How to Approach Lexical Variation in Sign Language Corpora

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

Abstract

Looking at lexical frequency and, by extension, lexical variation is often among the first objectives after compiling a sign language corpus, since the only prerequisite is existing sign gloss annotations. However, measuring lexical frequency in a theoretically and statistically meaningful way can be a challenge. In this paper, I provide an overview of how to approach lexical variation in sign language corpora. The aim is to show ways of tackle lexical variation from different angles, from data collection to statistics and visualization, and how to motivate choices based on the data available and the research goals, thus serving as a practical guide for sign language corpus research. Drawing from previous work by different sign language corpus project teams, various approaches to measuring lexical variation are illustrated with data from the Swedish Sign Language (STS) Corpus, with examples that can easily be adapted to any sign language corpus.

Keywords: sign language, corpus, lexical frequency, variation, sociolinguistics

1. Introduction

The number of available sign language corpora in the world is constantly increasing, and many corpora of individual sign languages are also growing in size (see, e.g., Kopf et al., 2021, 2022, 2023; Fenlon and Hochgesang, 2022). The first step of annotating a sign language corpora is often to segment and annotate individual *signs* in the data (Johnston, 2010). With annotation of individual lexical items (i.e. *signs*), an easy first exploration of the corpus data is to look at lexical frequencies – which signs are used the most, by whom and in what context? Lexical frequency has been studied for a number of sign languages already, with datasets of varying size (e.g., Morford and MacFarlane, 2003; McKee and Kennedy, 2006; Johnston, 2012; Fenlon et al., 2014; Börstell et al., 2016).

It is well known that the distribution of words in language(s) is extremely skewed, with a small number of words occurring frequently but most words occurring fairly rarely (Zipf, 1935). This skew in token frequencies needs to be taken into account when looking at lexical frequency, and makes it more challenging to look at lexical variation, especially in smaller corpora – and most sign language corpora are still relatively small. Thus, there are several aspects to consider when investigating lexical variation within individual sign languages, and I will in the following provide concrete examples of approaches taken in previous work, and opportunities and issues that come with them. While mostly illustrated with examples from the Swedish Sign Language (STS; *svenskt teckenspråk*) Corpus (Öqvist et al., 2020), the methods could be applied to any sign language corpus. Finally, the paper concludes with a summarized list of benefits and downsides to different approaches and metrics.

2. Data and Methods

For the examples in this paper, I use data from the STS Corpus (Öqvist et al., 2020) presented in different ways depending on the approach to investigating lexical variation.

The STS Corpus data (Mesch et al., 2012) was retrieved from *The Language Archive* (`https://archive.mpi.nl/tla/`) in July 2023 and consists of 189,679 sign tokens across 298 annotation files and 42 signers.

The data was retrieved, processed and visualized using R v4.3.2 (R Core Team, 2023) and the packages `patchwork` v1.2.2 (Pedersen, 2022), `scales` v1.2.1 (Wickham and Seidel, 2022), `signglossR` v2.2.4 (Börstell, 2022), `tidylo` v0.2.0 (Schnoebelen et al., 2022) and `tidyverse` v2.0.0 (Wickham et al., 2019).

Simulated example data and code for calculating and plotting frequencies and variation can be found at: `https://github.com/borstell/r_functions/blob/main/plotting_corpus_variation.R`

