Corpus linguistics as a research method

18\textsuperscript{th} Seminar on 12\textsuperscript{th} April 2016
Institute for Research in Digital Culture and Humanities
Open University of Hong Kong

Althea Ha
Centre for Applied English Studies
The University of Hong Kong
Outline

- What a corpus is
- Why we use corpora in linguistic research
- Different types of corpora
- Considerations when using/building a corpus
- Text analytical tools
- A corpus-based lexical study - Academic Word List (Coxhead, 2000)
- What corpus linguistics is
What a corpus is


- Text is “natural language used for communication, whether it is realised in speech or writing” (Biber & Conrad, 2009, p. 5).
Why corpora in linguistic research

- different aspects of linguistics such as lexis, registers, and genres
  - discipline-specific vocabulary
  - written vs. spoken
  - academic writing
- new constructions in English
  - new words / phrases
  - new senses attached to existing words
Why corpora in linguistic research

But how?

- intuition?
- existing prescriptive rules?
- a number of authentic texts?
Why corpora in linguistic research

- A corpus is “a more reliable guide to language use than native speaker intuition is” (Hunston, 2002, p. 20).

- “It is hard to imagine any area of vocabulary research into acquisition, processing, pedagogy, or assessment where the insights available from corpus analysis would not be valuable” (Schmitt, 2010, p.307).
Why corpora in linguistic research

“Corpora and EAP are perfect companions...corpora bring evidence of typical patterning and salient features to the study of academic discourse, providing data which represent a speaker’s experience of language in a restricted domain” (Hyland, 2009, p. 317).

OUHK - RIDCH 18th Seminar (April 2016) - Corpus linguistics as a research method
Why corpora in linguistic research

- corpora electronically processed in text analytical tools
- useful statistical information such as number of word types, frequency, co-occurrences
- comparison between corpora
Different types of corpora

- general English
- spoken English
- national varieties of English
- academic English
- languages other than English
  - parallel
  - monolingual

(Schmitt, 2010)
Major corpora available to date

- British National Corpus (BNC) (100 million tokens)
- Collins WordBanks Online (WordBanks) (553 million tokens)
- Longman Corpus Network (330 million tokens)
- Cambridge English Corpus (1.5 billion tokens)
  - Cambridge and Nottingham Corpus of Discourse in English (CANCODE)
  - Cambridge and Nottingham Spoken Business English (CANBEC)
- Cambridge Corpus of Business English
- Cambridge Corpus of Financial English
Considerations in using/building a corpus

Matching corpus data with the research purpose is the most crucial consideration in the design of a corpus (Koester, 2009; McEnery & Hardie, 2012).
Considerations in using/building a corpus

- Purpose of your research
- Criteria of the target corpus
- Survey of existing corpora
- Existing/self-built corpus
  - Categories of texts in the existing corpus
  - Structure and size of the self-built corpus
Sinclair’s (2004) criteria

- Mode
- Type
- Domain
- Language(s)/Language varieties
- Location
- Date
What corpus to use

- survey of existing corpora
  - accessible?
  - free / at a cost?
  - categorisation of texts
  - read descriptions and guidelines
  - built-in tools in the website
Sinclair’s (2004) steps towards a representative corpus

- structural criteria → principal corpus components;
- available text types for each component
- rank the text types in terms of importance
- estimate the size of the corpus
  - number of text types
  - number of texts
  - total number of words
- compare the self-built corpus with the original planned corpus
Caveat

- no perfect corpus
  - even though it is huge and carefully designed
  - never have exactly the same characteristics as the language itself
- as representative as possible
  (Sinclair, 2004)
Text analytical tools

- Existing web-based corpora
  - built-in tools in the website
    - search concordances of a word
    - search collocates of a word
    - search and compare the frequency of words and phrases in different genres

- Self-built corpora
  - WordSmith Tools (Scott, 2012)
  - Range (Heatley, Nation, & Coxhead, 2002)
Text analytical tools

- Web-based mega corpus
  - Wordbanks at [http://www.collins.co.uk/page/Wordbanks+Online](http://www.collins.co.uk/page/Wordbanks+Online)

