From Corpus to Lexical Database to Online Dictionary: Issues in Annotation of the BSL Corpus and the Development of BSL SignBank

Kearsy Cormier[1], Jordan Fenlon[2], Trevor Johnston[3], Ramas Rentelis, Adam Schembri[4], Katherine Rowley[1], Robert Adam[1] and Bencie Woll[1]

University College London[1], Gallaudet University[2], Macquarie University[3], La Trobe University[4]
Correspondence to: Deafness, Cognition and Language Research Centre, University College London
49 Gordon Square, London, WC1H 0PD, United Kingdom
Email: k.cormier@ucl.ac.uk, jordan.fenlon@gallaudet.edu, trevor.johnston@mq.edu.au, hipas8@mac.com, a.schembri@latrobe.edu.au, kate.rowley@ucl.ac.uk, r.adam@ucl.ac.uk, b.woll@ucl.ac.uk

Abstract

One requirement of a sign language corpus is that it should be machine-readable, but only a systematic approach to annotation that involves lemmatisation of the sign language glosses can make this possible at the present time. Such lemmatisation involves grouping morphological and phonological variants together into a single lemma, so that all related variants of a sign can be identified and analysed as a single sign. This lemmatisation process is made more straightforward by the existence of a comprehensive lexical database, as in the case for Australian Sign Language (Auslan). When annotation of data collected as part of the British Sign Language (BSL) Corpus Project began, no such lexical database for BSL existed. Therefore, a lemmatised BSL lexical database was created concurrently during annotation of the BSL Corpus data. As part of ongoing work by the Deafness Cognition & Language Research Centre, this lexical database is being developed into an online BSL dictionary, BSL SignBank. This paper describes the adaptation of the Auslan lexical database into a BSL lexical database, and the current development of this lexical database into BSL SignBank.

Keywords: corpus, lexicon, dictionary, lemmatisation, British Sign Language, ID gloss

1. Introduction

A systematic approach to corpus annotation that involves lemmatisation of glosses is required to make a sign language corpus into a true linguistic corpus in the sense intended by McEnery and Wilson (1996) – i.e., a finite, accessible, representative set of language recordings that is machine-readable. Such lemmatisation involves not only grouping together morphological but also phonological variants into a single lemma, so that all related variants of a sign can be identified and analysed as a single sign. This lemmatisation process is made more straightforward by the existence of a comprehensive lexical database, as in the case for Australian Sign Language (Auslan) (Johnston, 2001). When lexical annotation of data collected as part of the British Sign Language (BSL) Corpus Project (Schembri, Fenlon, Rentelis, & Cormier, 2011) began in 2011, no such lexical database for BSL existed. Publicly available BSL dictionaries (e.g., Brien, 1992) focused on translation equivalents and were not lemmatised in a way which would allow *ID glossing*, i.e., type-token matching (Johnston, 2010). In order to lemmatise the data for the purposes of the BSL Lexical Frequency Study, a lexical database for BSL was created concurrently during annotation. As part of ongoing work by the Deafness Cognition and Language Research Centre (DCAL, 2011-2015), this lexical database is being developed into an online BSL dictionary, BSL SignBank. Here we describe the adaptation of the Auslan lexical database (Johnston, 2001) into a BSL lexical database, and the current development of this lexical database into an online BSL dictionary.

2. BSL Lexical Database (BLD)

When planning annotation of the BSL Corpus data, we began by taking advantage of the fact that a lexical database for Auslan (a sign language variety closely related to BSL which shares much of the same lexicon) already existed (Johnston, 2001). The Auslan lexical database (ALD) was initially created as an offline database, first in tabular format in Microsoft Word and then later HyperCard, then FoxPro, then FileMaker Pro. As of 2004, the Auslan lexical database additionally exists as an online dictionary as Auslan SignBank (http://www.auslan.org.au). The dictionary contains approximately 7000 entries (4000 of which are publicly viewable) and is organised in an order based on phonological parameters (Johnston, 2003). This ordering aids in identifying signs that are homonyms (or near homonyms) as signs that are formationally the same (or similar) end up as entries that are adjacent to each other, so that decisions about whether these signs are homonyms or not can be made more easily.

