Visualizing *Lects* in a Sign Language Corpus: Mining Lexical Variation Data in Lects of Swedish Sign Language

Carl Börstell[1] & Robert Östling[2]
[1]Dept. of Linguistics, Stockholm University
S-106 91 Stockholm, Sweden
calle@ling.su.se
[2]Dept. of Modern Languages, University of Helsinki
FI-00014 Helsinki, Finland
robert.ostling@helsinki.fi

Abstract

In this paper, we discuss the possibilities for mining lexical variation data across (potential) *lects* in Swedish Sign Language (SSL). The data come from the SSL Corpus (SSLC), a continuously expanding corpus of SSL, its latest release containing 43 307 annotated sign tokens, distributed over 42 signers and 75 time-aligned video and annotation files. After extracting the raw data from the SSLC annotation files, we created a database for investigating lexical distribution/variation across three possible lects, by merging the raw data with an external metadata file, containing information about the age, gender, and regional background of each of the 42 signers in the corpus. We go on to present a first version of an easy-to-use graphical user interface (GUI) that can be used as a tool for investigating lexical variation across different lects, and demonstrate a few interesting finds. This tool makes it easier for researchers and non-researchers alike to have the corpus frequencies for individual signs visualized in an instant, and the tool can easily be updated with future expansions of the SSLC.

Keywords: Swedish Sign Language, sign language, corpus, lexical variation, data visualization, interface

1. Introduction

Lexical variation is a topic that has received a fair amount of attention in sign language linguistics (Lucas, 2006; Schembri and Johnston, 2012). However, it is only recently that sign language corpora have come about, meaning that the study of lexical variation now has access to a larger, more varied dataset than ever before. To date, sign language corpora are available for a number of sign languages (see Börstell et al. (2014b) for a non-exhaustive list) with more under way, but their size in terms of tokens is far from that of spoken languages. Although sign language corpora are not big by token count, they do require a substantial space for data storing, since sign language data is necessarily recorded in video format. Perhaps because of this, most sign language corpora are not easily accessible to non-researchers, seeing as they often require downloading of heavy bundles of video and annotation files, and mostly render corpus search results in a strictly numerical form (i.e. without any type of graphical visualization). Thus, with this study, we looked to mine and re-compile the data from a sign language corpus by adding signer metadata for sociolinguistic factors known to interact with lexical variation directly into a searchable database, but also create a simpler graphical user interface (GUI) that directly visualizes the output of any corpus search without depending on video files, in an attempt to make the corpus data more accessible in a lightweight format.

2. Background

2.1. Lexical Variation

Variation in sign language has been a topic researched since the early days of sign language linguistics (Lucas, 2006). The specific focus of the research has varied, with different studies looking at variation on levels ranging from sublexical to discourse units, and the explanations for which factors are responsible for the variation have included region, age, gender, and ethnicity (Bayley et al., 2015). A well-known work on the issue of lexical variation is the book *What's your sign for* PIZZA*?* (Lucas et al., 2003), which presents the findings of a large-scale project on lexical variation in American Sign Language (ASL) across the United States. More recently, with the advent of true sign language corpora, some studies have been conducted looking at variation in British Sign Language (BSL), such as Fenlon et al. (2013) investigating the contextual and sociolinguistic factors affecting the shape of the 1-hand configuration, and Stamp et al. (2014) investigating the regional variation of color signs. This second study made use of corpus data, but specifically a subset of corpus data consisting of lexical items elicited using word lists. For Swedish Sign Language (SSL), the only previous study concerning variation is Nilsson (2004), which looked at the form variations of the first-person pronoun PRO1 in discourse data, although not from a sociolinguistic perspective. However, the online dictionary of SSL (Björkstrand, 2008) does contain some information about sociolinguistic features of signs, such as regional distribution of particular signs, as well as signs seen as old-fashioned, but this dictionary is not linked to, or based on, corpus data (Mesch et al., 2012a).

