Building BSL SignBank: The lemma dilemma revisited

Jordan Fenlon\textsuperscript{a}, Kearsy Cormier\textsuperscript{a} and Adam Schembri\textsuperscript{b}

\textsuperscript{a}Deafness, Cognition & Language Research Centre, University College London, 49 Gordon Square, London, WC1H 0PD, United Kingdom

\textsuperscript{b}La Trobe University, Melbourne (Bundoora), Victoria, 3086, Australia

CORRESPONDING AUTHOR:

Jordan Fenlon
Deafness, Cognition & Language Research Centre,
University College London
49 Gordon Square
London, WC1H 0PD
England
\texttt{j.fenlon@ucl.ac.uk}

Phone: +44(0)20 7679 8679

Fax: +44(0)20 7679 8691

Co-authors’ email addresses: \texttt{k.cormier@ucl.ac.uk, a.schembri@latrobe.edu.au}
Abstract

One key criterion when creating a representation of any lexicon within a dictionary or lexical database is that the lexeme must be clearly distinguished from idiosyncratic or systematically modified variants. However, few researchers have attempted to outline these principles as they apply to sign languages. As a consequence, the vast majority of sign language dictionaries and lexical databases are mixed collections of unidentified phonetic, phonological, morphological, and lexical variants of lexical signs (e.g., Brien 1992) which have not addressed what may be termed as the lemma dilemma. In this paper, we outline the lemmatisation practices used in the creation of BSL SignBank (Fenlon et al. 2014a), a lexical database and dictionary of British Sign Language based on signs identified within the British Sign Language Corpus (http://www.bslecorpusproject.org). We argue that the principles outlined here should be considered in the creation of any sign language lexical database and ultimately any sign language dictionary and reference grammar.

Keywords: lemma, lexeme, lemmatisation, sign language, dictionary, lexical database.

1 Introduction

When one begins to document the lexicon of a language, it is necessary to establish what one considers to be a lexeme. Generally speaking, a lexeme can be defined as a unit that refers to a set of words in a language that bear a relation to one another in form and meaning. These words and their lexeme can be represented by a lemma, the canonical form from which other variants are derived. For instance, in English, walk, walks, walking, and walked are all considered to be variant forms of the lemma - walk. Spoken language dictionaries (e.g. Collins English Dictionary and the Oxford English Dictionary) have widely adopted the practice of listing words primarily by lemmas in the language (e.g. walk, walks, walking, etc. are not listed separately in the Oxford English Dictionary but under walk). In principle, the situation with signed languages need not be any different. However, before one can begin to list the signs in a sign language, one needs to develop clear criteria defining what are considered to be variant forms of the same lexeme and what might be considered separate lexemes (e.g., Johnston and Schembri 1999). Unfortunately, few sign language dictionaries have attempted to organise their entries in this manner, so lemmatisation for the purposes of lexical database building has not taken place for most documented sign languages.
This task is complicated by the need to group together phonetic and phonological variants without reference to a standard writing system.

This paper intends to address this issue in relation to the creation of BSL SignBank (Fenlon et al. 2014a), a lexical database and dictionary of British Sign Language (BSL) based on signs identified within the British Sign Language Corpus (Schembri et al. 2014). As no appropriate resources for BSL existed prior to the creation of BSL SignBank, signs found in the BSL Corpus needed to be lemmatised in a principled way so that they could be added to BSL SignBank. This was done in part building on the work of Johnston and Schembri (e.g. Johnston 1989; Johnston and Schembri 1999) for Australian Sign Language, which is historically related to BSL. In Section 2, we briefly describe the notion of lemmatisation in relation to sign languages and discuss previous attempts to develop lemmatisation principles for sign languages. Section 3 provides a brief overview of BSL SignBank and its association with the BSL Corpus. In Section 4, we outline the lemmatisation principles used in the creation of BSL SignBank building on previous work with other sign languages and enhancing our description with problematic cases that arose during our efforts. In Section 5, we argue that the principles outlined here should be considered in the creation of any sign language lexical database and discuss further issues involved in the development of sign language dictionaries.

2 Sign language lexicography

2.1 Challenges in sign language lexicography

Sign language research is still in its infancy compared to research on spoken languages. It was not until the mid twentieth century that linguistic descriptions of sign languages emerged (Tervoort 1953; Stokoe 1960). Since then, sign languages have attracted the attention of a growing number of researchers working across the language sciences and an investigations of many aspects of sign languages as living languages have emerged (see Pfau et al. 2012 for an extensive review). Despite major advances in our understanding of sign languages, however, there remains much to be explored. For many of the world’s sign languages, a dictionary that is intended to be representative of its core lexicon is notably absent. There are good reasons for the absence of such a resource. Firstly, there is no standard orthography or notation system for sign languages. Secondly, sign language dictionaries have been (until very recently) unable to exploit large corpora to obtain a collection of signs for inclusion in dictionaries. Instead word lists are often created simply by taking the spoken language of the surrounding community, listing key concepts from that language, and providing translation equivalents in the corresponding sign language. This practice has been widely criticised by sign language lexicographers as such an approach presents a misleading and limited
view of the sign language lexicon (e.g., Brien and Turner 1994; Johnston and Schembri 1999; Zwitserlood 2010). Not only does this practice imply that there is a one-to-one correspondence between English words and signs, it is likely that the resulting collection of signs will not be very representative of the sign language lexicon. The issue of how to order signs within a printed dictionary without a standard writing system is also problematic; lexicographers advocate the practice of listing signs according to formational criteria specific to sign languages but vary in the way they do this. This was a practice first adopted by Stokoe et al. (1965) for American Sign Language and replicated in, for example, the Auslan (Australian Sign Language) dictionary (Johnston 1989) and the Dictionary of British Sign Language/English (Brien 1992). Lastly, sign language dictionaries are inherently bilingual or multilingual dictionaries (we are not aware of any monolingual sign language dictionaries). Whilst signs may be presented in pictures, translation equivalents and definitions are presented in a written language (e.g. the language of the surrounding hearing community) and the opportunity to search for a sign via the spoken language may be a preferred option for many. Many sign language lexicographers have suggested ways to overcome these issues. In fact, with the advent of electronic dictionaries and sign language corpora, some of the issues may no longer need to hinder the development of sign language dictionaries (Kristoffersen and Troelsgård 2012), though there are still challenges to overcome (Zwitserlood 2010; McKee and McKee 2013). In this paper, we wish to focus on a specific problem that all sign language lexicographers must contend with: lemmatisation. This issue is important now that sign language corpora are becoming more widespread, since these corpora require lemmatised resources to aid with the process of annotation.

2.2 Lemmatisation and sign languages

The traditional notion of lemmatisation for spoken languages refers to the practice of grouping together morphological variants under a single lemma. For example, in English, the variants jump, jumps, jumping, and jumped are all considered variants of a single lemma jump. Generally speaking, lexicographers tend to follow the division between inflectional and derivational morphology as a guide when practising lemmatisation (Sterkenburg 2003). Inflectional variants (which arise out of a process that modifies a word) are not recognised as separate lexemes whilst derivational variants (arising out of a process that forms new words) are treated as such. Following this, jumps, jumped, and jumping are grouped together under a single lemma since each variant is derived by the addition of an inflectional suffix marking tense or aspect. In contrast, teach and teacher would be recognised as two separate lexemes because teacher is formed through the addition of an agentive suffix (i.e. a derivational morpheme) that changes the part of speech of the word from a verb to a noun. The change in meaning that results from the latter process is different
to the former; whilst the addition of a derivational morpheme changes the overall meaning of the sign, the addition of an inflectional morpheme only augments meaning. Instead, the inflectional morpheme functions as part of the grammar of the language (e.g. marking tense, number, agreement, etc.).

However, developing consistent lemmatisation principles along the same lines for sign languages can be difficult. Whilst some researchers argue that the division of inflectional and derivational morphology can be applied to sign languages (e.g., Sandler and Lillo-Martin 2006), others have argued otherwise. For example, Liddell (2003) has claimed that American Sign Language (ASL) is basically an inflectionless language, as there is no tense marking, aspectual modifications and plural marking are optional, and the so-called verb agreement system is best analysed instead as a reference tracking system. Liddell instead proposes that signs modified for aspect, plurality and person-marking are better understood as part of a derivational system. Similar claims have been made for other sign languages (e.g., Bergman and Dahl 1994; Johnston & Schembri, 2007). The categorisation of modified forms into inflectional and derivational variants is further complicated by the fact that it can be difficult to determine the part of speech of a sign (e.g. see Section 4.2.1) and since similar modifications may be regarded as inflectional in one instance but derivational in another (e.g. compare the aspectual modification in WALK to the movement seen in ALCOHOLIC described in Section 4.1.1). The ambiguity between the two categories and a lack of general consensus in the literature causes problems for the sign language lexicographer; it is not clear how principles of lemmatisation for sign languages can be derived using structural distinctions borrowed from spoken languages like English.

