A Lexical Frequency Analysis of Irish Sign Language

Robert G. Smith & Markus Hofmann

Technological University Dublin

Robert.smith@tudublin.ie, markus.hofmann@tudublin.ie

Abstract

As word frequency has a significant impact on language acquisition and fluency, it is often a point of reference for the teaching and assessment of a language and as a control for psycholinguistic studies. This paper presents the results of the first objective frequency analysis of lexical tokens from the Signs of Ireland corpus. We investigate the frequency of fully lexical (phonetically constrained and listed in the lexicon), partly lexical (phonetically unconstrained and listed in the lexicon) and non-lexical signs (not listed in the lexicon) in Irish Sign Language as they are presented in the corpus. We compare the accuracy of the lexical gloss frequency data with a supplementary corpus subset that is tagged for grammatical class and with results from previous lexical frequency studies conducted for American Sign Language, Australian Sign Language, British Sign Language, and New Zealand Sign Language. This study has found that, in the main, frequency statistics from Irish Sign Language are in line with previous studies and that the text type and annotation strategy can significantly impact results. We found that, without a formalised lexicon, lexical glosses fell short of the requirements for a lexical frequency analysis. However, supported by grammatical class data, frequency data may be reported for various symbolic units.

Keywords: *sign language, corpus annotation, grammatical class, frequency analysis, Irish Sign Language (ISL)*

1. Introduction

The core aim of this paper is to explore how the Signs of Ireland (SOI) corpus dataset functions with respect to the derivation of a word frequency list. In doing so, we highlight some of the limitations in the composition of the dataset as it currently stands, particularly with regards to annotation consistency and text type, ultimately motivating an argument for the future development of a lexical database. We illustrate that the deployment of a smaller, more comprehensively annotated subset of the corpus offers some additional insight and goes some way towards increasing the reliability of the frequency analysis results. In addition, we compare results from the SOI frequency analysis to results from similar studies for four different sign languages: American Sign Language (ASL), British Sign Language (BSL), Australian Sign Language (Auslan) and New Zealand Sign Language (NZSL). On the one hand, this side-by-side comparison offers some insight into how ISL compares with other sign languages at the lexical level. On the other hand, such a comparison serves as leverage to improve confidence in the findings of the frequency analysis.

Understanding the frequency of the lexical word in regular language exchange can be insightful in the development of L1 and L2 curricula and assessment. This word frequency effect is commonly considered in applied linguistics, psycholinguistic and diachronic studies with respect to spoken languages, its effects reported: *"in the processing of phonology, phonotactics, reading, spelling, lexis, morphosyntax, formulaic language, language comprehension, grammaticality, sentence production, and syntax"* (Ellis 2002, p. 143). The recognition and production of words, morphemes and even phrases is a function of their frequency of occurrence in a language. Thus, word frequency has a significant impact on language acquisition and fluency (Ellis, 2002; Conrad, 2005).

Studies investigating such phenomena in spoken languages may introduce controls for word frequency to ensure that results are interpreted appropriately. Numerous corpus-based resources and tools exist for use in such studies, such as the British National Corpus (100 million words) and Sketch Engine (>37 billion words). Comprehensive lexical databases with word frequency lists, such as Wordnet, DANTE and SubtLex, are also publicly available online. Such resources are regularly used in studies ranging from speech perception and production to aphasia and machine learning e.g., (Hoffman, et al. 2011; Maas et al., 2011). The same is not possible for most sign language studies, as objective frequency data simply does not exist for the vast majority of sign languages, including Irish Sign Language (ISL). As a result, all but the most recent linguistic studies of sign languages have been carried out with small subjective datasets e.g., (Vinson et al., 2008). It is generally accepted that, without the aid of computational algorithms, researchers relied on expert knowledge of the language to select data and design experiments, resulting in somewhat subjective datasets (Fenlon et al., 2014; 2015).

Driven by a need to examine claims and language descriptions based on small datasets, the past 16 years have seen substantial progress in the labour intensive development of machine-readable sign language corpora. Fenlon et al. (2015) note that user intuition or detailed expert analysis of a small dataset cannot reveal patterns of a language to the same extent that is attainable by computer-assisted statistical analysis of large datasets. The mid to late 2000s saw a conscious effort to collect and pain stakingly annotate machine-readable sign language corpus data in multiple parallel projects across Europe and the United States of America (Fenlon et al., 2015; Hochgesang & Fenlon, in press). Indeed, it is only in the past decade that it has been possible for researchers to gain linguistic insight based on these resources. Sign language corpora have the capacity to change and evolve with new theories, therefore, as a result, a corpus is never 'complete'. Today's sign language corpora are at varying stages of development, yet all may be considered to be in their infancy.

To date, only four objective lexical frequency studies have been published across all sign languages that we are aware of. The composition of these studies is summarised in Table 1. The contrasting size of the datasets is noteworthy as well as the language registers used. For instance, the ASL and NZSL datasets consist of data from the formal, conversational and narrative registers, while the BSL and Auslan principally use one register. The size of each dataset is measured by the number of sign tokens. The smallest unit in a corpus is typically referred to as a 'token'. For the purpose of this work, we assume that a token is equal to the lexical gloss tag of a sign. Figure 1 illustrates how tokens are presented along a timeline in the ELAN interface. Note that the top tier, labelled, 'Lexical Gloss' tier, lists 11 tokens, of which, the token *SLIP-OUT* is highlighted blue to exemplify the scope of a lexical gloss token. This token type is also referred to as a *sign token* in the literature (Fenlon, et al. 2014; Johnston 2011).

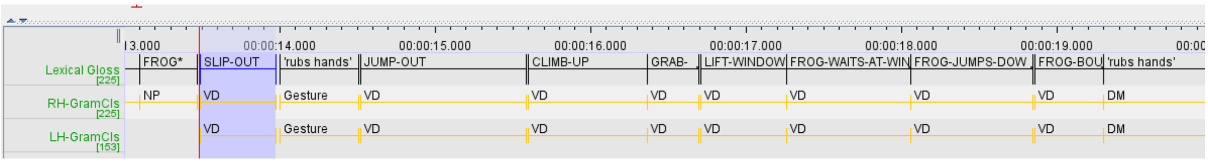

Figure 1: Lexical gloss tokens as they appear on the timeline of the SOI corpus – (source: C-32-M-27-PSN.eaf)

There are three points of note regarding the corpus data used in these studies:

1) The BSL dataset is based on 24,823 tokens collected from the BSL corpus, all of which originate from spontaneous conversational data. Although some of the other lexical frequency studies report on spontaneous/casual register data, only the BSL dataset is wholly comprised of spontaneous conversational data.

2) The Auslan corpus, like the SOI corpus, includes elicitation tasks which result in multiple recounts of the same narrative in the corpus. This has a direct effect on frequency count (Johnston 2011, p. 172).

3) The ASL dataset is reported to be an electronically formatted database, not machine-readable. This would reduce the opportunity for automated (computer-assisted) data extraction and analysis. Like the SOI corpus, the NZSL, BSL and Auslan corpora are stored and managed with the ELAN multimedia annotation application (ELAN, 2018).

Dataset	**Language**	**Sign token count**	**No. of participants**	**Register**	**Study**
Database	ASL	4,111	27	Varying	(Morford & MacFarlane, 2003)
Machine-readable corpus	NZSL	100,000	80	Varying	(McKee & Kennedy, 2006)
Machine-readable corpus	Auslan	63,436	109	Principally narrative	(Johnston, 2011)
Machine-readable corpus	BSL	24,823	249	Spontaneous conversational data	(Fenlon, et al., 2014)

Table 1: Composition of datasets used in previous sign language frequency studies

Given this context, our paper presents the fifth lexical frequency investigation of a sign language and is the first to consider lexical frequency data for ISL. The results of this exploratory work offer an insight into the frequency with which symbolic units[i] are used in ISL. In Section 2 we describe the methodological approach, providing some detail on the composition of the SOI corpus, the SOI corpus subset, and the symbolic units investigated in this study.

Considering signs at a number of symbolic levels, Section 0 first reports the lexical frequency distribution within the SOI corpus of function signs with respect to content signs. Section 3 then goes on to report the lexical frequency distribution of fully lexical (phonetically constrained and listed in the lexicon), partly lexical (phonetically unconstrained and listed in the lexicon) and non-lexical signs (not listed in the lexicon). Findings are discussed in the context of the four previous lexical frequency studies where comparable results have been reported (comparable statistics for NZSL are limited).

