Tikrit University

College of Education for Humanities

English Department

M.A Studies / Discourse Analysis

Corpus Approaches to Discourse Analysis

Dr. Muhammed Badea Ahmed

Corpus Approaches to Discourse Analysis

There are a number of advantages in using corpora to look at the use of language from a discourse perspective. As Biber, Conrad and Reppen (1998) point out, until recently many discourse studies have been based on comparatively small sets of textual data and have not typically been corpus-based. As a result it is often hard to generalize from these analyses. Larger sets of data analysed from a corpus perspective can make these findings of discourse studies more generalizable. Corpus studies can make an important contribution to our understanding of the characteristics of spoken and written discourse.

7.1 What is a corpus?

Before discussing corpus-based approaches to discourse analysis it is necessary to define what a corpus actually is. It is generally assumed that a corpus is a collection of spoken or written authentic texts that is representative of a particular area of language use, by virtue of its size and composition. It is not always the case, however, that the corpus is representative of language use in general, or even of a specific language variety, as the data set may be very specialized (such as material collected from the internet) and it may not always be based on samples of complete texts. The data may also be only of the spoken or written discourse of a single person, such as a single author's written work. It is important, then, to be aware of the specific nature and source of corpus data so that appropriate claims can be made from the analyses that are based on it (Kennedy 1998 , Tognini-Bonelli 2004). A corpus is usually computer-readable and able to be accessed with tools such as con�cordances which are able to find and sort out language patterns. The corpus has usually (although not always) been designed for the purpose of the analysis, and the texts have been selected to provide a sample of specific text-types, or genres, or a broad and balanced sample of spoken and/or written discourse (Stubbs 2004). Corpus studies draw on collections of texts that are usually stored and analysed electronically. They look at the occurrence and re-occurrence of particular linguistic features to see how and where they occur in the discourse. They may look at words that typically occur together (collocations) or they may look at the frequency of particular items. Corpus studies may look at language use in general, or they may look at the use of a particular linguistic feature in a particular domain, such as spoken academic discourse, or use of the item in a particular genre, such as university tutorial discussions. Corpus Approaches to Discourse Analysis 145

7.2 Kinds of corpora

General corpora

Corpora may be general or they may be specialized. A general corpus , also known as a reference corpus : aims to represent language in its broadest sense and to serve as a widely available resource for baseline or comparative studies of general linguistic features. (Reppen and Simpson 2004: 95) One use of a general corpus, for example, might be to examine words that collocate with girl and lady in English in general (Sigley and Holmes 2002) as opposed to words they collocate in particular domains of use, such as online personal ads. A further use of a general corpus might be to see to what extent hedges such as sort of and kind of are typical of English, in general, compared with what words these hedges typically collocate with in spoken academic discourse (Poos and Simpson 2002). A general corpus, thus, provides sample data from which we can make generalizations about spoken and written discourse as a whole, and frequencies of occurrence and co-occurrence of particular aspects of language in the discourse. It will not, however, tell us about the language and discourse of particular genres or domain of use (unless the corpus can be broken down into separate genres or areas of use in some way). For this, we need a specialized corpus .

Specialized corpora

A specialized corpus , as Hunston (2002 : 14) explains is: a corpus of texts of a particular type, such as newspaper editorials, geography textbooks, academic articles in a particular subject, lectures, casual conversations, essays written by students etc. It aims to be representative of a given type of text. It is used to investigate a particular type of language. Specialized corpora are required when the research question relates to the use of spoken or written discourse in particular kinds of texts or in particular situations. A specialized corpus might be used, for example, to examine the use of hedges in casual conversation or the ways in which people signal a change in topic in an academic presentation. It might look at an aspect of students' academic written discourse and compare this with use of the same features in published academic writing, or it may look at discourse features of a particular academic genre such as theses and dissertations, or a discourse level aspect of dissertation defences. 146 Discourse Analysis

The Michigan Corpus of Academic Spoken English

In contrast to a general corpus, then, a specialized corpus is usually designed with a particular research project in mind. An example of this is the Michigan Corpus of Academic Spoken English (MICASE) which has data from a wide range of spoken academic genres as well as information on speaker attributes and characteristics of the speech events contained in the data. This is an open access corpus and is available without charge to people who wish to use it (http://quod.lib.umich.edu/m/micase/). One study carried out using the MICASE corpus was an investigation of the uses of hedges such as sort of/sorta and kind of/kinda in spoken academic discourse. These were found to be more common in some disciplines such as the humanities, than in others such as science (Poos and Simpson 2002). Other MICASE studies have examined the ways in which new episodes are flagged in academic lectures and group discussions by the use of frame markers such as OK , so and now (Swales and Malczewski 2001) as well as other aspects of spoken academic discourse such as hedging in the discourse of academic lectures (Mauranen 2001). In the following example from Mauranen's study the hedging is in italics: okay. okay, um, let me get into sort of the more serious stuff, and, um, what i'm hoping to do with the remainder of of this first hour, is just give you some uh bit of perspective, show where biology fits into, sort of the rest of your education, and hopefully i can, um begin this framework that we're gonna fill in in the rest of the term. so i i have entitled this lecture, philosophy of science . . . or at least that's the point i'm talking about now. (Mauranen 2001 : 174) Findings from MICASE projects have been integrated into training courses for international teaching assistants and for the teaching of oral presentations (Reinhart 2002). The MICASE data has also been used in the development of English language tests (MICASE online).

