**Choice of Word Frequency Norms can Dramatically Affect Inference.**

Johesh Fruehwald

Some factors influencing variation are observable, and others must be estimated. Different estimates may be correlated, but are they interchangeable?

**Case study 1: TD Deletion**

**TD Deletion Data: Monomorphemes**

- west
  - [west]
  - [wast]

**Frequency Norms: Zipf Scaled**

\[
\log_{10}(\text{frequency per million words}) = 3
\]

**Model**

\[td = \text{zipfscore} + \left(1 \times \text{Word} + \text{zipfscore} \times \text{Speaker}\right)\]

**Regression Results:**

- **Buckeye Corpus**
  - 6,691 Tokens
  - Frequency Norm
    - Within Corpus: 0.29
    - Celex: 0.52
    - Sublex: 0.34

- **Philadelphia Neighborhood Corpus**
  - 18,236 Tokens
  - Frequency Norm
    - Within Corpus: 0.599
    - Celex: -0.006
    - Sublex: -0.302

**Discussion**

The three different frequency norms result in very different estimated frequency effects. The within-corpus frequency norm estimated a frequency effect twice to 100 times the size of the others.

**Case study 2: /ay/ raising**

**Data:** /ay/ Raising from the PNC

**Model:**

\[F1 = \text{decade} \times \text{zipfscore} + \left(\text{decade} \times \text{Word} + \text{zipfscore} \times \text{Speaker}\right)\]

**Regression Results:**

- **Within Corpus**
  - **Ceflex**
    - estimate: 0.64
    - CI: (0.57, 0.71)
    - decade: -0.12
    - CI: (-0.13, -0.10)
    - freq: -0.09
    - CI: (-0.15, -0.01)
    - decade/freq: -0.006
    - CI: (-0.01, 0.01)

- **Sublex**
  - estimate: 0.67
  - CI: (0.6, 0.74)
  - decade: -0.12
  - CI: (-0.13, -0.10)
  - freq: -0.05
  - CI: (-0.12, -0.02)
  - decade/freq: -0.003
  - CI: (-0.01, 0.01)

**Discussion**

This time, the within-corpus frequency norm estimates the smallest frequency effect, but two of the norms don’t have a reliable effect, while the remaining one does.

**Why the differences?**

The biggest difference between these norms is their estimates of low frequency words. Recommendation: Use the norms with the best low frequency word estimates.

**References**