Because of the ease of manipulating an offline database in FileMaker Pro (e.g. adding/deleting/editing entries, searching, sorting), as opposed to a bespoke online database which requires a programmer for manipulation, we began by cloning the offline FileMaker Pro version of Johnston's Auslan lexical database in early 2011. This was the beginning of the BSL Lexical Database (BLD).

Annotators began lexical annotation of the BSL Corpus data by first searching BLD for keywords linked to the meaning of each BSL sign in the corpus video. If the sign already existed in BLD (i.e., if it was an Auslan sign that had been carried over into BLD), annotators ensured that the sign was coded as a BSL sign if it had not been already, and used that entry to annotate the sign in question (either with the Auslan ID gloss, or with a different ID gloss if needed). For BSL signs that were not in BLD already (i.e., they were not Auslan signs from ALD), annotators added entries for these BSL signs. New entries included very basic lexical attributes: ID gloss, movie clip, and keywords (English translation equivalents). The BSL Corpus team met weekly to discuss lemmatisation issues. (See §3.1 for more on lemmatisation.)

The BSL Lexical Frequency Study (LFS) (Cormier, Fenlon, Rentelis, & Schembri, 2011) was based on approximately 25,000 lemmatised sign tokens from the conversational data in the BSL Corpus, annotated using BLD. These 25,000 sign tokens represented 2506 signs, including 'partly-lexical' signs (e.g., pointing signs and classifier constructions) and 'non-lexical' signs (e.g., constructed action). (Annotations were carried out following Johnston's guidelines for the annotation of the Auslan corpus, www.auslan.org.au/about/annotations/.) Roughly 16,000 sign tokens from the LFS (representing roughly 1500 sign types) were lexical signs, and all of these signs are represented in BLD. Preliminary annotation of an additional 25,000 sign tokens from the conversation data and also concurrent ID glossing of sign tokens from the lexical elicitation task resulted in the inclusion of approximately 1800 sign types in BLD as of mid-2011.

3. From BLD to BSL SignBank

Although work on the LFS was completed with the end of the BSL Corpus Project in June 2011, further development of BLD continued, as part of DCAL's plan to create a corpus-based online dictionary and reference grammar (2011-2015).

The first step in adapting BLD into an online dictionary was to check form-meaning pairings between similar signs within the database. This initially entailed fitting the newly added signs (approximately 700 of the 1800 BSL lexical signs in BLD) into the numbering system outlined in Johnston (2003). This numbering system has signs ordered by the handshape of the dominant hand, following an order that roughly follows the order of numeral signs (and thus, the number of extended fingers) in Auslan from zero upwards. Within each handshape, one-handed signs are first, followed by signs made with two hands that have the same handshape (*double-handed signs* in Johnston's terminology), followed by signs with two hands that have different handshapes (*two-handed signs* in Johnston's terminology). Within this, signs were then ordered by primary location, from the top of the head downward. Ordering beyond these features (handedness, handshape, and location) then roughly followed a series of other phonological parameters (e.g. symmetry, orientation, location on non-dominant hand, and contact). However, as Johnston (2003:456) notes:

"The Auslan dictionaries only partially implemented the finer decision schema… because, in practice, discrimination beyond three or four levels within the decision schema has not been necessary in order to sequence most lexical signs. The reason for this is simply that the data contain few exemplars of more finely discriminated lexical signs. Indeed, even in those handshape sections that contain hundreds of distinctive signs, often no need arose to adhere to any strict sequencing beyond the major and minor features and secondary tabulation."

Attempting to add 700 BSL signs into this numbering system quickly proved to be problematic, particularly for dense phonological neighbourhoods. For example, Auslan and BSL both have many double-handed signs in neutral space with unmarked handshapes (e.g., with the 1 handshape or 5 handshape). Because there was no strict sequencing for Auslan signs via Johnston's (2003) system beyond the major parameters and a few minor parameters (because, as noted above, it was not needed for Auslan signs), it became difficult to find only one appropriate position within the numbering system where these signs belonged. After attempting to add in a few hundred BSL signs into the Johnston numbering system, we found that we ended up with several clusters of phonologically similar signs scattered throughout these dense phonological neighbourhoods, which made it increasingly difficult to find homonyms, near-homonyms, minimal pairs, and near minimal-pairs (which was meant to be one of the purposes of the numbering system in the first place – to easily identify these similar signs to check lemmatisation).