The issue of lemmatisation is complicated further as there are no standard writing systems widely used for any sign language. For spoken languages with written orthographies, the lexicographer is able to ignore phonetic and phonological variants, which makes distinctions between potentially related lexical variants straightforward.² For example, in English, variable pronunciations of bath such as [bæθ] and [baθ] would be paired together automatically as they both are represented identically in standard English orthography. With sign languages, variation can also exist at the phonetic and phonological level but without a standard writing system, the sign language lexicographer must also contend with the task of grouping together phonetic and phonological variants in a principled way.³ One can therefore see that what must be done with sign languages is an extension of the traditional notion of lemmatisation because phonetic and phonological variants, as well as morphological variants, must also be grouped together under one lemma. Lemmatisation in this sense is necessary to conduct any linguistic research that relies on lexical contrast including studies of phonology, morphology and the lexicon.

There have been few attempts to create sign language dictionaries following principles based
on this extended notion of lemmatisation. One such attempt is the Auslan lexical database, available initially as a lemmatised print dictionary (Johnston 1989; Johnston and Schembri 1999), later as an offline lexical database (Johnston 2001b), and most recently as Auslan SignBank, an openly publicly accessible online dictionary and as a restricted access lexical database (http://www.auslan.org.au). Other sign language dictionaries described as lemmatised resources include the Danish Sign Language dictionary (Centre for Tegnsprog 2008), the online New Zealand Sign Language dictionary (McKee et al. 2011), and a dictionary of German Sign Language (Deutsche Gebärdensprache, DGS) currently in progress (http://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/dictionary.html) based on Konrad et al.’s (2012) lexical database. Sign language lexicographers working on these dictionaries have referred briefly to issues regarding lemmatisation (particularly regarding the treatment of polysemy and in distinguishing phonological and lexical variants) in their respective publications (e.g., Kristoffersen and Niemelä 2008; Troelsgard and Kristoffersen 2008; McKee and McKee 2013; Zwitserlood et al. 2013). However, the most comprehensive attempt to define lemmatisation practices in the context of sign languages can be found in Johnston and Schembri (1999). Within this paper, principles of lemmatisation developed as part of the work on the Auslan dictionary are outlined and Johnston and Schembri advocate that the lemmatisation principles set out in this paper should be followed when compiling a dictionary or lexical database of any sign language.

Importantly, whilst there are some lemmatised dictionaries of sign languages currently, the vast majority of sign language dictionaries do not appear to have addressed what may be called the lemma dilemma (cf. Brien & Turner, 1994). Indeed, this appears to be the case for previous dictionaries of BSL and this is a major motivation for the development of a lexical database of BSL using the criteria set out in this paper, building on the work of Johnston and Schembri and others. Within the dictionary of BSL published by Brien (1992), as noted by Johnston and Schembri (1999), modified variants of a single lemma are often listed separately without any indication of their relation to one another. For example, one-handed and two-handed variants of a sign meaning ‘healthy’, ‘health’, ‘well’, ‘fine’, ‘fit’ are presented separately (see Figure 1) as are variants of the same lexeme (see Figure 2).
Submitted 14th November 2014 (check before citing)


**Figure 1.** Sign number 1494 and sign number 1489 from Brien (1992).

<table>
<thead>
<tr>
<th>Sign number 1195</th>
<th>Sign number 1196</th>
</tr>
</thead>
</table>

**Figure 2.** Sign number 1195 and Sign number 1196 from Brien (1992).

Our own examination of Brien (1992) also revealed that signs which have more than one unrelated meaning (i.e. manual homonyms) are also listed as a single entry. In the following example from Brien (1992), the sign form 989 has a range of English words associated with it covering the meanings of ‘pain’, ‘expensive’, ‘wow’ and the place name ‘Brighton’.

<table>
<thead>
<tr>
<th>Sign number 989</th>
</tr>
</thead>
</table>

**Figure 3.** Sign number 989 in Brien (1992).

Since some modified variants are presented separately and manual homonyms are conflated in Brien (1992), it is very difficult to gain a clear understanding of the core lexicon of BSL from this dictionary. Johnston and Schembri (1999) note that if consistent lemmatisation principles were applied to this dictionary, then this would very likely have the overall effect of reducing the number.
of unique signs listed. In other words, not following lemmatisation practices presents a distorted view of the core lexicon.

It is important to note that the increase in sign language corpora which have only begun to be available in recent years have highlighted the absence of a lemmatised resource for many sign languages. This has led sign language researchers to re-address the issue of lemmatisation and to develop such resources in parallel with corpus annotation. Prior to this, there was little attention given to the matter since many dictionaries were targeted at learners and thus were biased towards providing translation equivalents. Instead, sign language corpora require a system that identifies lexemes in a consistent way with a corresponding identifying gloss (i.e. an ID gloss, see below) so that more effective and reliable searches can be conducted.\(^6\) In other words, one can search for and retrieve tokens of a given lexeme quickly since it has been consistently labelled throughout the corpus regardless of how it might have been modified (for frequency lists based on lemmatised sign language corpora see Johnston 2012; Fenlon et al. 2014b). Without such a system, reports of lexical frequency of signs within a corpus (or any dataset) are inherently inaccurate and searching the corpus for all tokens of a given lexeme becomes a practically impossible task. In the following section, we describe BSL SignBank and its association with the BSL Corpus and discuss the notion of an ID gloss in further detail before outlining our principles of lemmatisation.

3 British Sign Language and BSL SignBank

3.1 British Sign Language

BSL is the language of the British deaf community. Whilst accurate statistics are difficult to obtain, a recent census report suggests that the number of signers in the United Kingdom is 15-20,000 (Office for National Statistics 2011). This figure is not without controversy and others have suggested that the number of signers is likely to be higher; the British Deaf Association report a figure of 120,000 extrapolated from a census report on the number of signers based in Scotland (http://www.bda.org.uk/British_Sign_Language_(BSL)). As with most sign languages in urban and industrialised countries, BSL’s emergence is associated with the establishment in 1760 of the first deaf school in the UK. Since deaf children are very rarely born to deaf parents who sign, deaf schools play an important role in language transmission and act as a child’s first point of exposure to a large community of signers. Although BSL has been the focus of linguistic research since the 1970s, there is still much about its structure and use that is poorly understood. As mentioned above, we currently do not have a dictionary that is representative of its lexicon. Recently, work has commenced on the creation of a BSL Corpus (Schembri et al. 2014) which contains a mix of naturalistic and semi-spontaneous data from 249 signers from around the UK. It is hoped that the
BSL Corpus will serve as a reference corpus and will enhance our understanding of BSL at all levels of the language. However, the process of annotation at the lexical level is dependent on the availability of a lemmatised resource to enable sign forms to be identified in a consistent way. As no such resource was available at the time of annotation, BSL SignBank was created to serve as a lemmatised resource including the core lexical signs in BSL; this is the focus of our paper here.

3.2 BSL SignBank

BSL SignBank (Fenlon et al. 2014a) is an electronic resource publicly available online that has a dual purpose as a lexical database for sign language researchers and as a dictionary for the wider sign language community. In this paper, we focus primarily on its role as a lexical database (though see section 6 for its role as a dictionary). Access to the lexical database is available to researchers who register as university staff or as research students (http://bslsignbank.ucl.ac.uk/accounts/register/).

At the time of writing, BSL SignBank consists of 2528 lexical signs of BSL (signs that are highly conventionalised in both form and meaning across the sign language community). Of these, roughly 1700 sign types occur in (a) 50,000 sign tokens from the BSL Corpus conversation data, annotated as part of a lexical frequency study (Fenlon et al. 2014b) and a project on directional verbs (Cormier et al. 2014) and (b) 295 sign types for colours, numbers and countries from the BSL Corpus lexical elicitation data annotated as part of a study on lexical variation (Stamp 2013; Stamp et al. 2014). Additionally, (c) roughly 700 sign types in BSL SignBank are from Brien (1992); these signs were added to BSL SignBank (and, where necessary, re-lemmatised with existing signs in SignBank) in order to ensure that the online dictionary would contain core vocabulary of BSL (as entry into BSL SignBank prior to this point was contingent on a sign appearing in the portion of the BSL Corpus that had been annotated to date). BSL SignBank is a living resource in that it will continue to grow as further annotation work on the BSL Corpus is carried out (i.e. new signs that are encountered in the corpus, following (re-)lemmatisation, are being added to BSL SignBank continuously).