2. Methodology

For the purpose of producing data comparable to that of previous studies, we leverage three datasets. The first dataset consists of all 11,161 lexical gloss tokens of the SOI corpus. It is this dataset from which the lexical frequency list was generated. The second dataset, a subset of the first, consists of the 100 most frequent lexical glosses in the corpus (a total of 3,528 sign tokens ranked by frequency). The third subset consists of 2,971 lexical gloss tokens which are annotated for grammatical class (henceforth, the 'grammatical class subset'). The timeline illustrated in Figure 1 (above) presents a snapshot of this subset. As outlined earlier, Figure 1 depicts the lexical gloss tier along a timeline as presented in the ELAN interface. This tier appears in all SOI corpus files. Figure 1 also illustrates two grammatical gloss tiers, labelled 'RH-GramCls' and 'LH-GramCls', which contain tokens for the right-hand grammatical class and left-hand grammatical class respectively. It is these tiers that differentiate the grammatical class subset from the rest of the corpus. With the data organised in this time-aligned fashion, one may deduce that the *SLIP-OUT* lexical gloss token, as it appears in Figure 1, is a two-handed verbal depicting (VD) sign.

2.1. SOI Corpus

The SOI corpus was developed as part of the Languages of Ireland programme at the Centre for Deaf Studies, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, with the simple goal of capturing a snapshot of authentic ISL as it was used at the time, with the resulting corpus used to support research. The original SOI team produced a demographically representative sample of ISL as it was used in the summer of 2004. We offer a description of the corpus forthwith but the reader is directed to Leeson et al. (2006) and Leeson and Saeed (2012) for a detailed demographic breakdown of the SOI corpus content and a listing of contributors.

The SOI corpus was one of the first digital, machine-readable, sign language corpora developed in the world (Fenlon et al. 2015). At the time, it was the most richly annotated sign language corpus in Europe, and, today, it remains one of the most richly annotated corpora internationally.

Demographic	**SOI Content**	**Grammatical Class Subset**
Text type	6 x PSN[a] & 39 x SSN[b]	4 x PSN[a] & 7 x SSN[b]
Number of videos	45	11
Participants	41	9
Age	18 - 79	20 - 79
Geographic spread	5 Counties	4 Counties
Gender	16 Male / 25 Female	6 Male / 5 Female
Lexical gloss tokens	11,161	2,971
Total tokens (all tiers)	51,753	18,232
Total duration	91 minutes 78 seconds	26 minutes 43 seconds

[a] *Picture Story Narrative*

[b] *Self Selected Narrative*

Table 2: Demographic breakdown of the SOI corpus and the grammatical class subset

The SOI corpus project ran from 2004 to 2007, at which point all of the self-selected narratives (SSN) and many of the picture sequence narratives (PSN) had been annotated. Currently, approximately 50% of the elicited data has been richly annotated. This equates to 11,161 tokens on the lexical gloss tier or a total of 51,753 annotation tokens on all tiers (as of 23rd July 2019) as illustrated in Table 2. Since completion of the SOI project, more annotations have been added to the corpus as required by various research projects. Such projects include studies by Mohr-Militzer (2011), Leeson & Saeed (2012), Fitzgerald (2014), Napoli & Leeson (2020), Ferarra, et al., (2020) and in ongoing work by the authors.

Of the 11,161 glosses in the corpus, 4,045 are unique gloss tokens – that is, a unique instance of a lexical gloss. For example, the table in Appendix 1 shows that the unique lexical gloss *HOME* appears 57 times in the corpus. Of the unique tokens, 66% (2,344) are *hapax legomena* (occur once in the corpus). This equates to 22% of all lexical glosses in the corpus (See Table 2).

	Unique Gloss Tokens	Hapax Legomena	Percent of Corpus
Before Cleaning	4,045	2,702	24%
After Cleaning	3,761	2,344	22%

Table 3: The count of lexical gloss tokens before and after the cleansing process

The SOI corpus project was ahead of its time in it's efforts to annotate linguistic phenomena such as manual simultaneity, role shift (constructed action/discourse) and gesture, which were not to be investigated until a number of years later. As such, a high number of *hapax legomena* seems indicative of the pioneering work carried out during this project. In an effort to reduce the number of *hapax legomena*, variations of lexical glosses were combined in a data cleansing process. Variations include phonological variants, morphological variants and alternate versions of a sign which carry the same meaning.

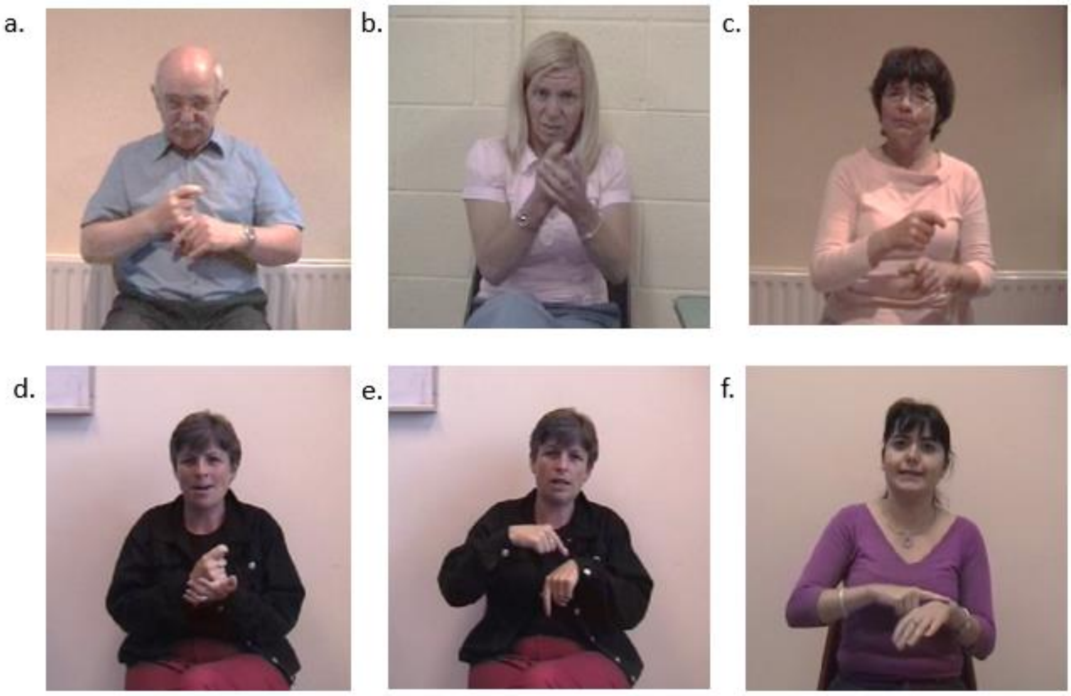

Figure 2: Illustration of variants in the lexical gloss TIME

Without the benefit of being informed by earlier approaches, it was no doubt challenging to determine strategies for tagging lexical glosses. According to Leeson et al. (2006), the annotation team aimed to standardise lexical gloss tags by reducing the citation form of a sign (the form accepted by the ISL community to be 'correct') and its variants to a single glossed form, which are today called ID-Glosses (Johnston 2008). The illustration of variants for the

English word TIME in Figure 2 exemplifies the difficulties of such an undertaking. For many fully-lexical signs, this is not a straightforward prospect, yet it becomes more challenging with partly lexical and non-lexical signs. In these, there is evidence that the team's efforts towards consistency and accuracy were less successful. Attempts to capture context-specific data in the lexical gloss resulted in more unique lexical gloss tags and a substantial number of *hapax legomena*. Figure 3 and 4 illustrate how annotators sought to capture context in their annotations. Figure 3 depicts the signer holding a tree branch in each hand. The lexical gloss for this was annotated as HOLD-BRANCHES. However, the annotator may have been less contextual and annotated HOLD-UPRIGHT-OBJECT-BILATERAL. Such an annotation strategy would be less reliant on context and would allow the annotation to be stored and used in all instances were that signer was holding an upright object in each fist e.g., depicting a signer holding ski poles. In Figure 4, a second example is annotated as OWL FLY but the fact that the signer is depicting an owl is only clear from the context of the utterance. This sign, alone, represents a flying entity. The sign could be annotated as ENTITY-FLY or similar. Such abstraction can broaden the scenarios for which the annotation can be used. Of course, glossing has conventions (Pizzuto & Pietrandrea 2001), but annotations are not stable and are often subjective. The SOI data suggests that this instability is most prevalent amongst depicting signs which are more likely to be partly-lexical.

