Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 21–26
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Measuring Lexical Similarity across Sign Languages in Global Signbank

Carl Börstell[1], Onno Crasborn[1], Lori Whynot[2]
[1]Radboud University / [2]Northeastern University
Erasmusplein 1, 6525 HT Nijmegen, The Netherlands / 360 Huntington Ave., Boston, MA 02115, USA
c.borstell@let.ru.nl, o.crasborn@let.ru.nl, l.whynot@northeastern.edu

Abstract

Lexicostatistics is the main method used in previous work measuring linguistic distances between sign languages. As a method, it disregards any possible structural/grammatical similarity, instead focusing exclusively on lexical items, but it is time consuming as it requires some comparable phonological coding (i.e. *form* description) as well as concept matching (i.e. *meaning* description) of signs across the sign languages to be compared. In this paper, we present a novel approach for measuring lexical similarity across any two sign languages using the Global Signbank platform, a lexical database of uniformly coded signs. The method involves a feature-by-feature comparison of all matched phonological features. This method can be used in two distinct ways: 1) automatically comparing the amount of lexical overlap between two sign languages (with a more detailed feature-description than previous lexicostatistical methods); 2) finding exact form-matches across languages that are either matched or mismatched in meaning (i.e. true or false friends). We show the feasability of this method by comparing three languages (datasets) in Global Signbank, and are currently expanding both the size of these three as well as the total number of datasets.

Keywords: lexical similarity, linguistic distance, cross-linguistic comparison, lexicostatistics, false friends, mutual intelligibility

1. Introduction

Glottolog 4.1 (Hammarström et al., 2019), one of the (if not *the*) most comprehensive language databases to date, lists 194 sign languages of the world. However, we know very little about the possible genealogical relationships between different sign languages, and many such claims are based solely on historical records of language contact and influences. Due to the scarcity of historical documentation and the fact that all sign languages should still be considered under-studied, little is known about linguistic distances between sign languages, a metric which could be used to estimate possible phylogenies. However, there are few methods for calculating linguistic distances that could be applied to sign languages, considering the format and quantity of available data. Previous work in this domain has mainly used *lexicostatistics*, a method of comparing form overlap between lexical items across languages based on concept lists with translations into the languages in question. For sign languages, such studies have mostly been undertaken on an areal basis, with the intention of using lexical overlap as a metric for the likelihood of two languages being related (Woodward, 1991; Woodward, 1993; Woodward, 2000; McKee and Kennedy, 2000; Guerra Currie et al., 2002; Johnston, 2003; Bickford, 2005; Al-Fityani and Padden, 2010). These studies have in common that they compare the form similarity of two signs with the same meaning (i.e. concept-matched) from different sign languages, although the exact method for comparing sign forms across languages has varied between studies. In general, these studies consider the four basic form parameters of a sign (see Figure 1) and count two forms with all parameter values equal as *identical*, forms with one parameter value differing as *similar*, and more differing values as *different* forms.[1]

Parameter	Value
Location	neutral space
Handshape	B hand
Orientation	palm forward
Movement	ipsilateral movement

Figure 1: The NGT sign NEE-E ('no') with form parameter descriptions (Crasborn et al., 2020b).

This methodology has proven valid in the sense of finding greater similarity across sign languages known to be related (Johnston, 2003), but it can also be a somewhat crude measure that finds similarity that is purely incidental, and some studies have thus either tried to include iconic motivation as an additional factor in such measures (Ebling et al., 2015), or introduced a more fine-grained method for comparing sign forms across languages by separating form parameters into more detailed (sub)features (Yu et al., 2018). Here, we follow a path more similar to the latter, by using the uniformly coded cross-linguistic sign language lexical database *Global Signbank* (Crasborn et al., 2020a) to automatically measure lexical similarity across sign languages.

An ultimate goal with this method is to predict communicative success in cross-signing contexts (Zeshan, 2015; Byun et al., 2018) and mutual intelligibility across sign languages (Sáfár et al., 2015). The hypothesis is that languages with similar phonologies may show overlap in sign forms, which may or may not encode the same meaning. If the meaning

[1]Some studies would conflate parameters and thus look at three rather than four parameters.