For each lexeme within BSL SignBank, a set of keywords and a video showing the sign in its citation form is provided. Decisions regarding which variants were considered citation forms are beyond the scope of this paper but are covered in some detail in Cormier, Fenlon, et al. (2012a). The set of keywords are translation equivalents for the meaning of the sign and are designed to act as a proxy for definitions and to cover its full range of meaning. Additionally available in the lexical database is the ID gloss and a phonological description of each entry. The ID gloss can be seen as analogous to the headword or lemma for spoken languages (i.e. it is associated with the
unmodified citation form of a given sign). For each lexeme, the ID gloss is a unique English-based translation used primarily as an annotation tag in the corpus for all occurrences of that lexeme regardless of how it might be modified. It is important to note that the choice of the English word as an ID gloss for a particular lexeme is not meant to indicate the sign’s core meaning or grammatical function. It is merely a label to uniquely identify each lexeme, to be used in annotation of sign language data, in lieu of any standardised orthography for the language. For the purposes of annotation, given the current state of technology with annotation software tools such as ELAN, it is much more useful to use ID glosses that have some meaningful connection to the lexeme, e.g. via one of the translation equivalents, since annotation is done by typing in the ID gloss. In the future, it may be practical to have ID glosses which are meaningless codes, such as numbers or symbols. The important points here are: (1) it does not matter what particular label is chosen as an ID gloss for a given lexeme, since meaning and grammatical information is/will be provided separately in both SignBank and in annotation files for each lexeme and (2) ID glosses as labels for lexemes are necessary only for annotation purposes and therefore are not visible in the public dictionary view of BSL SignBank. A major motivation for making the ID gloss only made available to researchers as part of the lexical database is to avoid giving the impression that the ID gloss represents the meaning of that sign or that it has some kind of presiding status over the other associated English words. Researchers can use the ID gloss as a means to search and retrieve tokens of a particular sign in the BSL Corpus in order to conduct further analysis if they wish. The lexical database includes a powerful search tool which can be used not only for searching by English keyword but also by any other field available within the database, including phonological fields.

In Figure 4, an example of a BSL lexeme with its video, associated keywords, and ID gloss is provided. One can see that the keywords associated with the lexeme ID glossed as EXCITED cover a range of meanings represented by different English translation equivalents and that the citation form associated with this sign is considered to be the two handed variant (variants which are likely to be older, particularly those that can be assumed based on documented phonological processes, are often selected as citation form; see Cormier et al. 2012a).

<table>
<thead>
<tr>
<th>ID Gloss: EXCITED</th>
</tr>
</thead>
</table>
Figure 4. EXCITED in BSL SignBank

In the following sections, we frequently refer to signs within BSL SignBank using the ID glosses assigned to them and in some cases also by the set of keywords associated with each lemma. In each case, it is important to remember we are referring to the lemma that the ID gloss represents. The lemmatisation principles that we describe in the next two sections are those that emerged in the course of developing BSL SignBank, building on Johnston & Schembri (1999), and are intended to serve as a guide to others undertaking similar tasks. Such development requires regular discussions and consultations between sign language linguists and language consultants from the local deaf community.

4 Principles of lemmatisation

In the following sections, we outline both straightforward and problematic cases that we encountered when lemmatising BSL SignBank whilst outlining our basic criteria for lemmatisation. This section is divided as follows: first we describe criteria for distinguishing between morphological variants (both sign modification and sign formation), then we describe criteria for distinguishing between phonological variants with similar or related meanings based on phonological form alone, and then finally we focus on meaning/semantic criteria alone (e.g. determining when the meaning of a given token is different enough to warrant a separate lemma). Note that given the issues involved in attributing alterations in the form of a sign as inflectional or derivational (see Section 2), we use the alternative terms sign modification and sign formation here (respectively) instead (cf. Johnston and Schembri 2007). Finally, although we discuss form and meaning criteria separately, it must be noted that it is always necessary to refer to both the sign’s form and meaning to determine whether a given sign is a token of an existing lemma within the database or a separate lemma.

4.1 Morphology: sign modification

Our practice of lemmatisation involves grouping together morphological variants that are the result of sign modification – in other words, modifications of existing/known conventional signs that augment/change the meaning of the stem in predictable ways, both formationally and semantically, in sets of signs. In the case of BSL, variants marked for number, aspect, and what has come to be known as agreement in the literature (Padden 1983) are all assigned to the same lemma.
(representing the unmodified form). Examples of each are provided below; in some cases, exceptions are noted. Additionally, the use of sign modification with specific lexemes can assist in distinguishing the lexeme in question from other formationally similar signs (e.g. see CHILD vs. LOW below).

4.1.1. Aspect marking. Aspect marking is typically used with verbs or other predicates to exemplify how an action, event or state unfolds over time. For example, two variants of the sign WALK each mean ‘to walk’ and ‘to walk for a long time’ respectively (the latter having a larger circular motion associated with durational aspect).

In some cases, a token may appear to be an aspectual modification of a lexeme but instead has a specific meaning that cannot be predicted on the basis of the modification alone. For example, DRINK, with a single movement toward the lips, has the meaning ‘to drink any type of beverage’ whether it is a glass of water, juice, or alcohol. Durational aspect marking, e.g. with a large, repeated, inward motion, can be added to this sign to indicate drinking over a period of time. A related variant with slightly different movement (i.e. smaller, outward circular motion along the horizontal plane) means ‘alcohol’, ‘to drink alcohol excessively’, or ‘one who drinks alcohol excessively’ rather than ‘to drink any type of beverage excessively’. It is this specific meaning, which cannot be predicted from the addition of durative aspect to DRINK alone, that warrants a separate lexeme status for this variant, ALCOHOLIC. Lexemes like ALCOHOLIC demonstrate that not all tokens should be grouped together based on modifications to underlying forms alone but careful attention should be paid to the resulting change in meaning. If the change in meaning is unpredictable (as is the case in ALCOHOLIC), then separate lexeme status may be justified.

4.1.2. Directional modification. Some verbs can be spatially modified so that they are directed towards the physically present referents of their arguments, or locations associated with absent referents. These verbs are known as agreement verbs (Padden 1983), directional verbs (Baker-Shenk and Cokely 1978), or indicating verbs (Liddell 2000). There is no justification in assigning separate lexeme status to each directional variant as the resulting change in direction does not usually change the overall meaning of the sign (i.e., whatever its direction, the overall verbal meaning of the sign remains the same). In other words, these variants are all considered to constitute a single lexeme.

4.1.3. Number and distributive marking. Sign languages have a number of options available to mark number. For nouns, this includes repetition of the sign in question with a short sideways path movement (primarily for signs produced in neutral space). Verbs (such as the verb ASK, shown in Figure 6 in citation form) can also be modified for number/distribution using two different types of marking: the multiple and the exhaustive. Both types differ in movement and meaning: the former involves multiple repetitions of the verb towards several locations in space with a distributive
reading (e.g. ‘ask each’); the latter involves a single sweeping movement across the signing space and has a general plural or exhaustive reading (e.g. ‘ask more than one’ or ‘ask all’). Since the difference between these variants is always predictable, we do not consider these variants as separate lexemes.

![ASK](image)

**Figure 5.** ASK

Another way that signers may mark multiple referents is by using both hands instead of one (Johnston and Schembri 1999). Some one-handed signs (usually verbs) can be produced using both the dominant and non-dominant hands to indicate more than one referent (and, in some cases, to give a reciprocal meaning to the verb). For example, the sign LOOK2 is entered into BSL SignBank as a one handed sign which can often be produced with both hands in spontaneous signing to mean ‘two people looked ahead’. In such cases, both hands may be seen to be acting as independent signs as both appear to represent the action of different entities (in contrast to signs like WORK where the two hands produce a single sign). Alternatively, the hands may be oriented towards each other to give the meaning ‘two people looked at each other’ as in **Figure 6**. In each case, the resulting articulation is not treated as separate lexeme as there is no significant change in meaning.