3. Approaches to Lexical Variation

In order to look at lexical variation in any language, one needs to have enough data, such that it covers the relevant variables involved in variation – whether, e.g., age, gender or geographic belonging (Bayley et al., 2015). While variation can be studied separately from a corpus, through interviews and elicitation with the signing community directly (Lucas et al., 2009; Fisher et al., 2016; Safar, 2021) or indirectly through distributed surveys online (Kimmelman et al., 2022), the focus in this paper is data collected within a sign language corpus project. However, even within corpus projects, similar alternative data collection approaches have been used. For example, several projects have included a targeted lexical elicitation task as part of the corpus data collection – i.e. tasks alongside the collection of naturalistic conversational data. The targeted interview/elicitation approach facilitates comparisons of signs in domains known for variation, such as color terms in British Sign Language (BSL) (Stamp et al., 2014) and German Sign Language (DGS) (Langer, 2012), as it results in a larger target sample. Some corpus projects have also adopted a method of crowdsourcing signs and lexical variation as well as perceptions about variation and usage of already documented variants through direct or online community involvement (Kankkonen et al., 2018; Wähl et al., 2018; Hanke et al., 2020). Targeted elicitation tasks are suitable for comparing variation between different groups with regard to specific items/domains since it results in a higher number of data points per item and a better coverage with many signers being represented (cf. Section 3.4). However, elicited data will not be directly comparable to other items/domains found only in the conversational portion of the corpus data, as the distribution of occurrences will look very different.

In the following sections, I will mainly focus on how to approach and measure lexical variation in naturalistic, conversational corpus data.

3.1. Counts: "How Many Have You Got?"

As was mentioned in the introduction, the Zipfian distribution of lexical items in a corpus means that token frequencies will be extremely skewed: some items are very frequent whereas most items are very infrequent. Thus, raw counts of frequencies are often quite uninformative as they are only meaningful for a particular corpus (or, corpus size) and will have a huge range between items in the upper vs. lower end of the frequency span. For example, saying that there are 10,846 occurrences of PRO1 (first-person pronoun), 414 occurrences of TYP@b ('kinda'; fingerspelled) and 7 occurrences of ÄLG(Jbt) ('moose') in the STS Corpus is quite meaningless unless they are compared to the total number of tokens in the corpus (n=189,679) or possibly to each other. Nonetheless, in the online STS Dictionary (teckenspråkslexikon, 2023), the only currently available information about corpus frequencies of dictionary entries is raw corpus frequencies, available for those entries that have been linked to the corpus (cf. Mesch et al., 2012). This was why we in Börstell and Östling (2016) developed a search tool for exploring meaningful lexical frequencies and variation in the STS Corpus by rather focusing on *relative* frequencies within and across groups of signers or text types, which is discussed further in Section 3.2.

3.2. Proportions: "It's All Relative!"

One way of approaching *relative frequencies* in a corpus is to simply say how many times an item occurs relative to the total, usually rescaled to arrive at a more interpretable number, e.g., occurrences per 100,000 tokens. This means that we could reformulate the frequencies in Section 3.1 and say that PRO1 occurs 5,718 times per 100,000 tokens, TYP@b 218 times per 100,000 tokens and ÄLG(Jbt) about 4 times per 100,000 tokens. This metric is more intuitive and more useful as it is comparable across corpora or subcorpora of different sizes. However, it does not address the issue of variation, as it does not differentiate where the tokens come from within the corpus.

In Börstell and Östling (2016), we identified the need to obtain relative frequencies of signs in the STS Corpus with attention to sociolinguistic variation. Thus, we developed an online search tool[1], parallel to the STS Corpus, that would display relative frequencies within different grouping variables that were likely to exhibit variation in lexical frequency distribution: age, gender, region and text type. Thus, frequencies were relative to the total number of tokens by subgroup. This allowed for comparisons across groups of different sociolinguistic variables very easily. For example, there was anecdotal evidence of the sign TYP@b ('kinda'; fingerspelled) being more frequent among younger signers, and this was corroborated with our search tool illustrating relative frequencies, showing that the sign is much more frequent among younger age groups. Figure 1 shows the same pattern in the current version of the STS Corpus, with over twice the number of tokens annotated compared to what was reported in Börstell and Östling (2016).

One potential feature that was not available in the search tool by Börstell and Östling (2016) was directly comparing relative proportions between multiple forms for the same meaning. Many sign languages exhibit variation in specific domains (e.g., numerals and color terms), such that the same meaning may be expressed by multiple forms. Such variation may consist of either completely different lexical items or phonological variants of a similar base (or iconic mapping), sometimes with sociolectal differences in their distribution (see, e.g., McKee et al., 2011; Langer, 2012; Stamp et al., 2014; Wähl et al., 2018; Safar, 2021; Lutzenberger et al., 2021, 2023). A rather straightforward way of comparing differences in the distribution of sign variants for the same meaning is to compare the

[1] The tool, *SSL-lects*, has been offline for a few years due to server replacements and anonymization concerns with the raw STS Corpus data, but there have been plans to integrate a similar tool directly in the online corpus and dictionary resources.

