COR: Corpus Linquistics

Lecture 11
Representativeness and Balance

Francis Bond
Department of Asian Studies
Palacký University
https://fcbond.github.io/
bond@ieee.org

https://github.com/bond-lab/Corpus-Linguistics

Overview

- Representativeness and balance
- Copyright and Ethics

Representativeness and Balance

This section is based on: "Corpus and Text: Basic Principles" by John Sinclair (2004) in *Developing Linguistic Corpora: a Guide to Good Practice* Martin Wynne ed, University of Oxford

Representativeness

To define representativeness we need to consider the following questions about the users of the language we will represent:

- What sort of documents do they write and read, and what sort of spoken encounters do they have?
- How can we allow for the relative popularity of some publications over others, and the difference in attention given to different publications?
- How do we allow for the unavoidable influence of practicalities such as the relative ease of acquiring public printed language, e-mails and web pages as compared with the labour and expense of recording and transcribing private conversations or acquiring and keying personal handwritten correspondence?
- How do we identify the instances of language that are influential as models for the population, and therefore might be weighted more heavily than the rest?

There are no definite answers to these questions.

How to be representative

1. decide on the structural criteria that you will use to build the corpus, and apply them to create a framework for the principal corpus components;

2. for each component draw up a comprehensive inventory of text types that are found there, using external criteria only;

3. put the text types in a priority order, taking into account all the factors that you think might increase or decrease the importance of a text type — the kind of factors discussed above;

4. estimate a target size for each text type, relating together (i) the overall target size for the component (ii) the number of text types (iii) the importance of each (iv) the practicality of gathering quantities of it;

5. as the corpus takes shape, maintain comparison between the actual dimensions of the material and the original plan;

6. (most important of all) document these steps so that users can have a reference point if they get unexpected results, and that improvements can be made on the basis of experience.

The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.

Balance

- for a corpus to be balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgments.

- Most general corpora of today are biased towards text: estimates of the optimal proportion of spoken language range from 50% to 90%
  because speech is so expensive to collect this imbalance is likely to remain.

- Balance can conflict with representativeness
  - Consider popular magazines in English
  - there are a large number of them and most use a highly specialised language
  - It is an important text type, but it is almost impossible to select a few texts which can claim to be representative
  - Can magazines for fly fishermen, personal computers and popular music really represent the whole variety of popular magazines (as is the case in The Bank of English)?

- Specialised corpora are constructed after some initial selectional criteria have been applied, for example a blog corpus or a patent corpus. More delicate criteria are used to partition them, but the issues of balance and representativeness remain cogent and central in the design.

The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.

Other issues

- For large web corpora one issue is that of who wrote something
  - Was it a native speaker?
  - Was it a text-mill?
  - Was it translated?
  - Was it created by a script?
  - Was it created by a language model?
- As more text is written by machine, we get problems of data-pollution
  - Machine generated text misses rare constructions
  - If you train a new model, you loose even more, ...
- If you use corpora as knowledge sources, other problems arise
  - Accuracy
  - Bias

Summary

- The extent to which conclusions from a test can be generalized depend on both the nature of the evaluation function and the size and representativeness of the test set
- The extent to which conclusions drawn from a corpus study can be generalized to all language depend crucially on the design of the corpus
- In general more **representative** data is better data
- But we need to be aware of limitations:
  - Data sparsity
  - Out of vocabulary items
  - Over fitting
  - Domain adaption

Copyright and Licensing

Copyright

- Governments grant certain rights to authors of creative works, typically called copyrights in order encourage them to produce more
  - The most basic right is the right to forbid people from copying it without permission
  - Any work produced is by default copyright of the author
- Some or all of these rights can be waived or transferred
  - An author may sell the rights for a manuscript to a publisher
  - A blogger may place their postings in the public domain
  - A publisher may give permission to an author to post their paper on their website
  - A work may be distributed under a license that allows copying only under some conditions

- Copyright laws are national laws, although they may be harmonized by treaties
  - A text may be illegal to copy in one country, but legal in another
- Copyright laws change over time
  - E.g. in the U.S. originally 14 years for books only
  - Now 70 years after the death of the author for almost everything (but not recipes)
- New technology complicates things
  - Sending email involves making multiple copies on different servers
  - Recording speech can happen without the creator's knowledge

Copyright issues are very complicated

Some Rough Guidelines

- Copying something which is under copyright is illegal unless specific permission is granted or it falls under **fair dealing**, such as for the purpose of research or education

- How can you get permission?
  - You can buy it (for some works)
  - You can get signed permission from the copyright holder (or recorded permission for preliterate speakers)
  - You can get implicit permission (e.g. for email or web pages)
  - It can be permitted by a license
    - **CC-by** allows you to copy and redistribute if you acknowledge
    - **CC-by-nconc** allows you to copy and redistribute if you acknowledge and it is for non-commercial use

- The following factors will be considered to decide if it is fair dealing (in Czechia)
- All exceptions must pass the three-step test under the Berne Convention
  - any exception to copyright may only be used in special cases
  - it must not conflict with a normal exploitation of the work
  - does not unreasonably prejudice the legitimate interests of the author
- Examples of exceptions are
  - Quotation
  - Government and journalistic use, civil and religious ceremonies and school performances
  - Exhibitions and Catalogues
  - Freedom of Panorama (copyrighted works in public exhibits)
  - Fair use: caricature and parody
    Less broad than US

Copyright for Corpora

- Arguments for restrictive licensing
  - Competitive advantage (common for speech corpora)
  - Compensation for the effort of creation
  - Minimize effect on the value of the original work
- Arguments for open licensing
  - Annotation is expensive, making the data open gets the best return on this investment
  - Annotation is typically ongoing, opening the data gets you more feedback
  - Researchers are evaluated by the impact that their work has. Open data generally has more impact.
  - Language data is part of our shared heritage

Choice of License

- Should be considered early on (before you start compiling your corpus)
- May depend on the funding body
- Depends on the source data
- General trend is to open licensing
  - Open Science Project
  - Open Access Journals
  - Open Source Software
- Try to chose a standard license (such as Creative Commons)

UPOL's policy on Data Sharing

- Directive (EU) 2019/1024 on open data and the reuse of public-sector information

  The directive is based on the general principle that public and publicly funded data should be reusable for commercial or non-commercial purposes.

  - **Dynamic and real-time data** — made available as API and bulk download
  - **Research data**
    - Publicly funded research data must be openly available
    - Following the principle of 'open by default'
    - Data should be findable, accessible, interoperable and reusable (the FAIR principle)
  - **Fair trading and non-discrimination**
    - The reuse of documents is open to everyone in the market
    - Any applicable reuse conditions should be non-discriminatory

Creative Commons Licenses

License	Derivative Works	Same License	Commercial Use
CC-BY	+	-	+
CC-BY-SA	+	+	+
CC-BY-NC	+	-	-
CC-BY-ND	-	-	+
CC-BY-NC-SA	+	+	-
CC-BY-NC-ND	-	-	-

BY Attribution (all licenses)

SA Share Alike (requires copies to have the same license)

NC Non-Commercial (Not Open)

ND No Derivatives (allows only exact copies) (Not Open)

Many, many other license also exist (GPL, MIT, BSD, Apache, ...)

The Open Definition

- The Open Definition sets out principles that define **openness** in relation to data and content.
- It makes precise the meaning of **open** in the terms **open data** and **open content** and thereby ensures quality and encourages compatibility between different pools of open material.
- It can be summed up in the statement that:

  "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)."
- Put most succinctly:

  "Open data and content can be freely used, modified, and shared by anyone for any purpose"

Conclusions

Be careful of copyright
Make your data open by default