+

Corpus Linguistics

CS140b
NL Annotation for ML

January 16, 2018
Professor Meteer

+ Corpora for linguistic research

- It is quite typical for researchers to use *any collection of texts* for linguistic analysis.
  - Often proceed opportunistically: whatever data comes in handy is used.
- However, the term *corpus* usually implies the following characteristics:
  - sampling/representativeness
  - finite size
  - machine-readable form
  - a standard reference
  - (time-bound)

Sampling and representativeness

- Sampling is a fundamental characteristic of any empirical work.
  - It is impossible to study every single instance of a phenomenon of interest.
  - With language, this is even more difficult: languages change continuously.
  - A corpus is a “snapshot” of the language at a specific time.

+ Finite size

- Usually, corpora have a fixed size.
  - E.g. BNC is 100 million words

- But not always. Some corpora keep growing over time.
  - Example: COBUILD Corpus built at Birmingham university is periodically updated.
  - Very useful for lexicographic work: if the corpus is updated regularly, it remains a good source of new words and usages.

+ Static and non-static

- Sample corpus:
  - a corpus which represents a sample of a language within a specific period
  - BNC is a good example of this, covers 1960-1993
- Monitor corpus:
  - a dynamic sample
  - normally covers a relatively brief span of time (i.e. decades, not centuries)
  - updated regularly to keep track of changes within the language