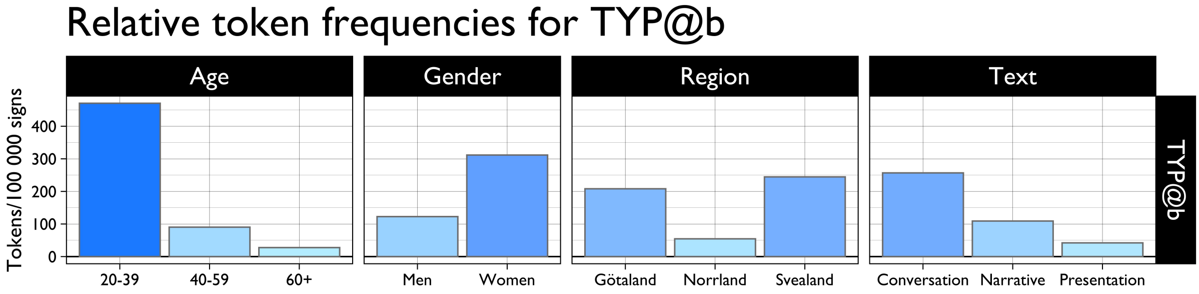

Figure 1: Relative frequencies of the sign TYP@b ('kinda'; fingerspelled) across sociolinguistic groupings in the STS Corpus.

proportion of tokens they each have relative to their combined total, distributed across the sociolinguistic groupings of interest. For example, Figure 2 shows the relative proportions between a one- and two-handed (phonological) variant of the sign for '(an)other' in STS. Based on the relative proportions alone, it is quite clear that the one-handed variant is more common overall but that the oldest signers have a slight preference for the two-handed variant.

Searching for lexical variants or any signs with related meanings is, however, not necessarily straightforward. Glosses are often selected on the basis of a written word with similar meaning, but semantic extension and polysemy may mean that signs are related without sharing a similar gloss (cf. Johnston, 2010; Ormel et al., 2010). Because of this, searching for variants or related signs may already require some knowledge about the language as well as the annotation conventions of the corpus (e.g., how glosses are used).[2]

With these approaches, one issue is that they mainly target specific signs (individually or paired) that we already suspect may display some type of sociolectal variation in their distribution. In Section 3.3, we will see how other metrics can be used to identify interesting distributional variation directly from the data.

3.3. Ratio: "What Are the Odds?"

Looking at frequencies relative to sociolinguistic groupings made it possible to visualize variation differences for items suspected to exhibit variation. However, in Börstell and Östling (2016), we also wanted to find ways of *identifying* potential variation-exhibiting items without necessarily knowing about them through previous – often anecdotal – evidence. Thus, we applied a Bayes factor approach, calculating distributions relative to token counts among the same sociolinguistic groupings and could identify certain signs that were overrepresented in some subgroup. While this metric was not available in the search and visualization tool itself, it could be an interesting addition since it is possible to see both positive and negative values, and as such the directionality of frequency: higher or lower than expected. In Figure 3, a similar implementation is used in a visualization, but with weighted log odds using a Bayesian prior estimated from the data itself, which accounts for differences in sampling variability (see Monroe et al., 2008; Schnoebelen et al., 2022). With this approach, we can confirm that age is a major factor in the distribution of tokens, with TYP@b being skewed towards younger age groups. The gender distribution here is less informative, seeing as the STS Corpus has more women in the younger age groups and more men in the older age groups. Somewhat surprisingly, the text type distribution in Figure 3 is switched compared to Figure 1, which is a consequence of the informative prior taking the sampling variability into account – using an uninformative prior will instead correspond more closely to the relative frequencies in Figure 1, albeit on a different scale.