![LOOK vs. LOOK-two-handed](image)

**Figure 6.** LOOK (‘look’, ‘see’, ‘watch’, ‘observe’, ‘gaze’, ‘stare’, ‘notice’) vs. LOOK with two hands

4.1.4. *Intensification.* The movement in some signs can be modified so as to intensify the meaning in some way. For instance, the signs RED and QUICK can be modified so that the beginning of the sign is held for longer than is usual to mean something is ‘very red’ or is ‘very
quick’. Generally speaking, this type of modification is not considered a separate lexeme (i.e. forms corresponding to ‘red’ and ‘very red’ will come under the same lemma RED as the two variants are formationally similar (i.e. they differ in only one parameter) and the change in meaning is predictable.

In some cases, a sign’s meaning can be intensified by using both hands (cf. the process noted above where a one-handed sign becomes a two-handed sign) – see Figure 7. As mentioned previously, two-handed variants generally do warrant a separate entry into a lexical database; and the use of both hands as a marker of intensification is treated no differently here.

Figure 7: WHAT (‘what’, ‘which’, ‘why’) vs. WHAT with two hands
4.1.5. Lemmatisation based on different sign modification patterns. If we have two variants with related meanings that are similar in form, the argument for recognising two separate lexemes may be based on whether one variant can take different morphological modifications. This is the case for POUND and POUND2; the lexical variant POUND cannot incorporate number but POUND2 can. This is also the case for HOUR and HOUR2. It is assumed that HOUR is the underlying sign in number incorporated tokens because in citation form it has a handshape which is already a number (i.e. 1) while HOUR2 does not. HOUR2, however, is considered to be a separate lexeme since it can be modified for aspect to mean ‘hours’/‘a long time’.

<table>
<thead>
<tr>
<th>POUND</th>
<th>POUND2</th>
</tr>
</thead>
</table>

Figure 8. POUND and POUND2

Likewise, both CHILD and LOW are formationally similar in citation form and have a similar general meaning of (something of) a low-level height. However, these signs can be modified differently. CHILD can be modified to indicate number (with a movement to the side) whilst LOW cannot (see Figure 10). Additionally, LOW can be modified for intensification with a longer downward movement (i.e. to mean something is very low); this kind of modification is not observed for CHILD. The fact that both take different types of modification is indicative of the fact that they represent different parts of speech: CHILD typically functions as a noun and LOW typically functions as an adjective. Similar claims can be made for the pair ADULT/HIGH.
Similarly, WOOD3 and HARD are identical in form (repeated movement of the thumb against the palm of the other hand; see Figure 15) and have related meanings. However, WOOD3 can be used to refer to any wood regardless of degree of hardness. This distinction in meaning justifies recognising these two forms as separate lexemes. Additionally, they are also considered to be separate lexemes because HARD can be modified in a way to mean ‘very hard’ via a change in the movement (i.e. single movement of both hands downward) and this does not apply to WOOD3.

In this section, we discuss criteria for determining lexemes in relation to sign formation. In contrast to forms that take sign modification, forms that result in sign formation are treated differently. By sign formation, we mean sign forms which appear transparently related to other sign forms but the effect of this assumed modification is to express a meaning which may not be fully predictable (as it would be with sign modification). In addition, this sign form could itself serve as a ‘stem’ which could be modified. In these cases, it is often justified to recognise two different
lexemes on this basis; the difference between these related forms may correspond roughly to what some consider to be derivational morphology in the literature.

4.2.1. Noun vs. verb alternation. One well-known example of sign formation commonly referred to in the sign language literature is the movement alternation observed in related noun and verb pairs in some sign languages (Supalla and Newport 1978; Sutton-Spence and Woll 1999). In these pairs, the movement associated with the noun is often shortened and repeated whilst the movement associated with the verb is longer and not repeated, as with signs like ASL FLY taken from Supalla and Newport (1978) (Figure 12).

![ASL FLY vs. ASL AEROPLANE](image1.png)

**Figure 12:** ASL FLY and ASL AEROPLANE

Despite claims by Sutton-Spence and Woll (1999), the extent to which this distinction exists in BSL is unclear (cf. Johnston & Schembri, 2007, for Auslan). Some related noun and verb pairs appear to differ systematically based on movement alone. These include pairs like KEY and LOCK, and DOOR and OPEN-DOOR. Some expected pairs however do not appear to have a systematic difference such as AEROPLANE, TEACH and FINGERSPELLING (i.e. there is no repeated movement when the sign functions as a noun). As a rule, only related pairs where the movement alternation appears to be systematic and not predictable based on iconicity alone are listed separately in the database (as is the case for KEY/LOCK where LOCK iconically involves a single forearm rotation movement as in a locking action, but KEY involves a repeated forearm rotation which is not how keys are typically used). In BSL, as in Auslan, it is often difficult to determine whether this alternation appears to be systematic for a given pair as there is rarely a clear consensus amongst signers as to whether one exists (Schembri et al. 2000; Johnston 2001a). Further analysis using the corpus data together with elicited data will determine this and current decisions may need to be revisited.

4.2.2 Number incorporation. Some signs in BSL can incorporate number signs. In some cases, the underlying sign is considered a bound morpheme which must be combined with a number sign in order to be a fully articulated sign. One example of such a sign is POUND2. Here, the bound morpheme is represented by the location (i.e. all number signs produced at the chin indicate the British currency) and the number is represented by the handshape (i.e. in this case, the number ONE...
is indicated by the extended index finger). In each case, all related tokens are represented by the same lemma. In other words, POUND2 always represents the resulting combination of the two morphemes. Other examples of such tokens include AGE and HOUR. These lexemes differ significantly from POUND2 because they have a base form that does not indicate number. All three examples are presented in Figure 13.

Figure 13. POUND2 (‘pound’, ‘pounds’, ‘sterling’) and POUND2-THREE, AGE (‘age’) and AGE-TWO, HOUR (‘hour’, ‘hours’) and HOUR-FOUR

In this case, we do not recognise each possible form as a different lexeme. One of the main reasons for doing so is because, in some cases, the number of possible combinations would be numerous and it is simply not economical to expect each individual form combination to be entered into the database (this would rapidly exaggerate the number of unique entries listed). Instead, a single entry (e.g. POUND2, AGE, and HOUR in Figure 13) is added to the database along with information that this form could incorporate number.

4.2.3 Negative incorporation. Another type of sign formation that occurs in sign languages is negative modification. In BSL, the negative marker is often an open 5 hand with forearm rotation movement. Signs that can be modified for negation include WILL, WANT, AGREE, and TRUE. For lemmatisation, these variants are recognised as separate from the underlying base variant and are recognised as unique lexemes (i.e. they are entered separately into the database); the examples listed above are all listed as WILL-NOT, WANT-NOT, AGREE-NOT, and TRUE-NOT in BSL SignBank; see Figure 14 for examples. The approach taken here is in contrast to the approach taken for number incorporation since which forms take negative incorporation is not predictable, and the
number of possible combinations involving negative incorporation appears to be much smaller than the number of possible combinations involving number incorporation.

<table>
<thead>
<tr>
<th>WILL</th>
<th>WILL-NOT</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>TRUE</th>
<th>TRUE-NOT</th>
</tr>
</thead>
</table>

**Figure 14.** WILL, WILL-NOT, TRUE, TRUE-NOT

4.2.4. **Compounds and collocations.** Compounding refers to the process of creating new signs from two or more free morphemes and is a common process attested in sign languages documented to date (Klima and Bellugi 1979; Johnston and Schembri 1999; Sutton-Spence and Woll 1999). In each case, compounds are considered unique lexemes and are entered into the database as such. Examples of compounds in BSL include PROMISE (SAY+TRUE), BLOOD (RED+SPREAD) and CHECK (SEE+MAYBE) as in **Figure 15**. The challenge for the sign language lexicographer is to determine when two signs represent a lexicalised compound (and therefore a unique lexeme in SignBank) as opposed to when they represent a collocation (i.e. two signs that appear next to one another frequently but do not represent a lexeme). Fortunately, the compounding process has been described in the sign language literature and developing criteria for compound status based on this work is straightforward (Johnston and Schembri 1999; Sutton-Spence and Woll 1999). Firstly, the
compounding process may alter the production of the sign phonologically so that they become a monosyllabic sign (Brentari 1998). Secondly, the meaning derived from the combination of these two morphemes is often not predictable. For example, BSL CHECK (a compound of SEE and MAYBE) has a range of English translation equivalents not predicted by the combination of SEE and MAYBE (see Figure 15). In some cases, the resulting compound may have taken on a broader or more specific meaning. For example, BLOOD is a compound of RED and SPREAD. Separately these two signs could refer to anything red that is spreading (e.g. spilled ketchup, red lava, etc.) but as a compound this sign refers only to blood. The fact that the resulting compound has an unpredictable meaning is an important point in support of the fact that these signs constitute separate lexemes (this is also standard practice in lemmatised lists representing words from English, cf. Sterkenburg 2003).