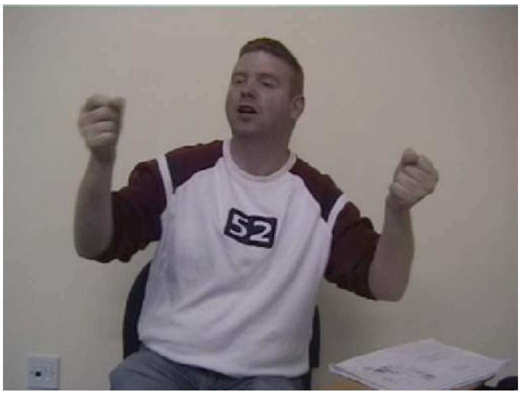

Figure 3: HOLD-BRANCHES (Source: C-32-M-27-PSN)

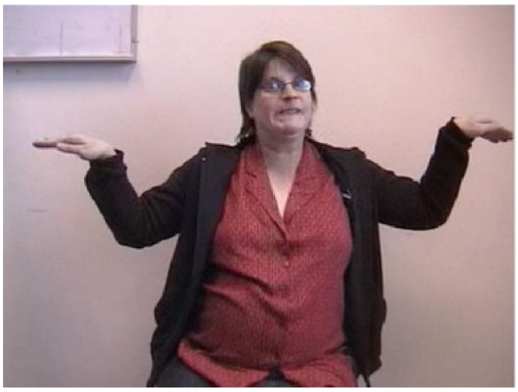

Figure 4: OWL FLY (Source: D-5-F-37-PSN)

Corpus projects such as the Auslan corpus and BSL corpus have adopted this strategy i.e., they endeavour to keep the gloss as non-contextual and as formalised as possible such that they may be analysed in a productive manner. Projects such as these have had the benefit of prolonged funding which have afforded the opportunity to develop annotation strategies and workflows over time. Indeed, the Auslan corpus annotation guidelines (Johnston & De Beuzeville, 2016) offer a most informed and in-depth guide to corpus development and annotation, yet this guide has evolved since the first version eleven years prior. The Auslan corpus, like the BSL corpus, draws on a lexical database for unique lexical gloss (ID-gloss) tagging, which allows for a consistent approach across all annotators. This approach is now prominent across sign language corpora.

2.2. SOI Annotation and the ISL Lexicon

The lexicon of ISL has never been documented formally as a lexical database. The only artefact resembling such a development was the process of lexical gloss annotation during the SOI corpus project. Prior to this, various short paper-based dictionaries, glossaries and teaching aids had been published (Foran S. J., 1979; NAD & SLTAI, 1992; Matthews, 2010). The most substantial digital record of ISL preceding the SOI corpus was an interactive application distributed on CD-ROM, titled "ISL Dictionary" (Micro Books, 1997). Boasting a glossary of 3,500 signs (3,700 in the revised edition), the ISL Dictionary CD-ROM was commercially developed and marketed as a tool for learners of the language, with the lexicon predicated on the listing that had been adopted for British Sign Language. A more recent glossary developed for Science, Technology, Engineering and Mathematics (STEM) terminology (Mathews & Mahon, 2018) was developed to aid students, teachers and parents. None of these resources can be considered a lexicon, rather a collection of glossaries developed as learning tools, not as a formal record of lexemes. None of the above document the grammatical class of a sign, define lexical items based on theories of grammar and lexicography or define symbolic units such as content or function signs, and none define what constitutes a sign lemma with associated morphological relationships (i.e. free and bound morphemes).

With respect to the SOI corpus, there has not yet been a consistent methodical approach to assigning tags for lexical gloss tokens in the corpus. Certainly, there are examples in the corpus of consistency or attempts at consistency across some lexical gloss tokens. However, it is widely accepted that to perform this task accurately, one must develop a dictionary for the language (Johnston, 2008). Such a large body of work is outside the scope of this study.

Without an established lexical database of lemmata, it is difficult to be consistent in tagging a large dataset. Although there are inconsistencies, there was a clear undertaking to identify a sign's canonical form and subsequently apply morphological modifiers such as movement direction, manner and degree during annotation of the SOI corpus. These efforts afford the possibility of carrying out a limited frequency analysis of lexical gloss tokens in the SOI corpus (e.g. see Leeson and Saeed, 2012). Indeed, that analysis and its limitations are the focus of this paper. A further comparison with data found in previous corpus studies of Auslan, ASL and NZSL serves to provide context for interpreting the findings of this analysis. As a method of gaining further insight, this study utilises a subset from the SOI corpus. In this manner, the work presented here uses a three-pronged approach to identify frequency statistics for ISL: 1) lexical gloss data from the SOI corpus, 2) the grammatical class subset, and 3) a comparison with results from previous studies of lexical frequency analysis in other sign languages.

2.3. Annotation of the Grammatical Class Subset

As part of a broader piece of research, a subset of the SOI corpus was tagged for grammatical class. The dominant and non-dominant hands were tagged separately with the controlled vocabulary tags listed in the Auslan annotation guidelines (Johnston & De Beuzeville 2016, p. 65). Additional tags were added by the author in order to categorise sign names, fingerspelling and gestures in the corpus. The grammatical class tagging of 2,971 lexical gloss tokens has been confirmed for accuracy by multiple rounds of visual inspection of each annotation by multiple researchers familiar with the linguistics of ISL. Later, 300 random samples were inspected by an independent subject matter expert who found the grammatical class tags to be inaccurate in only a single sample. In this study, we discuss only grammatical gloss tags that originate on the dominant hand[ii] as this data captures the grammatical class of the lexical gloss in the vast majority of instances.

Over the course of approximately one year, a random sample subset, demographically representative of the entire corpus (metadata was viewed but no content), was annotated for grammatical class. The subset is gender-balanced with 55% male and 45% female adult signers, across eight different age brackets from 20–79. Four of the five counties represented within in the SOI corpus are included, and geographic balance has also been retained. Finally, registers represented in the subset can be described as 36% PSN and 64% SSN. See Table 4 for further details, including duration and a count of lexical gloss tokens for each file.

File Name	Age	County	Gender	Task Code	Video Duration	Lexical Gloss Count
D-13-M-22-PSN	20-24	Dublin	M	PSN	00:00:35	184
C-32-M-27-PSN	25-29	Cork	M	PSN	00:02:03	225
WX-27-F-32-SSN	30-34	Wexford	F	SSN	00:02:11	245
D-5-F-37-PSN	35-39	Dublin	F	PSN	00:03:44	326
W-36-F-37-SSN	35-39	Waterford	F	SSN	00:03:38	445
W-36-F-37-PSN	35-39	Waterford	F	PSN	00:02:29	241
WX-22-M-42-SSN	40-44	Wexford	M	SSN	00:01:42	210
D-16-F-47-SSN	45-49	Dublin	F	SSN	00:02:48	376
D-17-M-62-SSN	60-64	Dublin	M	SSN	00:02:39	296
D-17-M-62-PSN	60-64	Dublin	M	PSN	00:04:01	377
W-39-M-77-SSN	75-79	Waterford	M	SSN	00:00:53	46
				Total:	00:26:43	2971

Table 4: Demographic details of the grammatical class subset

Including four accounts of the same PSN will certainly affect the lexical frequency count, nevertheless, the data was included for a number of reasons. Firstly, by including the PSN data, the subset remains representative of the SOI corpus, as the SOI corpus data also includes a number of PSNs. Secondly, the various portrayals of this PSN should not be considered a duplication of text akin to adding the same book numerous times in a written corpus. The style of signing used for each narrative differs across signers, with signers using different narrative strategies which would individuate the ratios of grammatical classes on a signer-by-signer basis. For example, Figure 5 illustrates 4 seconds from the timeline of a PSN (D-13-M-22-PSN). Here, the signer chooses to name the dog "*MO*" as a narrative strategy, despite the fact that no name for the dog was provided in the elicitation materials. This strategy does not occur in any of the other PSNs. Further, we note that the duration of the PSNs in Table 4 range from 35 seconds to just over 4 minutes, a further indication of diverse approaches to delivering a narrative. Certainly, we expect some lexical items such as *DOG*, *BOY* and *FROG* to have a higher frequency count but we may account for this when interpreting the results. This takes us to the third rationale, the tag value of a lexical gloss token, such as *DOG*, *BOY* or *FROG,* is not considered during frequency counts conducted with the grammatical class subset. Instead, only the data from the RH-GramCls or LH-GramCls tiers, such as VD, NP or Adj are used in such counts, depending on which is the dominant hand of the signer. In Figure 5, for example, the tag **MO* is simply counted as a *SNAME* (Sign Name) just as the tag *DOG* is counted as an

NP. This level of abstraction will allow us to render both the PSN and SSN to the same, albeit a more abstract, narrative type. The final and rather prosaic rationale for including the PSN in the dataset is simply that the grammatical class dataset was not originally developed for lexical frequency analysis. Only eleven files have been tagged for grammatical class. The time it would take to tag more files can be measured in months while removing the PSN files would significantly deplete the dataset.