- Web-based text analytical tool

- Desktop application
  - WordSmith Tools 6.0 (Scott, 2012)
inside his stomach was knotting. "The stuff you did on chimps, decoding the body of
addressed only at a regional level, by showing, in the course of reading key texts in
device for communication, Saussure did not dwell on the political implications of his
prevented best in Derrida's essay. The supplement of copula philosophy before
saw that Saussure had decided to exclude 'the thing itself, the reality' from his
by professionals in psychopathology, biology, behavioural paediatrics, cognition, research would be significantly advanced if greater attention were given to structural linguistics.
the early 1970s. The originator was John Grinder, then an Associate Professor of
promiscuously consort with a range of academic disciplines from econometrics to structural linguistics
that he would willingly switch from ology to study linguistics and the science of
before. And since meeting owls from other species, I've become quite interested in
likely assumptions and goals in order to appreciate the purpose of this type of
'). Volume 1 of Political Economy of Human Rights. (Nottingham: Spokesman, 1979). Linguistics: At the root of Chomsky's innovative impact on linguistics lies his philosophy of
in left-libertarian Jewish politics that Chomsky met Zellig Harris, a teacher of
and two daughters. Chomsky's entire working life has been as a scholar in linguistics
field in the late forties it was a small and specialized one, the high profile that
been largely due to him. In the fifties and sixties, Chomsky became renowned within
in other related fields as the founder and foremost exponent of a new approach to linguistics
promoting the war. Since that time, Chomsky has written many books and articles about linguistics.
great demand as a speaker, travelling widely and often talking - separately - about linguistics
the world. Pressure of time has meant that he has been unable to devote himself to linguistics
analyses of the extent he would have wished. It isn't just Chomsky's involvement in linguistics
makes this extraordinary man tick. In this book I shall look separately at Chomsky's
themselves (see the guide to further reading at the end of the book). We begin with
the basic questions which Chomsky is concerned with in this side of his work. What
in this side of his work. What linguistics isn't Chomsky's basic aim is to make linguistics
order to understand what this means, we shall need to look carefully at the terms
of primates - that's decent, solid, focused research. It's why we bought you. But
anthropology, philosophy, psycholinguistics, politics and literature, that the pretensions
focusing rather on the second consequence of his formulation of arbitrariness.
first published in 1971. In this essay, Derrida uses Benveniste's critique of
But he thought Saussure had been naive: When he spoke of the difference between
(Miami, Miami University Press). Bernstein, Richard J. (ed.) (1985), Habermas and
and sociology, featured over a dozen papers on sibling issues. However, the
that laterization, the separating out of the distinct functions and relations
which is concerned with the forms (processes) of language. For Foerster, awareness
at the University of California, and Richard Bandler, a mathematician, computing
has lent media studies an enormous vitality, but at the expense of the subject
if a place could be found for him under the ancient Language Owl. 'One would
successor was chosen at the pejorative meeting of the council.' But if first struck me as a worthwhile intellectual challenge when I was an undergraduate.
I also felt that an account of Chomsky's political work was needed, especially
At the root of Chomsky's innovative impact on linguistics lies his philosophy of
lie his philosophy of science and his epistemological approach to behaviour, both
under whom he studied at the University of Pennsylvania. After a short spell at
When he entered the field in the late forties it was a small and specialized one
has achieved since then has been largely due to him. In the fifties and sixties
and in other related fields as the founder and foremost exponent of a new approach
called Generative Grammar. Chomsky energetically criticized some of the main assumptions
, and many about politics. He is in great demand as a speaker, travelling widely
and politics to two different audiences. Typically, he will speak to a prestigious
to the extent he would have wished. It isn't just Chomsky's involvement in linguistics
and politics which has made people sit up and think: the quality of his work in
and his politics, and then examine what links there may be between the two. The
the rest of this chapter looks at the basic questions which Chomsky is concerned
isn't Chomsky's basic aim is to make linguistics a science. He has consistently
a science. He has consistently held to this aim throughout his work, and it lies
and science. Even before we do this, though, one thing is clear. Chomsky didn't
<table>
<thead>
<tr>
<th>word</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>linguistics</td>
<td>123</td>
</tr>
</tbody>
</table>

