The Role of Data in the Analysis of Academic Discourse

John Flowerdew
Lancaster University and Birkbeck, University of London
• Will start by looking at some books with the title *Academic Discourse*
• Will review text-focussed and contextually-focussed research
• Will look at two studies of my own to illustrate each approach
• Will briefly talk about data-driven learning (if time)
• Will conclude
Bourdieu et al. (1994) (1964)

• Like most work on academic discourse, the context is the university and students, teachers, and researchers
• A series of surveys conducted with university students in France in 1963-64
• Tested the students knowledge of academic terms such as epistemology, methodology, and phenomenon
• Students had difficulty defining these words out of context
Bourdieu et al.

• “Academic language is a dead language for the great majority of French people, and is no one’s mother tongue, not even that of the children of the cultivated classes.”
• Nevertheless, middle class students more familiar with the vocabulary than working class students
• “What is in question here is less a language as such than a relationship to language.”
• Middle class upbringing predisposes middle class students to achieve better in the academic context
• “Our survey reveals two fundamental facts: the importance of linguistic misunderstanding in higher education and the determining role of linguistic inheritance in academic success.”
Approach to data

- On the surface, one might think that “Discourse” in the title of the volume is reduced to academic terminology.
- But underlying this academic vocabulary is prior ethnographic and discourse analytic work.
- A “loop in which ethnographic material was tested statistically and then brought into a new ethnographic round of inquiry” (Blommaert, 2015, 7).
- This becomes more evident in Bourdieu’s (1986) later work, where he develops the concepts of habitus, and cultural and linguistic capital.

Approach to data

• (Like Foucault) Bourdieu puts more emphasis on the social context of discourse than on the actual substance of discourse (text)

• He emphasises the political/ideological dimension of discourse

• Power is attained by the powerful because they are better able to master the academic discourse, have greater knowledge, and are thereby able to exercise control

• “Surface” text is relatively less important
Flowerdew (2001)

• A collection of specially commissioned articles on the theme of academic discourse
• Situated in the context of the spread of academic English worldwide:
  • “English is now well established as the world language of research and publication … and more and more universities and institutes of learning are using English as the language of instruction (p.1).”
Flowerdew (2001)

- Research in academic discourse has an applied goal, the teaching of academic English:
  - "The study of academic discourse in English has throughout its short history been closely tied in with developments in the preparation of non-native speakers (NNSs) for study in English, English for Academic Purposes (EAP) (p.1.)"
Flowerdew (2001)

- Divided into sections covering the main approaches:
  - Genre analysis, corpus linguistics, contrastive rhetoric, and ethnographically inspired (naturalistic) approaches
- Each begins with a state of the art overview of the approach and continues with exemplificatory empirical studies
- Various academic genres are analysed: research articles, PhD theses, textbooks, argumentative essays, and business cases
- The first three sections are text-based, while the fourth (ethnography) is more contextually focussed
- The book is more empirical than Bourdieu et al.
- Each of the empirical studies is heavily focussed on the data (data-driven)
Hyland (2009)

• “My aim is to provide an overview of the key genres of the academy and on the way to say something about the nature of knowledge, of communication, and of the practices of those who work and study in universities” (p. viii)
• Emphasises the global nature of contemporary academic discourse
• Contextualised within the global massification of higher education
• Academic careers tied to publication and academic discourses are concomitantly taking on a growing impact
Hyland (2009)

- Discusses discourse communities and disciplines as important theoretical concepts
- Highlights the text/context dichotomy of approaches
- Textual approaches include genre analysis, corpus linguistics, and multi-modal approaches
- Contextual approaches include sociology of science, sociohistorical, ethnographic, and critical approaches
• Chapters 4-7 consider a range of academic discourse types
  • Research discourses
  • Instructional discourses
  • Student discourses
  • Popular discourses
• Final chapter returns to the global context and economic power
Text/Context Dichotomy

Text ←------------------------------------------→ Context
Text-focussed approaches
Text-focussed approaches