PROMISE

‘promise’, ‘honest’, ‘swear’, ‘truth’

CHECK


**Figure 15. PROMISE and CHECK**

In contrast, collocations are treated differently from compounds. Collocations refer to the frequent production of words next to one another. For example, most English speakers are likely to respond with the word ‘white’ or ‘blue’ when asked to complete the following fragment: ‘black and…’ because the words ‘black’ and ‘white’ or ‘black’ and ‘blue’ frequently appear next to one another in that specific order. Similar examples can be provided from BSL such as the signs MAKE
These pairs of signs represent concepts that frequently collocate in English (i.e. ‘make sure’: PT:PRO1SG WANT MAKE^TRUE, ‘I want to make sure’).

<table>
<thead>
<tr>
<th>MAKE</th>
<th>TRUE</th>
</tr>
</thead>
</table>

**Figure 16.** Citation forms for MAKE and TRUE

The case against granting MAKE^TRUE separate lexeme status (independently of its component signs MAKE and TRUE, as in Figure 19) is complicated somewhat when one observes their use in connected signing. That is, tokens can display a high degree of assimilation that appears to be conventionalised across signers. For example, MAKE followed immediately by TRUE is often articulated as a single monosyllabic sign with the handshape and orientation of the non-dominant hand in the sign TRUE anticipated at the start of the sign (see Figure 17). The processes by which this form is derived are identical to the ones attributed to the compounding process which can lead one to assume that what we are looking at is a compound.

**Figure 17.** MAKE^TRUE (high degree of assimilation)

However, there are important points that one should bear in mind. Whilst collocational pairs behave phonologically like compounds, their meaning remains predictable. That is, the meaning
that is derived from the combination of the signs MAKE and TRUE is not novel and can easily be predicted. Secondly, one should look at the degree to which a sign form is conventionalised across the sign language community. The production of the pairs MAKE^TRUE appears to be highly linked to signing speed; signers may delimit the pair more precisely when signing very slowly and precisely to their conversational partner. This line of reasoning leads us to conclude that there is no justification in recognising collocational pairs as a single lexeme.

In this section, morphological criteria for lemmatisation have been outlined. As noted previously, sign language lexicographers must also attempt to group together phonetic and phonological variants in a principled way without reference to a standard writing system. Therefore, additional phonetic and phonological criteria specific to sign languages also need to be developed. These criteria are outlined in the next section.

4.3 Phonetic and phonological criteria

In this section, we outline basic phonetic and phonological criteria for distinguishing between two related variants. Here it is often necessary to look at the extent to which two variants differ from each other by referring to the parameters for which each sign is specified. These parameters are: handshape, location, movement and orientation. Members within each parameter are often justified on the basis of minimal pairs. For example, the lexemes AFTERNOON and NAME differ in location whilst all other parameters are identical: AFTERNOON is produced with the same handshape, movement and orientation as NAME but is produced on the chin whilst NAME is produced on the forehead (see Figure 18) (e.g. handshape, location, movement and orientation).

![Figure 18: Phonological contrast along the location parameter in BSL](image)

It is important to remember that it is rarely the case that decisions regarding separate lexemes are made with reference to form alone. On a basic level, decisions often must be made based on form and meaning. In this section, we outline our criteria using straightforward examples from BSL SignBank and then discuss various possibilities with similar/different forms and meanings.
4.3.1. Phonological variants vs. lexical variants. If sign tokens A and B differ in only one phonological parameter (e.g. handshape, location, movement, orientation), and the meanings are the same or similar, then A and B are likely to be phonological variants. For example, the two variants of MOTHER shown in Figure 19 have the same meaning and differ only in handshape (i.e. the M-hand in which the index, middle and ring fingers are extended vs. the B-hand in which all fingers are extended). These two phonological variants are both related forms of MOTHER.

Figure 19. Phonological variants of MOTHER (‘mother’, ‘mum’, ‘mummy’)

In Figure 21, variants are distinguished via handshape. Related variants may also differ along the other parameters such as orientation (BOY), movement (DEBATE) or location (SLOW) (see Figure 20).
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 lexemes). For example, BSL \textit{NIGHT} is produced with two flat hands in neutral space, and \textit{NIGHT2} is produced with a bent-V handshape at the nose, as shown Figure \ref{fig:lex_var}. The ID glosses for these two lexical variants which have the same meanings (both have English translation equivalents of ‘night’, ‘tonight’, ‘evening’, ‘dusk’) are distinguished in BSL SignBank with numbers.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig20.png}
\caption{Example of related variants based on orientation, movement or location}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig21.png}
\caption{Lexical variants}
\end{figure}

4.3.2. **Phonological processes.** Straightforward instances of phonological variants that refer to a single lexeme include variants that are derived via documented phonological processes. One phonological process that is well documented is that some two-handed signs can be reduced to a one-handed sign, and/or that some one-handed signs can become two-handed so that a symmetrical second hand is added (Brennan et al. 1984; Johnston 1989; Brentari 1998). Earlier, we discussed additional use of the non-dominant hand where the resulting two-handed production had an effect on the overall meaning of the sign (e.g. it gave the sign a plural meaning or added intensification). Frequently, signs which allow this process may be produced as one-handed or two-handed variants in connected signing without any modification to the sign’s overall meaning and the two variants may exist in free variation (some research indicates that the presence or absence of doubling the additional non-dominant hand may be conditioned by the immediate phonological environment, e.g., Crasborn 2011). In BSL SignBank, one-handed and two-handed variants such as \textit{CAT(1-}}
handed) and CAT(2-handed) and RIGHT(1-handed) and RIGHT(2-handed) are always considered to constitute a single lexeme (Figure 22). This contrasts with common practice in many sign language dictionaries where one-handed and two-handed variants are often listed separately (e.g. see Figure 1 from Brien 1992).

![Figure 22](image)

**Figure 22.** CAT (‘cat’, ‘pussy cat’, ‘whiskers’) and RIGHT (‘right’, ‘correct’, ‘accurate’, ‘proper’): One hand versus two hand variants

Another phonological process relates to sign location. Some signs may be produced on the body or the non-dominant hand in citation form but can move to be located on other locations on the body or in space. Some of these involve predictable changes in meaning and are discussed in 4.4 below (e.g. EMIT or OPERATE). Other examples do not involve a change in meaning. For example, the sign KNOW may be produced on the head, or lowered in connected signing. Both are considered to be phonological variants of one lexeme, KNOW, as it is well known that signs can be lowered in connected signing and that this is can be conditioned by a number of linguistic factors (e.g., Schembri et al. 2006; Schembri et al. 2009). Another example of a locatable sign is PERSON which can be located in different areas of space to refer to a non-first referent but is not considered a different lexeme in each instance.

![Figure 23](image)

**Figure 23.** PERSON (‘person’, ‘character’, ‘client’, ‘identity’, ‘personality’, ‘individual’): first person and non-first person

4.3.3. *Phonetic modification/embellishment.* Some signs can be modified phonetically with no
resultant change in meaning. One well-known example of such modification in BSL is the change in handshape to a 5 hand in the final phase of the sign. For example, SAME and KILL can be modified so that the underlying handshape in the citation form changes to a 5 handshape in the final phase (see Figure 24). We use the term *phonetic embellishment* to refer to this type of modification.\(^{11}\)

![Figure 24](image)

**Figure 24.** Citation form and embellished forms for BSL SAME (‘sameness’, ‘same’, ‘alike’, ‘like’, ‘likewise’, ‘similarly’) and KILL (‘kill’, ‘murder’)

With other signs, phonetic modification may appear to result in a change in meaning but close attention must be paid to how consistent the form/meaning relationship is. The BSL sign VEHICLE has the keywords ‘car’, ‘automobile’, ‘van’, ‘vehicle’, ‘lorry’, ‘truck’, ‘driver’, and ‘bus’ associated with it. The sign iconically represents holding the steering wheel of a vehicle as if driving. This sign may be produced with the hands closer together or further apart (Figure 25). Signers may produce the sign with hands further apart for ‘lorry’ or ‘bus’ (i.e. vehicles associated with larger steering wheels), but there is no clear and consistent difference between these two forms in BSL (e.g. signers may use a form with a larger distance between the hands for any kind of vehicle). Because the distance between the hands does not constitute a phonological parameter but is a phonetic difference that varies gradually, and there is no evidence of consistent form/meaning differences, these are considered variants of one lexeme, VEHICLE.
Submitted 14th November 2014 (check before citing)


4.3.4. Repetition. Particular care must be taken when considering signs which may vary in terms of number of movement repetitions. If the meanings of these variants are the same or similar, then they are likely to be phonological variants. Examples of signs which may be repeated include HOT, PAST, NIGHT and CLEAR (see Figure 26). Each of these signs has a variant with repeated meaning that has the same meaning. In some cases, the repeated variant might be expected to co-occur more with mouthings corresponding to keywords that are multisyllabic (e.g. ‘heating’, ‘recent’, ‘evening’, ‘clarify’), but it is not clear to what extent this difference is consistent, there are likely counterexamples, and this is not enough to warrant separate lexemes.