Showing page 1
the last page
Corpus linguistics as a research method

... field. She smiled and poured some roofing tar into a polystyrene cup. It was and her growing interest in ancient history. A detour was made to find onceform at the new university in Dorpat, and by her mid-twenties she had gathered to her and all the philologists from the Pole to Constantinople would speak her name with one foot in science and the other in art, examining the intimate relationship his academic specialty. Tolkien had derived Gondor fictively from an older ancient geography, chess - in the course of a single evening. "Avalo was sweet-faced and at the ancient and esteemed University of Heidelberg. And may I ask what you're experts found anomalies in the language used. "The person who wrote the inscription University College London. In the first patient, a 19-year-old woman who received expert from the Salk Institute in California, US. "It tells us about the way language circles, the loss of a language is mourned because it deprives us of alternative at Manchester University remarked: 'Every language is the repository of the culture. The Swiss philologist had remarked that all meaning was not inherent in terms (still experts which allegedly cast doubt on the evidence of four detectives who went opinion polls, experts, politicians and spin doctors ground away on the airwaves professor visited him after hearing him was the definitive expert. "Academics ca at school. At University College, Oxford, where he studied classics. Daniel was expert at their tribunal hearings, in a bid to help establish the authorship of at the University of Washington, but this week he and his wife are homesteading from the University of Pennsylvania.1957. His book Syntactic Structures revolutionises begins teaching at MIT 1964. Active against the Vietnam War, including organis and philosophy at the Massachusetts Institute of Technology, and the book is 9-11 and literature at Hebrew University in Jerusalem. She was a pacifist, committed professor maintains that while our evolving vocabulary betrays a ontological shift at the University of Pennsylvania. "I'm a dual citizen and chameleons," he says, and child-development research. Psychopathic H� showed the world's image of artificial terms, it is the third side of what is known as the semantic triangle. There's andsmartening up deficient. "That's enough," a tart little voice behind them said has to offer. globecareers@globeemail.ca Employees: Absenteeism: Surveys professor at the University of Toronto who is gay. "That's the seeds of language philosophy, law and architecture. Derrida focused his work on language, showing
## Collocation candidates

<table>
<thead>
<tr>
<th>Attribute</th>
<th>In the range from:</th>
<th>Minimum frequency in corpus</th>
<th>Minimum frequency in given range</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>-1</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

**Show functions:**
- T-score
- MI
- MI3
- log likelihood
- min. sensitivity
- salience

**Sort by:**
- T-score
- MI
- MI3
- log likelihood
- min. sensitivity
- salience

[Corpus: WordBanks Online: English conc description]
Collocations

<table>
<thead>
<tr>
<th></th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>p/n classificatory</td>
<td>16.650</td>
</tr>
<tr>
<td>p/n Neuro</td>
<td>16.387</td>
</tr>
<tr>
<td>p/n Computational</td>
<td>15.942</td>
</tr>
<tr>
<td>p/n Anthropological</td>
<td>15.491</td>
</tr>
<tr>
<td>p/n computational</td>
<td>14.006</td>
</tr>
<tr>
<td>p/n Applied</td>
<td>13.726</td>
</tr>
<tr>
<td>p/n anthropological</td>
<td>13.199</td>
</tr>
<tr>
<td>p/n Retired</td>
<td>11.202</td>
</tr>
<tr>
<td>p/n comparative</td>
<td>10.615</td>
</tr>
<tr>
<td>p/n professor</td>
<td>10.340</td>
</tr>
<tr>
<td>p/n structural</td>
<td>10.202</td>
</tr>
<tr>
<td>p/n applied</td>
<td>9.154</td>
</tr>
<tr>
<td>p/n expert</td>
<td>9.146</td>
</tr>
</tbody>
</table>
Corpora have emerged as indispensable sources of language data for inquiry in different types of linguistic features and patterns. The main goal of this talk is to introduce corpus linguistics as a research approach. The talk will first explain what a corpus is and why we use corpora in linguistic research, followed by an overview of different types of corpora. Considerations in obtaining a suitable corpus when undertaking a corpus study will then be discussed. Data in a corpus need to be analysed and retrieved before serving as empirical evidence in research results, and to this end a few text analytical tools for organising and analysing corpus data will be shown. The talk will be concluded by using a prominent corpus-based lexical study that created the widely-used Academic Word List (Coxhead, 2000) to exemplify a corpus study.
<table>
<thead>
<tr>
<th>Category</th>
<th>Families</th>
<th>Types</th>
<th>Tokens</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1 Words (1-1000)</td>
<td>42</td>
<td>47</td>
<td>90</td>
<td>64.75%</td>
</tr>
<tr>
<td>Function:</td>
<td></td>
<td></td>
<td>(50)</td>
<td>(42.45%)</td>
</tr>
<tr>
<td>Content:</td>
<td></td>
<td></td>
<td>(31)</td>
<td>(22.30%)</td>
</tr>
<tr>
<td>&gt; Anglo-Sax</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>~Not Greco-Lat/Fr Cogn.</td>
<td></td>
<td></td>
<td>(12)</td>
<td>(8.63%)</td>
</tr>
<tr>
<td>K2 Words (1001-2000)</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4.32%</td>
</tr>
<tr>
<td>&gt; Anglo-Sax</td>
<td></td>
<td></td>
<td>(2)</td>
<td>(1.44%)</td>
</tr>
<tr>
<td>1k+2k</td>
<td></td>
<td></td>
<td></td>
<td>(60.07%)</td>
</tr>
<tr>
<td>AWL Words (academic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; Anglo-Sax</td>
<td></td>
<td></td>
<td>(3)</td>
<td>(2.16%)</td>
</tr>
<tr>
<td>Off-List Words:</td>
<td>2</td>
<td>11</td>
<td>21</td>
<td>16.11%</td>
</tr>
<tr>
<td></td>
<td>64+?</td>
<td>82</td>
<td>139</td>
<td>100%</td>
</tr>
</tbody>
</table>