A log odds approach was also taken by Stamp et al. (2014), who looked at larger groups of signs in specific domains (e.g., numerals and color terms) to see differences in the use of traditional (often regional) signs for concepts in these domains, finding that age was an important factor, with older signers being more likely to use the traditional signs with regional variation, while younger signers exhibit less variation, pointing to dialectal leveling.

3.4. Spread & Coverage: "The One with All the Tokens"

As has been mentioned earlier, lexical variation in corpus data can be a challenge due to the low token frequency of most lexical items even in large corpora, which means it is difficult to find items that occur across, e.g., sociolinguistic groupings in spontaneous, conversational data. This is why several corpus projects have opted to include an explicit lexical elicitation task as part of the data collection – this is, however, not the case for the STS

[2] I thank a reviewer for raising this point.

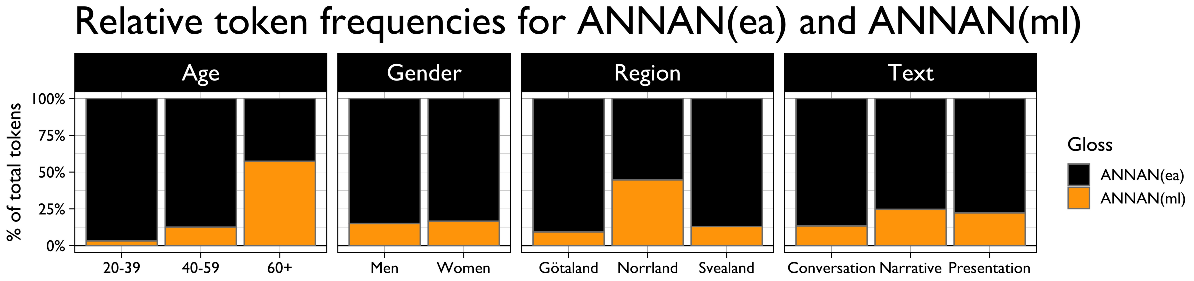

Figure 2: Relative proportions of the signs ANNAN(ea) (‘(an)other’; one-handed) and ANNAN(ml) (‘(an)other’; two-handed) across sociolinguistic groupings in the STS Corpus.

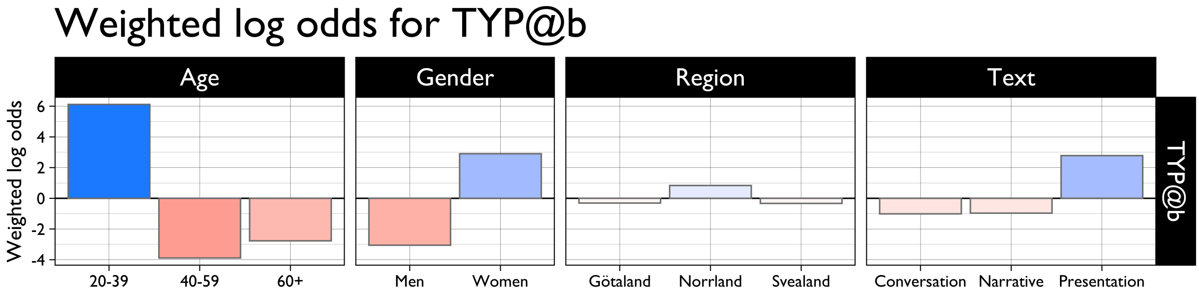

Figure 3: Weighted log odds of the sign TYP@b (‘kinda’; fingerspelled) across sociolinguistic groupings in the STS Corpus.

