Full title:
Distribution and duration of signs and parts of speech in Swedish Sign Language

Short running head:
Distribution and duration of signs and PoS in SSL

Carl Börstell[a], Thomas Hörberg[a], and Robert Östling[a,b]
[a]Stockholm University, [b]University of Helsinki

Abstract
In this paper, we investigate frequency and duration of signs and parts of speech in Swedish Sign Language (SSL) using the SSL Corpus.

The duration of signs is correlated with frequency, with high-frequency items having shorter duration than low-frequency items. Similarly, function words (e.g. pronouns) have shorter duration than content words (e.g. nouns). In compounds, forms annotated as reduced display shorter duration. Fingerspelling duration correlates with word length of corresponding Swedish words, and frequency and word length play a role in the lexicalization of fingerspellings.

The sign distribution in the SSL Corpus shows a great deal of cross-linguistic similarity with other sign languages in terms of which signs appear as high-frequency items, and which categories of signs are distributed across text types (e.g. conversation vs. narrative).

We find a correlation between an increase in age and longer mean sign duration, but see no significant difference in sign duration between genders.

Keywords: Swedish Sign Language, lexical frequency, part of speech, distribution, duration, sign language corpus, grammaticalization

1. Introduction

Frequency effects in language have been accounted for in a variety of domains. For instance, frequency has been shown to play a role in grammaticalization (Bybee 2003; Bybee 2007) and language acquisition and processing (Ellis 2002; Lahey & Ernestus 2014). In terms of phonology and phonetics, the frequency of words in a language is associated with articulatory reduction (Browman & Goldstein 1992; Bybee & Scheibman 1999; Gahl 2008). For spoken language, the interplay between word frequency, distribution, and length has been investigated with the help of frequency statistics from corpus data, and one observation is that token frequency correlates with word length, such that high-frequency items tend to have shorter length (cf. Zipf 1935; Zipf 1949), which is accounted for by articulatory reduction and economy in language. Although the relationship between word frequency and length has been extensively investigated for spoken language (in spoken and written form), it has, to the best of our knowledge, never been investigated to any larger extent for signed language, making this study the first of its kind. In this study, we seek to investigate the interaction between frequency and duration in a signed language using the naturalistic corpus data from a corpus of Swedish Sign Language (SSL). Looking at the possible interaction between frequency and duration in a signed language widens the perspective in terms of linguistic diversity. Only if the frequency phenomena accounted for in spoken language are present in signed language as well is it possible to call them universal, since human language appears in two distinct modalities: spoken and signed.

The field of sign language linguistics is still young, and the introduction of corpus methods to the field is even younger. With the advent of technical tools for collecting, storing and handling video data alongside linguistic annotations, the possibility of creating corpora of sign languages is now a reality. One existing sign language corpus is the Swedish Sign Language Corpus (SSL Corpus). After a number of years of compiling and editing the material, a work still in progress, we now have the opportunity to extract data from it to present some initial corpus linguistic findings from Swedish Sign Language.

In this paper, we examine some aspects of the interaction between frequency, sign type and duration in SSL using data from the SSL Corpus. Although some previous studies have already looked at sign frequency and distribution in other sign languages (Morford & MacFarlane 2003; McKee & Kennedy 2006; Johnston 2012; Fenlon et al. 2014), our study spans across linguistic phenomena of phonetics and grammaticalization, and is unique in the sense that it covers the interaction between frequency and duration, not only for signs but also for parts of speech, and that we investigate lexical frequency and frequency phenomena in a sign language that has not been considered in the previous research in this domain. The paper starts out with a description of the distribution of signs, parts of speech, and sign types in SSL, as well as a comparison of this distribution to that found in previous corpus studies on sign languages. We then relate our frequency findings to sign and part of speech durations based on general phenomena in corpus linguistics, and in what ways text type (e.g. conversation vs. elicited narrative) affects sign distribution. We also present two sub-studies on compounds and fingerspelling, respectively. For compounds, we look at the connection between sign duration and two types of compound signs, namely whether they are seen as "reduced"/"non-reduced" types. Regarding fingerspelling, we investigate the interaction between type and token frequency and duration on the one hand, and the word length of the (written form of the) target word in number of written characters on the other. These sub-studies are especially important for our understanding of grammaticalization patterns and lexicalization preferences in the signed modality. Lastly, we investigate differences in general sign duration across signer groups, to see whether factors such as age and gender play a role in the articulation rate of signs.