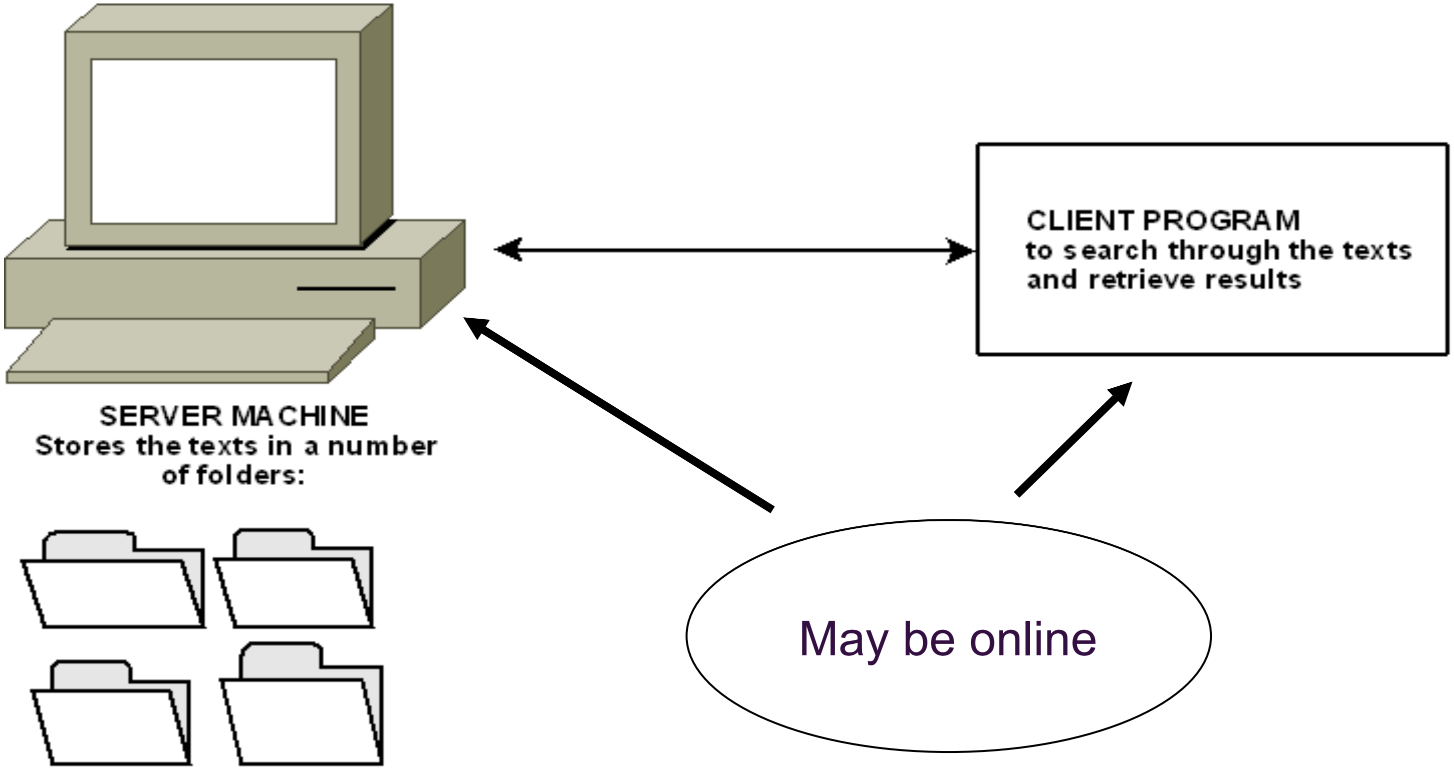
+ Time

- Unless the corpus is a monitor corpus, the sampling will inevitably mean that we' re restricted to a period of time.

- Can have interesting consequences:
  - Do you think the English language has changed since 1993? What aspects will have changed? Lexicon? Syntax?
  - *I'm down with that!*

+ Machine readability

- Very rare for a corpus nowadays to be in print.
- We've seen some advantages of machine-readability before
- What "machine readable" really means


SERVER MACHINE
Stores the texts in a number of folders:
CLIENT PROGRAM
to search through the texts and retrieve results
May be online