• Scientific discourse traditionally seen as the impersonal transmission of knowledge
• Basis in Aristotelian logic, and 18th century Cartesian rationalism
• Later, a realisation that academic discourse is more than the transmission of facts
• Emphasis now on the persuasive and attitudinal functions of academic text
  • “the rhetorical perspective [with] authors aiming … to convince the reader that the issue is indeed worth addressing, that their conclusions are well supported by data, that the analysis is methodologically flawless and that they have the necessary expertise in the field to take a stand on the matter (Warshal, 2020:12)
  • [research reports are] “complexly distanced reconstructions of research activities, at least part of this reconstructive process deriving from a need to anticipate and discountenance negative reactions to the knowledge claims being advanced.” (Swales, 1990:175)
Some notable approaches of text-focused work

• Origins in register analysis in the 1960s (frequency counts of linguistic features to characterise particular registers) (done by hand on very small corpora before concordancing software available) (Barber, 1962; Halliday, McIntosh and Strevens, 1964)

• Developed into more functional approaches, including rhetorical function analysis (Trimble, 1985), genre analysis (Swales, 1990), and Systemic Functional Linguistics (Halliday, 1994)
Some notable examples of text-focussed work (cont.)

- In the 1980s, larger corpora began to be used with the advent of digitalisation and corpus linguistics techniques
- Some notable highlights:
  - Hyland’s work on disciplinary discourses, stance and engagement and, more recently, with Jiang (2019), a diachronic approach
  - Biber’s multi-dimensional analysis (MD, or MDA) (based on factor analysis of linguistic features) (involved, narrative, argumentative, abstract)
  - MICASE, MCUSP, BASE, and BAWE
  - Mauranen’s work on lingua franca academic English (ELFA, WrELFA)
Application of findings

• Data analysed and fed back to learners in pedagogic materials (e.g. Swales and Feak, 2004)


• The direct use of corpus linguistic techniques by learners, referred to as data-driven learning (DDL) (Johns, 1990,1991)

An example of the text-focussed approach

Characterisation of signalling nouns

• SNs are abstract nouns which function in particular ways in discourse
• Examples are advantage, case, evaluation, possibility, procedure, situation, solution
• An SN refers to a general area of meaning of which the specific meaning is found elsewhere in the clause or text
• Identity of reference between SN and specifics/realisation
• Can be viewed as nominalized processes (many derived from verbs)
• A functional, not a formal, category
Realisation/specification patterns (1)

• Within Clause

• The idea that one could take these men out of society and cure them in hospitals appealed to citizens who feared these offenders and to psychiatrists who wanted to change them.

• Their role is to carry out the depolarizing phase of an action potential.
• Across clause

• For many years the study of life's origins was not considered worthy of serious speculation by biologists because, it was argued, the absence of a geological record made the course of events resulting in the appearance of life unknowable. This situation has changed.

• Resources are not unlimited. Shortages, temporary or permanent, can result from several causes. Brisk demand may bring in orders that exceed manufacturing capacity or outpace the response time required to gear up a production line…
Corpus

- matched sub-corpora of lectures, textbook chapters and research articles across 10 natural science and social science disciplines
- 613,514 words in the corpus
- annotation by automated techniques and by hand
- a small corpus because manual analysis required in order to make sure all types and tokens in the corpus are identified
- other studies use larger corpora, but many patterns are missed out
- a small corpus can be adequate if you are looking at a frequent feature, which is the case here
- it also allows you to become intimately familiar with the data
Overall frequency

- 845 SN types in the corpus

- 16,058 SN tokens in the corpus
  - 2.63 SNs per 100 words
  - one SN per 38 words
  - 26,300 SNs per million words
15 Most Frequent SNs in the corpus

