Representation in an academic corpus

Divisions and revisions
Richard Forsyth, CELTE

- R.S.Forsyth@warwick.ac.uk
- 07949-451290
- 024-7657 5729
- BAWE team
Hilary Nesi, CELTE

- H.J.Nesi@warwick.ac.uk
- 07765-410300
- 024-7657 5729
- Bawe team
First, some quotations:

"A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language." (Biber et al., 1998: 246.)

"A corpus is a body of text assembled according to explicit design criteria". (Atkins et al., 1992: 5.)
How to start compiling a corpus?

- Look backwards
- Monkey see; monkey do...
Brown Corpus:

"The selection of material to be included followed a two-phase procedure: an initial subjective classification and decision as to how many samples of each category would be used, followed by random selection of the actual samples within each category." (Francis & Kucera, 1982: 5)
2 phases:

- First define categories & target numbers in each category
- Then perform random selection, within categories
Defining our categories

We are interested in:

- the similarities and differences between *genres* of assessed writing produced in different *disciplines*
- the similarities and differences between *genres* of assessed writing produced at different *stages* of university study
Phase 1: Disciplines and domains

- The tree of knowledge?

- Fourfold split:
  - Arts & Humanities
  - Life Sciences
  - Physical Sciences & Engineering
  - Social Sciences
“before Linnaeus, systems of classification were often highly whimsical. Animals might be categorized by whether they were wild or domesticated, terrestrial or aquatic, large or small, or even whether they were thought handsome or noble or of no consequence. Buffon arranged animals by their utility to man. Anatomical considerations barely came into it. Linnaeus made it his life’s work to rectify this deficiency by classifying all that was alive according to its physical attributes. Taxonomy -- which is to say the science of classification -- has never looked back.” (Bryson, 2004: 434.)
“Discipline, however, is not a neat category” (Becher 1990, 335).

- The module as access unit
  - CS231 Human Computer Interaction
  - EC221 Mathematical Economics 1B
  - MA235 Introduction to Mathematical Biology
  - PS351 Psychology & the Law
  - PX308 Physics in Medicine

- (the field awaits its Linnaeus....)
Phase 2 : Selection/Sampling

- Random sampling?
  - 2 views on this
"Random sampling simply ensured that, within the stated guidelines, the selection of individual texts was free of the conscious or unconscious influence of personal taste or preference." (Hofland & Johansson, 1982: 3.)
A Selection sayings on sampling:

- "Unfortunately, the standard approaches to statistical sampling are hardly applicable to building a language corpus." (Atkins et al., 1992: 4.)

- "For language studies, however, proportional samples are rarely useful." (Biber et al., 1998: 247.)
Some sampling schemes

- Random sampling
- Stratified sampling
- Cluster sampling
- Quota sampling
- Opportunistic sampling
- Judgemental sampling
Sampling units: problems of access

- Population = assignments (scripts)
  - students
  - modules
  - departments
  - disciplines
What sort of sampling scheme have we chosen?

- Closest to cluster sampling
  - but with Judgemental/Opportunistic intrusions

- (not "strata", as clusters don't jointly cover population)
The 4-by-4 matrix

- Four years of study (undergraduate and taught postgraduate)
- Four broad disciplinary groupings (life sciences, physical sciences, social sciences, humanities)
The sampling grid (= 3072):

<table>
<thead>
<tr>
<th></th>
<th>AH</th>
<th>SS</th>
<th>PS</th>
<th>LS</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>64 X 3</td>
<td>64 X 3</td>
<td>64 X 3</td>
<td>64 X 3</td>
</tr>
<tr>
<td>2nd</td>
<td>64 X 3</td>
<td>64 X 3</td>
<td>64 X 3</td>
<td>64 X 3</td>
</tr>
<tr>
<td>3rd</td>
<td>64 X 3</td>
<td>64 X 3</td>
<td>64 X 3</td>
<td>64 X 3</td>
</tr>
<tr>
<td>4th</td>
<td>64 X 3</td>
<td>64 X 3</td>
<td>64 X 3</td>
<td>64 X 3</td>
</tr>
</tbody>
</table>
The departmental grid

<table>
<thead>
<tr>
<th>university discipline</th>
<th>Brookes</th>
<th>Reading</th>
<th>Warwick</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH</td>
<td>• English</td>
<td>• Archaeology</td>
<td>• History</td>
</tr>
<tr>
<td></td>
<td>• History of Art</td>
<td>• Classics</td>
<td>• Theatre Studies</td>
</tr>
<tr>
<td>SS</td>
<td>• Health &amp; Social Care</td>
<td>• Agriculture</td>
<td>• Biological Sciences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Food Science</td>
<td>• Medicine</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Zoology</td>
<td></td>
</tr>
<tr>
<td>PS</td>
<td>• Computing</td>
<td>• Meteorology</td>
<td>• Engineering</td>
</tr>
<tr>
<td></td>
<td>• Planning</td>
<td>• Cybernetics</td>
<td>• Physics</td>
</tr>
<tr>
<td>LS</td>
<td>• Anthropology</td>
<td>• Business</td>
<td>• Law</td>
</tr>
<tr>
<td></td>
<td>• Hospitality</td>
<td></td>
<td>• Sociology</td>
</tr>
<tr>
<td></td>
<td>• Tourism</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>