Submitted: 21st July 2014

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

j.fenlon@ucl.ac.uk

Phone: +44(0)20 7679 8679

Fax: +44(0)20 7679 8691

Co-authors’ email addresses: k.cormier@ucl.ac.uk, a.schembri@latrobe.edu.au
Building BSL SignBank: The lemma dilemma revisited

Abstract

One key criterion when creating a representation of any lexicon within a dictionary or lexical database is that the lemma 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 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, lemmatisation practices used in the creation of BSL SignBank, a lexical database and dictionary of British Sign Language based on signs identified within the British Sign Language Corpus (http://www.bslcorpusproject.org), are outlined. The paper concludes 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, lemmatisation, sign language, dictionary, lexical database.

1 Introduction

When one begins to list all the words of a language in a dictionary, it is necessary to establish what one considers to be a lemma. Generally speaking, a lemma can be defined as the canonical form of a set of lexemes. For instance, in English, walk, walks, walking, and walked are all considered to be variant forms of the same 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). In principle, the situation with signed languages need not be any different. However, before one can begin to list all the signs in a sign language, one needs to develop clear criteria defining what they consider to be variant forms of the same lemma and what they consider to be separate lemmas (e.g. Johnston and Schembri 1999). Unfortunately, few sign language dictionaries have attempted to organise their entries in this manner so lemma identification for the purposes of lexical database building has not taken place for most 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, a lexical database and dictionary of British Sign Language (BSL) based on signs identified within the British Sign Language Corpus (Schembri et al. 2011). 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 foreign 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 Lemmatisation and sign languages

The traditional notion of lemmatisation for spoken languages refers to the practice of grouping together morphological variants as a single lemma. For example, in English, the variants jump, jumps, jumping, and jumped are all considered variants of a single lemma represented by its citation form (or the head 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 lemmas 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 as a single lemma since each variant is derived by the addition of an obligatory inflectional suffix marking tense or aspect. In contrast, teach and teacher would be recognised as two separate lemmas as the lemma 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 noun to verb. 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 to establish. 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, aspectral modifications and plural marking are optional, and the so-called verb agreement
system is best analysed as a reference tracking system that lacks agreement marking. 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). 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 such distinctions as they have for 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. 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. 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.

There have been few attempts to create sign language dictionaries following principles based on this extended notion of lemmatisation. Since the advent of sign language research, many dictionaries of sign languages (typically aimed at sign language learners) have been created simply by taking the spoken language of the surrounding community, listing key concepts from that spoken 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). Instead, these lexicographers advocate the practice of listing signs according to formational criteria specific to sign languages, a practice first adopted by Stokoe et al., (1965) for American Sign Language and replicated in the Auslan (Australian Sign Language) dictionary (Johnston 1989) and the Dictionary of British Sign Language/English (Brien 1992). However, Johnston & Schembri (1999) note that whilst this is the best approach, many sign language
dictionaries today have not really grappled with what one may call the lemma dilemma – i.e. lemmatisation. One such example is Brien (1992). Within this dictionary of BSL, Johnston and Schembri (1999) note that 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 modified variants of the same lexeme (see Figure 2 and Figure 3).


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’.

![Sign number 989](image)


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. Currently, there are very few lemmatised lexical databases/dictionaries available for any sign language. One 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 2001a), and most recently as an openly publicly accessible online dictionary and as a restricted access lexical database (http://www.auslan.org.au). The principles of lemmatisation which have been developed as part of this work on Auslan have been published in Johnston and Schembri (1999), and this output is the most comprehensive attempt yet to define lemmatisation practices in the context of sign languages. Johnston and Schembri advocate that the lemmatisation principles set out in their paper should be followed when compiling a dictionary/lexical database of any sign language.3

An additional (and separate) issue in the creation of a sign language dictionary is outlining clear criteria for distinguishing lexical signs from other components of the sign language lexicon. Building on previous work (Johnson and Liddell 1984; McDonald 1985; Brennan 1992), some researchers have proposed models of the sign language lexicon which divide the lexicon into the following components: the core native component, the non-core native component and the non-native component (e.g., Padden 1998; Brentari and Padden 2001). The core native component consists of fully lexical signs that are highly conventionalised in form and meaning and typically abide by a set of phonological constraints (Brentari and Padden 2001). It is this component that we
are primarily concerned with because fully lexical signs are the main candidates for inclusion in a sign language dictionary. Signs in the non-core native component (e.g. classifier signs) or the non-native component (i.e. fingerspelled sequences) are not typically listed in a dictionary since, unlike fully lexical signs, they are not standardised within a sign language community. However, tokens from these components may become lexicalised to such a degree that they can be recognised as lemmas. Determining criteria for when a given token is a core lexical sign is an issue that needs to be clearly outlined by the sign language lexicographer from the start (e.g. when can a fingerspelled sequence be considered a possible lemma?). This is an important issue and, although we do not discuss it in detail in this paper, we return to it briefly at the end.

