Corpus usage

Beáta Megyesi

Uppsala University
Department of Linguistics and Philology
beata.megyesi@lingfil.uu.se
Last time

- Empirical language studies
- Corpus linguistics
- Corpora: definition
- Some corpora
- Corpus archives and distribution
Content today

- Using available corpora
- Corpus types
- Corpus representativeness
- Corpus sampling
- Corpus usage in language studies:
  - concordances
  - collocations
  - frequency lists
  - keywords
Using available corpora

- A corpus is always designed for a particular purpose (whereas an archive is not)
- Thousands of available corpora; many of them are created for specific research projects.
- Most of them are English but there are many resources for other languages today: for European languages, Arabic, Chinese, Hindi, etc.
- Corpus choice depends on the research question, which plays a crucial role!
General corpora

- A general corpus is balanced with regards to genres and domains that typically represent the language under consideration.
- The corpus may be written or spoken data.
- Examples: British National Corpus, American National Corpus
- Several general corpora for a particular language allow contrastive language studies
Specialized corpora

- A specialized corpus is specialized relative to a general corpus.
  - domain or genre specific
  - designed to represent a sub-language
- Corpus of Professional Spoken American English: academic discussions (e.g. meetings) and White House Press conferences (question-answer sessions)
- valuable resources for investigations in the relevant domains and genres
- can be extracted from general corpora
Written corpora

- The first modern corpus of English: Brown corpus (Brown University Standard Corpus of Present-Day American English) (Kucera and Francis, 1967)
  - 500 chunk of approximately 2000 words of written text sampled from 15 categories in 1961
- Number of corpora that follows the Brown model:
  - The Lancaster-Oslo-Bergen Corpus of British English (LOB) (Johanson, Leech and Goodluck, 1978)
  - Australian Corpus of English,
  - Wellington Corpus of Written New Zealand English,
  - Kolhapur Corpus for Indian English
- Subsamples of these corpora are annotated:
  - Lancaster Parsed Corpus is a parsed corpus, a sub-sample of 133000 words from the LOB
  - Susanne Corpus is parsed, a sub-sample of 128000 words taken from the Brown
Spoken corpora

- Spoken part of general corpora (BNC)
- London-Lund Corpus (LLC): spoken British English from 1960s to 1970s
- Lancaster/IBM Spoken English Corpus (SEC)
- Cambridge and Nottingham Corpus of Discourse in English (CANCODE)
- Santa Barbara Corpus of Spoken American English
Synchronic corpora

- comparing varieties of a language
- International Corpus of English (ICE) designed for the synchronic study of world Englishes
  - collection of 20 corpora of 1 million words each
  - composed of written and spoken English produced after 1989
  - countries/region in which English is the first or major language
- few corpora for regional dialects
- Longman Spoken American Corpus
Diachronic corpora

- Texts from the same language gathered from different time periods
- used to track changes in language evolution
- typically written language
- Helsinki Diachronic Corpus of English Texts
  - 1.5 million words taken from 400 text samples
  - dating from eight to eighteenth centuries
  - wide range of genres and divided into 3 periods (Old, Middle and Early Modern English)
Learner corpora

- Collection of the writing or speech of learners acquiring L2
- Data collected are the L2 production of the learners
  - International Corpus of Learner English (ICLE)
- Developmental corpus: data produced by children acquiring their L1
  - Child Language Data Exchange System (CHILDES)
Monitor corpora

- Sinclair (1991)
- constantly supplemented with material, keeps increasing in size
- the proportion of text types included in the corpus remains constant
- typically larger than sample corpora
- Bank of English: 525 million words at present
- Global English Monitor Corpus: world’s leading newspapers in English expected to grow to billions of words
Corpus representativeness

