



Corpus Approaches to Discourse Analysis

Ribut Wahyudi, UIN M2I Malang, 18-04-2013.

Email: ribut@bsi.uin-malang.ac.id

Why do we need corpus ?

- Biber et al (1998) there are relatively small sets of textual data and have not typically been corpus based. Therefore it is often hard to generalize the analyses (in Paltridge 2006, p.155)
- Larger sets of data, using corpus based, can make findings in discourse studies more generalizable.
- Corpus studies provide insights of the characteristics of spoken and written discourse.

What is corpus ?

- Corpus studies draw on “collection of texts stored and analyzed electronically”
- They look at “occurrence and re-occurrence of particular linguistic features to see how and where they occur in discourse” e.g. collocation, the frequency of particular features.
- May look at “language use in general or in particular domain” e.g. spoken academic discourse & university tutorial discussions.

Kinds of corpora

- General corpora
- Specialized corpora
- The Michigan Corpus of Academic Spoken English (MICASE)
- The British Academic Spoken English (BASE)
- The British Academic Written English (BAWE) Corpus
- The TOEFL Spoken and Written Academic Language Corpus

Design and Construction of Corpora

- Such established corpora can be used for doing corpus-based discourse studies
- If the data you are looking for are NOT available there, then you need to create your own corpus. e.g. Hyland's (2002a) study of personal pronouns such as I, me, we, and us in Hongkong student's academic writing. (In Paltridge, 2006).

Issues to consider in constructing corpus

- what to include in the corpus
- the size of the corpus and categories along with the texts to be included
- the way in which data can be collected
- kinds of questions that will be examined using the data (McCarty and Carter, 2001).
- Other issues: nationality, gender, age, occupation, education, native language, dialects etc

More issues

- Authenticity, representativeness and validity of the corpus
- Kinds of texts to include in the corpus
- Size of the texts in the corpus
- Sampling and representativeness of the corpus

The Longman Spoken and Written English Corpus

- Made up of 40 million words
- It has 4 major discourse types: conversation, fiction, news and academic prose.
- + Two additional categories: non-conversational speech and general written non-fiction prose.

Discourse characteristics of conversational English

- Non-clausal units in conversational discourse
- Personal pronouns and ellipsis in conversation
- Situational ellipsis in conversation
- Non-clausal units as elliptic relies in conversation
- Repetition in conversation
- Lexical bundles in conversational discourse

Performance phenomena of conversational discourse

- Silent and filled pauses in conversation
- Utterance launchers and filled pauses
- Attention signals in conversation
- Response elicitors in conversation
- Non-clausal items as a response forms
- Extended coordination of clauses.

Constructional principles of conversational discourse

- # Keep talking, limited planning, qualification of what has been said.
- Prefaces in conversation
- Tags in conversation

Corpus studies of the social nature of discourse

- the use of discourse means in wider social terms
- Swales (2003) use MICASE corpus, asks whether the use of spoken language in academic settings can help us understand whether the university is a single community of practice and a set of ‘tribalized coteries’ of communities of practice. (In Paltridge, 2006).

Conts...

- The area of academic speaking (in contrast to academic writing) has fewer differences between disciplines
- Academic speaking across the university tend to be informal and conversational
- guarded rather than evaluative
- deferential rather than confrontational
- He concludes that there are fewer barriers to cross disciplinary oral communication than perhaps in written academic communication.
- Hyland's (2004b) study of generic structure of second language student's acknowledgement => discourse structure and social role. (In Paltridge, 2006).

Collocation and corpus studies

- Hyland and Tse (2004) study of dissertation acknowledgement of “special thanks”, “sincere thanks”, “deep thanks” etc.
- Ooi (2001) studies about the language of personal ads on internet sites in the US and Singapore. He uses concordance program WordSmith Tools to examine word frequency and lexical and grammatical collocation.
- Brithiaux (1994) studies on the writers frequently used of *personal chaining* and *hyphenated items* in personal advertisement. e.g. artistic, athletic, adorable 18-32 year old, good-looking etc.

Criticism of corpus studies

- Flowerdew (2005) provides a summary of and response to some criticisms
- One criticism is that computer based orientation of corpus studies leads to atomized, bottom up investigation of language use.
- Another criticism is that corpus studies do not take account of contextual aspect of text.
- Tribble (2002) counters these views by providing a detailed discussion of contextual features such as: social context of text, communicative purpose of text, role of readers and writers of text, shared cultural values required by readers and writers of texts and knowledge of other relevant texts. (See Paltridge, 2006, p. 175).

Reference

- Paltridge, B. 2006. Discourse Analysis. London: Continuum.

Thank You

Kingsoft Office

Make Presentation much more fun