- **Words in text (tokens):** 139
- **Different words (types):** 82
- **Type-token ratio:** 0.59
- **Tokens per type:** 1.70
- **Lex density (content words/total):** 0.58
- **Pertaining to onlist only**
  - **Tokens:** 118
  - **Types:** 71
  - **Families:** 64
  - **Tokens per family:** 1.84
  - **Types per family:** 1.11
  - **Anglo-Sax Index:** %
  - **Greco-Lat/Fr-Cognate Index:** (Inverse of above) %
<table>
<thead>
<tr>
<th>N</th>
<th>Keywords</th>
<th>Texts</th>
<th>%</th>
<th>% Texts</th>
<th>% N</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>THIS</td>
<td>28,438</td>
<td>0.38</td>
<td>727</td>
<td>100.00</td>
</tr>
<tr>
<td>18</td>
<td>MILLION</td>
<td>25,724</td>
<td>0.34</td>
<td>674</td>
<td>92.71</td>
</tr>
<tr>
<td>29</td>
<td>SO</td>
<td>24,147</td>
<td>0.32</td>
<td>651</td>
<td>90.92</td>
</tr>
<tr>
<td>30</td>
<td>OTHER</td>
<td>23,199</td>
<td>0.31</td>
<td>700</td>
<td>95.29</td>
</tr>
<tr>
<td>31</td>
<td>WAS</td>
<td>22,590</td>
<td>0.30</td>
<td>721</td>
<td>95.17</td>
</tr>
<tr>
<td>32</td>
<td>BUSINESS</td>
<td>22,181</td>
<td>0.30</td>
<td>705</td>
<td>95.97</td>
</tr>
<tr>
<td>33</td>
<td>YEAR</td>
<td>22,129</td>
<td>0.29</td>
<td>722</td>
<td>95.31</td>
</tr>
<tr>
<td>34</td>
<td>WHICH</td>
<td>21,612</td>
<td>0.28</td>
<td>717</td>
<td>98.62</td>
</tr>
<tr>
<td>35</td>
<td>QUARTER</td>
<td>20,957</td>
<td>0.28</td>
<td>690</td>
<td>94.91</td>
</tr>
<tr>
<td>36</td>
<td>NOT</td>
<td>20,088</td>
<td>0.27</td>
<td>700</td>
<td>95.29</td>
</tr>
<tr>
<td>37</td>
<td>AN</td>
<td>18,807</td>
<td>0.25</td>
<td>720</td>
<td>90.04</td>
</tr>
<tr>
<td>38</td>
<td>FINANCIAL</td>
<td>18,550</td>
<td>0.25</td>
<td>697</td>
<td>95.87</td>
</tr>
<tr>
<td>39</td>
<td>CAPITAL</td>
<td>18,791</td>
<td>0.25</td>
<td>707</td>
<td>97.25</td>
</tr>
<tr>
<td>40</td>
<td>BUT</td>
<td>18,396</td>
<td>0.25</td>
<td>690</td>
<td>94.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text File</th>
<th>File Size</th>
<th>Tokens</th>
<th>Tokens Used for Word List</th>
<th>Sum of</th>
<th>Types (Distinct Tns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>b_AR_UK_HSBC_2012.txt</td>
<td>93,751,267</td>
<td>7,477,331</td>
<td>5,451,265</td>
<td>43,422</td>
<td></td>
</tr>
<tr>
<td>b_AR_US_BofA_2012.txt</td>
<td>2,579,026</td>
<td>208,677</td>
<td>167,931</td>
<td>5,253</td>
<td></td>
</tr>
<tr>
<td>b_AR_US_Citi_2012.txt</td>
<td>2,315,170</td>
<td>194,571</td>
<td>166,346</td>
<td>6,156</td>
<td></td>
</tr>
<tr>
<td>b_AR_US_SMP_2012.txt</td>
<td>2,071,704</td>
<td>198,995</td>
<td>174,518</td>
<td>6,991</td>
<td></td>
</tr>