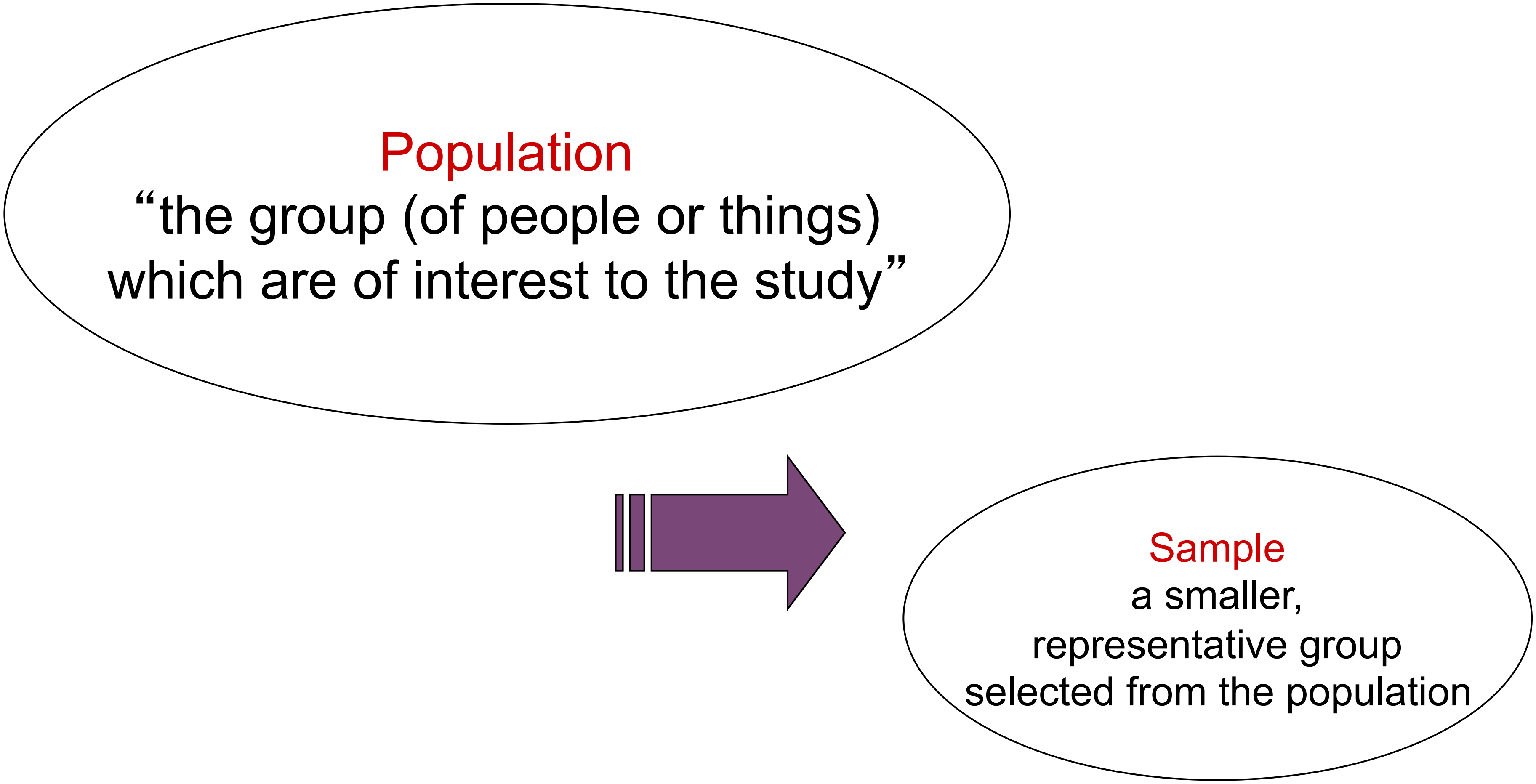
+ Client programs for corpus search

- Tools for searching through large collections of plain text (with/out annotation). E.g.
  - WordSmith
  - MonoConc Pro
  - Very useful to build frequency lists etc…
- Corpus-specific clients E.g.:
  - SARA
    - program created for the BNC
    - sensitive to the specific annotations in the BNC
    - allows search for patterns such as DETERMINER+NOUN
- Online servers with web-based client
  - SketchEngine, etc
  - Increasingly popular

+ A standard reference

- This is not an essential aspect of a corpus, but it is useful.
- It presupposes:
  - wide availability
  - broad coverage
- If a corpus is a standard reference, then it becomes:
  - a common source of data, hence studies are replicable
  - a yardstick against which to measure other, newer corpora

Samples and populations



Population
“the group (of people or things) which are of interest to the study”

Sample
a smaller, representative group selected from the population

+ Sampling to avoid skewness

- Chomsky’ s criticism about the skewness of corpora:
  - any sample of the language will be biased, including some things but not others
- This is rather like sampling from the human population:
  - psychologists who select samples of people for experiments know that skewness is a risk
- A good sample should capture the variability in a population.

+ # Prerequisites for sampling

1. definition of the boundaries of the population
   - written part of the BNC: English published within the UK between 1960 and 1993
   - Brown Corpus: written English published in the US in 1961

2. definition of the sampling unit
   - books, periodicals, radio broadcasts…

3. sampling frame = the list of sampling units
   - Brown Corpus: the list of books and periodicals in the Brown University Library and the Providence Athenaeum.
   - BNC: more sophisticated, considered who wrote what and who was the target audience

+ Defining the language population

1. language production
2. language reception
   - Both of these are demographically-oriented.
   - focus on characteristics of the producer or receiver
     - sex, age, social class…
   - typical of the approach in the BNC
3. language as product
   - starting point is “what’s out there”, irrespective of who produced it and for whom
   - typical of the approach in the Brown Corpus

+ Sampling in the BNC

- Population definition looked at both production and reception
- Sources for production (who publishes what?):
  - Catalogues of books published per annum
  - Lists of books in print
- Sources for reception (what is read by whom?):
  - bestseller lists & prizewinners
  - library lending statistics

+ Sampling techniques

- Once population is defined and sampling frame identified, actual sampling can proceed in several ways:

1. simple random sampling: identify a subset randomly from the total set of sampling units in the frame
   - □ may omit rare items in the population, because if X is more frequent than Y, X’s chances of being selected are higher

2. stratified random sampling:
   a. split population into relatively homogeneous groups or strata
   b. sample each stratum randomly

+ Sampling of written text in the BNC

- After sources were selected based on production/reception criteria, they were classified on the basis of 3 main features:
  - domain (“subject”)
    - imaginative, arts, belief and thought, …
  - time (when published)
    - 1960 – 1974; 1975 – 1993
  - medium
    - book, periodical, written-to-be-spoken, etc
- These then determine the strata for sampling in the BNC.

