

Corpus Linguistics (L615)

Basics of Corpus Linguistics

Markus Dickinson

Department of Linguistics, Indiana University
Spring 2009

Representativeness: the extent to which a sample includes the full range of variability in a population

- ▶ distinguishes corpora from archives
- ▶ allows findings to be generalized to a particular variety of language

A corpus is a sample of language use (i.e., from a particular population)

- ▶ balance: types of genres
- ▶ sampling: how the text is selected

Different kinds of criteria:

- ▶ **external criteria:** defined situationally (cf. genres, registers)
- ▶ **internal criteria:** defined linguistically (text types)
 - ▶ Internal criteria often used to define representativeness
 - ▶ e.g., distributions of words or grammatical constructions

But this distributional argument is problematic: we shouldn't pre-define a distribution when the corpus should tell us this

- ▶ External criteria seem more justified: linguistic characteristics are independent of selection process
- ▶ But internal analysis can help define what texts should next be added

Non-linguistic & linguistic information

Corpora can be evaluated by the degree to which they capture the following (Biber 1993):

1. range of text types in a language
2. range of linguistic distributions in a language

Biber claims: “if a corpus does not represent the range of text types in a population, it will not represent the range of linguistic distributions”

- ▶ Do you agree with this statement?

Should corpora be updated regularly?

- ▶ And if not, do they become un-representative?

Two general types of corpora:

- ▶ **sample corpus**: static corpus
- ▶ **monitor corpus**: dynamic corpus which grows

Multiple sample corpora can also provide a view of language change (e.g., Helsinki, LOB corpora)

- ▶ **General** corpora = designed to provide an overall description of a language
 - ▶ The British National Corpus (BNC) is supposed to represent modern British English *as a whole*
 - ▶ Such corpora need to be balanced w.r.t. different text types
- ▶ **Specialized** corpora = domain or genre specific
 - ▶ Also need to be balanced within the particular domain
 - ▶ Representativeness represented by 'closure' or 'saturation': is the concept represented completely?
 - ▶ Divide corpus into equal segments: if a new, equally sized segment has the same number of lexical items, corpus is saturated

What should be covered in a balanced corpus?

- ▶ **balanced** = covers a range of text categories
 - ▶ Definition depends upon the intended uses
 - ▶ No real objective measure of balance
 - ▶ Usually based on proportional sampling (discussed later)
- ▶ Balance can be based on a **text typology**, a classification of text types

Different axes that are important in designing a balanced corpus:

- ▶ written and spoken language
- ▶ genres
- ▶ production variables (gender, age, social class, ...)
- ▶ reception variables (large audience, small audience, level of formality, ...)
- ▶ temporal slices

The **British National Corpus (BNC)**, a corpus of 100 million words, is generally balanced

- ▶ 90% written, 10% spoken
 - ▶ Written texts selected based on: domain (subject field), time, & medium (books, periodicals, etc.)
 - ▶ Spoken texts selected based on demographics and context (meetings, lectures, radio broadcasts)

Aim is to represent contemporary British English

Language production and reception

Language can be categorized as:

- ▶ **Language reception:** language people hear & read
 - ▶ newspapers, novels, media, etc.
- ▶ **Language production:** language people speak & write
 - ▶ essays, letters, recorded conversations, etc.

Texts which reflect reception are easier to acquire ... but they do not accurately reflect production

Final thoughts on balance

- ▶ Corpus design should be well-documented so the user can decide if it's appropriate and balanced enough.
 - ▶ Subcorpora should be easily obtainable, for particular research questions
- ▶ We can't wait until there exists a perfect definition of balance.

To achieve representativeness & balance, we have to **sample** language (since we cannot exhaustively describe it)

- ▶ The sample should be representative of the larger **population**

To properly sample, need to define:

- ▶ **sampling unit**: book, periodical, newspaper, articles, chapters, ...
- ▶ **sampling frame**: list of all possible units, from which that actual ones are selected
 - ▶ Brown corpus sampling frame: list of books & periodicals in Brown University Library & Providence Athenaeum
 - ▶ Supposed to represent written English text published in 1961

Defining a target population

Difficulties of defining a target population for language corpora (Atkins, Clear, & Ostler 1992)

- ▶ Hard to rigorously delimit target population
- ▶ Hard to demonstrate that all features of the population are represented
- ▶ No obvious unit of language to sample & define population (texts, sentences, words)

Aspects of defining the target population (Biber 1993):

1. boundaries of the population (which texts to include/exclude)
2. hierarchical organization within the population: categorizations of texts and their definitions

- ▶ Ways to sample:
 - ▶ **simple random sampling**: sample from all sampling units
 - ▶ **stratified random sampling**: organize units by category and then sample
 - ▶ Can organize by demographic parameters or whatever is desirable for representativeness
 - ▶ Sampling proportions: should the categories be weighted?
- ▶ Sample size: e.g., full text units or text chunks?