2. Background

2.1 Frequency and duration across modalities

A well-known phenomenon in linguistics is the power law distribution of words in language, for which the token frequency (i.e. the number of times a specific word occurs in a text/corpus) of words is inversely proportional to their frequency rank (i.e. the placement of the word in a ranking from the most frequent to the least frequent, in terms of occurrences), commonly known as "Zipf's law" after George K. Zipf (see Zipf 1935; Zipf 1949). A closely related phenomenon is, as noted by Zipf himself, that frequency and word length tend to be correlated as well, with high-frequency items being shorter than low-frequency items. This has been investigated in a number of subsequent studies from phonetic and psycholinguistic research (e.g. Wright 1979; Pierrehumbert 2002; Gahl 2008; Diessel 2007) to probabilistic distribution research (e.g. Jurafsky et al. 2001; Sigurd, Eeg-Olofsson & van de Weijer 2004). This phenomenon has also inspired a line of work within the domain of grammaticalization (Bybee, Perkins & Pagliuca 1994; Bybee & Scheibman 1999; Bybee & Hopper 2001a; Bybee 2002; Bybee 2003; Bybee 2007), explaining the evolution of grammatical items out of the reduction of frequently used lexical items. The common denominator in these areas of research when it comes to the question of word frequency and length is basically that economy in the production of language is associated with frequency, with high-frequency items normally being more reduced than low-frequency items. However, this has not been thoroughly investigated in the visual-gestural modality of signed language (for general discussions on lexicalization and grammaticalization in the signed modality, see Johnston & Schembri 1999; Pfau & Steinbach 2006).

What has been investigated to some extent for signed language is instead duration of signs without explicitly taking frequency into consideration. An early study by Bellugi and Fischer (1972) compared the articulation rate in spoken and signed language based on a small dataset from bilingual users of English and American Sign Language (ASL). They found that while articulation rate in spoken language is generally higher, pauses tended to constitute a higher percentage of the total production time (i.e. lowering the speech rate). However, they also found that the mean duration per proposition in the production was more balanced between the modalities. They consider the grammatical differences between English and ASL a potential factor behind the differences between the modalities, such that the complex simultaneous morphology of signed language—for instance incorporation/agreement elements (cf. Emmorey 2003; Vermeerbergen, Leeson & Crasborn 2007)—allows ASL do without certain words/morphemes that are required in English. In another cross-modal comparison, Grosjean (1979) found some differences connected to production rate between sign and speech. He observed that a change in global rate meant that signers change the time they spend articulating, while speakers change the time spent pausing. Also, the physiological necessity of breathing was found to have consequences for production patterns, such that speakers tend to breathe at syntactic breaks, whereas signers (with articulators separate from the breathing apparatus) breathe at locations independent from syntactic patterning. For spoken language, differences in articulation rate have been found between different groups of speakers, pointing to factors such as region, age, and gender affecting the rate of articulation, for instance showing that younger adults have higher articulation rate than older adults, and that men exhibit faster speech than women (Byrd 1994; Ramig 1983; Jacewicz et al. 2009). Such factors affecting production rate have, to the best of our knowledge, not been researched for any sign language.