<table>
<thead>
<tr>
<th>Rank</th>
<th>Lemma</th>
<th>Norm. Freq</th>
<th>singular</th>
<th>plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>case</td>
<td>890</td>
<td>685</td>
<td>205</td>
</tr>
<tr>
<td>2</td>
<td>way</td>
<td>668</td>
<td>504</td>
<td>165</td>
</tr>
<tr>
<td>3</td>
<td>problem</td>
<td>619</td>
<td>438</td>
<td>181</td>
</tr>
<tr>
<td>4</td>
<td>result</td>
<td>572</td>
<td>292</td>
<td>280</td>
</tr>
<tr>
<td>5</td>
<td>thing</td>
<td>538</td>
<td>342</td>
<td>196</td>
</tr>
<tr>
<td>6</td>
<td>theory</td>
<td>494</td>
<td>430</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>model</td>
<td>474</td>
<td>425</td>
<td>49</td>
</tr>
<tr>
<td>8</td>
<td>idea</td>
<td>427</td>
<td>370</td>
<td>57</td>
</tr>
<tr>
<td>9</td>
<td>example</td>
<td>421</td>
<td>315</td>
<td>106</td>
</tr>
<tr>
<td>10</td>
<td>effect</td>
<td>393</td>
<td>235</td>
<td>158</td>
</tr>
<tr>
<td>11</td>
<td>point</td>
<td>393</td>
<td>346</td>
<td>47</td>
</tr>
<tr>
<td>12</td>
<td>reason</td>
<td>321</td>
<td>217</td>
<td>104</td>
</tr>
<tr>
<td>13</td>
<td>equation</td>
<td>316</td>
<td>248</td>
<td>68</td>
</tr>
<tr>
<td>14</td>
<td>question</td>
<td>313</td>
<td>227</td>
<td>86</td>
</tr>
<tr>
<td>15</td>
<td>process</td>
<td>300</td>
<td>258</td>
<td>42</td>
</tr>
</tbody>
</table>
Frequency by discipline

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Tokens (Norm. Freq)</th>
<th>% of SNs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>1328</td>
<td>5.1%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>1257</td>
<td>4.8%</td>
</tr>
<tr>
<td>Ecology</td>
<td>1477</td>
<td>5.6%</td>
</tr>
<tr>
<td>Engineering</td>
<td>1997</td>
<td>7.6%</td>
</tr>
<tr>
<td>Physics</td>
<td>2000</td>
<td>7.6%</td>
</tr>
<tr>
<td>Natural sciences</td>
<td>8058</td>
<td>30.8%</td>
</tr>
<tr>
<td>Business</td>
<td>2926</td>
<td>11.2%</td>
</tr>
<tr>
<td>Economics</td>
<td>3059</td>
<td>11.7%</td>
</tr>
<tr>
<td>Law</td>
<td>3969</td>
<td>15.8%</td>
</tr>
<tr>
<td>Political science</td>
<td>4143</td>
<td>15.2%</td>
</tr>
<tr>
<td>Sociology</td>
<td>4018</td>
<td>15.4%</td>
</tr>
<tr>
<td>Social sciences</td>
<td>18115</td>
<td>69.2%</td>
</tr>
<tr>
<td>Total</td>
<td>26174</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
### Frequency by genre

<table>
<thead>
<tr>
<th>Genre</th>
<th>SN tokens (Norm. Freq)</th>
<th>% of SNs in genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journals</td>
<td>96979</td>
<td>37.9%</td>
</tr>
<tr>
<td>Textbooks</td>
<td>94239</td>
<td>30.5%</td>
</tr>
<tr>
<td>Lectures</td>
<td>68965</td>
<td>31.6%</td>
</tr>
<tr>
<td>Total</td>
<td>260183</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
SN types that are disproportionately frequent in one genre