2.2. The SSL Corpus

The SSL Corpus (SSLC) is a corpus of naturalistic, dyadic signing of Swedish Sign Language. The SSLC data were collected over three years (2009–2011), and comprises 300 video recordings distributed over 42 signers (Mesch et al., 2012b), with the signers selected in order to approximate a balanced and representative sample in terms of age groups,

genders, and regional distribution (Mesch, 2012; Mesch et al., 2012a; Wallin and Mesch, 2015).[1] To date, 75 (i.e. 25%) of the video files have been edited, glossed, and translated (Mesch et al., 2015). The video files are annotated using the ELAN software, producing annotation files (.eaf) that are underlyingly XML files, allowing for multiple annotation tiers time-aligned to a media file (Wittenburg et al., 2006). Currently, the SSLC annotation files consist of two main tier types: sign gloss annotations; and Swedish translations. The only segmentation that has been done for the SSL data is on the lexical level, with sign glosses being entered into annotation cells corresponding to the duration of individual signs on the time-axis, though the possibility of introducing a syntactic/prosodic segmentation has been investigated (Börstell et al., 2014a). Apart from the sign glosses—i.e. the labels uniquely identifying each sign in the corpus (Mesch and Wallin, 2015; Wallin and Mesch, 2015)—the SSLC has also recently been tagged with parts of speech, using a semi-automatic tagging procedure (Östling et al., 2015).

3. Methodology

3.1. Aim

In the SSLC, the participants are grouped according to three different variables, as provided by the signer metadata documented during the collection of the primary (i.e. sign language) data. These three group variables are: (a) **Region**, the regional affiliation of the signers based on the *landsdelar* (lit. 'country parts') of Sweden—Norrland, Svealand, and Götaland; (b) **Age group**, the categorization of signers into six age groups; and (c) **Gender**, female or male.[2] Furthermore, the individual files in the SSLC are categorized into three different text types—conversation, narrative, and presentation, respectively. However, the signer metadata and the text type information are not available directly in the SSLC annotations to be used with ELAN as the user interface. The raw metadata files themselves contain information about individual signers and are thus not publicly available. In this project, we used the metadata files to match the anonymous signer-IDs to each group variable, such that the resulting database does not contain neither personal details about individual signers, but rather sign frequency data for groups of signers (or text types).

The aim of this work was two-fold: firstly, we wanted to link the group variables of the signer metadata directly to the lexical data in the SSLC, storing it as a type of database; secondly, we wanted to create methods for mining interesting data, either by using computational search methods for research purposes, or as an custom-built, easy-to-use interface for which researchers and non-academics alike could search this database and get instant visual representations of the lexical frequency distributions across all group variables.

In this paper, we also make a short evaluation of the data and our search interface, and provide a few examples of how the tool can be used for quick visualizations of lexical distributions.

3.2. Data

For this study, we used the data from the latest version of the SSLC. This version comprised 75 annotation files, consisting of 43 307 sign tokens. However, many tokens are tagged with any of the suffixes `@x` or `@z`, marking that the sign gloss is uncertain or the sign unidentifiable (Wallin and Mesch, 2015), hence such signs were excluded from our dataset. Thus, we arrived at a dataset of 39 733 sign tokens, distributed over 4 676 sign types. However, since the SSLC is still being annotated, the corpus is not (yet) balanced in terms of the distribution of annotated tokens within each group variable in the metadata. In order to account for the imbalance in token frequency across groups, we based all results on relative frequencies (see 3.2.1. and 3.3.). The distribution of sign tokens within each of the three group variables is given in Tables 1, 2, and 3, and the distribution of sign tokens across text types is given in Table 4.

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

Table 1: Distribution of signers and tokens according to region.

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

Table 2: Distribution of signers and tokens according to age.

Gender	Signers	Tokens
female	20	15 862
male	22	23 871

Table 3: Distribution of signers and tokens according to gender.