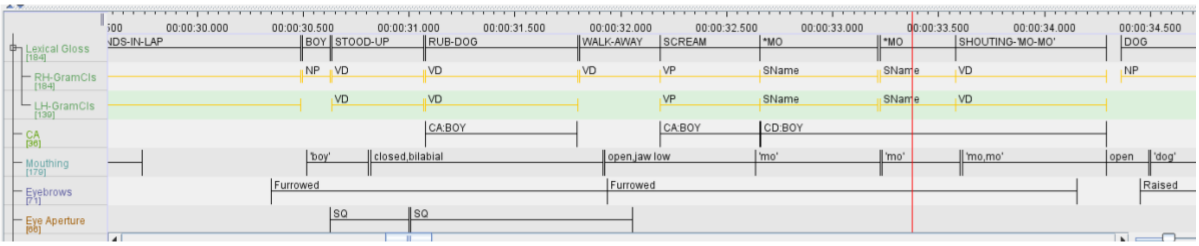

Figure 5: Timeline data from D-13-M-22-PSN – the narrative strategy introduces the name "MO" for the dog persona

2.4. Symbolic Units

Symbolic units are used here to categorise, in a way that is consistent with previous studies, the function of a sign or its degree of conventionality. For instance, function signs are considered to be grammatical units, i.e., signs that perform a syntactic function in a sentence (*do, but, in, or, if*), while content signs are considered to be conventionalised units from a language's lexicon that convey meaning as opposed to a grammatical function. Such examples from the SOI corpus include DEAF, DOG and JEEP-ZOOMS-UP-BEHIND.

Symbolic Unit	Grammatical Class
Function Sign	*Pro, Aux, Det, Conj, Prep, Wh-ProQ, WH-Rel, DM, Loc*
Content Sign	*NP, NLoc, ND, VP, VD, VIDir, VILoc, Adj, Adv, Num, Neg, Salutation*
Partially Lexical	*NLoc, ND, Pro, Loc, VD, VIDir, VILoc*
Fully Lexical	*NP, VP, Adj, Adv, Aux, Num, Det, Conj, Prep, WH-ProQ, WH-Rel, Neg, DM, Salutation*
Non-Lexical	*Interact, Gesture*
Finger Spelling	*x-fs (where 'x' is the part of speech)*
Sign Name	*SName*
Pointing Sign	*NLoc, Pro, Loc, VIDir, VILoc, Det*
Depicting Sign	*ND, VD*
Other	*Buoy, Fragment, Title, Unsure*

Table 5: Grammatical classes and the symbolic units that they constitute

Content and function signs may be fully lexical or partly lexical (Johnston & Schembri, 2010). Fully lexical signs like DOG and DEAF are considered to be 'frozen' or fully conventionalised in their form and meaning. These are phonologically constrained such that semantic meaning may only be deciphered if all phonemic components are present. Phonemic components of sign languages are well established in the literature and include handshape, movement, location, orientation and non-manual features. Signs may be considered partly lexical if one or more phonemic components relies on context to convey meaning. For example, indicating and depicting verbs are considered partly lexical because both types have some variability in relation to context (Johnston, 2011). Indicating verbs must specify agreement between subject and object by identifying the manner of movement between loci in the signing space (Johnston, 2011). Depicting signs, in contrast, use a specified handshape with a contextualised movement, orientation, direction and non-manual features to convey the nature of motion, size and shape, handling, and location (Cormier et al., 2012). All fully and partly lexical signs may be recorded in a language's core lexicon. Glosses which are excluded from the lexicon, such as gestures and incomplete sign fragments, are classified as non-lexical signs. Table 5 illustrates the grammatical class clusters that are used to identify symbolic units in the grammatical class subset (Johnston & De Beuzeville, 2016).

3. Analysis and Findings

3.1. Lexical Frequency Distribution

As discussed in Section 2.1, the data cleansing process saw the removal of sign variants. The same process saw the removal of duplicate tokens which are the result of inconsistent tagging. Examples of such duplicates may be seen in items 1, 5, 10 and 17 of Table 6, which illustrates the top twenty most frequent lexical glosses as they appear in the list before the cleansing process. The listed items appear multiple times in Table 6, yet they refer to the same sign. When such occurrences are removed from the frequency list, the top 100 most frequent ISL signs account for 32.6% of all signs in the corpus. This figure is lower than those presented in findings from the BSL (57%), and Auslan (53%) studies. As with spoken languages, however, sign languages differ typologically, and findings in one language do not necessarily apply to another. As with all languages, sign languages can be grouped by family. Contemporary ISL is descended from French Sign Language (LSF), as is ASL, while BSL, Auslan and NZSL are members of the BSL family (Leeson & Saeed, 2012; Hammarström, Forkel, & Haspelmath, 2019; Leonard & Conama, this volume).

Similar statistics were not reported in the NZSL and ASL studies. Undoubtedly the high frequency of *hapax legomena* in the SOI corpus frequency list is the cause of this inaccuracy. As such, it is not currently possible to generate a lexical frequency list from the SOI corpus. The data, nevertheless, may still offer some important insights which are discussed directly.

An annotation schema for the SOI corpus was never published. As a result, we must infer such a schema from the annotations as they are presented in the corpus. Consider, for instance, the gloss tokens listed in Table 6, these are the twenty most frequent lexical glosses that occur in the SOI corpus, before the cleansing process. Note that there are 8 forms of INDEX on the list. These are deictic (pointing) signs, referred to as indexical signs that typically use the index finger or other parts of the hand(s) for pointing. Pointing signs may be used as determiners, pronouns or locatives and may be one or two handed signs (Engberg-Pedersen, 2003; Padden, 1988; Zimmer, Patschke & Lucas, 1990; Emmorey, 2002). The annotation schema quite clearly allows for a movement type (i.e., pointing) and then a directional modifier. In this manner, INDEX+me denotes the signer is pointing to 'self'. Prototypically, this sign points to a locus near the centre of the signing space, positioned directly in front of the signer's chest area. INDEX+f indicates the signer is pointing to the *F* locus (directly to the front of the signing space). While INDEX+c indicates the signer is pointing to the centre locus of the signing space. The *+c* modifier and the *+me* modifier both point to the *C* locus. A search of the dataset shows that, with very few exceptions, INDEX+me and INDEX+c are not used in the same files, suggesting that the use of one tag over the other is a matter of annotation style and is not a choice made of morphological necessity. Adding to this irregularity, there are no less than twenty different variations of the token INDEX+me in the frequency list, seven of those occurring in the most frequent 100 signs (provided in Appendix 1). These include variations of uppercase/lowercase, use of the minus symbol in place of the plus symbol, and instances where an asterisk character is attached to the gloss. The various other variations of INDEX+me that occur outside of the most frequent 100 signs are similarly the result of inconsistent annotating. Section 3.2 offers further discussion on pointing signs.

Rank	**Lexical Gloss Token**	**Frequency** (*n* = 11,161)	**Per 1,000**
1	INDEX+me	261	23.4
2	BUT	99	8.9
3	INDEX+c	96	8.6

4	SEE	85	7.6
5	INDEX-Me	82	7.4
6	BOY	78	7.0
7	INDEX+f	78	7.0
8	INDEX+fl	73	6.5
9	*DOG	72	6.5
10	INDEX+ME*	72	6.5
11	HAVE	69	6.2
12	INDEX+fr	65	5.8
13	MY	56	5.0
14	TO	56	5.0
15	HOME	55	4.9
16	AND	54	4.8
17	INDEX-Me*	54	4.8
18	FROG	51	4.6
19	ONE	51	4.6
20	'hands up'	47	4.2

Table 6: An ineffectual list of the 20 most frequent signs in the SOI corpus - before cleansing

Rank	**Lexical Gloss Token**	**Frequency** (*n* = 11,161)	**Per 1,000**
1	INDEX (1st person, inc variants)	633	58.5
2	INDEX (2nd & 3rd person, inc variants)	620	57.3
3	BUT	106	9.8
4	BOY (inc variants)	101	9.3
5	SEE (inc variants)	93	8.6
6	HAVE (inc variants)	80	7.4
7	*DOG (fingerspelled)	78	7.2
8	TO	62	5.7
9	AND	58	5.4
10	MY	57	5.3
11	HOME	57	5.3
12	LEAVE/leave object (inc variants)	55	5.1

13	hands up' (inc variants)	53	4.9
14	FROG	51	4.7
15	ONE	51	4.7
16	ABOUT (inc variants)	51	4.7
17	THINK (inc variants)	48	4.4
18	FOR (inc variants)	47	4.3
19	MAN (inc variants)	46	4.3
20	IN	45	4.2

(inc variants) denotes the inclusion of multiple sign variants in the count

Table 7: The 20 most frequent signs in the SOI corpus - after cleansing

Appendix 1 lists the top 110 most frequently used lexical glosses in the SOI corpus after the data has been cleansed of variants and duplicates. For quick reference, ***Error! Reference source not found.*** lists the top twenty most frequent glosses. Clearly, variants have been resolved in this list, such variants include as DOG* and *S*EE++ *which* occur in the pre-processed frequency list exported by ELAN. Alternate variants were deleted and their frequency count added to the dominant variant. This brought the frequency count of BOY from 78 to 101 and SEE++ from 85 to 93.