Language	Sign entries	Coded signs	% coded
NGT	4,026	3,531	88%
CSL	2,248	568	17%
IS	200	200	100%

Table 1: Language datasets and number of coded signs in Global Signbank (Crasborn et al., 2020a).

overlaps (true friends), the prediction is that mutual intelligibility is higher; if not (false friends), this could be an impeding factor for cross-signing. As an example, the NGT (*Nederlandse Gebarentaal*; Sign Language of the Netherlands) sign WAT-A ('what'; Figure 2a) is identical to the ASL (American Sign Language) sign WHERE, and the NGT sign WAAR-A ('where'; Figure 2b) is identical to the ASL sign WHAT. This overlap in form but mismatch in meaning may disrupt cross-signing, since the addressee recognizes the form but associates it with a different meaning. Disruption in comprehension due to these types of false friends were indeed found in a study on comprehension of International Sign (IS) for signers of Japanese Sign Language and Auslan (Australian Sign Language) (Whynot, 2015).

2. Data and Methodology

2.1. Global Signbank

In our current data, we have a number of sign languages stored in an online lexical database called *Global Signbank* (Crasborn et al., 2020a). The languages – each represented as a separate *dataset* – are accessed in a graphical user interface (Figure 3) in which signs can be searched by translation keywords (e.g. in Chinese, Dutch, English), sign glosses (unique labels for signs), and are displayed as video files (`.mp4`), animated images (`.gif`), and still images (`.png`), together with fields containing phonological form-descriptions of signs.

We use data from three languages, in order of size of the datasets (see Table 1): NGT – 4,026 signs; 3,531 (88%) of which have phonological coding (Crasborn et al., 2020b); Chinese Sign Language – CSL, Shanghai variety; 2,248 signs; 568 (17%) of which have phonological coding (Crasborn et al., 2020c); and International Sign – IS; 200 signs; 200 (100%) of which have phonological coding (Whynot, 2020). NGT and CSL are two urban, unrelated languages; IS is a sign system based on mainly European-derived sign languages, used primarily as a form of communication at international deaf events, not used as an L1 in any community (Hiddinga and Crasborn, 2011; Whynot, 2015).

The relevant form-description fields in Global Signbank included in our sign similarity comparison are listed below:

- Handedness
- Strong Hand
- Weak Hand
- Handshape Change
- Relation between Articulators
- Location
- Relative Orientation: Movement
- Relative Orientation: Location
- Orientation Change
- Contact Type
- Movement Shape
- Movement Direction
- Repeated Movement
- Alternating Movement

2.2. Concepticon

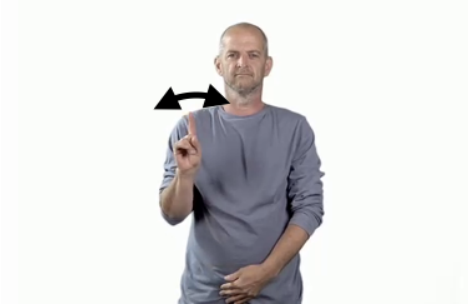
Since we want to look at lexical similarity across languages, we need a way to map form to meaning uniformly across datasets. We use the Concepticon concept list database (List et al., 2019) for this purpose. Concepticon is a database of collected concept lists from a diverse set of linguistic studies, compiled into one master list with links to individual lists collected – one list being the ECHO Swadesh list for sign languages (Woll et al., 2010). We use a crude method of mapping the English keywords/glosses in Global Signbank sign entries (see Figure 3) to concepts in Concepticon through string-matching. By doing so, we can compare signs not only from form to meaning (by manually looking at form-matches and evaluating their meaning-correspondence), but also meaning to form (by comparing those forms that are mapped to the same concept). The matching, mapping, and comparison steps are described in the following section.

2.3. Similarity Measure

After matching all language datasets to Concepticon as described above, we proceed to the automatic comparison of sign forms. Here, we compare any two signs on each form-description field and compute the number of overlapping fields. Since not all fields are relevant to all signs, we calculate the differences only for fields that have a value listed for both signs. This means that the comparison of a one-handed and a two-handed sign will result in a different value for the *Handedness* field (1 vs. 2 hands), but the field *Weak Hand* will be skipped altogether as it is only relevant for two-handed signs. Thus, we get a binary value (0 = different; 1 = same) for each relevant field, and divide the total by the number of fields compared to arrive at a sign similarity score between 0 and 1. The comparison is done with an automated script.