Before we outline principles for lemmatisation, it is necessary to clarify the advantages that come with a lemmatised resource. Firstly, as demonstrated earlier using Brien (1992), not following lemmatisation practices will present a distorted view of the core lexicon of a sign language. This is not to say that modified variants should not be included; instead, their association with other related variants (e.g. the head lemma and other variants derived from the same head lemma) should be acknowledged and consistently represented in the dictionary/lexical database. Secondly, identifying lemmas and glossing them accordingly in a systematic way ensures that tokens representing a lemma can be quickly retrieved from large computerised datasets (or corpora). This was a major motivation for lemmatising BSL SignBank given its provenance from the BSL Corpus. Sign language researchers have traditionally 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 transcription is so slow that it is practically inappropriate as a first step towards enriching corpus data (Johnston 2010, in press). Using a system that identifies lemmas in a consistent way (e.g. ID glossing, see below) ensures that the corpus in question will be machine-readable – i.e. one can search for and retrieve tokens of a given lemma 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. 2014). Without such a system, reports of lexical frequency of signs within a corpus (or any dataset) are inherently inaccurate.

3 BSL SignBank

BSL SignBank was created to serve as a lemmatised resource listing all the core lexical signs in BSL. Its inception was motivated by the fact that, as mentioned earlier, existing BSL dictionaries had not been lemmatised consistently and such resources were required in order to adequately annotate the BSL Corpus at the lexical level (using a system of ID glossing) to ensure machine
readability and to determine lexical frequency. At the time of writing, the BSL SignBank consists of 2528 fully lexical signs of BSL. 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. 2014) 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). For each sign that is added to BSL SignBank, at a minimum, an ID gloss, keywords (i.e. English translation equivalents), and a video representing the sign in its citation form is also provided. Within SignBank, the ID gloss can be seen as analogous to the head lemma for spoken language (i.e. it is associated with the unmodified citation form of a given sign). For each sign, the ID gloss is a unique English-based gloss that is always used with that sign in annotation (regardless of how it might be modified). Each entry is also accompanied with a set of keywords designed to function as translation equivalents for the meaning of the given sign. Lastly, we also provide a video of each sign in unmodified citation form. 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). In Figure 5, one can see that the keywords associated with the form 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 (older variants, 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>
<tbody>
<tr>
<td>Keywords (translation equivalents): excitement, interest, keenness, excited, interested, keen, exciting, interesting, excite, interest, eager, enthusiasm, enthusiastic, stimulate, zeal, zest, hobby</td>
</tr>
</tbody>
</table>
In the following sections, we refer to signs within BSL SignBank using the ID glosses assigned to them as well as the set of keywords associated with each sign. The 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 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 lemmas and morphological variants (both sign modification and sign formation), then we describe criteria for distinguishing between lemmas and 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 – i.e. 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 the same ID gloss, which represents the unmodified form (i.e. citation form). Examples of each are provided below; in some cases, exceptions are noted. Additionally, the use of sign modification with specific lemmas can assist in distinguishing the lemma 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 an underlying lemma but instead has a specific meaning that cannot be predicted on the basis of the modification alone. For example, the lemma 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 lemma status for this variant (subsequently ID glossed as ALCOHOLIC). Lemmas 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 lemma status may be justified.

4.1.2. Grammatical agreement/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 some debate as to whether this modification is best considered morphological or gestural (i.e. a blend of linguistic units and pointing gestures) (e.g. Lillo-Martin and Meier 2011). However, whatever its status, there is no justification in assigning separate lemma 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). In other words, these variants are all considered to constitute a single lemma.

4.1.3. Number and distributive marking. Sign languages have a number of options available to mark the plural. 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; the latter involves a single sweeping movement across the signing space and has a more non-specific plural
meaning than multiple modification. Since the difference between these variants is always predictable, we do not assign separate lemma status to these variants.