It became clear that the only way to check phonologically similar signs to ensure proper lemmatisation (e.g., that homonyms had been distinguished) was to code phonological information for each of the entries in the database first, on the assumption that these would represent tentative lemmata until proper lemmatisation could be done. There were several options for phonological coding of the lexical entries in BLD. One was to use a standard notation system like HamNoSys (the Hamburg Notation System). The Auslan lexical database contained HamNoSys transcriptions for each entry. However, HamNoSys is a phonetic transcription system, a much greater level of phonetic detail than was needed for organisation/sorting of the database. Furthermore, we needed the ability to search for/sort by various combinations of phonological parameters. HamNoSys transcriptions consist of a string of symbols, and sorting via parameters representing the symbols in the middle of the string would not have been straightforward. It is for this reason that the Auslan lexical database contains fields that redundantly encode information about the major phonological parameters for each Auslan entry (handedness, handshape and location). Thus the next step was coding for these major phonological attributes for the 1800 BSL signs from the BSL Corpus Project. Fields for other phonological parameters (e.g., movement) will be added after a first attempt at lemmatisation via searching/sorting, to see what kinds of parameters will be needed to distinguish signs at a detailed level.

Before such searching/sorting for lemmatisation purposes can take place, the database needs to contain a certain core vocabulary. If this is not the case, entries would need to be re-lemmatised after core vocabulary is added. It thus helps to try to ensure that core vocabulary is included before this process takes place. There is no easy way to systematically determine what "core" signs might be missing from BLD, which was based largely on spontaneous conversational data. However, the lexicon of BSL has been documented to a degree in previous dictionaries. The only such dictionary based on linguistic principles similar to those in the ALD is Brien (1992), which contains just under 1800 lexical entries. Thus, one way to ensure that the lexical database contained important core vocabulary was to check if signs in Brien (1992) were in BLD and if they were not, to add them to the database. Based on previous work by Johnston and Schembri (1999), we were aware that signs in Brien (1992) had not been systematically lemmatised, but the degree to which this was true quickly became apparent once we began including lexical items from the BSL/English dictionary in the BLD. Homonyms in Brien (1992) are typically combined into one entry[1], while signs that are clearly phonological variants are sometimes listed as separate variants for no apparent reason. Thus the process of including signs from Brien (1992) in the BLD required us to lemmatise and/or re-lemmatise those entries (e.g. by considering the relationship between the Brien (1992) signs and potential phonological/lexical variants that already existed in BLD).

3.1 Lemmatisation

Here we outline the principles and procedures that we used in lemmatising signs that were added to BLD as part of the Lexical Frequency Study under BSLCP, and subsequently in lemmatising (and re-lemmatising) signs from Brien (1992) into/with signs from BLD.

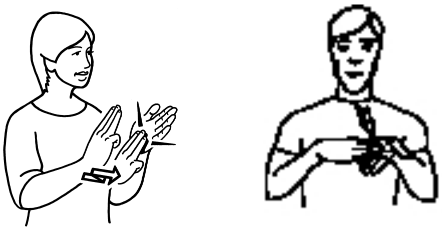
On a basic level, decisions about lemmatisation during annotation were made based on form and meaning. Two sign tokens A and B with the same form and the same meaning were considered to constitute a single lemma, with one ID gloss attributed to them. It is important to note that an ID gloss is not "the meaning" of the sign; it is simply a unique label given to a lexical item in order to aid in consistent identification of lexical items during annotation (Johnston, 2010). The meaning (via definitions) and/or English translation equivalents are stored in the lexical database. English mouthing was ignored for the purposes of lemmatisation, although of course mouthing can be used in determining some elements of meaning.

Lemmatisation involves not only grouping phonological variants but also morphological variants into a single lemma. Therefore, morphological modifications used in particular tokens such as directionality/agreement marking, number marking, aspect marking, etc were not used to distinguish lemmas.

Two sign tokens A and B with clearly different lexical meanings were considered to constitute two different lemmas, with a different ID gloss given to each one. This was the case regardless of whether the phonological forms were completely different, similar, or identical.