<table>
<thead>
<tr>
<th>Genre</th>
<th>Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journals</td>
<td>model</td>
<td>86.3%</td>
</tr>
<tr>
<td></td>
<td>policy</td>
<td>57.3%</td>
</tr>
<tr>
<td></td>
<td>method</td>
<td>57.0%</td>
</tr>
<tr>
<td></td>
<td>result</td>
<td>54.1%</td>
</tr>
<tr>
<td></td>
<td>argument</td>
<td>52.7%</td>
</tr>
<tr>
<td></td>
<td>effect</td>
<td>51.9%</td>
</tr>
<tr>
<td></td>
<td>approach</td>
<td>51.7%</td>
</tr>
<tr>
<td></td>
<td>difference</td>
<td>51.2%</td>
</tr>
<tr>
<td>Textbooks</td>
<td>solution</td>
<td>78.1%</td>
</tr>
<tr>
<td></td>
<td>strategy</td>
<td>60.3%</td>
</tr>
<tr>
<td></td>
<td>equation</td>
<td>50.5%</td>
</tr>
<tr>
<td></td>
<td>theory</td>
<td>50.5%</td>
</tr>
<tr>
<td>Lectures</td>
<td>thing</td>
<td>91.5%</td>
</tr>
<tr>
<td></td>
<td>idea</td>
<td>79.0%</td>
</tr>
<tr>
<td></td>
<td>procedure</td>
<td>65.3%</td>
</tr>
<tr>
<td></td>
<td>point</td>
<td>55.6%</td>
</tr>
<tr>
<td></td>
<td>way</td>
<td>52.2%</td>
</tr>
</tbody>
</table>
Some conclusions

• The approach is text (corpus)-based
• Automatic analysis of the data is combined with manual analysis
• Insights for semantics, syntax, corpus linguistics, and discourse analysis
• Also broader contextual insights for academic disciplines and genres
• (Analysed) data can feed into specific purposes language teaching
Contextual approaches
• Rather than focussing attention on the texts as data, contextual approaches take in the whole situation in which the texts are located as the data

• Methods used to generate the data:
  • Participant and non-participant observation (field notes)
    • (including analysis of texts encountered and created by participants)
  • Reflective/in-depth interview
  • Questionnaire survey
  • Focus group discussion
  • Biographical history
  • Diaries
  • Etc
Multiple perspectives on data

• Drawing on Cohen et al. (2011), Paltridge (2020:29) describes different kinds of multi-perspective research, or approaches to data:
  • methodological perspectives (different data sources and methods)
  • theoretical perspectives (a range of different theories)
  • investigator perspectives (the use of two or more researchers examining the same matter)
  • time perspectives (the collection of data on more than one occasion (longitudinal design))
  • space perspectives (the collection of data on more than one case (cross-sectional design)).
Considerable work using a contextual approach to data on how scholars go about writing for publication

Will look an early study of mine (Flowerdew, 2000)

If time, another study (Lillis and Curry, 2010):
The imperative to publish in English

• The numbers of scholars seeking to publish internationally is growing exponentially:
  • Massification of tertiary education
  • Marketisation of universities
  • Globalisation of scholarship
  • National prestige
  • University league tables
  • Career development closely related to publication ‘output’

• Most scholars writing for publication in English do not have English as their first language
Writing for publication in Hong Kong

• Problem-driven approach
• Research question: how do Cantonese L1 academics go about writing for publication?
• Research data:
  • Writers’ drafts, final texts, and feedback from editors and reviewers;
  • Interviews and email communication with the writers and editors;
  • Reviews of submitted manuscripts
Data sources and methods in a multi-perspective approach on writing for publication
Writing for publication in Hong Kong

- a single case study study of a PhD student struggling with the process of getting published
- it combines data sources as follows:
  - analyses of the student’s drafts and final text
  - interviews and email communication with the student
  - transcripts of email communications between the student and the journal editor, reviewers, and an in-house editor
  - field notes
  - discussion with an in-house editor who helped the student with his writing

Multiple perspectives on the data

• different data sources
• different methods of analysis
• different theories applied to the data (LPP, discourse community)
• different investigator perspectives (researcher and a senior research assistant)
• different time perspectives (the paper is part of a project running over many years) (longitudinal design)
• different space perspectives (included a range of campuses and disciplines; later extended by my PhD student, Yongyan Li, into Mainland China) (cross-sectional design)
Theories and findings