3.2. Indexical (Pointing) Signs

The deictic word and an indexical sign differ in form as opposed to function (Coppola & Senghas, 2010). In spoken languages, deixis are phonologically constrained. This is not the case for indexical signs. Indexical signs point to a locus in the signing space as defined by context, which means the phonemes for direction, orientation and often movement, must be unconstrained (Fenlon et al., 2019). There are instances in which indexical signs use the whole hand, the fist or fingers other than the index to point. Many lexical gloss tokens in the SOI corpus do not specify these and therefore all indexical handshapes are included in the index counts.

Indexical signs are represented by no less than 271 different lexical gloss tokens in the SOI corpus. We have already identified some of the frequently used index glosses. Most of the remaining tokens reflect the use of additional modifiers such as +*fl* ‘front and left’, and +*hi*+*fr* ‘high and front right’ for movement direction or ++ for repetition of movement. However, some annotators seem to have tagged for person by using INDEX-1, INDEX-2 and INDEX-3 to identify the first, second or third person. Other annotations stand out in their idiosyncratic nature, such as INDEX+me_WANT++ and CL-INDEX 'HOSE' +c+f wriggle. Inconsistencies in categorising glosses and/or the high level of detail contained in glosses are problematic for analysis.

	ISL	**Auslan**	**ASL**	**BSL**	**NZSL**
	n=11,161	*n=63,436*	*n=4,111*	*n=24,823*	*n=100,000*
Hapax Legomena	216.7	70[a]	-	-	-
INDEX+me (1st Person)	58.5	50.8	56.4	69.8	67.2
INDEX (2nd & 3rd Person)	57.3	32.5	79.3	55.5	47.9
All INDEX	115.8	123	138	-	-
Gesture [tokens containing the word 'Gesture']	18.7	35.7	2	55.3	-

[a] *Based on a subset of 55,859 tokens as reported in* (Johnston, 2011)

Results are displayed per 1,000 words

Table 8: Corpus distribution comparison of sign type in ISL, Auslan, ASL, BSL and NZSL

There are obvious limitations to what may be discovered from the dataset, both before and after the cleansing process. However, it is possible to identify all of the glosses that represent the first person pronoun, by identifying all variations of *INDEX+me*, *INDEX+c* and *INDEX-1* for example. When all of these variations are considered they account for 5.9% of the corpus. This increases to 17.9% when all pronouns, locatives and determiners are considered. These figures are generally in line with previous sign language lexical frequency studies. Every effort has been made to categorise indexical signs as they occur in the frequency list, nevertheless, it must be assumed that some occurrences may have been omitted. As such, it is useful to compare these findings with other studies on this topic. As there are no such studies for ISL, Table 8 presents data for ISL alongside comparative data from studies carried out in Auslan, ASL, BSL and NZSL.

The data suggests that ISL, like ASL, BSL, Auslan and NZSL, is a lexically dense language. This is contrary to spoken English which typically has a low lexical density (Ure, 1971). The primary indicator being that the list of the 100 most frequent signs contains a higher ratio of lexical signs to functional signs. This is the conclusion from lexicon frequency studies of all five sign languages which, as stated, are from two distinct language families. Each of these studies qualify their findings by drawing attention to a need for increased sample size and a broadening of the registers considered; the results from this study should be qualified in the same manner. While, across the five datasets, there are a total of 195,954 gloss tokens, even this total is considered a small dataset in the context of spoken language corpora. Nonetheless, this data is substantial given that it accounts for reports on five different languages via five independent studies, with consistent findings. In addition to sample size, data composition is a consideration. All of the aforementioned corpora (with the exception of BSL) used similar elicitation tasks, specifically, PSN and SSN, with resulting corpus data skewed towards storytelling. Johnston (2011) suggests that a more complete picture of the core lexicon can be gained with the inclusion of spontaneous conversational data while Fenlon et al. (2014) report that the entirety of the BSL corpus consists of spontaneous conversation data, with a higher frequency of indexical signs identified reflecting text type.

An interesting point of note is that all of the five sign language studies considered here share a common spoken language neighbour, English. Ure (1971) found that the English language, and particularly spoken English, has a low lexical density. It is interesting then, that the sign language studies, all of which are in contact with English in some way, have found a high level of lexical density in their respective sign language. This suggests that high lexical density can be attributed to the modality of sign languages. This finding supports the notion that many grammatical functions of a sign language occur in the performance of simultaneous non-manual features and spatial modifiers (Leeson & Saeed, 2007). Indeed, function signs account for only 16% of the most frequent 100 lexical gloss tokens in the SOI corpus investigated here. Further analysis with the grammatical class subset found that 25% of glosses were function signs. These figures are comparable to the BSL and Auslan studies which found that 22% and 25% (respectively) of their 100 most frequent signs are function signs.

3.3. Fully Lexical Signs

Of the 100 most frequent signs in the SOI corpus, 87% are fully lexical. This figure is comparable to the ASL study which reports a figure of 73% based on all 4,111 tokens. The Auslan and BSL studies report 65% and 65.4% respectively. What is of significance is the data on which these figures are based. While the ISL findings are based on the most frequent 100 signs, the ASL, BSL and Auslan findings are based on the whole dataset. Further analysis of fully lexical signs with the 2,971 tokens from the ISL grammatical class subset found a significantly smaller percentage of signs (46%) to be fully lexical (Table 3 provides a listing of all grammatical classes that are fully lexical). The diverging figures as presented for the ISL datasets serve to highlight the dangers involved in interpreting and indeed comparing these findings. Although the overwhelming majority of the 100 most frequent ISL signs were found to be fully lexical, this was not supported by the grammatical class subset. The cause of this is a matter of variation in glossing practices across studies (Fenlon et al. 2014, p. 31) and data presentation. The 100 most frequent signs is a listing of 4,375 tokens categorised by type (lexical gloss token) and ordered by frequency of occurrence while the grammatical class subset is a simple random sample, consisting of 2,971 sign tokens selected from the SOI corpus. One would expect the frequency list to offer a better insight into the most commonly occurring signs in the corpus, however, as previously noted, the lexical frequency list generated by the SOI corpus may not be accurate. As a result, the finding of 87% fully lexical signs, discussed at the top of this section, may be considered inaccurate. The grammatical class subset data then offers some further insight and a more accurate result. The lower figure of 46%, for fully lexical signs, is in line with what might be expected from the narrative text type. Indeed, the relatively high number of partly lexical signs (45.5%) and depicting signs 23.4% is a further indication of the text type.