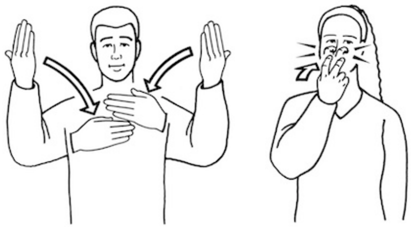
Beyond this basic level, there are various possibilities with similar/different forms and meanings. These are the primary criteria we considered:

Phonological variants. If sign tokens A and B differ in only one phonological parameter, and the meanings are the same or similar, then A and B are likely to be phonological variants of one lemma. For example, BSL MOTHER(M-hand) and MOTHER(B-hand), shown in Figures 1a and 1b, differ only in handshape and have the same meaning. These two phonological variants are both part of the lemma represented by the ID gloss MOTHER.



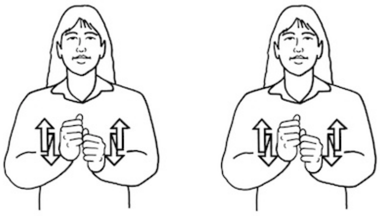
Figures 1a and 1b: Phonological variants of lexeme MOTHER: MOTHER(M-hand) and MOTHER(B-hand)

Lexical variants. If sign tokens A and B differ in more than one phonological parameter, and the meanings are the same or similar, then A and B may be lexical variants (separate lemmas). For example, BSL NIGHT1 is produced with two flat hands in neutral space, and NIGHT2 is produced with a bent-V handshape at the nose, as shown in Figure 2. These two lexical variants which have the same meaning (both have English translation equivalents of 'night', 'tonight', 'evening', 'dark') are distinguished in the ID gloss with numbers.



Figures 2a and 2b: Lexical variants NIGHT1 and NIGHT2

Homonyms. If sign tokens A and B differ in meaning but have the same phonological form, these forms are homonyms (separate lemmas). For example, both BSL BROTHER and MARCH-MONTH are produced with two A-hands in neutral space brushing against each other with alternating movement, as shown in Figure 3.

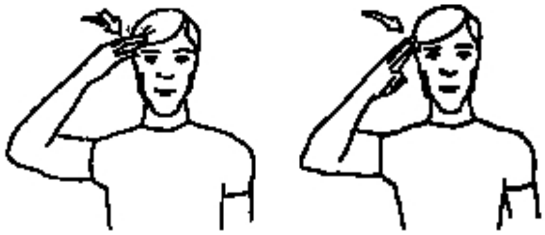


Figures 3a and 3b: Homonyms BROTHER and MARCH-MONTH

[1] The combination of homonyms into a single entry is actually not uncommon within lexicography, as distinguishing similar versus different meanings can be difficult even for spoken languages (Atkins & Rundell, 2008).

There were also additional criteria that were considered during lemmatisation beyond form and meaning:

Association of variant with social factors. Even if two variants A and B have the same meaning and differ only in one parameter, if one of the variants has a strong association with a particular social group (e.g. region, age, gender) or particular register (e.g. child-directed signing), this may be enough to lemmatise it separately. For instance, in addition to MOTHER as in Figure 1 above, there are other variants meaning 'mother' with similar handshapes to MOTHER (as seen in Figure 1 above) but produced at the forehead. However, these additional variants (shown below in Figure 4) were judged to constitute a separate lexeme from MOTHER since they are thought to be found in child-directed signing (i.e., they are associated with English translation equivalents 'mummy' and 'mum' in addition to 'mother'). In addition to this, the sign MOTHER is clearly a single manual letter sign (derived from two-handed fingerspelled M), whereas the relationship between MUM and the two-handed manual alphabet is less clear (we assume that the M-hand variant is a post-hoc initialisation of the original sign), as described below in §3.2.



Figures 4a and 4b: Lexeme MUM (two phonological variants, MUM(B-hand) and MUM(M-hand)

Morphological differences in variants. If variant A can take different morphological modifications compared to variant B (e.g. agreement/directionality, aspect marking, number marking), this may be enough to lemmatise them separately even if they are phonologically similar.

For each pair or set of sign tokens in question, all of the above criteria were considered when determining whether variants belonged to the same or different lemmas. Often these criteria compete with each other, and sometimes decisions have to be made on the basis of competing criteria that may be of equal importance. This means that it can be a considerable challenge maintaining consistency in principles of lemmatisation across all the data.