+ Sampling of spoken discourse in the BNC

- The features defining the sampling frame differ for spoken language:
  - demographic component
    - informal conversation recorded by 124 volunteers
    - selected by age, sex, social class, geographical region
  - context-governed component
    - more formal encounters
    - meetings, lectures, etc

+ Balance and representativeness

- Balance:
  - refers to the range of types of text in the corpus
  - e.g. the BNC's construction was based on an *a priori* classification of texts by *domain*, *time* and *medium*
- Representativeness:
  - refers to the extent to which the corpus contains the full range of variation in the language.
- Representativeness depends on balance as a prerequisite.

+ When is a corpus representative?

- Biber (1993):
  - “Representativeness refers to the extent to which a sample includes the full range of variability in a population”.
- What variability?
  - variability of text types (different genres, different registers)
  - variability of linguistic phenomena (lexical, syntactic)
    - Not all linguistic features are distributed in the same way

+ Variability in distributions

- Active, declarative clauses are probably more frequent overall than passives.
  - But passives become very frequent in certain types of text (e.g. academic discourse).
- Certain word orders are “marked”, hence probably less frequent than the unmarked.
  - *cf.* SVO vs other orders in Maltese

+ Variability in distributions

- Some words may be completely absent in everyday usage, but highly frequent in specialised registers.
  - neutrino, morpheme, palato-alveolar…

- The same is true of word senses:
  - *qoxra* (MT) = *shell* – probably the most frequent sense
  - *qoxra* can also mean “seafaring vessel” (*qoxra tal-baħar*)
    - more likely to be used in this sense in the fishing/sailing register

+ The need for a priori criteria

- Problem:
  - before we begin to sample for representativeness, we need a notion of what the range of variability is.
- Therefore some criteria need to be defined a priori.

+ Linguistic variability and text type

- It is likely that genre or register or text type is a determining factor of linguistic variability.

- All the foregoing examples were made with reference to text type.

- Two plausible views:
  1. sample based on text type to capture linguistic variability (as in the BNC)
  2. sample based on a predefined model of what linguistic variability there is

+ External (situational) criteria

- Define sampling frames by the social and communicative contexts in which a particular sample of text/speech is produced.
- Biber (1993) suggested external criteria should determine the sampling frame to ensure representativeness.
- Under this view, texts are selected to cover a predefined range of uses/ purposes/contexts. This is the BNC approach.

+ External criteria

- Sampling based on situational criteria would proceed as follows:
  1. identify the range of types / genres/registers
  2. identify the units within each type
     - NB: The size of each category will reflect how widespread or common the type is
  3. sample from the units within each type

+ Internal (“linguistic”) criteria

- Define sampling frames on the basis of linguistic features (e.g. lexico-grammatical) that distinguish texts.
  - Example: “to be representative our corpus should contain the majority of (word) types in the language, as defined in some standard dictionary”

- Potential problems with internal criteria
  - Internal criteria risk becoming circular:
    - you need a good linguistic resource (such as a corpus) to study the distribution of relevant features
    - but you’re need the features to design the corpus!

+ Balance between text types

- We’ve noted that representativeness depends on balance:
  - language variation is captured in the sample if it comes from the same sources that determine the variation
- But balance is very difficult to assess.
  - Depends on an agreed-upon definition of what the range of text types is.

+ The notion of "domain" in the BNC

- imaginative (21.91%)
- arts (8.08%)
- belief and thought (3.40%)
- commerce/finance (7.93%)
- leisure (11.13)
- natural/pure science (4.18%)
- applied science (8.21%)
- social science (14.80%)
- world affairs (18.39%)
- unclassified (1.93%)

- Why represent commerce/finance separately?
- Why is commerce/ finance more represented than arts?
- Why not have a separate category for "poetry"?

+The notion of “medium” in the written BNC

- book (55.58%)
- periodical (31.08%)
- misc. published (4.38%)
- misc. unpublished (4%)
- to-be-spoken (1.52%)
- unclassified (0.4%)

- Why more books than periodicals? Aren’t periodicals more numerous?
- Why not more “unpublished”? Most written discourse remains unpublished.

+ Summary

- Sampling (in general)
  - inclusion of a subset of the relevant units in a population, to ensure representativeness of relevant features
- Balance
  - ensuring that the range of types of text is represented correctly in the sample
- Representativeness
  - ensuring that interesting variation of linguistic features is captured

+ Summary

- To achieve representativeness, we need to ensure balance.

- Balance is usually achieved through external criteria.
  - These are used to determine the sampling frame.