In the first step, we want to compare *all* signs in one language to *all* signs in another language. This means that we disregard meaning in this first automatic comparison stage, and let our script iterate through all possible sign pairs across datasets and store the similarity score for each such pair. This step of our cross-linguistic sign comparison is illustrated in Figure 4 using NGT and CSL.

In the second step, we want to compare only those pairs of signs across languages that are matched to the same Concepticon concepts. Since some concepts may be matched to several sign entries within a language dataset (due to form variations), the script iterates through each variant for a concept in one language and compares it to each variant for the same concept in the other language, and subsequently return the sign pair with the highest lexical similarity. This is illustrated schematically in Figure 5 in which only signs matched to the concept 'no' are compared to each other. In



(a) The NGT sign WAT-A ('what').

(b) The NGT sign WAAR-A ('where').

Figure 2: The NGT signs WAT-A (a) and WAAR-A (b) (Crasborn et al., 2020b).

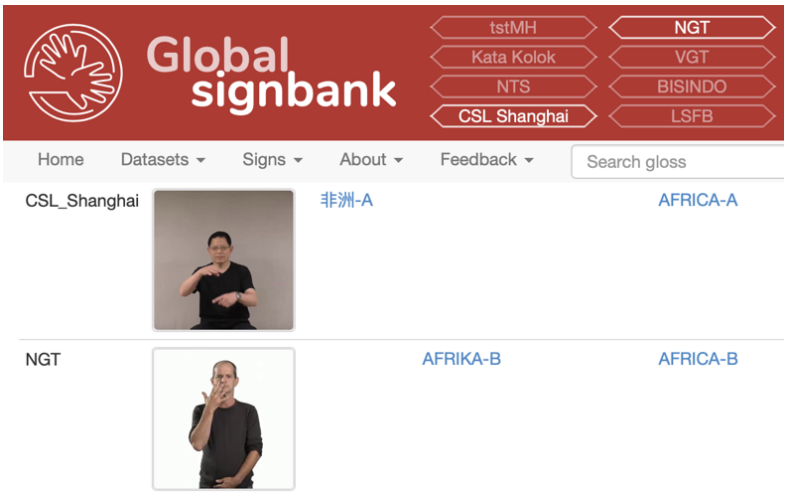


Figure 3: The graphical user interface of Global Signbank, showing the search results for *Africa* in two datasets (languages): CSL and NGT. Glosses are available in English for both datasets, as well as Chinese for CSL and Dutch for NGT.

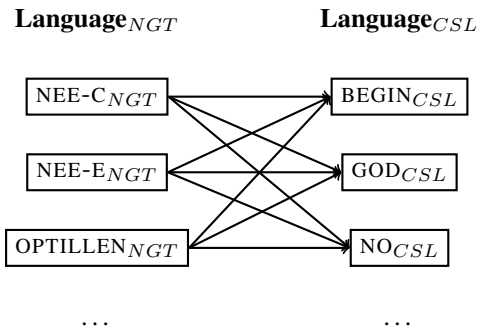


Figure 4: Cross-linguistic sign form comparison, all combinations.

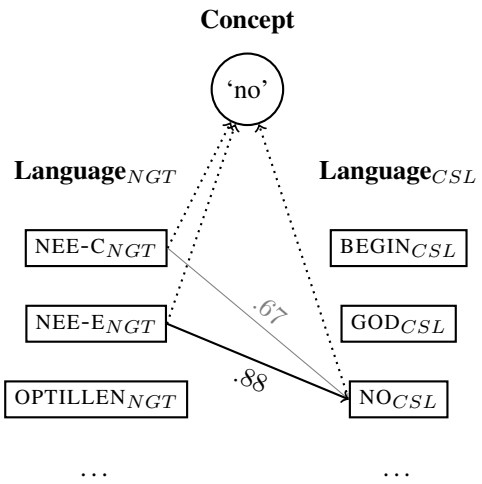


Figure 5: Cross-linguistic sign form comparison, only concept-matched combinations (dotted lines) compared. Highest similarity sign pair (thick black line) returned.

this minimal example, the sign variants NEE-C ('no'; Figure 6a) and NEE-E ('no'; Figure 6b) in NGT are both compared to the CSL sign NO ('no'; Figure 6c), after which only the pair NEE-E$_{NGT}$ and NO$_{CSL}$ is kept as it has the highest degree of overlap (.88), only differing in the CSL sign having a repeated movement.