- A corpus is a sample of a language or language variety (population).
- Representativeness is determined by:
  - balance, i.e. the range of genres in the corpus, and
  - sampling, i.e. how text chunks for each genre are selected
- Biber (‘Representativeness in corpus design’, 1993)
  - define the population (“sampling frame”)
  - decide the hierarchical structure in population (“strata”) - genre, text types, origin
  - decide the optimal size
Corpus representativeness

- How can we know that the sample we are using is representative of the language or language variety?
Corpus representativeness

- We must consider balance and sampling to ensure representativeness.
- "A corpus is thought to be representative of the language variety it is supposed to be represent if the findings based on its contents can be generalized to the said language variety.” (Leech, 1991).
Corpus representativeness

- General corpora serve as basis for an overall description of a language or language variety, e.g. ANC, BNC, SUC and should cover proportionally as many text types as possible and depend on sampling from a broad range of genres.

- Specialized corpora tend to be domain or genre specific but should be balanced by including a wide range of types so that it can be claimed to represent the particular variety of language.
Corpus balance

- Representativeness depends upon how balanced the corpus is, i.e. the range of text categories.
- The acceptable balance is determined by the intended usage of the corpus.
- A balanced corpus (general or specialized) covers a wide range of text categories.
- Corpus builders adopt an existing corpus model when building their own corpus.
- Example: BNC serves as example for ANC, SUC, Korean National Corpus, Polish National Corpus.
BNC

- 100 million words: 90 % written, 10 % transcripts of spoken data
- written texts criteria:
  - domain: content type of the text
  - time: period of text production
  - medium: type of text publication (books, periodicals, unpublished manuscripts)
- spoken data criteria:
  - demographic: age group, sex, social class, geographical region
  - context-governed: meetings, lectures, radio broadcasts
## Composition of the written BNC

<table>
<thead>
<tr>
<th>DOMAIN</th>
<th>%</th>
<th>TIME</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imaginative</td>
<td>21.91</td>
<td>1960-74</td>
<td>2.26</td>
</tr>
<tr>
<td>Arts</td>
<td>8.08</td>
<td>1975-93</td>
<td>89.23</td>
</tr>
<tr>
<td>Belief and thought</td>
<td>3.40</td>
<td>Unclassified</td>
<td>8.49</td>
</tr>
<tr>
<td>Commerce/finance</td>
<td>7.93</td>
<td>MEDIUM</td>
<td></td>
</tr>
<tr>
<td>Leisure</td>
<td>11.13</td>
<td>Book</td>
<td>58.58</td>
</tr>
<tr>
<td>Natural/pure science</td>
<td>4.18</td>
<td>Periodical</td>
<td>31.08</td>
</tr>
<tr>
<td>Applied science</td>
<td>8.21</td>
<td>Misc. published</td>
<td>4.38</td>
</tr>
<tr>
<td>Social science</td>
<td>14.80</td>
<td>Misc. unpublished</td>
<td>4.00</td>
</tr>
<tr>
<td>World affairs</td>
<td>18.39</td>
<td>To-be-spoken</td>
<td>1.52</td>
</tr>
<tr>
<td>Unclassified</td>
<td>1.93</td>
<td>Unclassified</td>
<td>0.40</td>
</tr>
</tbody>
</table>
## Composition of the spoken BNC

<table>
<thead>
<tr>
<th>REGION</th>
<th>%</th>
<th>CONTEXT-GOVERNED</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>45.61</td>
<td>Educational</td>
<td>20.56</td>
</tr>
<tr>
<td>Midlands</td>
<td>23.33</td>
<td>Business</td>
<td>21.47</td>
</tr>
<tr>
<td>North</td>
<td>25.43</td>
<td>Institutional</td>
<td>21.86</td>
</tr>
<tr>
<td>Unclassified</td>
<td>5.61</td>
<td>Leisure</td>
<td>23.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unclassified</td>
<td>12.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INTERACTION</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monologue</td>
<td>18.64</td>
</tr>
<tr>
<td>Dialogue</td>
<td>74.87</td>
</tr>
<tr>
<td>Unclassified</td>
<td>6.48</td>
</tr>
</tbody>
</table>
Corpus balance

In selecting texts for inclusion in the corpus, account was taken of both production, by sampling a wide variety of distinct types of material, and reception, by selecting instances of those types which have a wide distribution.