Stratified random sampling

“Stratified samples are almost always more representative than non-stratified samples” (Biber 1993)

- ▶ Variance between groups (i.e., text types) is larger than variance within a group
- ▶ So, being sure to include 100% of different groups will ensure better variability of language

A handful of issues for creating corpora when sampling texts:

- ▶ combining texts in different formats into a streamlined format
- ▶ taking different encodings and putting into one
- ▶ converting speech to text
- ▶ obeying all copyright laws for distribution

More on representativeness

Biber (1992)

Corpus Linguistics

Basics of Corpus
Linguistics

Representativeness

Balance

Sampling

Representativeness
(redux)

Representativeness tries to capture the full range of variability in a corpus

- ▶ But what kind of variability?
 - ▶ If cultural variability, we'll want to view text as a *product* and collect published works
 - ▶ If linguistic variability, we'll want spoken texts, etc.

And if we want linguistic variation, it's not clear that proportional sampling of genres will help

- ▶ e.g., many conversations (maybe 90% of our language) will have the same linguistic variety

Situational parameters for stratified sampling

1. Primary channel: written/spoken/scripted
2. Format: Published/not published
3. Setting: Institutional/other public/private-personal
4. Addressee
 - 4.1 Plurality: unenumerated/plural/individual/self
 - 4.2 Presence: presence/absence
 - 4.3 Interactiveness: none/little/extensive
 - 4.4 Shared knowledge: general/specialized/personal
5. Addressor
 - 5.1 Demographic variation: sex, age, occupation
 - 5.2 Acknowledgement: Acknowledged individual/institution
6. Factuality: Factual-informational/intermediate or indeterminate/imaginative
7. Purposes: Persuade, entertain, ...
8. Topics: ...

Sample size

Let's say you want to build a corpus that allows for generalizations about linguistic structure

- ▶ How big does this corpus need to be?
- ▶ Part of that, as we'll see, depends upon how frequent the individual constructions of interest are

There's no magic formula for determining a good sample size, but a useful notion is to calculate the standard error

$$(1) s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

- ▶ s = standard deviation of some variable
- ▶ N = sample size

Using the standard error

$$(2) s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

The standard error tells us how far a sample mean might be from the true population mean

- ▶ To have a lower $s_{\bar{x}}$, we need a higher N

It's difficult to use this equation:

1. We need to determine what's a tolerable error rate
2. We use the standard deviation of *some variable*, and that will change for different variables (cf. linguistic features)
3. Circularity: We need to know the standard deviation to get a good sample size estimate . . . which comes from having done an analysis

Lengths of text samples

Linguistic corpora present challenges for sampling since there isn't a good a priori definition of an "observation"

- ▶ e.g., observations can (but may not) spread over several words

Biber compares 10 different linguistic features across texts and computes a reliability coefficient

- ▶ Reliability coefficients "assess the stability of frequency counts across the 100-word samples"
- ▶ Common features (e.g., nouns) are relatively stable, whereas rare features (e.g., *wh* relatives) are less stable
 - ▶ Less stability means that we need *longer* texts to ensure its presence
 - ▶ e.g., Any 200-word sample will likely contain the same number of nouns; not true of *wh* relatives

Sample size (again)

Recall the calculation of standard error:

$$(3) s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Sample size depends upon our tolerable error (te)

$$(4) te = t \times s_{\bar{x}}$$

Solving for N , we arrive at:

$$(5) N = \frac{s^2}{(te/t)^2}$$

See table 17.4 to see how the estimated sample size varies depending upon the linguistic feature

- ▶ Nouns: 59.8 words required
- ▶ Conditional clauses: 1,190.0 words required

Variation across registers

Assume we have:

- ▶ pilot corpora of different registers
- ▶ a set of linguistic features of interest
- ▶ calculations of standard deviation & variance for each feature

Then, we:

1. Calculate the average normalized deviation for each register (e.g., 0.37, 0.39, 0.49)
2. Decide how many texts we want overall
3. Solve an equation to determine the number of texts for each
 - ▶ $0.37x + 0.39x + 0.49x = 140$
 - ▶ This allots more texts to the register with higher variance