How to Approach Lexical Variation in Sign Language Corpora

Carl Börstell
University of Bergen
carl.borstell@uib.no

Intro

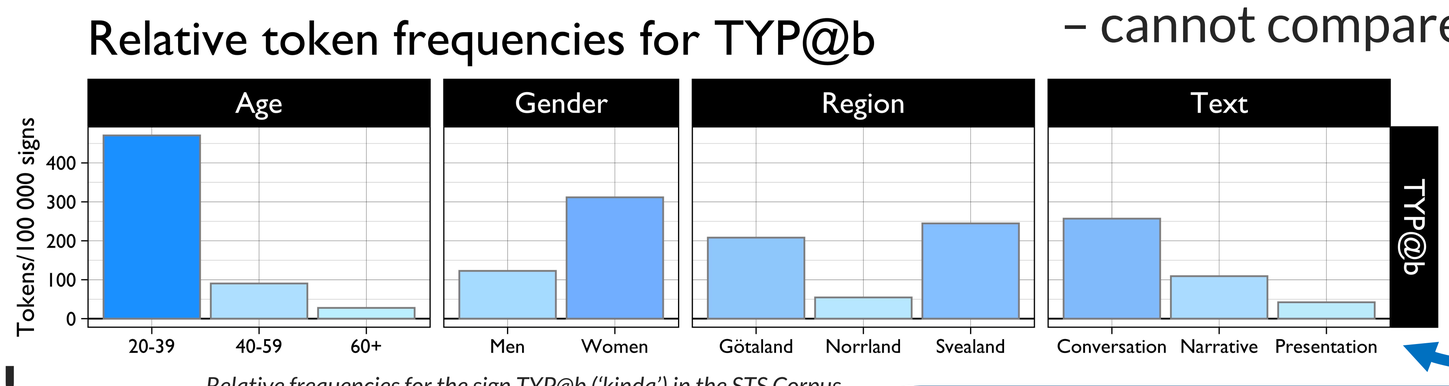
Lexical frequency can be a first step into exploring a corpus

– ***but how to do it?***

Here's a short guide!

Relative frequency

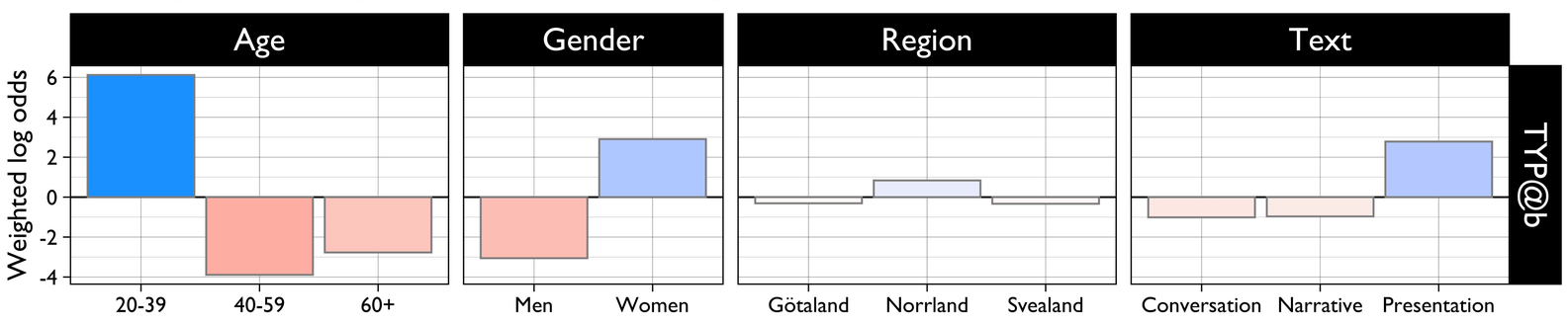
Raw frequency is <u>not</u> a useful metric – cannot compare it across corpora!



Relative frequencies for the sign TYP@b ('kinda') in the STS Corpus teckenspråkskorpus.su.se

Solution: Use ***relative* frequency** – e.g., tokens per 100,000 signs![1]