It should be noted that the crude division of regions into *landsdelar* does not correspond to Deaf schools, for which there have traditionally been seven: one in Norrland; four in Svealand; and two in Götaland (see Figure 1).[3]

[1] `http://www.ling.su.se/teckenspraksk orpus`

[2] Though additional metadata such as educational background and age of onset for sign language acquisition have been documented during the data collection, this information was not available to us for each signer as the other metadata, thus restricting our study to the selected variables.

[3] NB: Some cities had more than one Deaf school.

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

Table 4: Distribution of files and tokens according to text type.

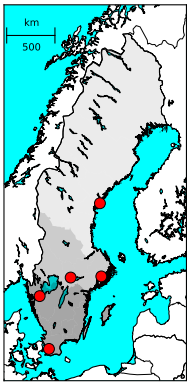

Figure 1: The *landsdelar* of Sweden—Norrland (light gray), Svealand (gray), Götaland (dark gray)—with the locations of the deaf schools (red dots).

3.2.1. Extracting and reading the relevant data

All sign data were extracted from the ELAN annotation files and then matched to the external metadata on signers, so that we end up with a 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:

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

3.3. Identifying Unevenly Distributed Signs

Rather than just obtaining the social and geographic distribution of particular signs, we are also interested in *finding* the signs that are used significantly more often by some groups than by others.

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 $\alpha = 1$.

3.4. Constructing a Visual Interface

For the visual interface, we wrote a program that took the input sign objects read from the datafile and waited for a user input, in this case asking for a specific sign gloss to be plotted. When a sign gloss was entered into the interface, the program would plot it using the Matplotlib module (Hunter, 2007). A bar chart was subsequently created for each of the group variables—region, age group, and gender—as well as one for text type, presenting the sign's relative frequencies in tokens per 100 signs. This interface was implemented as a web script and made accessible online.[4]

4. Results and Evaluation

4.1. Evaluating the Data Visualization

The obvious problem with the SSLC data is its small scale. Even after balancing out the skewed token distribution within variables, the fact remains that ≈40 000 tokens is insufficient for estimating reliable statistics for anything but the most high-frequent items. The most frequent sign in the SSLC is PRO1 (Börstell et al., Submitted). The graphs in Figure 2 show the distribution of relative token frequencies for PRO1 across each group variable.

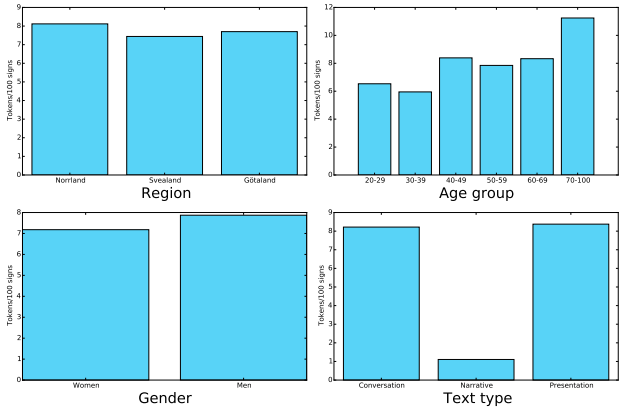

Figure 2: The distribution of the sign PRO1 (n = 3 018).

As is visible from these graphs, the relative frequencies are more or less even for each group variable. This is to be expected from a sign that is highly frequent. Unsurprisingly, it is for text type that the sign PRO1 shows a skewed distribution, with the sign being relatively uncommon in the narrative texts, which in the SSLC are mainly elicited narratives (as opposed to self-experienced narratives). However, we also wanted to see if specific items do exhibit a distribution that reflects *lectal* lexical variation.

