27/09/12

§2.2 Proto-Structure

"Formally Sampling Language" 2.2

- ~~Describe which samples are possible w// language~~
  - ~~Desirable properties of corpora~~ // as they are currently be used
    - ~~Size~~
    - ~~Randomness~~
    - ~~Validity~~
    - ~~Limits on sampling method [practical]~~

- Desirable Properties of a corpus
  - Size
  - Randomness
  - Validity (internal, for external see §2.1)
  - Replicability / Reliability
  - Flexibility

Based on current techniques, but not restricted by method

- ~~Description of current best practice in sample~~ [2.2.1]
- Sampling as applied to language, (ie what is applicable)

A - Population Definition
B - Selection of a Sampling Frame
C - Sampling Methods
D - Determination of size // Power
E - Post-hoc / Reweighting
F - Bias Restriction methods

2.2.2

(NB. Perhaps these should be done in-place)

Comparison of Current Methods to A-F

A - Population Definition	∞, specialist + GP compare.
B - Sampling Frame	⇩ why no speech etc
C - Sampling Methods	trad. Web to come next section avail. sample
D - Size / Power	ling - estimates + Big data findings
E - Post-hoc / Reweighting	
F - Bias Estimation.	- Literature lum ling.

A- Population Definition (external validity, ~~focus~~ replicability focus)

- Largely linguistic as a problem
  - What was the aim of corpora used today?
  - What is the basis of external validity in §2.1?
  - Can these be improved?
  - Literature on population estimates for languages
  - ~~Extent? Database? Users~~
  - Causal systems, I vs. E language etc. w/ ↩

- Statistical Measures of specific cases (GP is too inf.)
  - Methods for pop. estimation (proportions)
  - ~~Application~~
  - Application to language
    - Estimates of web size
    - " Books/Publications
    - " Spoken word
    - " All. T(ratios)

perhaps this belongs higher up? 2-P

↓add more re. 2.1 linguistic angle.

B- Sampling Frame (Internal validity, ~~replicability~~ reliability focus)

- Enumeration or Bounding
  - Special-purpose corpora
  - General Purpose
    - Personal corpora
    - ~~Restrict/research~~
  - ~~Web~~

- Subsampling and (ab) use of GP corpora

- Auxiliary information: forms and availability

C - Sampling Methods

Non Random, compare to current practice

- Snowball Sampling - web crawler comparison
- Purposive - compare to GtP corpus Building, see notes.

Random, ideal case

- Simple Random - Focus on problems with no population, practical issues in establishing PO, practical + legal issues with selection
- Stratified Sampling - How to select strata? (with ling. relevance). Multydimensional Strata. Inverse Prob. Weighting. PPS possibilities with web/offline.
- Multistage Sampling ~and especially~ - IPW, how to balance samples for general purpose use?
- Cluster Sampling - Website-wide samples. Relate to the structure and legality of publishing the web.
- Adaptive Sampling - Perhaps this, guided by multistage, offers a good approximation of WaC methods as they are currently?

D - Sample Size // Power

- Criteria for selecting sizes (Linguistic)
  - Common models "should be big enough for x"
  - Sufficient representation "should contain n xs"
  - Breadth "should contain n types of x"

- Existing corpora - focus on the rise of big data.

- For multistage designs, how large should the initial sample be?
  - is this small enough to be guided by the user?

should move.

E - Reweighting (perhaps should be merged with Bias methods)

- Establishing weights using auxiliary data
  - Possible sources of valid aux data (focus on possible uses, perhaps too dependent)
  - Existing manual resampling in Ling (eg. selecting categories from BNC)

F – Bias Estimation

NB: Most of a work in progress for now.

- Resampling (Focus on generality)
- ~~Data~~ Comparison to our data

§2.2 notes

- External Validity of Corpora
  -> also transferability when qualitative.
- Should 'Big Data' be mentioned more
- Is current practice basically nonrandom. If so, is it valid to base new corpora on the balance of old ones? [no] ↘ 4WD
- Do we wish to have <u>representative</u> or <u>comparable</u> results?
- Is linguistic data ever simple enough for ratio or regression estimation to be used to guide samples?
- Adaptive sampling looks to be driven in a way similar to conventional corpus building
  - conditions of interest may be very useful for §5 - expressing a sampling policy in a repeatable, valid, and useful way
- It would be nice to separate "how to represent a sampling frame" into a separate discussion, but this probably belongs in §5.