Thus, having chosen to sample such things as popular novels, or technical writings, best-seller lists and library circulation statistics were consulted to select particular examples of them.

(Aston and Burnard’s (1998) summary of the design criteria of the BNC).
Corpus balance

- Balance is important for a static sample corpus but not for a dynamic monitor corpus
- When designing a corpus, think about:
  - balance by including a great variety of language samples
  - document corpus design criteria explicitly
  - make the documentation available to corpus users
- Interpret the results of corpus-based studies with caution and consider whether the corpus data used were appropriate.
Corpus sampling

- A corpus is a sample of a much larger population.
- A sample is representative if what we find for the sample also holds for the general population.
- The sample should reproduce the characteristics of the population.
- Text categories are sampled proportionally in order to achieve representativeness and balance.
Corpus sampling method

- Define the sampling unit and the boundaries of the population (book, newspaper, periodical...)
- Define the sampling frame, i.e. a list of sampling units
- Define the population by demographic distribution (age, sex, social class), receivers, text genres
Sampling methods

- **Sampling method:**
  - Simple random sampling: all sampling units within the sampling frame are numbered and the sample is chosen by use of a table of random numbers.
  - Stratified random sampling: first divide the population into relatively homogeneous groups (strata) and sample each strata random.

- **Sample size:**
  - Full texts/whole documents - useful for linguistics but problems with copyright issues
  - Text chunks (initial, middle, or end)
A corpus is considered representative if what we find on the basis of the corpus also holds for the language it is supposed to represent.

Representativeness is typically achieved by balancing, i.e. covering a wide range of frequent and important text categories that are proportionally sampled from the target population.

Caution: there is no objective way to balance a corpus or to measure its representativeness.

Always check if the corpus is suitable for your research question!
Building your own corpus

▶ How would you proceed to collect material and which problems can you see for each step?
Building your own corpus

▶ Be clear about your research question(s)!
▶ Determine the material you need to collect.
▶ Check that no corpus of such material exist.
▶ Find source of data.
▶ Decide the corpus size.
  ▶ not easy: availability of suitable data affects the corpus size
  ▶ preferably data in machine-readable form - often the main limiting factor in corpus building
  ▶ OCR-scanning: time-consuming, error prone
  ▶ Copyright issues: copyright holders might ignore you, time-consuming, much effort, trouble and frustration, different rules in different countries
Building your own corpus

- Size is not all-important. (Leech, 1991)
- Maybe a small corpus is enough for your research question: it depends on the frequency and distribution of the linguistic features under consideration.
- Corpora for lexical studies are much larger as we are interested in the frequency of the distribution of the word while corpora for grammatical studies are smaller.
- Corpora that need extensive manual annotation are necessarily small (semantic, pragmatic).
- Corpus tools might have a maximal number of extractions.
- The optimum size of the corpus is determined by the research question and practical considerations.
Building your own corpus

- Consider balance and representativeness issues
- Data capture:
  - Machine-readable data is preferable compared to paper-based materials
  - Internet - important source
    - web pages normally use Hypertext Mark-up Language (HTML):
      - convert the web pages into plain text files by removing the HTML tags, or
    - convert the HTML tags into another format (XML)
Building your own corpus

Corpus mark-up: contextual (genre, source) and textual (paragraph and sentence) information

Character encoding: important for multilingual data with different writing systems - Unicode

Corpus annotation: PoS tagging, parsing, semantic annotation
Building your own corpus

“A colleague of mine, Matti Rahkonen at the University of Jyväskylä, once claimed that he could subsume his experiences of large-scale corpus work in two words: ’Never more!’

