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

Kearsy Cormier¹, Jordan Fenlon¹, Trevor Johnston², Ramas Rentelis, Adam Schembri³, Katherine Rowley¹, Robert Adam¹ & Bencie Woll¹
DCAL, University College London¹, Macquarie University², La Trobe University³

Corpus annotation
A representative, accessible and machine-readable linguistic corpus requires annotation, which requires lemmatisation – i.e. grouping together phonological & morphological variants into one lemma. Dictionaries of few sign languages have been consistently lemmatised to allow for ID glossing (type-token matching), aside from Australian Sign Language (Auslan) (Johnston 2001, 2003).

Lexical database vs. dictionary: Social factors
For a dictionary not including phonological variants, social factors should be taken into consideration. Based on strictly linguistic principles, GREEna and GREEnb are phonological variants of one lemma. However, regional distribution of these variants in the BSL Corpus suggests they should both be listed in a dictionary.

Criteria for citation form/headword
★ Highest frequency (or assumed frequency)
★ Predictability based on documented phonological processes, historical changes (e.g. linked to iconicity), and/or nativisation processes (e.g. linked to fingerspelling)
★ Association with a priority social group
★ Listing in existing national dictionaries
All criteria considered but social factors and listing in other dictionaries are rarely considered alone