Corpus. It also means that any grouped metric, such as relative frequencies per age group, should also include a measure of spread across signers, at least for low-frequency items – that is, how many signers in the data use the sign at least once (i.e. signer coverage). As an example, in Börstell and Östling (2016) we discussed the known regional variation between two signs for ‘moose’ in STS: one that depicts the horns (considered the more general and widespread sign) and one that depicts the snout/muzzle (considered a northern variant). In our paper, we noticed that only the “northern” variant was present in the data, found in the northern (Norrland) region as expected. However, not only is it impossible to establish the source of variation, due to the lack of tokens for the other variant, the signer coverage was very poor, with all occurrences being produced by a single signer. In the current, larger STS Corpus dataset, the pattern is unfortunately still the same, with only one of the two variants being produced with 7 occurrences in the whole corpus, all produced by the same signer: an older man from Norrland. Since it is clearly impossible to generalize from a single signer, it can be wise to include signer coverage in a visualization or simply checking the distribution across signers when looking at any token frequencies, but particularly lower ones. Figure 4 shows an example of the signer coverage for three signs, PRO1, TYP@b and ÄLG(Jbt), with dots representing each of the 42 signers in the STS Corpus, where the blue ones represent signers with attested tokens (darker means a higher proportion of total tokens) and grey ones represent signers without attested tokens. As this figure shows, highly frequent signs such as PRO1 will have a large and fairly even spread across signers, whereas signs such as ÄLG(Jbt) cannot be generalized in their usage despite having more occurrances (n=7) than the global median number of tokens (n=1) in the whole corpus.

3.5. Topics & Representativeness: “What Are We Talking About?”

Small(er) corpora, such as most sign language corpora, are quite susceptible to idiosyncrasies skewing the data. For example, multiple sign language corpora have included the same elicitation tasks to elicit narrative texts. Because of this, it comes as no surprise that signs for concepts such as ‘snowman’ and ‘frog’ may be much more frequent than expected from any regular conversation within the deaf community, simply due to the influence of the contents in the elicitation stimuli. Specific topics, and consequently associated words/signs, will always be subject to sampling procedures in the data collection, regardless of the type of corpus. Since sign language corpora involve members of the deaf or signing community, it is expected that concepts such as ‘deaf’ and ‘hard-of-hearing’ may be orders

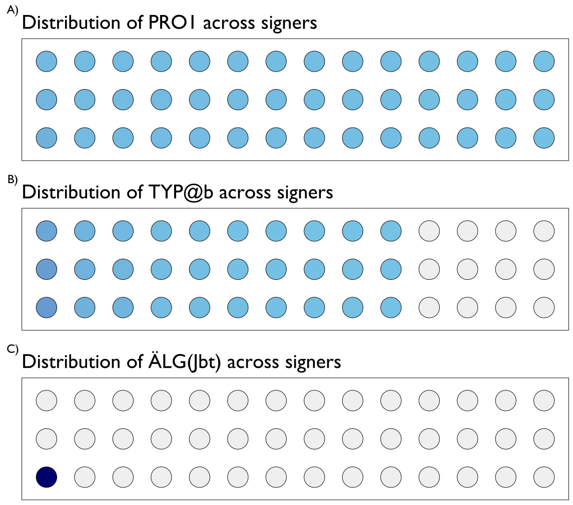

Figure 4: Distribution of tokens across signers for three signs : A) PRO1; B) TYP@b; C) ÄLG(Jbt). Each dot represents a signer; blue-filled dots show signers with attested tokens, with the darkness of the fill color representing proportion of total tokens.

of magnitude more frequent in a sign language corpus than any spoken language corpora. This is not a problem as it directly reflects themes and topics that are relevant in the community, but other topics that are introduced due to targeted tasks in the data collection procedure will often result in some lexical items being overrepresented in a way that is not representative of issues of particular significance to the community at large.

While the use of similar topics/content across sign language corpora is a great resource for cross-linguistic work on, e.g., grammatical and discourse structure (cf. Ferrara et al., 2022), it inadvertently leads to a skew in particular lexical items, which should be taken into account when looking at lexical frequency and variation.

3.6. Conventions & Conventionalization: “That’s Not Even a Word!”

As discussed in more detail by Langer et al. (2016), not all tokens are necessarily representative of the regular usage of the individual signer who produced them. For example, some signs are used metalinguistically, in the sense that sign variants are produced i) to illustrate how *others* sign something, ii) as a direct copy of the interlocutor’s sign choice, or iii) to emphasize how the signer themself does *not* sign (Langer et al., 2016, 140). Similarly, signs may also be produced in a manner different from established lexical items in the language, such as being produced in a context showing, e.g., how non-signers or learners are attempting to sign or gesture (Langer et al., 2016, 141).