For region, we take the example of the sign ÄLG(Jb) ('moose'), which is listed as a regional northern sign in the SSL dictionary (Björkstrand, 2008).[5] Figure 3 shows the distribution of the seven tokens found for this sign, supporting the claim that this sign is associated with Norrland, with

[4] `http://mumin.ling.su.se/cgi-bin/sslects.py`

[5] Suffixed tags in round brackets indicate a specific form for meanings for which there are sign variations. The letters within the brackets describe the handshape.

all tokens coming from this region. As for the identification of unevenly distributed signs, the sign ÄLG(Jb) does in fact appear in the top (15th place) of signs with an uneven distribution across regions, showing that the method correctly identifies this sign as a sign with a skewed regional distribution (in this case, being associated with a specific region, viz. the north). Unfortunately, the non-northern sign for 'moose' (ÄLG(5)) is not yet attested in the SSLC.

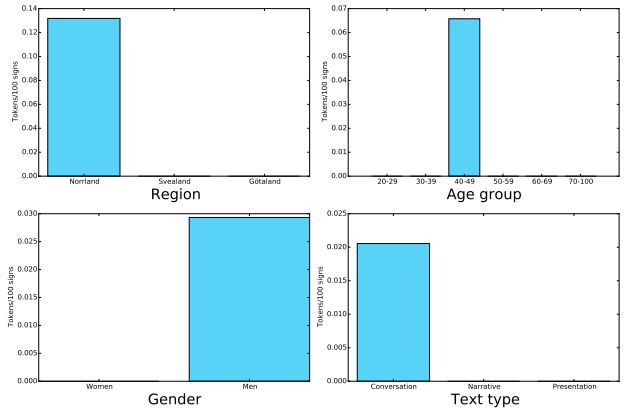

Figure 3: The distribution of the sign ÄLG(Jb) ('moose') (n = 7).

For age, there are not many signs marked as typical for younger or older signers in the SSL dictionary that also occur in the SSLC. However, there are signs generally perceived as more typical to a certain generation or age group. One such sign is TYP@b ('kind of', lit. 'type'), which is said to be more typical among younger signers, as it is a borrowing from spoken Swedish (where it is also associated with younger speakers).[6] Figure 4 appears to support this idea, with the 77 tokens of the sign being largely distributed over the younger age groups. Furthermore, the sign TYP@b appears in the very top (5th place) of signs with an uneven distribution across age groups, showing that the method again correctly identifies this sign as a sign with a skewed distribution (in this case, being associated with younger signers).

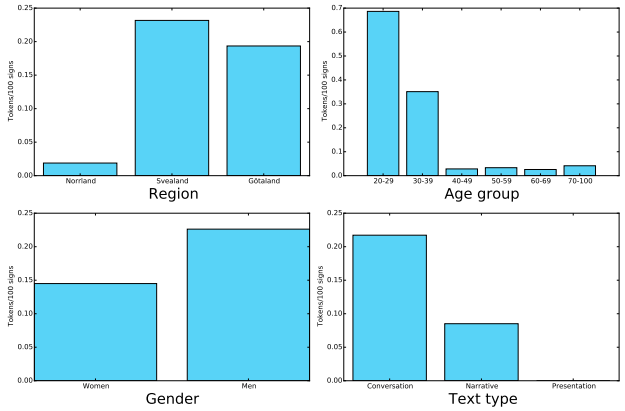

Figure 4: The distribution of the sign TYP@b ('kind of') (n = 77).

Finally, for gender, there is one pair of signs often claimed to be in a gendered complementary distribution, namely the signs SNYGG@b and SNYGG(H), both meaning 'attractive', but the former said to be used by women and the latter by men. Figures 5 and 6 seem to support this, although it should be noted that the graphs are based on very few absolute tokens (3 and 1, respectively)—also, the few tokens make these signs hard to identify statistically as showing an uneven distribution.