... I can wholeheartedly agree with Rahkonens description of the corpus-building process, but I am not quite so dejected. Rather, I could join in an old Louis Prima favourite that has given me consolation and encouragement: ’Next time...’” (Källgren, 1996)
Corpus analysis

- Problem with corpus usage is decontextualization; user does not know how the corpus was built but uses statistics, e.g. frequencies.
- Solution to decontextualization: Build your own corpus if you do not need a large one!
Corpus analysis

- Describe your corpus well!
- Base your claims on your corpus!
- Avoid unreasonable generalizations!
Corpus analysis

- Terms: token, type, hapax legomena, lemma
- Concordances
- Collocations
- Frequency list
- Key words
Key terms

Token: sequences of letters separated by spaces or punctuation
  ▶ There are 8 tokens in this sentence.

Type: uniq tokens, counting each repeated item once
  ▶ Malte and Nora like icecream and chocolate cake.
  ▶ 9 tokens och 8 types (Malte, and, Nora, like, icecream, chocolate, cake, ).

Hapax legomena: the word that occur only once
Key terms

**Lemma:** base word form, the canonical form of an inflected word; usually, for verbs: the infinitive or the present tense, first person singular (run-run, runs, ran, running), and for nouns: the nominative singular. Refers to the particular form that is chosen by convention to represent the lexeme.

**Lexeme:** refers to the set of all the forms that have the same meaning.
Concordances

- Search word in context (KWIC)
- An alphabetic list of words that appear in running text.
- Concordancer searches a corpus for a selected word/phrase and presents every instance of that word/phrase in the centre of the screen, with the words in context to the left and right
- Node word is the word in the centre
- Observe the most frequent meanings or collocates or phraseology of an individual word/phrase, meaning distinctions, and meaning and pattern
- Investigate semantic preference (words that often appear together) and discourse prosody
Konkordans

▶ http://sprakbanken.gu.se/konk
Collocations

- a sequence of words or terms which co-occur more often than would be expected by chance
- Specific combinations of words, lexical units, not whole context
- Ex.: strong vs. powerful tea, powerful vs. strong wizard
- Ex.: toy more often co-occur with children than with men or women
- Calculation: take a node word and count the instances of all words occurring within a particular span, e.g. a context window of 9 words, four to left, four to the right and present the frequency of each collocated word
- Words with similar meaning have similar context
A frequency list is a sorted list of words (word types) together with their frequency, where frequency here usually means the number of occurrences in a given corpus.

<table>
<thead>
<tr>
<th>rank</th>
<th>word</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>you</td>
<td>1222421</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>1052546</td>
</tr>
<tr>
<td>3</td>
<td>to</td>
<td>823661</td>
</tr>
<tr>
<td>4</td>
<td>the</td>
<td>770161</td>
</tr>
</tbody>
</table>

The list can be displayed in frequency order, in alphabetical order, or in the order of the first occurrence of the type in the corpus.

Comparing frequency list of two corpora (especially specialized) can give info about the differences between texts concerning the lexical items used.
Word frequencies

- Shows the number of times a particular token appears.
- Several ways:
  - Token – run ran running, house houses.
  - Type – run, house
  - All words, content words, PoS, some forms of a particular lemma, etc.
- Frequencies – the number of times a word occurs in a corpus.
  OK to use if you compare frequencies within one corpus.
## Frequencies: example

