can reveal important patterns in language use, while also pointing to areas within texts which would be interesting for more intensive, qualitative analysis (such as the rhetorical transition zones in research articles).

The examples in this chapter also showed that corpus-based techniques are useful for linguistic studies even when analyses cannot be done completely automatically or when information needs to be tracked throughout a text. Interactive computer programs and innovative output formats make it possible to investigate issues that have previously been considered intractable. In the future, therefore, as more studies exploit these techniques, we should be able to learn more about patterns of discourse that hold across texts and registers.

Notes

1. There is a fourth type of pronominal reference – cataphoric – in which the pronoun refers to a noun that occurs later in the text. For simplicity we exclude this type of reference from the sample analyses in this chapter.
2. A fuller presentation of related findings is given in Biber (1992). That study followed a somewhat different methodology, with some output files being coded by hand.
3. The overall distribution of referring expressions was confirmed by an analysis of all the texts in these four registers from the London–Lund and LOB corpora. The full analysis resulted in slightly higher counts for news reportage and academic prose, but the overall distributional pattern remained unchanged.
4. As noted earlier, "inferrable" is also an important category for given information. However, judgements about what is inferrable raise a number of methodological issues that go beyond the scope of this chapter, and we have chosen to exclude "inferrable" from the discussion of results. Interested readers can find a useful discussion in Prince 1981.

Further reading

Because discourse studies often involve detailed interpretive analysis, there have been relatively few corpus-based investigations in this area. The following selected studies are corpus-based to differing extents: Biber (1992), Burges (1996), Chafe (1991), Fox (1987), Fraurud (1990), Givón (1983), Morris and Hirst (1991), Stenström (1987), Thavenius (1983), Youmans (1991).