3.2 Citation form or headword status

Given a set of phonological variants, for the purposes of a lexical database and/or dictionary, one may want to ascribe headword (or citation form) status to one of these variants. This is not always necessary, as it is possible to have phonological variants listed in a lexical database with ID glosses that do not ascribe primary status to any single variant (e.g. with distinguishing phonological information as part of the ID gloss). However, because BLD had been created as part of a study on lexical frequency under BSLCP, it was only lexical variants that were important, not phonological variants. Thus phonological variants were not distinguished in the LFS annotations nor were they distinguished as separate entries in BLD. For each BLD entry with known phonological variants, one of those variants was chosen as the headword, or citation form – i.e. the form shown in the movie clip and the form for which phonological information is coded in BLD. Citation forms were decided based on these criteria:

Frequency (or assumed frequency). Given two phonological variants A and B, the variant with the highest frequency, or assumed frequency if there is no frequency information available, or the variant that is most widely used/understood across all social groups, could be considered the citation form or headword.

Phonological processes. Given two phonological variants A and B, if there is a known phonological process that could explain the change from A to B, then variant A could be considered the citation form or headword. Such phonological processes include change of sign location to one closer to centre of the body (Lucas, Bayley, Rose, & Wulf, 2002; Schembri et al., 2009), change in phonological parameter from more complex/marked to less complex/marked value (Battison, 1974, 1978), or distalisation of a variant from use of joints closer to the body to use of joints further away from the body (Mirus, Rathmann, & Meier, 2001). For example, the sign TOMORROW may be produced with movement of the elbow joint, wrist joint, and/or joint at the large knuckle of the index finger. The most distalised variant uses primarily the large knuckle joint only. The citation form as shown in Figure 5 includes the use of the more proximal elbow joint.

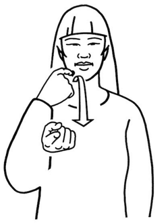


Figure 5: Citation form for TOMORROW

Iconicity. Given two phonological variants A and B, if A is more iconic than B, then A could be considered the citation form or headword, on the assumption that iconic signs become more arbitrary over time (Frishberg, 1975; Klima & Bellugi, 1979).

Nativisation processes. If A and B are both lexical signs with some association with fingerspelling (e.g. via initialisation or fingerspelled loan), but A is closer to the fully fingerspelled word, then A could be considered the citation form or headword, following nativisation processes of fingerspelled forms (Brentari & Padden, 2001; Cormier, Schembri, & Tyrone, 2008). For example, MOTHER(M-hand) as shown in Figure 1 above is considered the citation form for the lemma MOTHER, because as noted above the M-hand variant is clearly a single manual letter sign derived from two-handed fingerspelled M .

Prestige (or assumed prestige) status. Given two phonological variants A and B, if variant A but not variant B is strongly associated with a social group that is known or assumed to carry prestige (e.g., region, native signer language background, etc), then variant A could be considered the citation form or headword.

Listing in other dictionaries (e.g. Brien 1992). Given two phonological variants A and B, if variant A is listed in another national BSL dictionary, especially Brien (1992), then variant A could be considered the citation form.

As with lemmatisation, these criteria were considered together rather than in isolation, and each set of related variants is considered on a case-by-case basis. For example, although given the two phonological variants MUM(B) and MUM(M) shown in Figure 4 above differ in the same way that MOTHER(B) and MOTHER(M) differ (i.e., handshape), the citation form for MUM is considered to be MUM(B) rather than MUM(M), due to the (assumed) frequency of MUM(B) over MUM(M) and also the fact that the B-hand variant is less likely to have been derived directly from the fingerspelled letter M which is located on the non-dominant hand (as noted above). None of the criteria are given particular preference overall, although (assumed) prestige status and listing in other dictionaries are rarely considered unless none of the other criteria are useful in determining citation form or headword.

The challenges for determining citation form are similar to the challenges for lemmatisation as noted above. That is, criteria can compete with each other. For example, one-handed versus double-handed variants are complex. They could be explained via the phonological process of weak drop (Battison, 1974; Brentari, 1998) with the double-handed variant as citation form which can become one-handed. On the other hand, as Frishberg (1975) notes, one-handed signs can also become two-handed via a general process of signs tending towards symmetry, particularly for signs produced below the neck (outside the area of highest visual acuity), although Frishberg notes this also occurs with some signs above the neck as well. Thus phonological processes generally cannot be used to determine whether a one-handed or two-handed variant should be attributed headword/citation form status. Frequency (or assumed frequency) is often the main criterion for these decisions.

