Visualizing Lects in a Sign Language Corpus

Mining Lexical Variation Data in Lects of Swedish Sign Language

Carl Börstell & Robert Östling

calle@ling.su.se; robert.ostling@helsinki.fi

Stockholm University; University of Helsinki

Introduction

Corpora of sign languages are being developed around the world, giving researchers better data for conducting more diverse types of research. However, the corpus data is not always easily accessible, as they require special programs and downloading of heavy video files.

The SSL Corpus (SSLC) is no exception to this. Also, although metadata about sociolinguistic variables for the signers are available, they are not directly searchable in the SSLC, as they are stored externally. Because of this, we aimed to create a database containing signer and text metadata for each token in the SSLC and build a user interface that visualizes any sign's distribution across metadata variables based on relative distributions.

The SSLC

The SSL Corpus (SSLC)[1,2] used in this study comprises 39,733 tokens across 75 files and 42 signers. The signers are distributed across three **regions**, six **age groups**, and two **genders**. All corpus files are labeled as belonging to one of three **text types**. As the SSLC data are continuously being updated, the current version is not completely balanced across any of these categorizations or variables. The tables illustrate the token distribution across the variables.

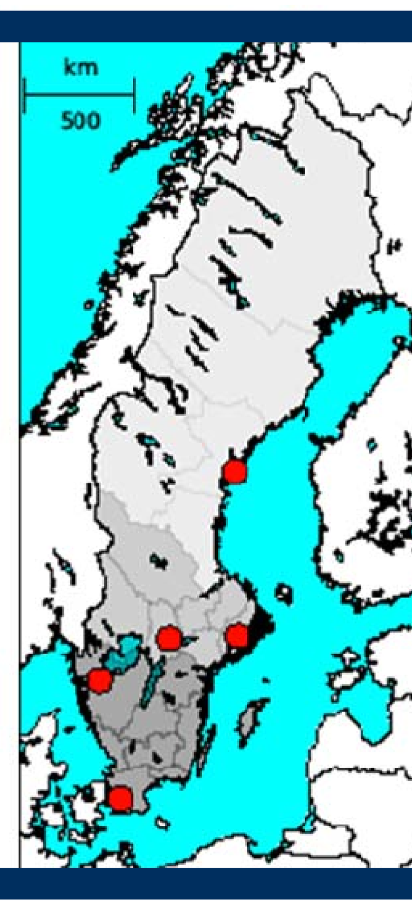
Distribution of tokens across variables

Region	Signers	Tokens
Norrland	4	5 310
Svealand	24	24 605
Götaland	14	9 818

Gender	Signers	Tokens
female	20	15 862
male	22	23 871

Age group	Signers	Tokens
20–29	9	4 225
30–39	6	11 680
40–49	7	10 646
50–59	8	3 007
60–69	8	7 756
70–100	4	2 419

Text type	Files	Tokens
Conversation	56	34 071
Narrative	14	3 525
Presentation	5	2 137



Map shows the location of the three regions: Norrland (light gray); Svealand (gray); Götaland (dark gray). The locations of the Deaf schools are shown as red dots.

Statistical method

We compute three rankings, one each for the categories of region, age, and gender. Signs are ranked by the Bayes factor between the hypothesis of separate categorical distributions versus an identical categorical distribution, assuming a Dirichlet prior for the categorical parameters:

$$b_s = \frac{B(x_s + \alpha)B(t - x_s + \alpha)}{B(t + \alpha)}$$

where x_s is a vector representing the distribution of the sign s and t is the distribution vector of all signs, and $B(x)$ is the multinomial Beta function:

$$B(x) = \frac{\sum_i \Gamma(x_i)}{\Gamma\left(\sum_i x_i\right)}$$

We use a uniform prior for the distributions, setting $\alpha = 1$.

Aim

The aim was three-fold:

1) Link **metadata** to each sign token and create a database of frequencies.
2) Construct an **interface** for visualizing lexical distribution across signer and text variables.
3) Testing a **statistical method** for automatically identifying signs that exhibit a skewed distribution.

Relative frequency

As the token distribution is not even across groups, we calculated a relative frequency for each sign based on the count $c_{s,g}$ representing the number of times sign s was used by any signer from group g. Then, we can compute the relative frequency among all the groups in a category G (e.g. age) using the maximum-likelihood estimate:

$$r_{s,g} = \frac{c_{s,g}}{\sum_{g' \in G} c_{s,g'}}$$

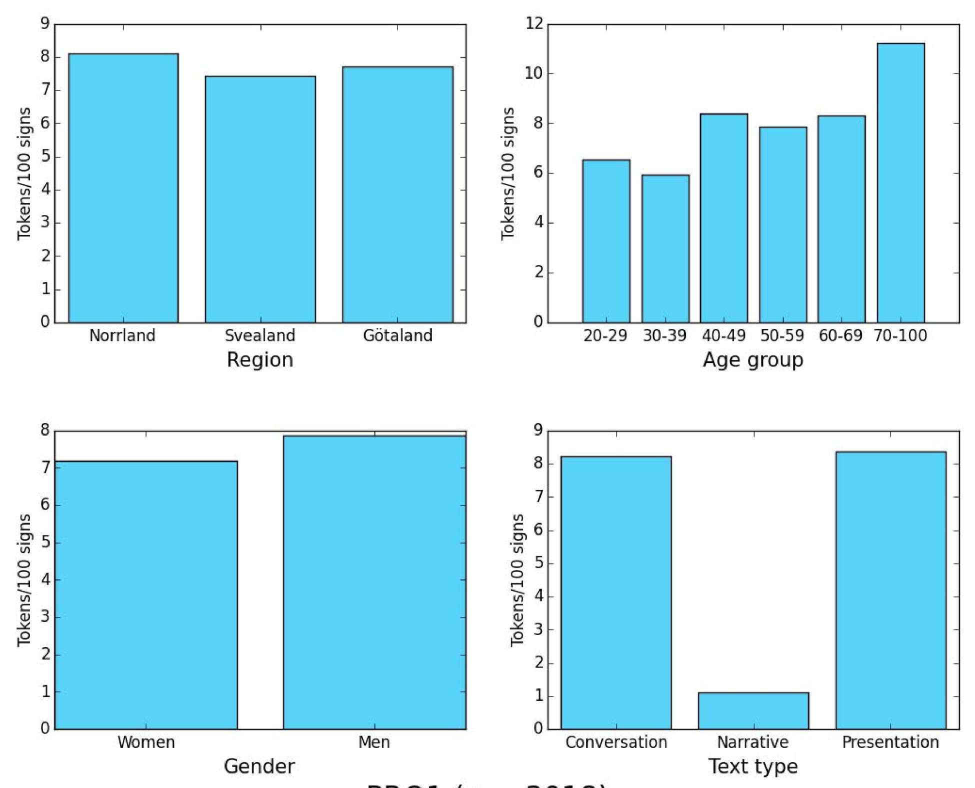
Interface

Using the Matplotlib package[3] in Python, we constructed a graphical user interface (GUI) that takes a sign gloss as input and outputs a graph for each of the four variables: region, age, gender, and text type. The graphs show the relative frequency of the sign across the groups within each variable. The interface is available online!

Try it yourself!

1) Enter website: mumin.ling.su.se/cgi-bin/ssllects.py
2) Enter sign gloss (e.g. "PRO1").
3) Get data visualization:

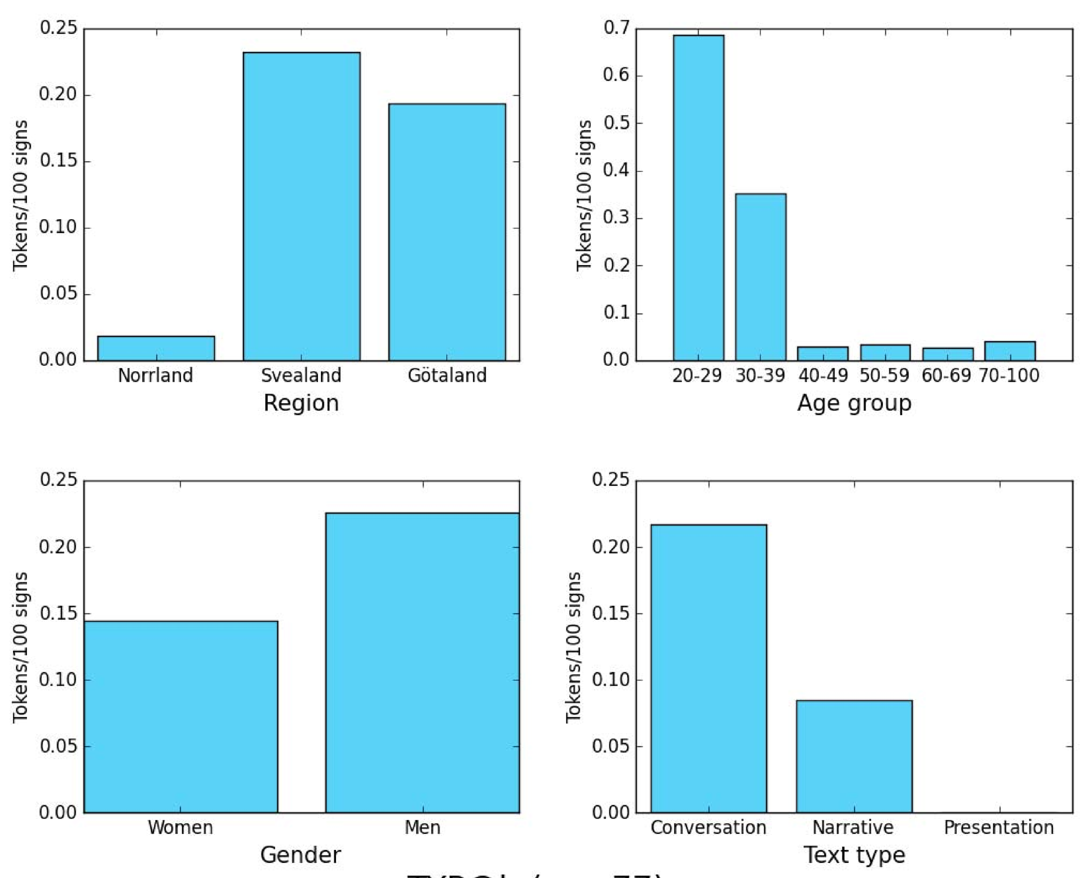
QR link to website:



PRO1 (n = 3018)

Evaluation & conclusion

Our statistical method correctly identifies several signs that are expected to show a skewed distribution, such as the sign TYP@b (a fingerspelled 'kinda'), which is used primarily by younger signers. We are confident that this will aid in future research on lectal lexical variation, as the SSLC expands. Also, the GUI makes the SSLC more accessible and data visualization quick and easy.



TYP@b (n = 77)

References

1. Mesch, J., Wallin, L., Nilsson, A.-L., and Bergman, B. (2012). Dataset. Swedish Sign Language Corpus project 2009–2011 (version 1).
2. Mesch, J., Rohdell, M., and Wallin, L. (2015). Annotated files for the Swedish Sign Language Corpus. Version 3.
3. Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. Computing In Science & Engineering, 9(3):90–95.

Poster presented at
The 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining
May 28, Portorož, Slovenia