• Theories applied to interpret the difficulties the writer experiences:
  • discourse community (e.g., Swales, 1990)
  • learning as peripheral participation (LPP) (Lave & Wenger, 1991)

• Key findings:
  • As an L2 writer, the participant went through (an ultimately successful) process that would have been a lot easier with a better command of English and non-discursive publishing skills
  • He was fortunate in working with an extremely helpful and understanding journal editor and copy editor
  • Writing for publication is a struggle to gain acceptance in an alien discourse community by means of LPP (or learning by doing)
Writing for publication in Europe

- Lillis and Curry further developed the approach described above on a considerably larger scale – use the term “text history”
- 50 professional scholars writing for publication, in 12 academic institutions, across four different European countries
- Similar multiple-perspectives used:
  - data sources, methods of analysis, theories, investigators, time, and space
Some key findings of Lillis & Curry (2010)

- Academic writing best viewed as a social practice rather than a textual phenomenon
- Academic texts are shaped by multiple parties – networked activity
- Role of literacy brokers and academic brokers in text creation
- Dichotomy of “staying local or going global”
- Problematic issues with the reviewing process
Data-driven learning (DDL)
Data-driven learning (DDL)

• “Data-driven learning is a learning approach (and in particular, an approach to learning foreign languages) in which learning is driven by research-like access to linguistic data.” (Wikipedia)

• “The task of the learner is to "discover" the foreign language, and … the task of the language teacher is to provide a context in which the learner can develop strategies for discovery - strategies through which he or she can "learn how to learn." (Johns, 1991:1)

• This is done through the use of corpora – on-line data-bases of language data
Frequency and concordancing data

- Students interact with the corpus through the use of software that provides frequency and concordancing data.
- Frequency data are lists of the words occurring in the corpus either in order of frequency or alphabetically.
- Concordancing data show how a given word or phrase is used in context.
- In this form of learning, learners are seen as “language detectives” (Johns, 1997:101).
- They try to solve linguistic problems by accessing the corpus data.
What can the concordancer tell you?

- **Word Level:**
  - What are the most frequent words in research papers in my discipline?
  - Which words can I use to boost or down-tone the strength of a point I want to make?
- **Phrase Level:**
  - What typical phrases does a given word appear in?
  - Can I use a certain phraseology I am not sure about, e.g. ‘discuss about’/’emphasise on’?
- **Section Level:**
  - What tenses should I use in the introduction/methods/results/discussion section of an RA?
  - What verbs can I use to report previous research?
- **Discourse Level:**
  - What sentence connectors are most frequently used in RAs in my discipline?
  - Which words follow *This* to refer back to the previous sentence/paragraph (SNs)?
Some advantages of DDL

• DDL can help writers:
  • Both compose and revise
  • Explore disciplinary discourse and genre variability
  • Acquire a resource for future independent use
• DDL can help teachers:
  • Individualise academic writing instruction
  • Foster student autonomy
• On-line free corpus software is readily available
• On-line corpora are free and easily accessed
• DIY corpora are free, relatively quick to build and easy to consult

(Charles 2012, 2014; Quinn, 2014)
• DDL not a research procedure in the sense described in previous sections above
• It’s a learning procedure
• But, on the other hand, learners of Academic English become researchers/discourse analysts themselves

Conclusion

• Have highlighted different approaches to data in analysis of academic discourse
• Contrast between the text-focussed and more socially-focussed approaches
  • language as “an object of study” vs. ‘language as a practice” (Bourdieu, 1980/90, 31)
• Findings resulting from the data in both approaches can feed into English for Academic Purposes (EAP) pedagogy
• Approach to data cannot be divorced from other dimensions of the research process:
  • research question, method, analysis, interpretation, and theory
• They are all inter-related
• We might posit a recursive cycle as follows:
  • research question – (theory) - data – method – analysis - interpretation – (theory)
Danke schön
References

Some books on Academic Discourse

- Bourdieu et al. (1994)

- Flowerdew (2001)

