Computational Methods in Linguistics

Bender and Wassink 2012
University of Washington
Week 7: 5/7-5/9

Overview

- Normalized frequency
- Best Practice: "It is usually considered good practice to report *both* raw and normalised frequencies when writing up quantitative results from a corpus" (McEnery & Hardie, 2012: p. 51)
- Practicum group work on term projects
- LSA Abstracts

Interim Project Reports

- Structure of data sets
- What your variables will be
- How you’re pulling down (or not) your data
- How you’re formatting your database
- How you’re structuring your data set for analysis
- How you’re checking your scripts to see if they are functioning as expected

Normalized Frequency

- Def: A frequency expressed relative to some other value as a proportion of the whole -- for example, a frequency of a word relative to the total number of words in the corpus.
- Can be compared even if they arise from datasets of different sizes.

Normalized Frequency

- $f_{(Lancaster)}$: 1103 occurrences of "Lancaster" in BNC
- *N*=corpus size: 87,903,571
- percentage: .0013% of written section
- *nf* = number of examples in corpus / total corpus x base of normalization
- *nf* represents relative frequency "how often might we assume we will see the word per *x* words of running text?"

Bases of normalization

- base of normalization might be set to 1,000,000 e.g., how often *Lancaster* appears on average, in each million words of the BNC
- percentage = occurrences per hundred
- occurrences per thousand
- occurrences per million (e.g., common notation is *career=1187x*), another:

```
48609/tune/270/24/15/1.176/15/19/15/1.176
48610/tune/24/11/1.04/187/1.04/11/7/5/.699
```

- *nf* = (1,103 / 87,903,571) x 1,000,000
- *nf* = 12.55

...over multiple corpora

- Corpus I: 1103 times in 87,903,571 words
- Corpus II: 10 times in 1,146,597 words
- more or less frequent in Corpus II?

Corpus	*Lancaster* Instantiations	*f/N*	*x100*	*nf (1m)*	*nf1/nf2* *"Corpus to corpus ratio"*
I (BNC)	1103	0.000013	0.0013	12.55	1.44
II (BE06)	10	0.000009	0.0009	8.72	

Ratio *nf*1/*nf*2 indicates how many times more often our token or type occurs in the corpus in which it is more frequent than in the one in which it occurs with lower frequency.

Your turn

- What is the total size of your corpus (by type)?
- What is your type count?
- What are the grouping variables that affect your count? (ensure counting same things if >1 corpora)
  - within constituent type?
  - within set of phonetic environments?
  - within a particular register?
- What is appropriate base of normalization?