3. Results

In the first step, we compared all sign forms in one language dataset against all sign forms in another language dataset, for each pairing across our three languages: NGT–CSL; NGT–IS; CSL–IS. Since the number of signs coded differs greatly across our three language datasets, the number of sign form matches are expected to differ accordingly. Indeed, we find most form overlaps with the pair-

(a) The NGT sign NEE-C ('no'). (b) The NGT sign NEE-E ('no'). (c) The CSL sign NO ('no').

Figure 6: The NGT signs NEE-C (a) and NEE-E (b) (Crasborn et al., 2020b), and CLS sign NO (c) (Crasborn et al., 2020c).

Pair	Matches	True	False
NGT–CSL	30 (5.3%)	12	18
NGT–IS	10 (5%)	6	4
CSL–IS	0 (–)	–	–

Table 2: Form-matches and number of true vs. false friends across all language pairings.

Pair	Matches	Mean	Median
NGT–CSL	194	.406	.375
NGT–IS	62	.436	.444
CSL–IS	43	.373	.333

Table 3: Concept-matches and their mean and median form-similarity scores across all language pairings.

ings that involve NGT – the largest of our datasets – and also more overlaps for the NGT–CSL pair than the NGT–IS pair, given that the CSL dataset is larger than the IS dataset. As shown in Table 2, 30 sign pairs are matched as form-identical across NGT and CSL. We look at each matched pair individually in order to evaluate whether they also match in meaning (true friends) or not (false friends). Of these 30 sign pairs, 12 pairs constitute true friends in that they have exact or similar meaning-matches: an example of an exact match in form and meaning is NGT and CSL signs for 'good' (Figures 8a–8b); an example of an exact form-match with a similar meaning is NGT JEZUS-A ('Jesus'; Figure 9a) and CSL GOD ('God'; Figure 9b). 18 pairs constitute false friends, sign pairs for which the forms are identical but the meanings are different: one example of this is the NGT sign OPTILLEN-A ('to lift'; Figure 10a) and the CSL sign BEGIN ('to begin'; Figure 10b).

For the NGT–IS pair, we find 10 sign pairs with identical forms, 6 of which are true friends and 4 of which are false friends, and for the CSL–IS pair we find no form-matches whatsoever. We find proportionally more true friends between NGT and IS than between NGT and CSL, which could be indicative of a general closer lexical similarity between the former languages than the latter. However, seeing as the datasets and the absolute numbers of form-matches are miniscule, this conclusion would be premature.

In the second step, we compared only those sign forms that were concept-matched to Concepticon. Again, we find more matches for the larger datasets, unsurprisingly as the number of potential matches is only as big as the smaller dataset (language) in any given pair (cf. Table 1). Thus, NGT–CSL has 194 concept-matched signs, NGT–IS has 62, and CSL–IS has 43. Concept-matched signs with mean and median similarity scores are shown in Table 3, and the distribution of similarity scores are shown in Figure 7.

These results point to NGT and IS generally having a higher

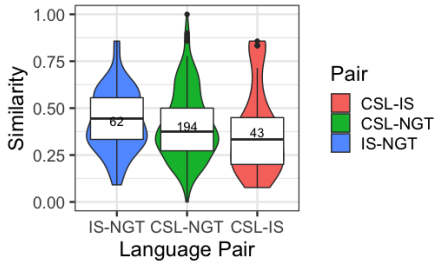


Figure 7: Distribution of sign form-similarity scores in concept-matched sign pairs across all language pairings.

similarity for signs denoting the same concept than either pairing including CSL. This, together with the higher proportion of true friends from the first step of the lexical comparison, may suggest a closer lexical distance between NGT and IS than any of the CSL pairings – and similar European vs. Asian sign language splits have been suggested (Yu et al., 2018). However, since our datasets are still small and also disproportionate in size, this is at best a preliminary suggestion in need of further examination.

4. Discussion

In this paper, we have described a method for comparing lexical similarity as an indicator of linguistic distance across sign languages represented as datasets in Global Signbank. Our method works in two directions: 1) from form to meaning (whether signs that overlap in form also overlap in meaning, i.e. are true or false friends); 2) from meaning to form (to what extent the phonological forms of signs for the same concept across languages are (dis)similar. With larger datasets (in terms of both languages and sign entries), we see the potential of this method to be used for lexicostatistics across a range of languages

(a) The NGT sign GOED-A (‘good’).