The British Academic Spoken English corpus

A similar spoken corpus to the Michigan corpus, the British Academic Spoken English (BASE) corpus (www2.warwick.ac.uk/fac/soc/al/research/collect/base/) was developed at the University of Warwick and the University of Reading in the United Kingdom. One study based on the British corpus looked at the relationship between lexical density and speed in academic lectures (Nesi 2001). This study drew on data from 30 undergraduate lectures and found there was a range of speeds in the spoken discourse of the people delivering academic lectures. The lectures that were faster tended to be less lexically dense and the lectures that were slower tended to be more lexically dense. Lecturers spoke more quickly or were more lexically dense if they did not expect students to take notes, or if they were not presenting new content in their lecture. They also spoke more quickly if they were telling an anecdote which was an aside to the main content of the lecture. Nesi found, in looking at published coursebooks on listening to lectures that this range of speeds and ways of talking were not Corpus Approaches to Discourse Analysis 147 included in the books that she examined. Observations of this kind then have important implications for the development of English for academic purposes courses which aim to prepare students to study in English medium universities.

The British Academic Written English corpus

Specialized corpora may also be based on written discourse alone. An example of this is the British Academic Written English (BAWE) corpus (Nesi 2011) developed at the University of Warwick, the University of Reading and Oxford Brookes University in the United Kingdom (www2.warwick.ac.uk/fac/soc/al/research/collect/bawe). This corpus examines students' written assignments at different levels of study and in a range of disciplines with the goal of providing a database for use by researchers and teachers to enable them to identify and describe academic writing requirements in British university settings. The BAWE corpus includes contextual information on the students' writing such as the gender and year of study of the student, details of the course the assignment was set for and the grade that was awarded to the piece of work so as to be able to consider the relationship between these variables and the nature of the students' written academic discourse (see wwwm.coven⍰try.ac.uk/researchnet/elc/Pages/corpora.aspx for information on other spoken academic English corpora).

The TOEFL Spoken and Written Academic Language Corpus

A specialized corpus may include both spoken and written discourse. An example of a corpus which does this is the TOEFL 2000 Spoken and Written Academic Language Corpus (a specialized corpus). This corpus aimed to provide a comprehensive linguis⍰tic description of spoken and written registers in US universities, although not, in this case, examples of student writing. The TOEFL corpus was made up of 2.7 million words and aimed to represent the spoken and academic genres that university students in the United States have to participate in, or read, such as class sessions, office hour con⍰versations, study group discussions, on-campus service encounters, text books, reading packs, university catalogues and brochures. The corpus data was collected across four academic sites, each representing a different type of university: a teacher’s college, a midsize regional university, an urban research university and a rural research university. The spoken data was mostly recorded by students, although academic and other staff recorded office hours material and service encounters. The spoken and written classroom material focused on the disciplines of business, education, engineering, humanities, natural and social science, at lower and upper undergraduate and graduate levels of study (Biber et al. 2002). A key observation of the TOEFL study was that spoken genres in US university settings are fundamentally different from written genres. The study found, however, that classroom teaching in the United States was similar in many ways to conversational genres. It found 148 Discourse Analysis that language use varied in the textbooks of different disciplines, but not in classroom teaching in different disciplines.

7.3 Design and construction of corpora

There are, thus, a number of already established corpora that can be used for doing corpusbased discourse studies. These contain data that can be used for asking very many questions about the use of spoken and written discourse both in general and in specific areas of use, such as academic writing or speaking. If, however, your interest is in what happens in a particular genre, or in a particular genre in a setting for which there is no available data, then you will have to make up your own corpus for your study. Hyland’s (2002a) study of the use of personal pronouns such as I , me , we and us in Hong Kong student’s academic writing is an example of a corpus that was designed to answer a question about the use of discourse in a particular genre, in a particular setting. The specific aim of his study was to examine the extent to which student writers use self-mention in their texts ‘to strengthen their arguments and gain personal recognition for their claims’ in their written discourse, as expert writers do (Hyland 2005a : 178). His question was related to issues of discourse and identity, and the place of this writing practice in a particular academic and social community. A corpus collected at another institution or in another country would not have told him what students at his institution did. He was, however, able to use an existing corpus to compare his findings with how published academic writers use personal pronouns in their writing as a reference point for his study. Thus, by using his own custom-made corpus and an existing corpus, he was able to compare the findings of his study with the practices of the broader academic community and make observations about the way the students position themselves in the discourse, in particular, on the basis of this. Harwood (2005) also compiled his own corpus for his study of the use of the personal pronouns I and we in journal research articles. For his study, Harwood selected research articles from electronic versions of journals as well as manually scanned articles and converted them to text format. His analysis of his data was both quantitative and qualitative. The quantitative analysis examined the frequency of writers’ use of I and we in the texts and the disciplines in which this occurred. The qualitative analysis examined the use of I and we from a functional perspective; that is, what the function was of these items in the texts, as well as possible explanations for their use. He then compared his findings with explanations of the use of I and we in published academic writing textbooks. A further example of a researcher-compiled corpus is Ooi’s (2001) study of the language of personal ads on the world wide web (discussed later in this chapter). Ooi had to make up his own corpus to see how people use language in this particular genre. A large-scale corpus of language use on the world wide web, in general, would not have told him this. ‘Off the Corpus Approaches to Discourse Analysis 149 shelf’ corpora and custom-made corpora, then, each have their strengths, and their limita□tions. The choice of which to use is, in part, a matter of the research question, as well as the availability, or not, of a suitable corpus to help with answering the question. It is not necessarily the case, however, that a custom-made corpus needs to be especially large. It depends on what the purpose of collecting the corpus is. As Sinclair (2001) has argued, small manageable corpora can be put together relatively quickly and can be honed to very specific genres and very specific areas of discourse use. They can also be extremely useful for the teaching of particular genres and for investigating learner needs.