4. Conclusion

Here we have described the process of adapting an existing lexical database for Auslan into a lexical database for BSL for the purposes of a study on lexical frequency, and the subsequent adaptation of this BSL lexical database into an online dictionary, BSL SignBank. The primary issues involved in preparing the lexical database for launch as an online dictionary involve systematic decisions about lemmatisation (in the course of checking existing lexical entries and adding new ones from other dictionaries) and also decisions about citation form based on sets of phonological variants. We have outlined the primary criteria used in making these decisions. Such criteria are tentative and always evolving as further work on the lexical database continues. Once a core set of lexical items within BLD has been amassed and lemmatised, this will be converted into BSL SignBank online, the initial launch for which is planned for 2013. This will initially contain at least 2000 entries. Eventually we expect BSL SignBank to have a number of entries similar to Auslan SignBank (i.e., 4000). It is clear that an online dictionary allows for growth and development over time in a way that was previously not possible with print dictionaries.

5. Acknowledgements

Support for this research was provided by grants RES-620-28-6002, Deafness, Cognition and Language Research Centre (DCAL) and RES-062-23-0825, British Sign Language Corpus Project (BSLCP) from the Economic and Social Research Council (ESRC), UK, and grant DP1094572 (2010-2013) Australian Research Council. We thank Frances Elton, Ellen Foote, Indie Johal, and Rosemary Stamp for assisting in lemmatisation, coding, and filming of signs for the BSL Lexical Database.

6. References

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Battison, R. (1974). Phonological deletion in American Sign Language. *Sign Language Studies, 5*, 1-19.

Battison, R. (1978). *Lexical borrowing in American Sign Language*. Silver Spring, MD: Linstock Press.

Brentari, D. (1998). *A prosodic model of sign language phonology*. Cambridge, MA: MIT Press.

Brentari, D., & Padden, C. A. (2001). Native and foreign vocabulary in American Sign Language: A lexicon with multiple origins. In D. Brentari (Ed.), *Foreign vocabulary: A cross-linguistic investigation of word formation* (pp. 87-119). Mahwah, NJ: Lawrence Erlbaum Associates.

Brien, D. (Ed.). (1992). *Dictionary of British Sign Language/English*. Boston: Faber & Faber.

Cormier, K., Fenlon, J., Rentelis, R., & Schembri, A. (2011). Lexical frequency in British Sign Language conversation: A corpus-based approach. In P. K. Austin, O. Bond, L. Marten & D. Nathan (Eds.), *Proceedings of the Conference on Language Documentation and Linguistic Theory 3*. London: School of Oriental and African Studies.

Cormier, K., Schembri, A., & Tyrone, M. E. (2008). One hand or two? Nativisation of fingerspelling in ASL and BANZSL. *Sign Language and Linguistics, 11*(1), 3-44.

Frishberg, N. (1975). Arbitrariness and iconicity: Historical change in American Sign Language. *Language, 51*, 696-719.

Johnston, T. (2001). The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics, 4*(1/2), 145-169.

Johnston, T. (2003). Language standardization and signed language dictionaries. *Sign Language Studies, 3*(4), 431-468.

Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics, 15*(1), 104-129.

Klima, E., & Bellugi, U. (1979). *The Signs of Language*. Cambridge, MA: Harvard University Press.

Lucas, C., Bayley, R., Rose, M., & Wulf, A. (2002). Location variation in American Sign Language. *Sign Language Studies, 2*(4), 407-440.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Mirus, G., Rathmann, C., & Meier, R. P. (2001). Proximalization and distalization of sign movement in adult learners. In V. Dively, M. Metzger, S. Taub & A. M. Baer (Eds.), *Signed languages: Discoveries from international research* (pp. 103-119). Washington, DC: Gallaudet University Press.

Schembri, A., Fenlon, J., Rentelis, R., & Cormier, K. (2011). British Sign Language Corpus Project: A corpus of digital video data of British Sign Language 2008-2011 (First Edition). London: University College London. http://www.bslcorpusproject.org.

Schembri, A., McKee, D., McKee, R., Johnston, T., Goswell, D., & Pivac, S. (2009). Phonological variation and change in Australian and New Zealand Sign Languages: The location variable. *Language Variation and Change, 21*(2), 193-231.