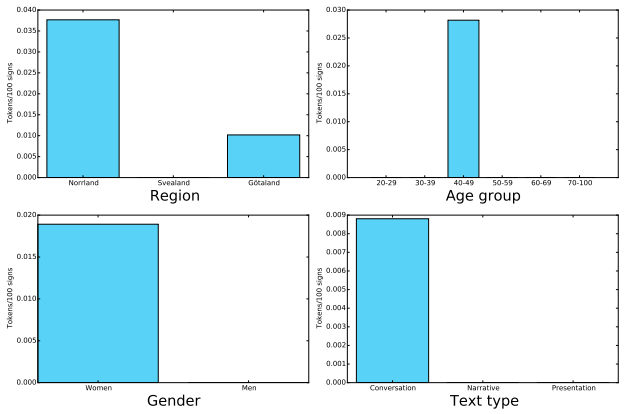

Figure 5: The distribution of the sign SNYGG@b ('attractive') (n = 3).

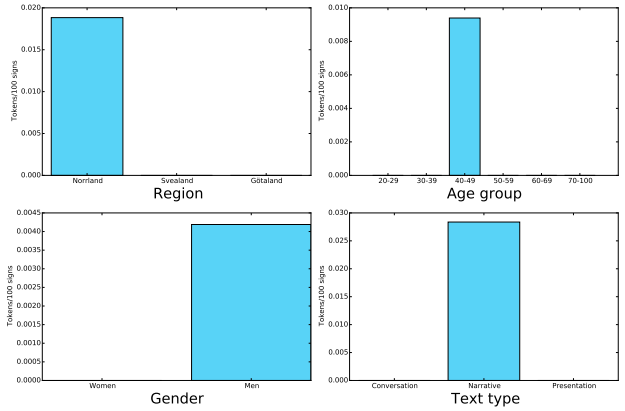

Figure 6: The distribution of the sign SNYGG(H) ('attractive') (n = 1).

4.2. Evaluating the Method Identifying Unevenly Distributed Signs

The output of the method identifying unevenly distributed signs (described in 3.3.) shows potential. Although the SSLC suffers from a quite limited amount of data in terms of token size—as do all sign language corpora—the method correctly identifies the signs that we selected from prior knowledge (albeit anecdotal, in some cases) about their lectal distribution. Thus, it shows potential as a method of automatically identifying signs with a skewed distribution based on lectal lexical variation. However, with the limited amount of data available in the current version of the SSLC, many signs identified as showing a skewed distribution are, as confirmed after a manual check, merely skewed due to conversation topics of individual signers rather than

[6] The tag @b indicates that the sign is fingerspelled.

as cases of lexical variation (i.e. a certain sign is skewed towards a specific group because of a single signer talking about a related topic and making it seem as though the group "overuses" the sign). In some cases, this points to interesting differences in conversation topics, as with the sign MAN(H) ('husband') being heavily skewed towards being used by female signers, whereas the sign FRU ('wife') is skewed towards male signers. Similarly, certain toponyms are, unsurprisingly, used more by signers from that region. Nonetheless, with an expansion of the corpus, we are optimistic of the possibilities that this method brings.

5. Conclusion

In this study, we have described the procedure of extracting data from raw corpus annotations, matching them to signer metadata, and constructing a database for investigating lexical distribution (and possible variation) based on the factors region, age, and gender, as well as the creation of a web-based data visualization tool that we have made publicly available, for researchers and non-researchers alike. We also utilize a method for automatically identifying uneven distributions, and find that it correctly identifies several signs that are expected to exhibit a skewed distribution based on lectal variation. Though the SSLC is still too small to do any large-scale investigations of lexical variation—simply based on the fact that the there are too few tokens as well as signers—we can still visualize some of the known or previously assumed cases of lexical variation in SSL, and more instantly than previously possible thanks to our database and GUI. With the expansion of the SSLC in terms of data, the database will get richer, and thus more adequate for research purposes on lexical variation. A larger corpus would also give the automatic identification of unevenly distributed signs a better dataset on which to conduct its calculations, for which we are confident it could serve as a useful tool for pinpointing interesting sociolinguistic variation. Also, making the web interface available online with direct access to and visualization of the SSLC data should make the corpus as a resource more available to the general public and more specifically the SSL community.