3.4. Non-Lexical Signs

A non-lexical sign typically refers to a gesture, although in the field of gesture studies and cognitive linguistics, there is a move away from categorising signs as gesture and non-gesture (Goldin-Meadow & Brentari, 2017). None of the four frequency studies identify gesture signs in the same way. The Auslan and BSL studies attempt to categorise gestures as manual, non-manual or constructed action (mimetic), while it is unclear how the ASL study classified gestures. At the time these datasets were developed, the literature had yet to sufficiently describe what constitutes a gesture in a sign language. Even so, the SOI corpus data shows evidence of gestural tagging on the lexical gloss tier. The SOI corpus annotation strategy for gestures was simply to insert the word gesture into the lexical gloss e.g., *gesture* 2X CL-5 PALMS UP, gesture 'leave it' and GESTURE 'NUDGE PERSON NEXT TO ME'. Such glosses account for 1.9% of the SOI corpus. There are, however, additional instances of signs which may be considered gestural in the SOI corpus, for example, the lexical gloss *'hands up'* occurs no less than 53 times in the corpus and is the 13th most frequent gloss in the dataset. When we include entries such as *'hands up'*, which have been annotated without context or function, the gloss count for gestures rises to 3.0%. Indeed, this figure would rise further still if contextual gestures which exclude the keyword 'gesture' were added, such as GET-ATTENTION-OF+sl. This variation in classifying gestural signs renders the dataset unusable in its current form. For a thorough count of gesture in the SOI corpus, the data would need to be marked up using the theoretical lens that applies for gestural analysis. Fenlon et al. (2015, p.31) suggest that variation between the corpora is to be expected and such variation will make a direct comparison of results problematic. In the case of the SOI corpus however, the variation in classification within the corpus must be resolved to account for current understanding of gesture before any comparison with results from other corpora.

In addition to grammatical class annotation, one of the authors (RS) has annotated gestures explicitly within the grammatical class subset. In identifying non-lexical signs for comparison, a count of the gesture tokens from the dominant hand tier is added to a count of tokens with the tag 'Interact' from the same tier, which provides a count of 4.2%. The 'Interact' tag identifies exclamatives and interjections. This count is multiples higher than the ASL value which is 0.2% but falls between the BSL and Auslan figure of 3.4% and 6.5% respectively (these figures are based on the most frequent 100 signs only).

3.5. Depicting Signs

Typically, depicting signs refer to verbs but depicting nouns may also depict the size shape and placement of an object (De Beuzeville, Johnston & Schembri, 2009). Examples from the SOI corpus include SMALL-HILL, BOULDER and SHAPE-OF-JAR. Depicting verbs are also referred to as classifier signs or classifier predicates. Examples of depicting verbs include BOY-GOES-ALONG, DRIVE-MOTORBIKE and LOOK-AROUND. Depicting signs are not specified within their lexical gloss in the corpus as the term did not exist when the SOI corpus was annotated. However, it is possible to identify some depicting signs in the corpus by their classifier handshape prefix on the lexical gloss tier. Examples include CL-C where the C handshape is mimetic of holding an object and 2X CL-C to CL-T ' PLACE-LADDER-AGAINST' where the handshape classifier changes from C to T while depicting the act of placing a ladder against a wall. It is clear from these two examples that tagging of handshapes is not consistent throughout the dataset. Many depicting annotations include the CL prefix but many do not. Many of the annotations include contextual detail and many do not. As such it is not feasible to accurately identify all instances of depicting signs in the frequency list.

Depicting signs can be quite easily identified in the grammatical class subset, however, as they are tagged VD (verb depicting) or ND (noun depicting). Depicting signs represent 23.4% of the subset. This figure is significantly higher than the 4.2% reported in the ASL corpus, signifying perhaps, that the ASL study defines depicting signs in a different manner. Still, a more convincing reason for this is presented by Morford and MacFarlane (2003); the ASL study identifies a subset of the narrative register, of which 17.7% are tagged as classifiers i.e. when the text type is narrative, the count is higher. Fenlon et al. (2014) suggest a higher count of depicting signs is indicative of a narrative text type. The BSL and Auslan studies report figures for depicting signs in the top 100 most frequent signs (2.3% and 11% respectively).

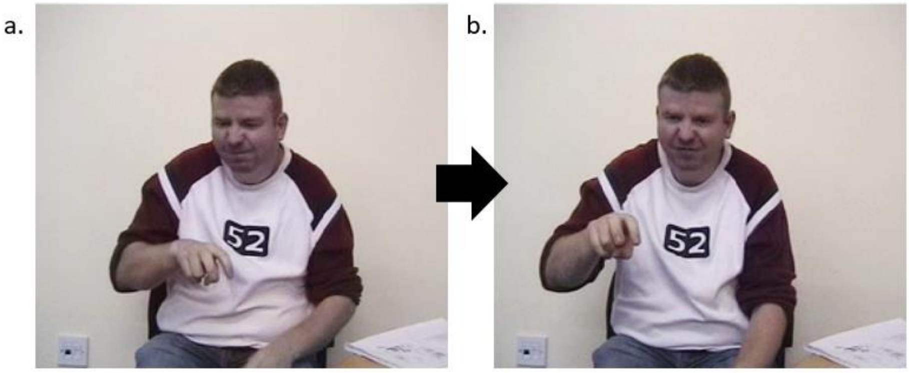

Figure 6: ‘walking’ with the ‘V’ handshape. Glossed as BOY-GOES-ALONG in the SOI corpus. a.) begin position, b.) end position (Source: C-32-M-27-PSN)

Unsurprisingly, due to both the highly contextual nature of the data and the variation in the annotation strategies applied, a large quantity of depicting signs in the SOI corpus are *hapax legomena*. It is impossible for any unique lexical gloss to appear in the top 100 most frequent ISL signs. The reason for this is best explained with an example. Consider a depicting verb which depicts ‘walking’. One might expect a signer to use a ‘D’ handshape and move it through the signing space from an initial start point to a destination point. A keyword search of the SOI frequency list for the word “walk” will find 35 hapax legomena which depict the act of walking plus various other iterations which occur more than once. Examples of such entries include CL+V+BENT+WALK-OVER, WALK-ALONG-CASUALLY, WALK-TO-WINDOW and WALK-WITH-SHOVEL. Of course, a search for the word “walk” will not find BOY-GOES-ALONG, the example illustrated in Figure 6. There are many such examples of depicting a walking action through the V classifier handshape and it would be extremely challenging to capture all such signs with confidence. In addition, these lexical gloss tokens are hapax legomena due to the fact that the context has been captured within the annotation. It is not a trivial task to separate the context from the action in such examples and as such, a single verb depicting a walking action will not appear in the top 100 most frequent signs.

4. Next Steps and Some Conclusions

This paper has presented ISL lexical frequency statistics based on two datasets, the SOI corpus as a whole, and a subset of same which was tagged for grammatical class, in what has been the first objective lexical frequency analysis for ISL. It became quickly apparent that, although the SOI corpus lexical gloss data can offer some useful insights, it is not, as currently stands, a reliable data source for frequency analysis. As a result, this work leveraged the grammatical class dataset for much of the analysis.

We acknowledge that results at this stage are indicative due to limitations with the size and composition of the subset. This study found that deictic pointing signs appeared at a lower frequency to that of the BSL corpus analysis. In addition, depicting signs appeared at a higher frequency. This appears to support findings from Fenlon et al. (2014), i.e., that a high frequency of pointing signs is indicative of a conversational text type while a higher frequency of classifier signs is indicative of a narrative text.

The data was presented in the context of the four previous lexical frequency studies carried out on sign language data. Accounting for limitations which arose with regard to annotation inconsistencies between corpora, we found that similar patterns of grammatical class usage occur across all five datasets. Each language was found to be dense in that the frequency of function signs was low in comparison to content signs. Each study reported similar ratios with regards to the frequency of fully lexical, partly lexical and non-lexical. Fully lexical signs are frequent while non-lexical signs occur with the lowest frequency. The relatively high frequency of partly lexical signs found in the ISL grammatical class dataset was attributed to the narrative text type.

Comparative data from the ASL study was found to diverge significantly from the ISL findings. This was found to be a result of differences in data composition and annotation strategies. Several strongly aligned results from the BSL and Auslan (and NZSL dataset, where available) studies suggest that more comparative datasets may yield a higher degree of correlation between ISL and ASL. Indeed, differences in annotation practice and text type was observed across all of the studies presented here. For a more comprehensive analysis, a consistent annotation strategy must be implemented across the lexical gloss tier in the SOI corpus. Examples of annotation inconsistency have been demonstrated in the case of fully lexical signs but were found to be most acute in the case of partly lexical and non-lexical signs as they are inherently more subjective. The high volume of *hapax legomena*, which result from such inconsistencies, serve to distort the frequency list. This issue was ultimately sidestepped through analysis of the grammatical class subset which afforded an analysis of the data from the grammatical class tiers in the place of the data from the lexical gloss tier. It was the grammatical class dataset that was used in the analysis of the various symbolic units discussed in this paper, including the derivation of frequency data for fully lexical, partly lexical and non-lexical signs as well as functional signs, content, deictic pointing signs and depicting signs. Future work will provide further frequency counts on all parts of speech tagged in the dataset.