- Hyland (2009)
In AD, key constructs that characterize Bourdieu's work, such as habitus, field, and cultural and linguistic capital, are not yet explicit, though an attentive reader can see them lurking just around the corner (e.g., the notion of linguistic capital is foreshadowed in the words 'he [the student] pays with words, because only words pay' (p 21)).
Flowerdew (2001)

• A collection of specially commissioned articles on the theme of academic discourse
• Situated in the context of the spread of academic English worldwide:
  • “English is now well established as the world language of research and publication … and more and more universities and institutes of learning are using English as the language of instruction (p.1).”
• Research in academic discourse has an applied goal, the teaching of academic English:
  • “The study of academic discourse in English has throughout its short history been closely tied in with developments in the preparation of non-native speakers (NNSs) for study in English, English for Academic Purposes (EAP) (p.1).”
Flowerdew and Forest (2015)

• emphasis on textual function
• Another way of referring to the relation between the SN and its lexical realization are the notions of “encapsulation” and “prospection”, which the authors introduce as discourse-based criteria for SN membership. Where prospection opens up a cataphoric relationship, encapsulation is anaphoric but “not necessarily ‘backward pointing’” (p. 48) meaning that the SN takes up some prior text elements and carries their meaning forward, which is an important textual function.
SNs disproportionately frequent in one disciplinary area

• **Natural sciences**: *equation* (87.6%), *solution* (87.0%), *method* (73.8%), *problem* (52.9%), *model* (52.2%), *result* (51.9%), *thing* (50.3%)

• **Social sciences**: *right* (100.0%), *policy* (99.1%), *issue* (96.6%), *theory* (95.1%), *argument* (93.3%)
Hyland (2009)

• “My aim is to provide an overview of the key genres of the academy and on the way to say something about the nature of knowledge, of communication, and of the practices of those who work and study in universities” (p. viii)
• Emphasises the global nature of contemporary academic discourse
• Contextualised within the global massification of higher education
• Academic careers tied to publication and academic discourses are concomitantly taking on a growing impact

Some books on Academic Discourse

• Bourdieu et al. (1994)

• Flowerdew (2001)

• Hyland (2009)
  • https://www.goodreads.com/book/show/7064929-academic-discourse
The multi-perspective approach

• “multi-perspective research; that is, research which incorporates multiple perspectives into its study design, data collection, and analysis.” (Paltridge, 2020:29)

• “A further study which combines data sources is Flowerdew’s (2000) study of a PhD student struggling with the process of getting published. The data sources for his study included analyses of the student’s draft and final texts; interviews and email communication with the student; communications between the student and the journal editor, reviewers, and in-house editor; field notes; and discussion with an editor who helped the student with his writing.”
Bourdieu et al. (1994) (1964)


- Like most work on academic discourse, the context is the university and students, teachers, and researchers.
- A series of surveys conducted with university students in France in 1963-64.
- Tested the students knowledge of academic terms such as *epistemology*, *methodology*, and *phenomenon*.
- Students had difficulty defining these words out of context.
Different Units of Analysis

• single words (e.g. Coxhead, 2000)
• multi-word units (collocations, lexical bundles, n-grams) (e.g. Hyland, 2008; Simpson-Vlach & Ellis, 2010)
  • ‘extended collocations that appear more frequently than expected by chance’ (Hyland, 2012:abstract)
• rhetorical/generic moves (Biber & Connor, 2007; L. Flowerdew, 2008; Flowerdew and Forest, 2009)
Some frequency results

- Overall SN frequency:
- Most frequent SNs
- SNs by discipline
- SNs across natural and social sciences
- Signalling nouns across genres
"Academic capital is in fact the guaranteed product of the combined effects of cultural transmission by the family and cultural transmission by the school (the efficiency of which depends on the amount of cultural capital directly inherited from the family)"

• “Academic capital is in fact the guaranteed product of the combined effects of cultural transmission by the family and cultural transmission by the school (the efficiency of which depends on the amount of cultural capital directly inherited from the family)”