6. Acknowledgments

We wish to thank Johanna Mesch for providing us with the metadata files from the SSLC.

7. Bibliographical References

Bayley, R., Schembri, A. C., and Lucas, C. (2015). Variation and change in sign languages. In Adam C. Schembri et al., editors, *Sociolinguistics and Deaf Communities*, pages 61–94. Cambridge University Press, Cambridge.

Björkstrand, T. (2008). Swedish Sign Language Dictionary online. `teckenspraksleksikon.su.se`.

Börstell, C., Mesch, J., and Wallin, L. (2014a). Segmenting the Swedish Sign Language Corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In Onno Crasborn, et al., editors, *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel [Language Resources and Evaluation Conference (LREC)]*, pages 7–10, Paris. European Language Resources Association (ELRA).

Börstell, C., Sandler, W., and Aronoff, M. (2014b). Sign Language Linguistics. In Mark Aronoff, editor, *Oxford Bibliographies Online: Linguistics*. Oxford University Press.

Börstell, C., Hörberg, T., and Östling, R. (Submitted). Distribution and duration of signs and parts of speech in Swedish Sign Language.

Fenlon, J., Schembri, A., Rentelis, R., and Cormier, K. (2013). Variation in handshape and orientation in British Sign Language: The case of the '1' hand configuration. *Language and Communication*, 33(1):69–91.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

Lucas, C., Bayley, R., and Valli, C. (2003). *What's your sign for PIZZA?* Gallaudet University Press, Washington, DC.

Lucas, C. (2006). Sign language: Variation. In Keith Brown, editor, *Encyclopedia of Language & Linguistics*, number 1993, pages 354–358. Elsevier, Oxford.

Mesch, J. and Wallin, L. (2015). Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, 20(1):103–121.

Mesch, J., Wallin, L., and Björkstrand, T. (2012a). Sign Language Resources in Sweden: Dictionary and Corpus. In Onno Crasborn, et al., editors, *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]*, pages 127–130, Paris. European Language Resources Association (ELRA).

Mesch, J., Wallin, L., Nilsson, A.-L., and Bergman, B. (2012b). Dataset. Swedish Sign Language Corpus project 2009–2011 (version 1).

Mesch, J., Rohdell, M., and Wallin, L. (2015). Annotated files for the Swedish Sign Language Corpus. Version 3.

Mesch, J. (2012). Swedish Sign Language Corpus. *Deaf Studies Digital Journal*, 3. `http://dsdj.gallaudet.edu/index.php?issue=4&section_id=2&entry_id=128`.

Nilsson, A.-L. (2004). Form and discourse function of the pointing toward the chest in Swedish Sign Language. *Sign Language & Linguistics*, 7(1):3–30.

Östling, R., Börstell, C., and Wallin, L. (2015). Enriching the Swedish Sign Language Corpus with part of speech tags using joint Bayesian word alignment and annotation transfer. In Beáta Megyesi, editor, *Proceedings of the 20th Nordic Conference on Computational Linguistics (NODALIDA 2015), NEALT Proceedings Series 23*, pages 263–268, Vilnius. ACL Anthology.

Schembri, A. and Johnston, T. (2012). Sociolinguistic aspects of variation and change. In Roland Pfau, et al., editors, *Sign language: An international handbook*, pages 788–816. De Gruyter Mouton, Berlin/Boston, MA.

Stamp, R., Schembri, A., Fenlon, J., Rentelis, R., Woll, B., and Cormier, K. (2014). Lexical variation and change in British sign language. *PLoS ONE*, 9(4).

Wallin, L. and Mesch, J. (2015). Annoteringskonventioner för teckenspråkstexter. Forskning om teckensprk (FOT-rapport) XXIV.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.