<table>
<thead>
<tr>
<th>AE</th>
<th>BA</th>
<th>KL</th>
<th>KR</th>
<th>ALLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>74 år</td>
<td>38 parti</td>
<td>63 gång</td>
<td>75 man</td>
<td>2714 år</td>
</tr>
<tr>
<td>48 match</td>
<td>37 år</td>
<td>61 hand</td>
<td>74 far</td>
<td>1381 tid</td>
</tr>
<tr>
<td>30 tränare</td>
<td>26 fråga</td>
<td>46 väg</td>
<td>70 häst</td>
<td>1234 del</td>
</tr>
<tr>
<td>29 minut</td>
<td>25 socialdemokrat</td>
<td>45 fråga</td>
<td>58 år</td>
<td>1035 gång</td>
</tr>
<tr>
<td>27 dag</td>
<td>25 land</td>
<td>45 bil</td>
<td>57 tid</td>
<td>1031 sätt</td>
</tr>
<tr>
<td>24 seger</td>
<td>24 riksdag</td>
<td>40 fall</td>
<td>50 dag</td>
<td>974 fråga</td>
</tr>
<tr>
<td>24 mål</td>
<td>24 regering</td>
<td>38 statsråd</td>
<td>48 gång</td>
<td>974 fråga</td>
</tr>
<tr>
<td>23 spel</td>
<td>23 dag</td>
<td>37 huvud</td>
<td>45 väg</td>
<td>851 människa</td>
</tr>
<tr>
<td>23 lag</td>
<td>22 väg</td>
<td>37 dörr</td>
<td>41 hand</td>
<td>780 barn</td>
</tr>
<tr>
<td>21 tävling</td>
<td>20 procent</td>
<td>35 rum</td>
<td>39 öga</td>
<td>777 ˘ g</td>
</tr>
<tr>
<td>21 plats</td>
<td>19 stat</td>
<td>32 år</td>
<td>39 sak</td>
<td>754 fall</td>
</tr>
<tr>
<td>20 tid</td>
<td>19 del</td>
<td>32 polis</td>
<td>36 del</td>
<td>748 land</td>
</tr>
<tr>
<td>20 gång</td>
<td>18 politiker</td>
<td>31 tid</td>
<td>35 mor</td>
<td>694 liv</td>
</tr>
</tbody>
</table>
Example

- Texts from SUC.

<table>
<thead>
<tr>
<th>File</th>
<th>AE</th>
<th>BA</th>
<th>KL</th>
<th>KR</th>
<th>ALLA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of words</strong></td>
<td>17092</td>
<td>16297</td>
<td>45186</td>
<td>14320</td>
<td>1166976</td>
</tr>
<tr>
<td><strong>Genre</strong></td>
<td>Sport</td>
<td>Politics</td>
<td>Thriller</td>
<td>Trivia</td>
<td>–</td>
</tr>
</tbody>
</table>

- Noun (excl. names)
- Lemma.
Relative frequencies

- We can compare frequencies between different corpora or genre of different size but normalization is necessary.
- Relative frequencies – the percentage of the word out of all words in the corpus
- Used when the word is compared in various corpora
- We extrapolate raw frequencies from corpora, and compare so that we can use the same factor, often in terms of thousand or million words.
Relative frequencies: example

- The pronoun *we* appears 2142 in a corpus consisting of 148,624 tokens and 2666 in another corpus of 483,913 tokens.
- Since the two corpora are not the same size, we have to normalize the values.
- Divide the number of the particular token by the total number of tokens and multiply by 1000 or million to get the occurrences per thousand or per million words.
  - $\frac{2142}{148624} \times 1000 = 14.41$ thus 14 occurrences per thousand word
  - $\frac{2666}{483913} \times 1000 = 5.5$ thus 6 occurrences per thousand word
- *we* is more than twice as common in the first corpus.
Keywords

- Words which are significantly more frequent in one corpus than another are called key words.
- They can be used to list significantly different lexis between two corpora.
- Keywords are calculated by carrying out a statistical test (e.g., loglinear) which compares the word frequencies in a text against their expected frequencies derived in a much larger corpus, which acts as a reference for general language use.
- Keywords can be lexical items that reflect the topic of a particular text, but they can be grammatical words sometimes...
Summary

- Use available corpora if possible.
- Think about the corpus representativeness and balance when you choose a corpus for your research question.
- If you do not find an appropriate corpus for your research question, you can build your own corpus.
- Corpora are useful in language studies!
Assignment

- Analyse corpora
  - concordances
  - collocations
  - frequency lists
  - keywords