Figure 26. Examples of repetition as a phonological variant within a lexeme

<table>
<thead>
<tr>
<th>HOT</th>
<th>PAST</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>NIGHT</th>
<th>CLEAR</th>
</tr>
</thead>
</table>
4.4 Meaning

In addition to the form of the sign, it is also necessary to refer to the sign’s meaning when determining whether two variants constitute the same lexeme or not. This is clearly required for homonyms; pairs of sign that differ in meaning but have the same phonological form. For example, both BSL BROTHER and MARCH-MONTH are produced with two A-hands in neutral space brushing against each other with alternating up-and-down movement (Figure 27).

![Figure 27. Homonyms – BROTHER and MARCH-MONTH](image)

In the case of BROTHER and MARCH-MONTH, the meanings of these signs are clearly different. As the meanings in each case are distinct, they are treated as homonyms and therefore separate lexemes. Another example is the form shown in Figure 3 (a single sign entry 989 from Brien 1992). This form is used in the homonyms BRIGHTON (‘Brighton’), EXPENSIVE (‘costly’, ‘dear’, ‘expensive’), PAIN (‘hurt’, ‘pain’, ‘painful’, ‘sore’), and WOW (‘wow’, ‘amazing’, ‘stunning’, ‘gosh’). Although they share the same form, the range of meanings of each of these four lexemes do not overlap so they have separate entries in BSL SignBank.

In some cases, the meaning represented by a single lexeme can also be very broad. For example, EXCITED is polysemous and displays a very broad range of meanings, including ‘excited’, ‘interested’, ‘motivated’, ‘eager’, ‘eagerness’, ‘ambitious’, ‘hobby’. These meanings are all related and they all share the same sign form (i.e. there is no difference in form when this sign is used to mean ‘excited’ or ‘interested’) as indicated in Figure 28 and Figure 29.

![Figure 28. Use of EXCITED to mean ‘interesting’](image)
The two entries from Brien (1992) presented in Figure 2 (sign number 1195 and 1196) are also an example of polysemy. These signs are each listed in Brien with keywords that denote a similar range of meaning. Additionally, both entries appear to be similar formationally and differ only in one aspect (Sign number 1195 is described as ending in both palms facing upwards). This difference in form is not enough to justify the identification of two separate lexemes and, within BSL SignBank, both are represented by the same lemma, ANGRY.

There are many cases in BSL SignBank where a single lexeme has several associated keywords demonstrating a broad range in meaning. TAKE includes the keywords ‘take’, ‘get’, ‘obtain’, ‘seize’, ‘grab’, ‘snatch’, ‘stealing’, ‘robbery’, ‘theft’, ‘burglary’, ‘robber’, ‘steal’, ‘rob’, ‘thieve’, ‘burgle’, ‘nick’, ‘pinch’, ‘shoplift’, ‘adopt’, ‘adoption’. BOY includes the keywords: ‘boy’, ‘boyfriend’, and ‘son’. Similarly BALL includes the keywords: ‘ball’, ‘round’, ‘sphere’, ‘spherical’, ‘football’, ‘world’, ‘globe’, ‘global’. In each case, there is no difference in the citation form of the sign used for each meaning and there is an obvious shared meaning between these keywords – e.g. the keywords for BOY and BALL all refer to a male human or a ball or something ball-shaped (respectively). Since these meanings are all related and they all share the same sign form, they are considered in each case to be part of one lexeme.

The broad range of meaning can also extend to include both abstract and concrete senses of a particular sign; often this is due to metaphorical extension of a concrete sense to an abstract sense. For example, the sign SPILL not only includes concrete senses linked to objects or people (‘spill’, ‘spread’, ‘disperse’) but also abstract senses such as ‘dissemination’. Similar examples can be found
Submitted 14th November 2014 (check before citing)

for ROAD where the keywords indicate both concrete senses (‘road’, ‘street’, ‘path’, ‘avenue’) and abstract senses (‘method’, ‘way’). ACCEPT can be used to mean ‘to accept an item from someone’, ‘to accept someone’s line of reasoning’ or ‘to catch an illness from someone’. Likewise, DESIRE is produced at the neck and can mean ‘thirsty’. As in the English word ‘thirst’, the use of DESIRE is not only applied to being thirsty for liquid nourishment but for other things as well, such as knowledge. Likewise, in BSL, it is also used to indicate desire of any kind. To represent this extension in meaning, DESIRE has a number of keywords such as ‘thirst’, ‘desire’, ‘thirsty’, ‘desirable’, ‘dry’, ‘fancy’, and ‘wish’. Again, as there is no difference in the form of the sign according to the senses displayed by the keywords, and because ‘desire’ is a metaphorical extension of the meaning ‘to be thirsty’, all of these keywords are assigned to one lexeme, DESIRE.

In many cases, the meanings exhibited by the keywords listed in BSL SignBank indicate words that have a superordinate/subordinate relationship to one another. A clear example of this can be seen with the sign BIRD which includes the keywords ‘bird’, ‘chicken’, ‘fowl’, ‘birdie’, ‘pigeon’. This list includes keywords that are co-hyponyms (e.g. ‘chicken’ and ‘pigeon’) as well as keywords that are superordinate/hyponyms such as ‘bird’. As there is no formational difference in the sign when used to refer to either a chicken or a pigeon or to any bird, there is no justification in acknowledging these items as separate lexemes.

In other cases, meanings of keywords in BSL SignBank have a metonymic relationship to each other. Examples include BADGE with keywords ‘badge’, ‘delegate’, ‘official’, ‘steward’, ‘supervisor’, ‘representative’, and also MEXICO with ‘Mexico’, ‘Mexican’, and ‘sombrero’. Our guiding principle is if the two forms are identical and the meanings are related, they refer to the same lexeme. Generally, this means that keywords which have a metonymic relationship with one another are grouped together under the same lemma. However, this is not always the case. For example, the sign COMMUNIST has keywords ‘communism’, ‘communist’, ‘feminism’, ‘feminist’, ‘militant’, ‘radical’, ‘socialism’, ‘socialist’, but is a separate lexeme from the identically formed sign RUSSIA which has the keywords ‘Russia’, ‘Russian’, ‘Soviet’, ‘Soviet Union’, ‘USSR’. One reasoning for this is because RUSSIA is a sign name which is highly institutionalised and therefore is recognised as a separate lexeme. Following this line of reasoning, one may argue that the keywords associated with MEXICO are indicative of two separate lexemes: MEXICO and SOMBRERO. However, this is not the case because there is little evidence that the sign which may be glossed as SOMBRERO is a conventionalised sign (i.e. is a conventionalised form/meaning pair) in BSL. For example, if a BSL signer were presented with this sign (without mouthing) and asked to tell us what it meant; it is unlikely that their first response would be ‘sombrero’ but most likely would be ‘Mexico’. A similar argument can be made for CHRISTMAS and BEARD2. Since the sign CHRISTMAS is an iconic depiction of Father Christmas’ beard, one could argue that the
Submitted 14\textsuperscript{th} November 2014 (check before citing)

separate lexeme status accorded to both CHRISTMAS and BEARD2 is not justified because the extension in meaning is similar to that amongst the keywords in MEXICO. However, although the sign CHRISTMAS is likely to have started as an extension of the sign BEARD to refer to Father Christmas as a sign name, it has now taken on a more general meaning (i.e. refers to Christmas, the holiday) and is highly conventionalised within the signing community (i.e. we would expect that signers would be just as likely to respond ‘Christmas’ and ‘beard’ when asked what this form means). There are many, many other examples like these with BSL SignBank and, in many cases, it can be difficult to decide when meanings are different enough to constitute separate lexemes (i.e. homonyms). In all cases, both form and meaning must be considered (see also section 4.4.1 below).