Furthermore, Langer et al. (2016, 141) also mention slips of the hand (i.e. errors in producing the target sign form). This is a question that very much concerns the annotation process in building a corpus, whether to mark accidental deviations/errors explicitly or to simply annotate target forms (if identifiable). In the Auslan Corpus, the procedure for fingerspelling has been to annotate both target form and actual realization in the same sign gloss (Johnston, 2019, 45). This way, the researcher could choose whether to focus on target forms or actual realization, which in itself would be relevant for lexical variation. In the STS Corpus, uncertain or interrupted glosses have been marked with special tags (“@z” and “@&”, respectively), but there is also a dedicated tag for so-called *home-made signs* (“@hg”), which are not considered established signs of the community as a whole (Mesch and Wallin, 2021, 25–26). While such signs make great candidates for a detailed analysis of lexical variation, they will not be generalizable to the larger community. Thus, a researcher interested in investigating lexical variation would need to know the annotation conventions of the specific corpus to be able to accurately match sign glosses to actual forms, and to motivate their reasons for including or excluding specific items.

4. Discussion & Conclusions

In this paper, I have given a brief introduction to the question of how to approach lexical variation in sign language corpora. The goal has been to provide anyone interested in doing research on a sign language corpus with concrete examples of issues to consider both theoretically and practically. How the data is annotated will directly influence what can be researched, and which analysis method is applied will affect the usefulness and interpretation of the results. For example, can related signs (e.g., lexical variants) be matched and compared based on glosses alone? Can glosses and search patterns easily distinguish phonological from lexical variants of the same meaning? Are we able to search lemma forms but still account for the frequency of different morphological forms (e.g., inflections) of that lemma? Can we easily attribute tokens to individual signers, and group signers and files by metadata features? These issues are concerns of the researcher using and searching the corpus as much as of the developer of the corpus resource itself, and require users to be familiar with both the language and the corpus conventions.

Unfortunately, few sign language corpora have integrated tools for directly querying a database and receiving a table or visualization of the search results in a meaningful way, such as regional variation visualized on a map (however, see Hanke, 2016; Hanke et al., 2023). Since lexical variation is an important part in applied areas such as language teaching and interpreting, it would be useful to incorporate simple search tools into the sign language corpus resources – see Isard and Konrad (2022) and Isard and Konrad (2023). Such tools could display not only raw search hits of sign glosses, but also relevant summaries of results presented as tables, graphs or maps, based on variables and metrics selected by the user. In the case of the STS Corpus (Öqvist et al., 2020), the current online interface with streamed videos and glosses is a great resource for teachers and students, but it unfortunately does not allow the user to query the database about relative frequencies or proportions between variants, nor export raw search results to be investigated externally, which renders it less accessible to the corpus linguist.

For the researcher who wants to approach questions of lexical frequency and variation in a sign language corpus, here are some points to consider when retrieving, interpreting and reporting the results:

- **Raw frequency:** Numbers will naturally be very skewed due to the Zipfian distribution of lexical items in any corpus and language. Logarithmic scaling can help for visualization purposes.

- **Relative frequency:** Metrics such as *occurrences per 100,000 tokens* will be more useful for comparisons across corpora/languages than raw frequencies, but will nonetheless be skewed across lexical items (i.e. *signs*).

- **Relative proportion:** A useful metric when comparing lexical or phonological variants for the same meaning, but will often suffer from a lack of data unless targeted lexical elicitation was part of the data collection.

- **Log odds:** Log odds are useful to show differences in frequency distributions based on some grouping variable (e.g., gender, region, text type) by accounting for imbalances in raw frequencies for different items, but will not distinguish form variation from differences in conversational content (i.e. topics). Note that the weighting and priors used will impact the results, so choose a method that suits your purposes.

- **Signer coverage:** Group-based variation (e.g., gender or region) in corpus data should preferably also account for signer coverage to ensure that the usage reflects the group as a whole rather than a single individual (signer) within it.