(b) The CSL sign GOOD (‘good’).

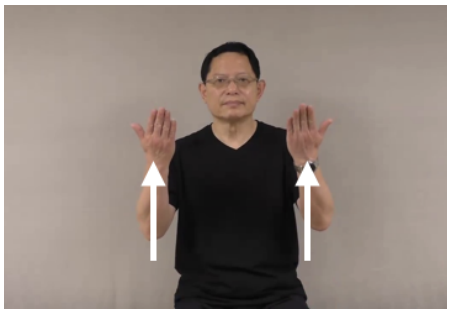
Figure 8: The NGT sign GOED-A (a) (Crasborn et al., 2020b) and CLS sign GOOD (b) (Crasborn et al., 2020c).

(a) The NGT sign JEZUS-A (‘Jesus’).

(b) The CSL sign GOD-A (‘God’).

Figure 9: The NGT sign JEZUS-A (a) (Crasborn et al., 2020b) and CLS sign GOD (b) (Crasborn et al., 2020c).

(a) The NGT sign OPTILLEN-A (‘to lift’).



(b) The CSL sign BEGIN (‘to begin’).

Figure 10: The NGT sign OPTILLEN-A (a) (Crasborn et al., 2020b) and CLS sign BEGIN (b) (Crasborn et al., 2020c).

through a (semi-)automated process, which would speed up the process compared to a purely manual comparison and allow for pairwise comparisons across a large set of languages which could be clustered along multiple dimensions (Bickford, 2005; Yu et al., 2018).

Furthermore, we hope to use some of these methods in order to quantify linguistic distances (focusing on lexical similarity) and apply the results to our ongoing project investigating cross-signing – that is, communication across different sign languages. When signers engage in cross-signing, they bring their individual sets of linguistic resources and skills, which include the use of material from their own primary language(s) as well as the adjustment and adaptation to the communicative context. Previous research has shown that deaf signers are able to communicate successfully without sharing any signed or spoken language, after only a short amount of time in the cross-signing context (Zeshan, 2015; Byun et al., 2018). Nonetheless, little is known about whether linguistic distance (as in a high degree of lexical similarity) influences the degree of communicative success in cross-signing contexts, though one could assume that cross-signing success is affected by lexical similarity, much like mutual intelligibility based on the amount of overlap in conventional lexical items (Sáfár et al., 2015). Such effects on comprehension have been shown in a study on IS, in which signers whose languages use signs similar to corresponding signs in IS would perform better on an IS lexical comprehension task (Whynot, 2015).

One of the unique features of the method outlined above is that it takes variation into account. Signers have in their linguistic repertoire not only their own preferred (e.g. dialectal, sociolectal) sign form for a concept, but are also familiar with other signs used in their language community. In our method, the best match sign pair is always used in cases of variants, which accounts for having passive knowledge of a sign form–meaning mapping without necessarily

producing it. In cross-signing interactions, these multiple variants constitute part of the communicative resources that a signer brings to the table, and in our measures of lexical similarity, we include this aspect of linguistic knowledge.

Using this method, we hope to establish a metric for linguistic distances not only for linguistic classification (in terms of lexical typology or genealogy), but also for the expected communicative success in cross-signing contexts. Historical connections between sign languages (based on, often scarce, historical records) may offer some explanation for potential cross-linguistic comprehension and mutual intelligibility. However, such cross-linguistic intelligibility may be possible without relatedness, by virtue of iconic motivation. That is, if the languages involved happen to recruit similar iconic patterns in sign formation, cross-signing comprehension may be more successful. Thus, although lexical form similarity is one metric that could easily be used to estimate cross-linguistic comprehension, including a more schematic perspective on iconicity mappings (Ebling et al., 2015) may prove to be necessary too.

5. Acknowledgements

This study was funded by the *Netherlands Organisation for Scientific Research* (NWO) grant number 277-70-014.

We would like to acknowledge our team: Tashi Bradford, Aurélia Nana Gassa Gonga, Maya de Wit, and Merel van Zuilen. A special thanks to Anique Schüller and Neil Ray for annotating Signbank datasets for this study.

6. Bibliographical References

Al-Fityani, K. and Padden, C. (2010). Sign languages in the Arab world. In Diane Brentari, editor, *Sign languages: A Cambridge language survey*, pages 433–450. Cambridge University Press, New York, NY.