![ASK](image)

**Figure 6. ASK**

Another way that signers may mark multiple referents is through the use of *doubling* (Johnston and Schembri 1999). For example, some one-handed signs (usually verbs) can be produced using the non-dominant hand to indicate more than one referent (and, in some cases, to give a reciprocal meaning to the verb). For example, the sign **LOOK**2 is entered into BSL SignBank as a one handed sign which can often be doubled 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 7**. In each case, the resulting articulation is not treated as separate lemma as there is no significant change in meaning.

![LOOK vs. LOOK doubled](image)

**Figure 7. LOOK vs. LOOK doubled**

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 assigned separate lemma status (i.e. forms corresponding to ‘red’ and ‘very red’ will come under the same ID gloss). A similar approach is taken for signs that correspond to the use of the comparative and superlative. For instance, **BETTER** is typically produced with a quick repeated movement in the comparative (i.e. ‘better’).
but with a single movement in the superlative (i.e. ‘best’). Within BSL SignBank, this type of modification is not considered to be a separate lemma as the two variants are formationally similar (i.e. they differ in only one parameter) and the change in meaning is predictable.

![Figure 8](image8.png)

**Figure 8.** BETTER (variant 1), BETTER (variant 2).

In some cases, a sign’s meaning can be intensified using doubling (as mentioned above, doubling refers to the process where a one-handed sign becomes a two-handed sign) – see **Figure 9**. As mentioned previously, doubled variants generally do warrant a separate entry into a lexical database; and doubling as a marker of intensification is treated no differently here.

![Figure 9](image9.png)

**Figure 9.** WHAT vs. WHAT-doubled
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 lemmas 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 lemma since it can be modified for aspect to mean ‘hours’/’a long time’.

![Diagram of POUND and POUND2](image1)

**Figure 10.** POUND and POUND2

![Diagram of HOUR, HOUR-TWO, HOUR2](image2)

**Figure 11.** HOUR, HOUR-TWO, HOUR2
Likewise, in the following example, 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 12). 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.

![Figure 12. CHILD and LOW](image)

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 lemmas. Additionally, these are also considered to be separate lemmas 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.

![Figure 13. WOOD3/HARD](image)

Finally, PERSON and PERSONAL are similar in form (the movement in PERSONAL is repeated) and have closely related meanings. These are however considered to be separate lemmas because PERSON (with a single movement downward) is a locatable sign which can be located in
different areas of space to refer to different referents, while PERSONAL (repeated version) cannot (i.e. it is body anchored).

![Figure 14. PERSON/PERSONAL (first person) cf. PERSON (non-first person).](image)

4.2  Morphology: sign formation

In this section, we discuss criteria for determining lemma status 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\(^4\) 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 assign a separate lemma status to two different variants on this basis; the difference between these related forms may correspond roughly to what some consider to be derivational morphology in the literature. In sign languages, however, the number of attested derivational morphemes is fewer than in spoken languages that have derivational morphology.

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 FLY taken from Supalla and Newport (1978) (Figure 15).
Despite claims by Sutton-Spence and Woll (1999), the extent to which this systematic distinction exists in BSL is unclear. 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 2001b). 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 marked using the same ID gloss. In other words, POUND2 always represents the resulting combination of the two morphemes. Other examples of such tokens include AGE and HOUR. These lemmas differ significantly from POUND2 because they have a base form that does not indicate number. All three examples are presented in Figure 16.
In this case, we do not recognise each possible form as a separate lemma. 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 dictionary (this would rapidly exaggerate the number of unique entries listed). Instead, a single entry (e.g. POUND2, AGE, and HOUR in Figure 16) is added to the dictionary 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 granted separate lemma status (i.e. they are entered into the dictionary); the examples listed above are all listed as WILL-NOT, WANT-NOT, AGREE-NOT, and TRUE-NOT in BSL SignBank; see Figure 17 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.
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 and frequent process attested in sign languages documented to date (Klima and Bellugi 1979). In each case, compounds are considered unique lemmas and are entered into the dictionary as such. Examples of compounds in BSL include PROMISE (SAY+TRUE), BLOOD (RED+SPREAD) and CHECK (SEE+MAYBE) as in Figure 18. The challenge for the sign language lexicographer is to determine when two signs represent a lexicalised compound (and therefore a unique lemma in the sign language dictionary) as opposed to when they represent a collocation (i.e. two signs that appear next to one another frequently). Fortunately, the compounding process is well documented in the sign language literature and developing criteria for compound status is straightforward. 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, including ‘check’, ‘checking’, ‘testing’, ‘trialing’, ‘trial’, ‘assessment’, ‘check-up’, ‘test’, ‘try out’, ‘assess’, ‘examine’, ‘examiner’, ‘experiment’, ‘inspector’, ‘inspect’, ‘survey’, ‘surveyor’. 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 lemmas (this is also standard practice in lemmatised lists representing words from English, cf. Sterkenburg 2003).
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 and TRUE (see Figure 19). These pairs of signs represent concepts that are also frequent collocates in English – i.e. ‘make sure’: PT:PRO1SG WANT MAKE^TRUE, ‘I want to make sure’.