- **Type of usage:** Some items may be used incorrectly (e.g., slip of the hand) or metalinguistically (e.g., commenting on how *others* sign (see Langer et al., 2016), and it is thus important to investigate how and why individual items occur in a specific context – especially for low-frequency items.

- **Annotation conventions:** Know the annotation conventions of the corpus you are using, as this directly impacts both what questions you can ask with the data and how to interpret the results.

5. Bibliographical References

Robert Bayley, Adam C. Schembri, and Ceil Lucas. 2015. Variation and change in sign languages. In Adam C. Schembri and Ceil Lucas, editors, *Sociolinguistics and Deaf Communities*, 1 edition, pages 61–94. Cambridge University Press, Cambridge.

Carl Börstell. 2022. Introducing the signglossR Package. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 16–23, Marseille, France. European Language Resources Association (ELRA).

Carl Börstell, Thomas Hörberg, and Robert Östling. 2016. Distribution and duration of signs and parts of speech in Swedish Sign Language. *Sign Language & Linguistics*, 19(2):143–196.

Carl Börstell and Robert Östling. 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*, pages 13–18, Portorož, Slovenia. European Language Resources Association (ELRA).

Jordan Fenlon and Julie A. Hochgesang, editors. 2022. *Signed Language Corpora*. Number 25 in Sociolinguistics in deaf communities. Gallaudet University Press, Washington, DC.

Jordan Fenlon, Adam Schembri, Ramas Rentelis, David Vinson, and Kearsy Cormier. 2014. Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua*, 143:187–202.

Lindsay Ferrara, Benjamin Anible, Gabrielle Hodge, Tommi Jantunen, Lorraine Leeson, Johanna

Mesch, and Anna-Lena Nilsson. 2022. A cross-linguistic comparison of reference across five signed languages. *Linguistic Typology*, 0(0).

Jami N. Fisher, Julie A. Hochgesang, and Meredith Tamminga. 2016. Examining Variation in the Absence of a 'Main' ASL Corpus: The Case of the Philadelphia Signs Project. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 75–80, Portorož, Slovenia. European Language Resources Association (ELRA).

Thomas Hanke. 2016. Towards a Visual Sign Language Corpus Linguistics. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 89–92, Portorož, Slovenia. European Language Resources Association (ELRA).

Thomas Hanke, Elena Jahn, Sabrina Wähl, Oliver Böse, and Lutz König. 2020. SignHunter – A Sign Elicitation Tool Suitable for Deaf Events. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 83–88, Marseille, France. European Language Resources Association (ELRA).

Thomas Hanke, Reiner Konrad, and Gabriele Langer. 2023. Exploring regional variation in the DGS Corpus. In Ella Wehrmeyer, editor, *Studies in Corpus Linguistics*, volume 108, pages 192–218. John Benjamins Publishing Company, Amsterdam.

Amy Isard and Reiner Konrad. 2022. MY DGS – ANNIS: ANNIS and the Public DGS Corpus. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 73–79, Marseille, France. European Language Resources Association.

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Trevor Johnston. 2012. Lexical frequency in sign languages. *Journal of Deaf Studies and Deaf Education*, 17(2):163–193.

Trevor Johnston. 2019. Auslan Corpus Annotation Guidelines.

Nikolaus Riemer Kankkonen, Thomas Björkstrand, Johanna Mesch, and Carl Börstell. 2018. Crowdsourcing for the Swedish Sign Language Dictionary. In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 171–176, Miyazaki, Japan. European Language Resources Association (ELRA).

Vadim Kimmelman, Anna Komarova, Lyudmila Luchkova, Valeria Vinogradova, and Oksana Alekseeva. 2022. Exploring Networks of Lexical Variation in Russian Sign Language. *Frontiers in Psychology*, 12:740734.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2021. Overview of Datasets for the Sign Languages of Europe. Publisher: Universität Hamburg Version Number: 1.0.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2023. The Sign Language Dataset Compendium. Technical report. Version Number: 1.3.