This study concludes that the lexical gloss tokens in the SOI corpus as they currently stand, are unsuitable for a lexical frequency analysis due, primarily, to the apparent absence of a lexical glossing schema for partly lexical and non-lexical items. Developing such a standard would be a substantial undertaking and would involve some degree of language standardisation which should be developed in partnership wtih members of the ISL community. Such an undertaking is outside the scope of this study. Future research should explore the potential of an ISL lexical database similar to the SignBank databases deployed for Auslan and BSL. Such a database has the potential to improve lexical gloss consistency and could be retrospectively deployed across the SOI corpus.

Despite issues of annotation consistency, the SOI corpus remains a significant resource for teaching and researching various aspects of ISL; its multiple tiers of annotation have resulted in a rich dataset. As with all sign language corpora however, it is minuscule in relation to spoken language corpora. As such, to form a dataset comparable to those spoken language corpora and one amenable to statistical analysis, it would be fruitful to further expand the SOI corpus in terms of the volume of lexical gloss tokens and the diversity of text types.

It is anticipated that the frequency data presented in this paper, although indicative, will be a useful resource for future work in areas such as applied linguistics, descriptive linguistics and psycholinguistics.

Acknowledgements

This research is supported by funding from TU Dublin. We are grateful to the work of the SOI corpus project team and the Deaf Studies students at the Centre for Deaf Studies, Trinity

College Dublin who aided in the grammatical class tagging. We also acknowledge the work carried out in the previous frequency studies and are grateful to have their data to leverage in this study. We would also like to acknowledge permission to use the SOI corpus data, and particularly, our thanks go to the SOI participants.

References

Auslan-Signbank (2019, July 30). *Dictionary*. Retrieved from Auslan Signbank: http://www.auslan.org.au/dictionary/

Brysbaert, M., & New, B. (2009) Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990.

Conrad, S. (2005) Corpus linguistics and L2 teaching. In E. Hinkel (Ed.) *Handbook of Rresearch in Second Language Teaching and Learning.* London: Routledge, pp. 393–410.

Coppola, M., & Senghas, A. (2010) Deixis in an emerging sign language. In D. Brentari (Ed.), *Sign Languages*. Cambridge: Cambridge University Press, pp. 543–569.

Cormier, K., Fenlon, J., Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R. & Woll, B. (2012) From corpus to lexical database to online dictionary: Issues in annotation of the BSL Corpus and the development of BSL SignBank. In O. Crasborn, E. Efthimou, S.E. Fotinea, T. Hanke, J. Kristoffersen & J. Mesch (Eds.) *5th Workshop on the Representation of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012.* Istanbul: ELRA, pp. 7–12.

Cormier, K., Quinto-Pozos, D., Sevcikova, Z. & Schembri, A. (2012) Lexicalisation and de-lexicalisation processes in sign languages: Comparing depicting constructions and viewpoint gestures. *Language & Communication*, 32(4), 329–348.

De Beuzeville, L., Johnston, T. & Schembri, A. C. (2009) The use of space with indicating verbs in Auslan: A corpus-based investigation. *Sign Language & Linguistics,* 12(1), 53–82.

ELAN. (2018, April 04). (Version 5.2) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/.

Ellis, N. C. (2002) Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2), 143–188.

Emmorey, K. (2002) *Language, Cognition, and the Brain: Insights from Sign Language Research*. Mahwah, NJ: Lawrence Erlbaum.

Engberg-Pedersen, E. (2003) From pointing to reference and predication: pointing signs, eyegaze, and head and body orientation in Danish Sign Language. In S. Kita (Ed.), *Pointing: Where Language, Culture, and Cognition Meet*. Mahwah, NJ: Erlbaum, pp. 269–292.

Fenlon, J., Cooperrider, K., Keane, J., Brentari, D. & Goldin-Meadow, S. (2019) Comparing sign language and gesture: Insights from pointing. *Glossa: a journal of general linguistics*, 4(1), 2 DOI: http://doi.org/105334/gjgl.499

Fenlon, J., Schembri, A., Johnston, T. & Cormier, K. (2015) Documentary and corpus approaches to sign language research. In E. Orfanidou, B. Woll, & G. Morgan (Eds.), *Research Methods in Sign Language Studies: A Practical Guide*. Oxford: Wiley and Sons, pp. 156–172.

Fenlon, J., Schembri, A., Rentelis, R., Vinson, D. & Cormier, K. (2014) Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua* (143), 187–202.

Ferrara, L., Anible, B., Hodge, G., Jantunen, T., Leeson, L., Mesch, J. & Nilsson, A.-L. (2020) A cross-linguistic comparison of reference across different signed languages. *SignCafé 2*. Ragusa, Italy.

Fitzgerald, A. (2014) *Mouthing and Mouth Gestures in Irish Sign Language*. Unpublished PhD dissertation. Trinity College Dublin.

Foran, S. J. (1979) *Irish Sign Language*. Dublin: National Association for Deaf People.

Foran, S. J. (1996) *Irish Sign Language: Revised Edition*. Dublin: National Assiciation for Deaf People.

Goldin-Meadow, S. & Brentari, D. (2017) Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and Brain Sciences*, 40(46). DOI:10.1017/S0140525X15001247

Hammarström, H., Forkel, R. & Haspelmath, M. (2019, 07 24) *Glottolog 4.0*. Retrieved from https://glottolog.org/

Hochgesang, J., & Fenlon, J. (in press) *Sign Language Corpora*. Washington DC: Gallaudet University Press.

Hoffman, P., Rogers, T. T. & Lambon-Ralph, M. A. (2011) Semantic diversity accounts for the "missing" word frequency effect in stroke aphasia: Insights using a novel method to quantify contextual variability in meaning. *Journal of Cognitive Neuroscience*, 23(9), 2432–2446.

Johnston, T. (2011). Lexical frequency in sign languages. *Journal of Deaf Studies and Deaf Education*, 17(2), 163–193.

Johnston, T. & De Beuzeville, L. (2016) *Auslan Corpus Annotation Guidelines*. Sydney: Centre for Language Sciences, Department of Linguistics, Macquarie University.

Johnston, T. & Schembri, A. (2010) Variation, lexicalization and grammaticalization in signed languages. *Langage et société* 1, 19–35.

Leeson, L. & Saeed, J. (2007) Conceptual Blending and the Windowing of Attention in Irish Sign Language. In M. Vermeerbergen, L. Leeson, & O. Crasborn (Eds.), *Simultaneity in Signed Languages – Form and function*. John Benjamins Publishing, pp. 55–73.

Leeson, L., Saeed, J., Macduff, A., Byrne-Dunne, D. & Leonard, C. (2006) *Moving Heads and Moving Hands: Developing a Digital Corpus of Irish Sign Language*. Information Technology and Telecommunications Conference. Carlow.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011) Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y. & R. Mihalcea (Eds.) proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. Association for Computational Linguistics, pp. 142–150.

Mathews, E., & Mahon, V. (2018) *Irish Sign Language Mathematics Glossary Project*. Irish Deaf Research 2018. Dublin: Unpublished.

Matthews, P. A. (2010) *Irish Sign Language Communication*. Dublin: Deaf Communications Ltd.

McKee, D. & Kennedy, G. (2000) Lexical comparison of signs from American, Australian, British and New Zealand sign languages. In K. Emmorey, & H. Lane (Eds.) *The Signs of Language Revisited: An anthology to honor Ursula Bellugi and Edward Klima*. Mahwah, NJ: Erlbaum, pp. 49–76.

McKee, D. & Kennedy, G. (2006) The distribution of signs in New Zealand Sign Language. *Sign Language Studies*, 6(4), 372–390.

Micro Books. (1997) *ISL Dictionary* CD-ROM. Ireland: Microbooks Limited.

Mohr-Militzer, S. (2011) *Mouth Actions in Irish Sign Language - Their System and Functions*. PhD dissertation. Universität zu Köln.

Morford, J. P. & MacFarlane, J. (2003). Frequency Characteristics of American Sign Language. *Sign Language Studies*, 3(2), 213–225.

NAD & SLTAI (1992) *Sign On*. Dublin: National Association for the Deaf and Sign Language Tutors Association of Ireland.

Napoli, D., & Leeson, L. (2020) Visuo-spatial construals that aid in understanding activity in visual-centered narrative. *Language, Cognition and Neuroscience*. DOI: https://doi.org/10.1080/23273798.2020.1744672.

Padden, C. (1988) *Interaction of Morphology and Syntax in American Sign Language*. New York: Garland Publishing, Inc.

Pizzuto, E. A., & Pietrandrea, P. (2001) The notation of signed texts: open questions and indications for further research. *Sign Language & Linguistics*, 4(1-2), 29–45.