Figure 18. PROMISE and CHECK

Figure 19. Citation forms for MAKE and TRUE
The case against granting MAKE\(^{\text{TRUE}}\) separate lemma 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, this sign 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 20). 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.

![MAKE\(^{\text{TRUE}}\)](image)

**Figure 20.** MAKE\(^{\text{TRUE}}\) (high degree of assimilation)

However, there are important points that one should bear in mind. Whilst collocational pairs behave phonologically like compounds, they remain predictable. For example, 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\(^{\text{TRUE}}\) appear 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 lemma.

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 (e.g. handshape, location, movement and orientation). It is important to remember that it is rarely the case that
decisions regarding separate lemmas 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 of one lemma. For example, the two variants of MOTHER shown in Figure 21 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 the lemma represented by the same ID gloss.

![Figure 21. Phonological variants of lexeme MOTHER: MOTHER(M-hand) and MOTHER(B-hand)](image)

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 22).
Figure 22. Example of related variants based on orientation, movement or location

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 NIGHT is produced with two flat hands in neutral space, and NIGHT2 is produced with a bent-V handshape at the nose, as shown Figure 23. 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.

Figure 23. Lexical variants

4.3.2. Phonological processes. Straightforward instances of phonological variants that belong to a single lemma also 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 undergo doubling so that a symmetrical second hand is added (Brennan et al. 1984; Johnston 1989; Brentari 1998). Earlier, we discussed instances of doubling where the resulting 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 may be doubled (or singled) in connected signing without any modification to the sign’s overall meaning and this doubling may exist in free variation (some research indicates that the presence or absence of doubling may be conditioned by the immediate phonological environment, e.g. Crasborn 2011). In BSL SignBank, one-handed and two-handed variants such as CAT(1-handed) and CAT(2-handed) and RIGHT(1-handed) and RIGHT(2-handed) are always considered to constitute a single lemma (Figure 24). 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).
Another phonological process relates to locatable signs. 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 lemma, 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).

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, the lemmas 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 25). We use the term phonetic embellishment to refer to this type of modification.

Figure 24. One hand versus two hand variants

Figure 25. Citation form and embellished forms for BSL SAME and KILL
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 assigning these variants separate lemma status in BSL. In contrast, this type of modification is reported to have an effect on meaning in Auslan. The effect is somewhat predictable as a kind of intensifier, but sometimes results in a semantic shift (sometimes, narrowing, sometimes extension). For example, in Figure 26, the insertion of a final phase of the sign KNOW in Auslan in which the hand moves away from the head and opens to a 5 hand is understood to mean ‘famous’ or ‘well known’ and the addition of handshape change to the final phase of the sign SAME in Auslan means ‘get used to’ (Johnston 1989). As a consequence, signs that have been modified in this way are likely to be assigned separate lemma status if there is a significant (ultimately unpredictable) change in meaning, as in Auslan with this opening 5-hand modification. This, however, does not appear to be the case for related pairs in BSL.

![Figure 26. Auslan KNOW/FAMOUS and SAME/GET-USED-TO](image)

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 27). 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 gradiently, and there is no evidence of consistent form/meaning differences, these are considered variants of one lemma, VEHICLE.

![Figure 27. Example of phonetic gradience within one lemma](image)

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 of a single lemma. Examples of signs which may be repeated include HOT, PAST, NIGHT and CLEAR (see Figure 28). 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 lemmas. This could be considered to be similar to echo phonology (Woll 2001) except in this case instead of the mouth mirroring the hands as in mouth gestures, the movement of the hands is mirroring the mouthing of the sign.