<tr>
<td>b_AR_US_WellsFargo_2012.txt</td>
<td>1,932,562</td>
<td>150,805</td>
<td>127,815</td>
<td>5,669</td>
<td></td>
</tr>
<tr>
<td>f_AR_IN_HDFC_2012.txt</td>
<td>961,276</td>
<td>78,377</td>
<td>61,393</td>
<td>6,643</td>
<td></td>
</tr>
<tr>
<td>f_AR_IN_AXIS_2012.txt</td>
<td>843,448</td>
<td>62,120</td>
<td>55,751</td>
<td>3,538</td>
<td></td>
</tr>
<tr>
<td>f_AR_US_BofA_2012.txt</td>
<td>1,366,294</td>
<td>96,706</td>
<td>85,334</td>
<td>4,102</td>
<td></td>
</tr>
<tr>
<td>f_AR_US_CapitalOne_2012.txt</td>
<td>2,214,986</td>
<td>152,416</td>
<td>129,901</td>
<td>4,888</td>
<td></td>
</tr>
<tr>
<td>f_AR_US_GS_2012.txt</td>
<td>1,558,448</td>
<td>110,079</td>
<td>104,471</td>
<td>7,903</td>
<td></td>
</tr>
<tr>
<td>f_AR_US_MS_2012.txt</td>
<td>2,230,476</td>
<td>199,888</td>
<td>132,859</td>
<td>4,836</td>
<td></td>
</tr>
<tr>
<td>f_AR_US_NVMelexon_2012.txt</td>
<td>1,585,628</td>
<td>121,497</td>
<td>106,045</td>
<td>4,926</td>
<td></td>
</tr>
<tr>
<td>f_AR_US_StateStreet_2012.txt</td>
<td>1,478,850</td>
<td>111,193</td>
<td>98,920</td>
<td>4,430</td>
<td></td>
</tr>
<tr>
<td>f_AR_US_Visa_2012.txt</td>
<td>963,354</td>
<td>69,915</td>
<td>64,278</td>
<td>4,133</td>
<td></td>
</tr>
<tr>
<td>f_AR_CH_Zurich_2012.txt</td>
<td>1,879,076</td>
<td>145,998</td>
<td>122,696</td>
<td>5,145</td>
<td></td>
</tr>
<tr>
<td>f_AR_HK_Alia_2012.txt</td>
<td>1,290,310</td>
<td>98,201</td>
<td>87,184</td>
<td>4,500</td>
<td></td>
</tr>
<tr>
<td>f_AR_HK_BING_2012.txt</td>
<td>3,215,406</td>
<td>254,976</td>
<td>228,106</td>
<td>7,714</td>
<td></td>
</tr>
<tr>
<td>f_AR_HK_Insurance_2012.txt</td>
<td>2,772,792</td>
<td>212,542</td>
<td>183,105</td>
<td>6,004</td>
<td></td>
</tr>
<tr>
<td>f_AR_HK_HSBC_2012.txt</td>
<td>2,933,530</td>
<td>191,143</td>
<td>169,690</td>
<td>5,974</td>
<td></td>
</tr>
<tr>
<td>f_AR_HK_Berkshire_2012.txt</td>
<td>975,772</td>
<td>65,630</td>
<td>57,322</td>
<td>5,005</td>
<td></td>
</tr>
<tr>
<td>r_AR_CA_BofA_2012.txt</td>
<td>795,435</td>
<td>78,763</td>
<td>64,185</td>
<td>3,742</td>
<td></td>
</tr>
<tr>
<td>r_AR_CA_Broadfield_2012.txt</td>
<td>1,128,466</td>
<td>86,350</td>
<td>75,692</td>
<td>4,402</td>
<td></td>
</tr>
<tr>
<td>r_AR_CA_Uddin_2012.txt</td>
<td>1,999,752</td>
<td>154,716</td>
<td>126,472</td>
<td>6,994</td>
<td></td>
</tr>
<tr>
<td>r_AR_HK_CheungKong_2012.txt</td>
<td>817,182</td>
<td>62,891</td>
<td>55,406</td>
<td>4,512</td>
<td></td>
</tr>
<tr>
<td>r_AR_HK_SHK_2012.txt</td>
<td>900,354</td>
<td>69,588</td>
<td>61,630</td>
<td>4,570</td>
<td></td>
</tr>
</tbody>
</table>
Academic Word List (Coxhead, 2000)