As has been well documented for many sign languages, new signs can be found that appear to be derived through a process known as loan translation (Battison 1978). These signs appear to be motivated by related meanings existing within BSL homonyms, where one or more of the signs involve a loan translation from English. For example, both the BSL signs HANGING and HANGOVER depict someone being hanged by the neck. With HANGOVER, this iconic depiction has no bearing on the meaning of being ill because of excessive drinking. Instead, what has happened here is that semantic extension has occurred with HANGING based on the English word ‘hang’ which comprises part of the word ‘hangover’. It is standard practice within BSL SignBank to recognise these entries (i.e. HANGING and HANGOVER) as separate lexemes provided they are considered to be in widespread use within the community (i.e. they are not restricted to one individual and occur several times across individuals within the corpus).

In addition, some signs can be relocated easily to different locations on the body or in the signing space with a slight change in meaning. For example, EMIT can be located and oriented in different ways for various meanings linked to emitting something (e.g. light or sound) as in Figure 30.

![Figure 30](image)

**Figure 30.** EMIT (‘radiate’, ‘emit’, ‘illuminate’) from different locations

All the examples in Figure 30 are considered to represent one lexeme, the general sign EMIT. This is because in each case, this token has not acquired a specific meaning. In other words, it continues to mean ‘something (a light source/sound source/etc.) emitted from somewhere’.
However, if the sign EMIT produced in high neutral space with the palm facing downwards (as in Figure 31) is presented to a BSL signer, it is highly likely that the signer will indicate that this sign has a specific meaning of ‘an overhead light’. Although this sign could feasibly refer to anything being emitted from that location (e.g. a speaker facing downwards), it has acquired a specific meaning and therefore is considered a separate lexeme.

![Figure 31. LIGHT](image)

Likewise, OPERATE can also be used productively in the same way as EMIT to refer to operating on any part of the body (e.g. by moving it to any body part such as the arm, neck, chest). However, when it is produced in certain locations, the overall form and meaning pairings appear more conventionalised. For example, when the sign is produced on the signer’s right side of the waist, it has the specific meaning of ‘appendectomy’. As a result, it is identified as a separate lexeme: APPENDECTOMY.

![Figure 32. OPERATE and APPENDECTOMY](image)

4.4.1. Multiply related variants: the case of CRUEL/KILL/MEAT/MURDER. Relationships between form and meaning across different variants can be quite complex. Here we describe one of the most complex set of relationships that we encountered in the course of annotating the BSL Corpus.
Submitted 14th November 2014 (check before citing)

There are seven BSL sign variants involving meanings linked to ‘killing’, ‘meat’, and ‘cruelty’, all produced at the neck, with overlapping/related forms and meanings (Figure 33).

<table>
<thead>
<tr>
<th>VARIANT 1: CRUEL/KILL</th>
<th>VARIANT 2: KILL</th>
<th>VARIANT 3: KILL/MEAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-hand with forearm twist</td>
<td>1-hand opening to 5-hand</td>
<td>1-hand touching neck</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VARIANT 4: MEAT</th>
<th>VARIANT 5: MEAT/MURDER</th>
<th>VARIANT 6: MURDER</th>
<th>VARIANT 7: MURDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-hand touching with repeated movement</td>
<td>A hand touching with repeated movement</td>
<td>A with thumb moving across throat</td>
<td>1-hand moving across throat</td>
</tr>
</tbody>
</table>

Figure 33. Seven sign forms linked semantically and phonologically in overlapping ways

The way we have lemmatised these variants is shown in Figure 34. These 7 variants form four separate lexemes (CRUEL, KILL, MEAT and MURDER), each with related phonological variants. At the endpoints of the chain, CRUEL (including only Variant 1) and MURDER (including Variants 5, 6 and 7) look quite different from each other and have different meanings. But by considering related lexemes KILL and MEAT (and their phonological variants, i.e. Variants 1 through 5), we can see how they all of these are related to each other, both in form and meaning. These multiply related variants highlight the importance of considering all possible related variants.
(in both form and meaning) when considering possible relationships between variants in the process of lemmatisation.

**Figure 34.** A chain of phonologically and semantically related variants

### 4.5 Mouthing and other non-manual features

In this section, we describe the effect of mouthing and non-manual features on decisions regarding lemmatisation. Phonological and morphological modifications are often accompanied by changes in non-manual features, but these changes are not enough to warrant separate lexemes. That is, lemmatisation is decided independently of non-manual features. For example, the English mouthing ‘forest’ frequently accompanies the sign for TREE when modified for number (i.e. it has a short sideways movement repeating the sign). However, the difference in form between the two variants is still considered to be grammatical. Although the number marked variant is frequently produced with a different mouthing, this is not considered to form a separate lexeme from TREE. Similar approaches are advised for other number marked variants accompanied with a specific mouthing (e.g. mouthing of ‘children’ with the plural marked version of CHILD is not enough to justify a separate lexeme for the plural form). Lemmatisation of these forms follows that of number and distributive marking as noted above in Section 4.1.3, regardless of mouthing. Strictly speaking, mouthing is only used as a cue for possible meanings in context (e.g. to develop the list of
submitted 14th November 2014 (check before citing)

associated keywords or definitions) rather than as justification for lemmatisation. This is true even if one mouthing is more likely to be used with a particular variant than another, as in the case of WORTH and WORTHb ‘worth’, ‘value’, ‘worthwhile’, ‘valuable’, ‘valued’, ‘appreciate’ (where the latter is initialised with the letter -W- and may be more likely to occur with mouthing ‘worth’ or ‘worthwhile’ than the other keywords).12

Figure 35. tree (‘tree’, ‘forest’, ‘wood’, ‘woods’)

Frequently, some signs can be modified for meaning using non-manuals features such as the mouth or facial expression. The type and number of mouth gestures that are attested in BSL is not well understood although Sutton-Spence & Woll (1999) describe 8 manner and degree mouth gestures that are important in BSL. For example, the verb DRIVE can be articulated with a specific adverbial mouth gesture which means ‘to drive carelessly’ (see Figure 36) (Liddell 1980; Lewin and Schembri 2011).

Figure 36. drive (‘driving’, ‘drive’) with th-mouth gesture

In cases such as these, it is not possible to recognise separate lexemes with every verb that has been modified with a mouth gesture – for example, the variants shown in Figure 36 are represented by the lemma DRIVE. (This and other mouth gestures which have a morphological function are thus treated the same as other sign modifications described in 4.1 above.) The same practice is applied to variants of signs that are formationally similar but differ in whether they express a negative or positive meaning based on facial expression/mouthing alone. For example, the sign EMOTION is understood to represent the full range of emotions. Signers may modify this sign

35
using a negative facial expression (accompanied with a specific mouth gesture) to convey an emotion or feeling described as disoriented. These variants are still considered to be modified variants of the lemma EMOTION.

Some signs when modified using facial expression may also modify the movement of the sign. For example, in the sign LOVELY (Figure 37.), the signer may use a negative facial expression to mean ‘not very nice’. In addition to this, the signer may also modify the path movement of the sign so the sign appears to be produced slower than usual and the final part of the sign is marked with a repeated downward movement. This type of modification has also been observed for other signs such as NICE (ending instead with a single twisting movement). In each case, the modification observed is best thought of as a prosodic modification to the production of the sign. That is, the type of modification described here is similar to an English speaker responding to something using a sarcastic tone ‘That was niiiice(!)’. In BSL SignBank, such variants are not recognised as separate lexemes. Both variants presented in Figure 37. are considered modified variants of the lemma LOVELY. This is because the difference in meaning is derived from prosodic modification rather than a major change in any of the underlying features of the sign (although the movement of the sign in LOVELY has been altered, the direction of the movement is still fundamentally the same).