![Figure 28. Examples of repetition as a phonological variant within one lemma](image)
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 lemma or not. This is clearly required for homonyms – i.e. 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 29).

Figure 29. Homonyms – BROTHER and MARCH-MONTH

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 each assigned a unique ID gloss. In some cases, the meaning represented by a single lemma can also be very broad. For example, the keywords associated with the ID gloss EXCITED display a very broad range of meanings: ‘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 30 and Figure 31 showing examples from the BSL Corpus (Schembri et al. 2011).

Figure 30. Use of EXCITED to mean ‘interesting’
I took my boy (to the football). (He was) very excited.

Figure 31. Use of EXCITED to mean ‘excited’

There are many cases in BSL SignBank where a single ID gloss has several associated keywords demonstrating a broad range in meaning. TAKE includes the keywords ‘take’, ‘get’, ‘obtain’, ‘seize’, ‘grab’, ‘snatch’, ‘stealing’, ‘robbery’, ‘theft’, ‘burglary’, ‘rober’, ‘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 lemma.

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 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 English with ‘thirst’, the use of this sign 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 lemma, 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/hypernyms 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 according these items separate lemma status.

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 are part of the same lemma. 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 lemma 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 lemma. Following this line of reasoning, one may argue that the keywords associated with MEXICO are indicative of two separate lemmas: 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 separate lemma 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 lemmas (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 ‘hung’ which comprises part of the word ‘hungover’. It is standard practice within BSL SignBank to recognise these entries (i.e. HANGING and HANGOVER) as separate lemmas 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 32.

![Figure 32](image-url)

**Figure 32.** EMIT from different locations
All the examples in Figure 32 are considered to be tokens of one lemma, 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 is produced in high neutral space with the palm facing downwards as in Figure 33 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 lemma.

Figure 33. LIGHT

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 lemma: APPENDECTOMY.

Figure 34. OPERATE and APPENDECTOMY
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 lemma relationships that we came across for BSL in the course of annotating the BSL Corpus.

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 35).

<table>
<thead>
<tr>
<th>VARIANT 1</th>
<th>VARIANT 2</th>
<th>VARIANT 3</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</th>
<th>VARIANT 5</th>
<th>VARIANT 6</th>
<th>VARIANT 7</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 35.** Seven sign forms linked semantically and phonologically in overlapping ways

The way we have lemmatised these variants is shown in Figure 36. These 7 variants form four separate lemmas (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 lemmas 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.
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 lexical status. That is, lemmatisation is decided independently of non-manual features. For example, the English mouthing ‘forest’ is frequently accompanied with the sign for TREE 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, a separate ID gloss is not created to accommodate this variant. 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 lemma 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 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).⁶

Figure 37. TREE (modified for number)

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 38) (Liddell 1980; Lewin and Schembri 2011).

Figure 38. DRIVE with th-mouth gesture

In cases such as these, it is not possible to assign separate lemma status to every verb that has been modified with a mouth gesture – e.g. the sign shown in Figure 38 is still ID glossed as DRIVE and not DRIVE-CARELESSLY. 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 using a negative facial expression (accompanied with a specific mouth gesture) to convey an emotion or feeling described as disoriented. These variants are still consistently referred to using the same ID gloss: EMOTION.

Some signs when modified using facial expression may also modify the movement of the sign. For example, in the sign LOVELY (Figure 39), 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 lemmas and instead are represented with the same ID gloss. Both variants
presented in Figure 39, are represented using the same ID gloss (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).

<table>
<thead>
<tr>
<th>LOVELY</th>
<th>LOVELY (prosodically modified)</th>
</tr>
</thead>
</table>

Figure 39. Example of prosodic modification

5 Lemmatisation of Brien (1992) entries revisited

The lemmatisation principles that we have outlined in Section 4 are principles that should be
prioritised in the creation of a sign language lexical database. If we were to apply the lemmatisation
principles outlined here to the examples presented in Figure 1 – 4 from Brien (1992), we would see
that this would present a clearer picture of the number of individual lemmas within those entries.
For example, the two separate entries in Figure 40 would be recognised as a single lemma
following Section 4.3.2 where one handed and two handed variants are often grouped together
under a single head lemma. The two entries presented in Figure 41 would also be seen as modified
variants of the same lemma since they do not differ significantly on a phonological level (i.e. only
one parameter is different) and they share the same set of keywords and meanings.
Figure 40. Sign nos. 1494 and 1489 from Brien (1992) as one lemma (HEALTHY) in BSL SignBank

Similarly, the signs presented in Figure 41 are also considered to be variant forms of one lemma. These signs are each listed with a number of 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 is not enough to justify the identification of two separate lemmas and both are identified within SignBank under the same ID gloss, ANGRY.