- Research purpose
  - to develop and evaluate a new academic word list
- Factors considered in building the Academic Corpus
  - Representation
  - Organization
  - Size
  - Word selection
Academic Word List (Coxhead, 2000)

**Representation**
- not only textbooks, but also a range of academic texts
  - 158 journal articles (print)
  - 51 edited journal articles (online)
  - 43 complete university textbooks or course books
  - 42 texts from the Learned and Scientific section of the Wellington Corpus of Written English (Bauer, 1993)
  - etc.
Academic Word List (Coxhead, 2000)

- Organization
  - 4 disciplines
    - arts, commerce, law, science
  - 28 subject areas
Academic Word List (Coxhead, 2000)

- **Size**
  - 3.5 million running words
  - so as to identify 100 occurrences of a word family
  - Coxhead referred to the data from Brown Corpus (Francis & Kucera, 1982)
### Composition of the Academic Corpus

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Arts</th>
<th>Commerce</th>
<th>Law</th>
<th>Science</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running words</td>
<td>883,214</td>
<td>879,547</td>
<td>874,723</td>
<td>875,846</td>
<td>351,333</td>
</tr>
<tr>
<td>Texts</td>
<td>122</td>
<td>107</td>
<td>72</td>
<td>113</td>
<td>414</td>
</tr>
<tr>
<td>Subject areas</td>
<td>Education</td>
<td>Accounting</td>
<td>Constitutional</td>
<td>Biology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>History</td>
<td>Economics</td>
<td>Criminal</td>
<td>Chemistry</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linguistics</td>
<td>Finance</td>
<td>Family and medicolegal</td>
<td>Computer science</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Philosophy</td>
<td>Industrial relations</td>
<td>International</td>
<td>Geography</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Politics</td>
<td>Management</td>
<td>Pure commercial</td>
<td>Geology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Psychology</td>
<td>Marketing</td>
<td>Quasi-commercial</td>
<td>Mathematics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sociology</td>
<td>Public policy</td>
<td>Rights and remedies</td>
<td>Physics</td>
<td></td>
</tr>
</tbody>
</table>
Academic Word List (Coxhead, 2000)

- Word selection
  - What a word is
  - morphologically different words (e.g. -s and -ed)
  - word types → word families
  - “[a] word family was defined as a stem plus all closely related affixed forms...” (Coxhead, 2000, p. 218)
Taking “analyse” as an example

- regular inflections
  - analysed, analysing, analyses
- derivations
  - analyser, analysers, analysis, analyst, analysts, analytic, analytical, analytically
- American spelling
  - analyze, analyzed, analyzes, analyzing
Academic Word List (Coxhead, 2000)

- Methods
  - Range (Heatley & Nation, 1996)
- Criteria for a member of a word family
  - Specialised Occurrence
    - excluding 2000 most frequent words
  - Range
    - occurs at least 10 times in each discipline
    - occurs in 15 or more subject areas (out of 28)
  - Frequency
    - occurs at least 100 times in the Academic Corpus
Academic Word List (Coxhead, 2000)

Results

- 570 word families
- 12% word coverage for commerce
- 9.3% word coverage for arts
- 9.4% word coverage for law
- 9.1% word coverage for science
- Average 10% word coverage for academic texts
Corpus linguistics

“Corpus linguistics is a research approach that has developed over the past several decades to support empirical investigations of language variation and use, resulting in research findings that have much greater generalizability and validity than would otherwise be feasible” (Biber, Reppen, & Friginal, 2010, p. 548).
Corpus linguistics

Corpus linguistics involves “dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions” (McEnery & Hardie, 2012, p. 1).

a methodological approach rather than a model of language (Biber, Conrad, & Reppen, 1998, p. 4)
Corpus linguistics

- it is *empirical*, analyzing the actual patterns of use in natural texts;
- it utilizes a large and principled collection of natural texts, known as *a corpus* as the basis for analysis;
- it makes extensive use of *computers* for analysis, employing both automatic and interactive techniques;
- it depends on *both quantitative and qualitative* analytical techniques.

(Biber, Conrad, & Reppen, 1998, p. 4)
List of References


Thank you for your attention.