Weighted log odds

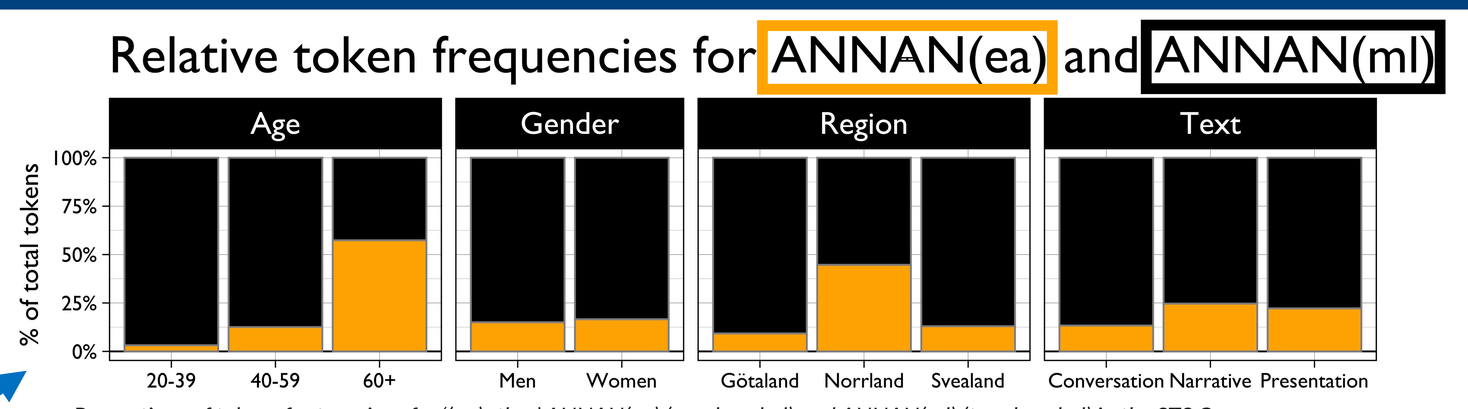


Weighted log odds for the sign TYP@b ('kinda') in the STS Corpus

If comparing frequencies for specific conditions – e.g., relative across age groups, regions, text – (weighted) **log odds** can help! Shows (over-)representation for each condition in a grouping.[3]

Proportions

If you want to compare e.g., lexical or phonological variants for some meaning, it can be sufficient to compare how many tokens they have each as **a proportion of the total**!

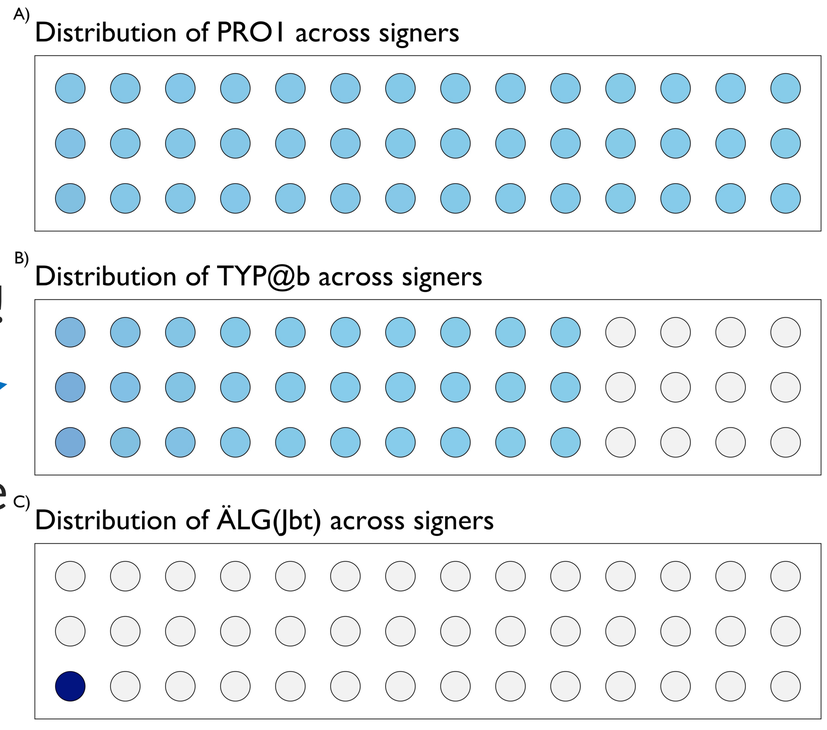


Proportions of tokens for two signs for '(an)other' ANNAN(ea) (one-handed) and ANNAN(ml) (two-handed) in the STS Corpus

Coverage/Dispersion

Whatever sign or construction you are looking for in a corpus, it is always wise to look at **coverage/dispersion**!

Even if a sign occurs multiple times, it may be used exclusively by a single signer or in a single text – i.e. it isn't very generalizable!



Signer coverage for three signs in the STS Corpus. Dots represent signers (n=42) and fill colors show signer's proportion of total tokens.

Checklist

- ☑ Use **relative or proportional frequency**: but know the raw!
- ☑ Use statistics like log odds to see **distributional differences**!
- ☑ Look at **coverage/dispersion**: how many signers/texts represented?
- ☑ **Know your data!** What do the annotations/tokens really mean?[2]

11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources, Turin, 25 May 2024

UNIVERSITY OF BERGEN

References

1. Börstell , C. & Östling, R. 2016. Visualizing Lects in a Sign Language Corpus: Mining Lexical Variation Data in Lects of Swedish Sign Language. In Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, 13–18, Portorož: ELRA.
2. Langer, G., Hanke, T., Konrad, R. & König, S. 2016. "Non-tokens": When Tokens Should not Count as Evidence of Sign Use. In Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, 137–142, Portorož: ELRA.
3. Tyler Schnoebelen, Julia Silge & Alex Hayes. 2022. tidylo: Weighted Tidy Log Odds Ratio.