Figure 41. Sign nos. 1195 and 1996 as one lemma (ANGRY) in BSL SignBank

The issue with the two entries in Figure 42 is also straightforward. The two entries can be considered related variants since one is modified for number. As a result, both entries are recognised as variants of one lemma – BOOKED and BOOKEDb both with keywords ‘book’, ‘signature’, ‘sign’, ‘register’, ‘note’, ‘minutes’, ‘list’, ‘petition’.

BOOKED (Sign number 720)  BOOKEDb (Sign number 721)
Finally, the single entry from Brien (1992) presented in Figure 43 would be separated into four ID glosses in BSL SignBank – i.e. as homonyms – since the keywords associated with this entry show a number of meanings they may or may not overlap to some extent.

In Figure 43, one can see that the range of keywords associated with this form includes ‘expensive’, ‘pain’, ‘Brighton’ and ‘wow’. Firstly, BRIGHTON is separated from the other meanings because it is a sign name (as is also the case with RUSSIA vs. COMMUNIST) and is a highly institutionalised place name used in the British Deaf community. Although the meanings are clearly different, one could argue that the meaning displayed by ‘expensive’ is an extension of the meaning ‘pain’ (e.g. the price of an object is so hard that it hurts the bank balance). More importantly, the separation of the two lemmas is further strengthened by the existence of well-known phonological variants associated with both signs. In Figure 44, one can see that the two handed variant associated with EXPENSIVE differs from the two handed variant associated with PAIN (which is symmetrical) and these variants do not share the range of keywords associated with the form presented in Figure 43. The fact that the one-handed variant of EXPENSIVE and one-handed variant of PAIN can be derived from different underlying citation forms is evidence for the separate lemma status of the two. A similar argument can be constructed for WOW (i.e. the two-handed variant of EXPENSIVE does not have the keyword ‘wow’ associated with it).
6 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 representation of the lexicon based on signs from Brien (1992) and the BSL Corpus (Schembri et al. 2011). 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. At a minimum, each entry within the lexical database should represent the lemma in citation form. Phonological and morphological variants may be added, provided that they are marked in such a way as to distinguish them from citation form. One way to do this is via double ID glossing within SignBank – i.e. one field that serves as head ID gloss (for head 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 head lemma or citation form (e.g. BOOKED), and a separate field called ID gloss which represents the particular phonological variant (e.g. BOOKED or BOOKEDb) – see Figure 42. 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. BETTER ‘better’, BETTERb ‘best’; LOVELY ‘lovely, nice’, LOVELYb ‘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 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 lemma (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 is linked to a particular social factor (in this case, region).

Additionally, it is clear that the non-core part of the lexicon and non-native part of the lexicon contain elements that become lexicalised. It is not always straightforward to identify whether a given token is indeed part of the non-core or non-native lexicon (e.g. a classifier sign, a token of constructed action, a pointing sign or a fingerspelled sequence) or whether there is enough evidence that this token has become lexicalised such that it warrants entry in the lexical database. In general, we have followed principles outlined in Cormier, Schembri and Tyrone (2008), Cormier, Quinto-Pozos, Sevcikova and Schembri (2012b) and Cormier, Schembri and Woll (2013) in making these decisions, but explicit criteria for determining lexicalisation status for signs from the non-core and non-native lexicon would be useful – we leave this for future research.

7 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, lemmatisation is crucial for achieving machine readability of sign language data which is important given the recent advances in sign language technology (e.g. 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.

8 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. We thank Sannah Gulamani, Radha Madjeswar, Mark Nelson, and Fiona Garfield for acting as models for the BSL
signs shown here. 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 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.
4 It should be noted that this assumed modification may or may not be diachronically true; evidence is often lacking.
5 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 lemmas by using the corresponding ID gloss (as in Figure 16 for POUND2-THREE, AGE-TWO and HOUR-FOUR)
6 It is important to remember that these keywords do not reflect the English words 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.
9 References

A. Dictionaries


B. Other literature


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