Stokoe, W. (1960) *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*. Studies in linguistics: Occasional papers. Buffalo, NY: University of Buffalo.

Ure, J. (1971) Lexical Density and Register Differentiation. *Applications of Linguistics*, 443–452.

Vinson, D. P., Cormier, K., Denmark, T., Schembri, A. & Vigliocco, G. (2008) The British Sign Language (BSL) Norms for Age of Acquisition, Familiarity, and Iconicity. *Behavior Research Methods*, 40(4), 1079–1087.

Zimmer, J., Patschke, C. & Lucas, C. (1990). A Class of Determiners in ASL. In C. Valli, & C. Lucas (Eds.), Linguistics of American Sign Language: An Introduction. Gallaudet University Press, pp. 201–210.

Appendix 1

Rank	**Lexical Gloss**	**Count**	**Percent of Unique Glosses**	**Percent of All Glosses**	**per 1000 Signs**
1	INDEX (1st person, inc variants)	633	17.94%	5.85%	58.5
2	INDEX (2nd & 3rd person, inc variants)	620	17.57%	5.73%	57.3
3	BUT	106	3.00%	0.98%	9.8
4	BOY (inc variants)	101	2.86%	0.93%	9.3
5	SEE (inc variants)	93	2.64%	0.86%	8.6
6	HAVE (inc variants)	80	2.27%	0.74%	7.4
7	*DOG (fingerspelled)	78	2.21%	0.72%	7.2
8	TO	62	1.76%	0.57%	5.7
9	AND	58	1.64%	0.54%	5.4
10	MY	57	1.62%	0.53%	5.3

11	HOME	57	1.62%	0.53%	5.3
12	LEAVE/leave object (inc variants)	55	1.56%	0.51%	5.1
13	hands up' (inc variants)	53	1.50%	0.49%	4.9
14	FROG	51	1.45%	0.47%	4.7
15	ONE	51	1.45%	0.47%	4.7
16	ABOUT (inc variants)	51	1.45%	0.47%	4.7
17	THINK (inc variants)	48	1.36%	0.44%	4.4
18	FOR (inc variants)	47	1.33%	0.43%	4.3
19	MAN (inc variants)	46	1.30%	0.43%	4.3
20	IN	45	1.28%	0.42%	4.2
21	WITH (inc variants)	45	1.28%	0.42%	4.2
22	KNOW (inc variants)	43	1.22%	0.40%	4.0
23	TIME (inc variants)	43	1.22%	0.40%	4.0
24	WANT (inc variants)	42	1.19%	0.39%	3.9
25	MOTHER (inc variants)	41	1.16%	0.38%	3.8
26	FATHER (inc variants)	41	1.16%	0.38%	3.8
27	WHERE (inc variants)	38	1.08%	0.35%	3.5
28	NOT (inc variants)	37	1.05%	0.34%	3.4
29	gesture	36	1.02%	0.33%	3.3
30	DRINK	35	0.99%	0.32%	3.2
31	DOG	38	1.08%	0.35%	3.5
32	TWO (inc variants)	34	0.96%	0.31%	3.1
33	FRIEND	33	0.94%	0.31%	3.1
34	DEAF (inc variants)	33	0.94%	0.31%	3.1
35	BROTHER (inc variants)	32	0.91%	0.30%	3.0
36	NEXT (inc variants)	32	0.91%	0.30%	3.0
37	TELL (inc variants)	32	0.91%	0.30%	3.0
38	pause (action, not sign, inc variants)	32	0.91%	0.30%	3.0
39	NO (inc variants)	32	0.91%	0.30%	3.0
40	GOOD	31	0.88%	0.29%	2.9
41	FINISH (inc variants)	31	0.88%	0.29%	2.9
42	LIKE	30	0.85%	0.28%	2.8
43	NOW	30	0.85%	0.28%	2.8
44	SAY (inc variants)	30	0.85%	0.28%	2.8
45	SAME (inc variants)	29	0.82%	0.27%	2.7
46	BEFORE (inc variants)	29	0.82%	0.27%	2.7
47	WORK (inc variants)	29	0.82%	0.27%	2.7
48	SMALL	28	0.79%	0.26%	2.6
49	NOTHING (inc variants)	28	0.79%	0.26%	2.6
50	TALK (inc variants)	27	0.77%	0.25%	2.5
51	WHAT (inc variants)	26	0.74%	0.24%	2.4

52	DRIVE (inc variants)	26	0.74%	0.24%	2.4
53	YEAR/YEARS (inc variants)	26	0.74%	0.24%	2.4
54	LOOK	25	0.71%	0.23%	2.3
55	THREE	25	0.71%	0.23%	2.3
56	SISTER (inc variants)	25	0.71%	0.23%	2.3
57	SELF/MYSELF	24	0.68%	0.22%	2.2
58	AGAIN (inc variants)	24	0.68%	0.22%	2.2
59	DAY (inc variants)	23	0.65%	0.21%	2.1
60	ALL (inc variants)	23	0.65%	0.21%	2.1
61	DRIVE-MOTORBIKE (inc variants)	23	0.65%	0.21%	2.1
62	FEEL (inc variants)	22	0.62%	0.20%	2.0
63	PLAY (inc variants)	22	0.62%	0.20%	2.0
64	FUNNY	21	0.60%	0.19%	1.9
65	MOTORBIKE	21	0.60%	0.19%	1.9
66	TREE	21	0.60%	0.19%	1.9
67	HAPPY (inc variants)	21	0.60%	0.19%	1.9
68	FROM (inc variants)	21	0.60%	0.19%	1.9
69	JOB (inc variants)	20	0.57%	0.18%	1.8
70	LOT	20	0.57%	0.18%	1.8
71	WATER (inc variants)	20	0.57%	0.18%	1.8
72	BACK (inc variants)	20	0.57%	0.18%	1.8
73	BECAUSE (inc variants)	19	0.54%	0.18%	1.8
74	LIVE	19	0.54%	0.18%	1.8
75	FIND (inc variants)	19	0.54%	0.18%	1.8
76	MAKE (inc variants)	19	0.54%	0.18%	1.8
77	OOPS (inc variants)	19	0.54%	0.18%	1.8
78	JASON	18	0.51%	0.17%	1.7
79	JUST (inc variants)	18	0.51%	0.17%	1.7
80	BEE (inc variants)	18	0.51%	0.17%	1.7
81	LOOK-AT	17	0.48%	0.16%	1.6
82	OUT	17	0.48%	0.16%	1.6
83	AMERICA	17	0.48%	0.16%	1.6
84	*IRAQ	16	0.45%	0.15%	1.5
85	DISAPPEAR	16	0.45%	0.15%	1.5
86	DO	16	0.45%	0.15%	1.5
87	FIVE	16	0.45%	0.15%	1.5
88	GO	16	0.45%	0.15%	1.5
89	MORNING	16	0.45%	0.15%	1.5
90	NEVER	16	0.45%	0.15%	1.5
91	NIGHT	16	0.45%	0.15%	1.5
92	THERE	16	0.45%	0.15%	1.5

93	BED	15	0.43%	0.14%	1.4
94	DIFFERENT	15	0.43%	0.14%	1.4
95	HUSBAND	15	0.43%	0.14%	1.4
96	MEAN	15	0.43%	0.14%	1.4
97	THING	15	0.43%	0.14%	1.4
98	WAIT	15	0.43%	0.14%	1.4
99	WAS	15	0.43%	0.14%	1.4
100	'rubs hands'	14	0.40%	0.13%	1.3
101	BREAK	14	0.40%	0.13%	1.3
102	HOUSE	14	0.40%	0.13%	1.3
103	IF	14	0.40%	0.13%	1.3
104	LOOK-AROUND	14	0.40%	0.13%	1.3
105	OTHER	14	0.40%	0.13%	1.3
106	SIX	14	0.40%	0.13%	1.3
107	SLEEP	14	0.40%	0.13%	1.3
108	STORY	14	0.40%	0.13%	1.3
109	gesture 2X CL-5 PALMS UP	14	0.40%	0.13%	1.3
110	BIG	13	0.37%	0.12%	1.2

Table 9: Ranked frequency of the top 110 signs in the SOI corpus - after processing for variants

[i] Symbolic units addressed in this paper include content signs, function signs, lexical signs, non-lexical signs, partly lexical signs, pointing signs and depicting signs.

[ii] Data was originally tagged for the right and left hand but with consultation of metadata regarding each signers handedness, this data was transposed to dominant and non-dominant hand for analysis.