Bickford, J. A. (2005). *The signed languages of Eastern Europe*. SIL International & University of North Dakota.

Byun, K.-S., de Vos, C., Bradford, A., Zeshan, U., and Levinson, S. C. (2018). First Encounters: Repair Sequences in Cross-Signing. *Topics in Cognitive Science*, 10(2):314–334.

Ebling, S., Konrad, R., Boyes Braem, P., and Langer, G. (2015). Factors to Consider When Making Lexical Comparisons of Sign Languages: Notes from an Ongoing Comparison of German Sign Language and Swiss German Sign Language. *Sign Language Studies*, 16(1):30–56.

Guerra Currie, A.-M. P., Meier, R. P., and Walters, K. (2002). A crosslinguistic examination of the lexicons of four signed languages. In Richard P. Meier, et al., editors, *Modality and structure in signed and spoken language*, pages 224–237. Cambridge University Press, Cambridge.

Hiddinga, A. and Crasborn, O. (2011). Signed languages and globalization. *Language in Society*, 40(4):483–505.

Johnston, T. (2003). BSL, Auslan and NZSL: Three signed languages or one? In Anne Baker, et al., editors, *Cross-linguistic perspectives in sign language research: Selected papers from TISLR 2000*, pages 47–69. Signum, Hamburg.

McKee, D. and Kennedy, G. (2000). Lexical comparisons of signs from American, Australian, British and New Zealand Sign Languages. In Karen Emmorey et al., editors, *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima*, pages 49–76. Lawrence Erlbaum Associates, Mahwah, NJ.

Sáfár, A., Meurant, L., Haesenne, T., Nauta, E., De Weerdt, D., and Ormel, E. (2015). Mutual intelligibility among the sign languages of Belgium and the Netherlands. *Linguistics*, 53(2):353–374.

Whynot, L. (2015). *Assessing comprehension of international sign lectures: Linguistic and sociolinguistic factors*. dissertation, Macquarie University.

Woodward, J. (1991). Sign language varieties in Costa Rica. *Sign Language Studies*, 73:329–346.

Woodward, J. (1993). The relationship of sign language varieties in India, Pakistan, and Nepal. *Sign Language Studies*, 78:15–22.

Woodward, J. (2000). Sign language and sign language families in Thailand and Viet Nam. In Karen Emmorey et al., editors, *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima*, pages 23–47. Lawrence Erlbaum Associates, Mahwah, NJ.

Yu, S., Geraci, C., and Abner, N. (2018). Sign Languages and the Online World Online Dictionaries & Lexicostatistics. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zeshan, U. (2015). "Making meaning": Communication between sign language users without a shared language. *Cognitive Linguistics*, 26(2):211–260.

7. Language Resource References

Onno Crasborn, et al., editors. (2020a). *Global Signbank*. Radboud University, Nijmegen. https://signbank.science.ru.nl.

Crasborn, O., van der Kooij, E., Zwitserlood, I., and Ormel, E. (2020b). Nederlandse Gebarentaal (NGT) dataset in Global Signbank. In Onno Crasborn, et al., editors, *Global Signbank*. Radboud University, Nijmegen. https://signbank.science.ru.nl.

Crasborn, O., van Zuilen, M., and Gong, Q. (2020c). Chinese Sign Language (CSL) dataset in Global Signbank. In Onno Crasborn, et al., editors, *Global Signbank*. Radboud University/Fudan University, Nijmegen/Shanghai. https://signbank.science.ru.nl.

Hammarström, H., Forkel, R., and Haspelmath, M. (2019). Glottolog database 4.1. https://glottolog.org.

List, J.-M., Greenhill, S. J., Rzymski, C., Schweikhard, N. E., and Forkel, R. (2019). Concepticon 2.1. https://concepticon.clld.org.

Whynot, L. (2020). International Sign (IS) dataset [Whynot 2015] in Global Signbank. In Onno Crasborn, et al., editors, *Global Signbank*. Radboud University, Nijmegen. https://signbank.science.ru.nl.

Woll, B., Crasborn, O., van der Kooij, E., Mesch, J., and Bergman, B. (2010). *Extended Swadesh list for signed languages*. http://www.let.ru.nl/sign-lang/echo/.