Figure 37. Example of prosodic modification: LOVELY (‘wonderful’, ‘lovely’, ‘beautiful’, ‘nice’)

5 Issues in the creation of a sign language dictionary

Thus far we have outlined some of the main issues and problems involved in lemmatising a sign language lexical database. Lexical databases include core lexical signs of the language. Decisions of what signs to include or exclude depend on resources available. BSL SignBank is a partial representation of the lexicon based on signs from Brien (1992) and the BSL Corpus (Schembri et al. 2014). As more annotation is undertaken on the BSL Corpus, more signs will be added, and further lemmatisation/re-lemmatisation will take place with existing signs in SignBank (and hopefully leading eventually to a resource that is completely representative of BSL’s lexicon).
At a minimum, each entry within the lexical database should represent the lexeme in citation form. This is not to say that phonological and morphological variants should not be included; instead, their association with other related variants (e.g. the lemma and other variants derived from the lemma) should be acknowledged and consistently represented in the dictionary/lexical database. One way to do this is via double ID glossing within SignBank; one field that serves as head ID gloss (for the lemma, or citation form), and a separate field that gives a unique variant gloss for every entry in the database. We have done this in BSL SignBank via a field called Annotation ID gloss which represents the lemma or citation form (e.g. LOVELY2), and a separate field called ID gloss which represents the particular phonological variant (e.g. LOVELY2 or LOVELY2b) – see Figure 37. This allows us to identify both unique lexical items (via Annotation ID gloss, e.g. for computing lexical frequency) and also unique entries within the database (via ID gloss).

A lexical database serves as a representation of the lexicon for researchers. It can also usefully serve a function as a dictionary for a variety of users including learners. However, it is important to recognise that a dictionary may need to serve different functions from a lexical database. For example, the inclusion of entries that are only in citation form might not be appropriate for a dictionary. There are cases where phonological/morphological variants might have consistent enough differences in definitions and/or translation equivalents (keywords) to warrant separate entries, to aid in searchability and linking of particular meanings with particular forms, particularly for learners (e.g. LOVELY2 ‘lovely, nice’, LOVELY2b ‘not very nice’, etc.). Additionally, particular phonological or morphological variants may be linked to social factors and thus may warrant a separate entry in a dictionary (so that learners are aware of these variants). For example, the sign GREEN in BSL is produced with a flat hand with a movement up the arm and has been identified in a number of regions across the UK. This sign has a phonological variant which is identical except that the movement is down the arm; this variant so far has only been identified in Belfast (Stamp 2013). Although these two variants clearly constitute one lexeme (they have the same meaning and differ only in one parameter), for the purposes of a dictionary it is useful to have the down-arm variant listed separately since it appears to be linked to a particular social factor (in this case, region).

An additional (and separate) issue in the creation of a sign language dictionary is outlining clear criteria for distinguishing lexical signs which belong in the lexical database from other types of signs which do not. These types of signs may be fingerspelled sequences or partly-lexicalised constructions which can differ from lexical signs at the phonological level. Importantly, these types of signs can change over time so that they become lexicalised. However, it is not always straightforward to identify whether a given token encountered in annotation is indeed part of the
non-core or non-native lexicon (e.g. a partly-lexicalised sign, a pointing sign or a fingerspelled sequence etc.) or whether there is enough evidence that this token has become lexicalised such that it warrants entry in the lexical database. We have followed principles outlined in Cormier, Schembri and Tyrone (2008) and Cormier, Quinto-Pozos, Sevcikova and Schembri (2012b) in making some lexicalisation decisions, but explicit criteria for determining lexicalisation status for all signs originating from the non-core and non-native lexicon would be useful – we leave this for future research.

6 Conclusion

Here we have outlined the lemmatisation practices used in the creation of BSL SignBank, a lexical database and dictionary of British Sign Language. In doing so, we have provided some criteria for addressing the lemma dilemma for sign languages. The principles we have outlined here should be considered in the creation of any sign language lexical database and ultimately any sign language dictionary and reference grammar. Lemmatised lexical databases are necessary in the annotation of any sign language data in order to make any kind of generalisations about the lexicon of that sign language. Additionally, the development of large datasets (e.g. corpora) that are consistently annotated in this way at the lexical level will also present a significant benefit to those working within the areas of sign language synthesis and automatic recognition of sign languages. These technologies cannot move forward until there are large, annotated, lemmatised corpora to feed into these systems to train them.

7 Acknowledgments

This work was supported by the Economic and Social Research Council of Great Britain (Grants RES-062-23-0825, British Sign Language Corpus Project; ES/K003364/1, Describing sociolinguistic variation in verb directionality in British Sign Language: A corpus-based study; and RES-620-28-0002, Deafness, Cognition and Language Research Centre (DCAL)). We thank Robert Adam, Ramas Rentelis, Katherine Rowley, Rose Stamp, and Bencie Woll for contributing to the development of the criteria outlined here for lemmatisation of BSL SignBank. Figures 4, 10, 11, 14, 15, 18, 21, 23, 26, and 27 are from BSL SignBank (Fenlon et al. 2014a); Figures 28 and 29 are from the BSL Corpus (Schembri et al. 2014). We thank Sannah Gulamani for acting as model for the remaining figures. We also thank Trevor Johnston for very helpful comments on an earlier draft of this paper. All remaining errors are our own.

1 The first two authors contributed to this paper equally. First author was determined by a coin toss.

2 Lemmatisation issues for sign languages are likely be more similar to those of spoken languages that lack standard writing systems (see e.g. Chebanne 2010) than that that do have standard
orthographies. However, codification of spoken languages is made easier by the fact that there is always at least the International Phonetic Alphabet to start with; sign languages lack even this.

3 See section 4.3 for more on sign language phonology.

4 Unless otherwise noted, use of the term ‘lemmatisation’ in the remainder of this paper will refer to this extended notion of grouping together of phonetic, phonological and morphological variants and distinguishing these from lexical variants.


6 Traditionally, sign language researchers have tended to use contextual glosses for signs or to transcribe signs at a phonetic/phonological level (Miller 2006; Frishberg et al. 2012). Enriching corpora with either or both contextual glosses and phonetic/phonological transcription as an initial step is problematic since contextual glosses do not represent lemmas, and because transcription is so slow that it is practically inappropriate as a first step towards enriching corpus data (Johnston 2010, 2014).

7 As noted by Felbaum (2014), the terms ‘dictionary’ and ‘lexical database’ are often used interchangeably, particularly in the context of digital or electronic lexical and lexicographic resources. Here, we follow Janssen (2005) in distinguishing dictionaries (lexicographic resources) from lexical databases (resources meant for computational exploitation). There are sign language lexical databases that are research tools but not dictionaries, for example, for Swiss German Sign Language (Deutschschweizer Gebärdensprache, DSGS) and for Spanish Sign Language (Lengua de signos o señas española, LSE) (Boyes-Braem 2001; Costello and Carreiras 2013).

8 Currently, different levels of access to BSL SignBank are offered depending on the user. Researchers have access to the lexical database which includes ID gloss, keywords, video, phonological information etc. The public view is intended as a dictionary for the wider sign language community (deaf people, learners, interpreters, etc.) and shows only the video and associated keywords for each lexeme.

9 In annotation, if one wishes to indicate the sum represented by a specific token then one can use the corresponding ID gloss for the number in question (e.g. AGE-THIRTEEN). This would acknowledge that the form identified is a combination of two lexemes by using the corresponding ID gloss (as in Figure 13 for POUND2-THREE, AGE-TWO and HOUR-FOUR).

10 It should be noted that this assumed modification may or may not be diachronically true; evidence is often lacking.

11 It is not clear under which circumstances these embellished variants appear but it appears to have little or no effect on the meaning of the sign in BSL. Since these variants do not differ from one another significantly in phonology or in meaning, there is no case for recognising these variants as separate lexemes in BSL. In contrast, this type of modification is reported to have an effect on meaning in Auslan, resulting in a semantic shift (Johnston 1989). As a consequence, signs that have been modified in this way are likely to be considered as separate lexemes in Auslan, even though they would not in BSL.

12 It is important to remember that these keywords do not reflect the English mouthings that are used when the sign is produced in everyday signing but are a list of keywords that are associated with the sign’s overall meaning.
8 References

A. Dictionaries


B. Other Literature


Submitted 14th November 2014 (check before citing)


Johnston, T. 2001b. ‘The Lexical Database of Auslan (Australian Sign Language.)’ Sign Language &
Submitted 14th November 2014 (check before citing)


Johnston, T. in press. ‘The Reluctant Oracle: Adding Value to, and Extracting Value from, a Signed Language Corpus through Strategic Annotations.’ Corpora.


Submitted 14th November 2014 (check before citing)


Supalla, T. and E. Newport. 1978. ‘How Many Seats in a Chair? The Derivation of Nouns and Verbs in
Submitted 14th November 2014 (check before citing)


