Using conversational data to determine lexical frequency in British Sign Language: The influence of text type

Jordan Fenlon [a,*], Adam Schembri [b], Ramas Rentelis [a],
David Vinson [a], Kearsy Cormier [a]

[a] Deafness, Cognition & Language Research Centre, University College London, 49 Gordon Square,
London WC1H 0PD, United Kingdom
[b] La Trobe University, Melbourne (Bundoora), Victoria 3086, Australia

Abstract

This paper presents findings from an objective lexical frequency study in British Sign Language (BSL) based on 24,823 tokens collected as part of the BSL Corpus Project. The BSL study is only the fourth objective frequency study involving sign languages to be reported and is also the first study for any sign language to be based on entirely on spontaneous conversational data. When compared to previous frequency studies (both spoken and signed), some similarities can be observed although differences that may be attributed to text type are also recorded. When compared with subjective frequency ratings collected for BSL, a positive relationship is reported (similar to what has been observed for spoken languages). This is in contrast to a previous study which suggested a much weaker relationship between the two; however, this conclusion was based on a frequency count derived from narratives. These differences highlight the importance of using frequency measures derived from natural and spontaneous data, an opinion that has been emphasised in the spoken language literature.
© 2014 Elsevier B.V. All rights reserved.

Keywords: Objective frequency; Subjective frequency; Sign languages; Text type; Conversation

1. Introduction

The effects of lexical frequency (i.e., how often a word occurs in a language) have broad consequences for grammaticalisation, language processing, sociolinguistic variation and change as well as first and second language acquisition (Bybee, 2002; Ellis, 2002) with such effects being well attested in studies of spoken language for over 60 years (e.g., Howes and Solomon, 1951). The lack of adequate frequency information for sign languages has created a significant barrier to the progress of research on these languages, with only three studies of objective frequency measures conducted to date: on American Sign Language (ASL), Australian Sign Language (Auslan), and New Zealand Sign Language (NZSL). As information regarding objective lexical frequency in British Sign Language (BSL) has not been readily available, researchers have sought to address this gap by collecting subjective frequency ratings for 300 BSL lexical signs from 20 deaf signers (Vinson et al., 2008). The current paper presents the findings from the first study of BSL

on objective measures of lexical frequency based on 24,823 sign tokens collected as part of the BSL Corpus Project (Schembri et al., 2011). The BSL study is also the first frequency study for any sign language to be based entirely on spontaneous conversational data. The results from this study underline the importance of text type when investigating lexical frequency in sign languages. Differences in the frequency of specific lexical items and sign categories across the sign language studies discussed here appear to be largely due to the type of texts contained within each dataset. Additionally, when we compare the subjective frequency ratings collected by Vinson et al. (2008) to the frequency of the same lexical items in the spontaneous conversational data reported here, we find that the link between subjective frequency ratings and objective frequency in BSL may be stronger than suggested by Johnston (2012). In Johnston's study, the same subjective frequency ratings were compared to a lexical frequency count consisting of a high proportion of narrative data collected from the closely related sign language variety, Auslan.

2. Background

2.1. Why is frequency important?

The need to consider lexical frequency has been emphasised by many researchers working within several sub-disciplines in linguistics. For example, proponents of a usage-based approach to language acquisition have underlined the crucial role that lexical frequency plays in the development of a child's understanding of syntactic structure and in forming grammatical generalisations (Kidd et al., 2010; Langacker, 2009). Researchers investigating language in use have described how frequently used words tend to be phonetically reduced over time (Bybee, 2002; Bybee and Scheibman, 1999; Diessel, 2007; Dinkin, 2007; File-Muriel, 2010; Guy, 2007; Philips, 1984) and how frequency interacts with grammaticalisation processes (Bybee, 2006; Bybee and Hopper, 2001). Within the field of language processing, online studies have repeatedly stressed that people's unconscious understanding of lexical frequency, formed from their past language experiences, is a major factor affecting the learning, recognition, comprehension and production of both words (e.g., Balota and Chumbley, 1984; Forster and Chambers, 1973; Gregg, 1976; Hall, 1954; Howes, 1954; Howes and Solomon, 1951; Savin, 1963) and sentences (Juliano and Tanenhaus, 1993; Jurafsky, 1996; Trueswell, 1996). Within applied contexts, lexical frequency is also important for lexicographers (McEnery and Xiao, 2011) and for those involved in curriculum design for second language learners (Nation, 2001). In sum, information about lexical frequency is vital for understanding the usage, processing and acquisition of the lexicon of any language.

Frequency data are typically obtained from studies of written or spoken language corpora[1] which have become increasingly available since the 1960s (e.g., Brown's Corpus of American English). With the advent of modern computing technology, it is now possible to obtain reliable frequency lists from extremely large corpora, such as the British National Corpus (100 million words) or monitor corpora such as the Bank of English (650 million words as of 2012). Frequency lists produced as a result of these corpora (e.g., Francis and Kucera, 1982; Leech et al., 2001) have in turn been used by researchers across the language sciences, as in the studies noted above.

Recent research examining frequency measures obtained from different corpora have revealed that it is the type of data that make up a corpus, as opposed to its size, that plays a key role in obtaining reliable frequency measures. Brysbaert and New (2009) conducted a detailed comparison of the relationship between different frequency measures and lexical decision latencies, inspired in part by the continuing usage of the relatively small and (now) dated frequency norms from Kučera and Francis (1967). Whilst increasing corpus size led to improved performance, only marginal gains were observed beyond a certain point. Instead the source of word frequency measures plays an important role, with text derived from speech, Internet discussion groups or subtitles from TV and film predicting performance far better than more traditional written sources such as books, journals and other such publications. Additionally, Brysbaert and New (2009) argue that data derived from natural and spontaneous circumstances accounts for lexical decision times better than prepared texts.

2.2. Objective lexical frequency in sign languages

There have only been three studies on lexical frequency in sign languages using objective measures, primarily because sign language corpora are still in their infancy. The first project to investigate the distribution of lexical items in a sign language was by McKee and Kennedy (2006) which, drawing on a dataset of 100,000 sign tokens in the Wellington Corpus of NZSL from 80 NZSL signers, remains the largest objective frequency study to date. This was followed by a

---

[1] We use the word 'corpus' to mean a large, representative, accessible and machine-readable dataset of language recordings (McEnery and Wilson, 2001).