Gabriele Langer. 2012. A colorful first glance at data on regional variation extracted from the DGS-Corpus: With a focus on procedures. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 101–108, Istanbul, Turkey. European Language Resources Association (ELRA).

Gabriele Langer, Thomas Hanke, Reiner Konrad, and Susanne König. 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*, pages 137–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Ceil Lucas, Robert Bayley, and Clayton Valli. 2009. *Sociolinguistic Variation in American Sign Language*. Gallaudet University Press.

Hannah Lutzenberger, Connie de Vos, Onno Crasborn, and Paula Fikkert. 2021. Formal variation in the Kata Kolok lexicon. *Glossa: a journal of general linguistics*, 6(1).

Hannah Lutzenberger, Katie Mudd, Rose Stamp, and Adam Charles Schembri. 2023. The social structure of signing communities and lexical variation: A cross-linguistic comparison of three unrelated sign languages. *Glossa: a journal of general linguistics*, 8(1).

David McKee and Graeme Kennedy. 2006. The distribution of signs in New Zealand Sign Language. *Sign Language Studies*, 6(4):372–391.

David McKee, Rachel McKee, and George Major. 2011. Numeral Variation in New Zealand Sign Language. *Sign Language Studies*, 12(1):72–97. Publisher: Gallaudet University Press.

Johanna Mesch and Lars Wallin. 2021. Annoteringskonventioner för teckenspråkstexter. Version 8. [Annotation guidelines for sign language texts].

Johanna Mesch, Lars Wallin, and Thomas Björkstrand. 2012. Sign Language Resources in Sweden: Dictionary and Corpus. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 127–130, Istanbul, Turkey. European Language Resources Association (ELRA).

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4):372–403.

Jill P. Morford and James MacFarlane. 2003. Frequency characteristics of American Sign Language. *Sign Language Studies*, 3(2):213–226.

Ellen Ormel, Onno Crasborn, Els van der Kooij, Lianne van Dijken, Ellen Yassine Nauta, Jens Forster, and Daniel Stein. 2010. Glossing a multipurpose sign language corpus. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 186–191, Valletta, Malta. European Language Resources Association (ELRA).

Thomas Lin Pedersen. 2022. *patchwork: The Composer of Plots*.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Josefina Safar. 2021. What's your sign for TORTILLA? Documenting lexical variation in Yucatec Maya Sign Languages. *Language Documentation & Conservation*, 15:30–74.

Tyler Schnoebelen, Julia Silge, and Alex Hayes. 2022. *tidylo: Weighted Tidy Log Odds Ratio*.

Rose Stamp, Adam Schembri, Jordan Fenlon, Ramas Rentelis, Bencie Woll, and Kearsy Cormier. 2014. Lexical variation and change in British Sign Language. *PLoS ONE*, 9(4).

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Hadley Wickham and Dana Seidel. 2022. *scales: Scale Functions for Visualization*.

Sabrina Wähl, Gabriele Langer, and Anke Müller. 2018. Hand in Hand - Using Data from an Online Survey System to Support Lexicographic Work. In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 199–206, Miyazaki, Japan. European Language Resources Association (ELRA).

George K. Zipf. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin, New York, NY.

Zrajm Öqvist, Nikolaus Riemer Kankkonen, and Johanna Mesch. 2020. STS-korpus: A Sign Language Web Corpus Tool for Teaching and Public Use. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 177–180, Marseille, France. European Language Resources Association (ELRA).

6. Language Resource References

Isard, Amy and Konrad, Reiner. 2023. *MY DGS – ANNIS*. Hamburg University.

Mesch, Johanna and Wallin, Lars and Nilsson, Anna-Lena and Bergman, Brita. 2012. *Dataset. Swedish Sign Language Corpus project 2009–2011 (version 1)*. Sign Language Section, Department of Linguistics, Stockholm University. PID https://hdl.handle.net/1839/b9b9c88a-f8df-4fa5-8eb0-53622108764d.

Svenskt teckenspråkslexikon. 2023. *Swedish Sign Language Dictionary online*. Dept. of Linguistics, Stockholm University.