When it comes to the issue of duration of the individual signs in sign languages, there have been a number of studies looking at the articulation of signs, often from a physiological perspective. For instance, reduction over time as part of lexicalization/grammaticalization has been investigated for compounds, showing that assimilation of handshape and location play a part in sign reduction, in that compound elements are reduced compared to their elements produced in isolation—mainly that initial elements are reduced more than final elements—and that previously iconic forms may be lexicalized as less iconic forms (Frishberg 1975; Woodward, Jr. 1976; Wallin 1982). For syllabic structure, the number of syllables with regard to stress has also been shown to be a factor, such that stressed syllables do not necessarily result in longer duration, but rather is associated with a higher number of syllables (Wilbur & Nolen 1986; Wilbur 1999). Tyrone and Mauk (2010) show that phonetic reduction in the shape of sign lowering (i.e. articulating the sign in a lower position relative to the body) is a feature of natural signing, and that factors such as production rate and phonetic context correlate with this feature. Looking at fingerspelling, a number of studies on ASL have investigated production and perception of fingerspelling with assimilation and phonetic reduction as properties of incorporating and adapting (sequential) written letters into the visual modality. These studies have shown that the duration of individual fingerspelled letters fall within the range 125–300 milliseconds (ms) per letter (Zakia & Haber 1971; Wilcox 1992; Jerde, Soechting & Flanders 2003; Quinto-Pozos 2010).

Parts of speech in sign languages have been investigated with regard to duration to some extent. For instance, Hunger (2006) claims that verbs have twice as long duration as corresponding nouns in semantically and phonologically related noun-verb pairs in Austrian Sign Language (ÖGS).

Several other studies have looked at other formational differences between nouns and verbs in noun-verb pairs, but duration has often not been explicitly stated as a distinctive feature, although other quantitative features, such as size and syllable repetition, have (see Tkachman & Sandler 2013 for an overview). Looking at context, Liddell (1978) and Grosjean (1979) found that signs in sentence-final position are longer than those mid-sentence, and several studies have shown that verbs often show up in a sentence-final position, particularly when they are also modified morphologically, such as with reduplication (cf. Fischer & Janis 1990; Bergman & Dahl 1994; Liddell 2003). Thus, there might be general differences in duration between parts of speech that are in fact the result of context. This study will, however, not take context into consideration, since the SSL Corpus does not feature any complete prosodic or syntactic segmentation (see section 3).

2.2 Sign language corpus linguistics and sign distribution

Applying corpus methods to sign languages is not a simple task, seeing as there have been a number of technical obstacles to solve. The most substantial technical issue is, of course, that the signed modality requires different means of recording than do spoken languages, which means that the recording and storing of sign language data need to be done using video rather than audio formats (cf. Crasborn et al. 2007; Johnston 2010). Additionally, sign linguistics generally does not have any standardized tool in terms of transcribing signed language equivalent to the International Phonetic Alphabet (IPA) for spoken language, although a number of systems for phonemic and/or phonetic notation do exist (cf. Miller 2006; Frishberg, Hoiting & Slobin 2012; Crasborn 2015). Since the common practice for writing down sign language data has instead been using *sign glosses* based on words from a written (i.e. spoken) language, we have the possibility to render sign language data in an easily machine-readable form. However, this practice requires strict conventionalization in the labels used for specific sign types, which has led to the development of lexical databases of sign glosses with fixed label–meaning relationships such that signs may be adequately and consistently referred to using the same label, which is a prerequisite for the construction of a corpus (cf. Johnston 2010; Schembri & Crasborn 2010; Mesch & Wallin 2015).

There have been four main lexical frequency studies on sign languages, to date, and each of these has dealt with a different language. These studies are: Morford and MacFarlane (2003) on American Sign Language (ASL); McKee and Kennedy (2006) on New Zealand Sign Language (NZSL); Johnston (2012) on Australian Sign Language (Auslan); and Fenlon, Schembri, Rentelis, Vinson and Cormier (2014) on British Sign Language (BSL) (based on Cormier et al. 2011).

The ASL study used a small-scale database (even by sign language standards), comprising only 4,111 tokens collected from video recordings. The NZSL study had the largest dataset, with over 100,000 tokens in their corpus. The Auslan and BSL corpora are similar in the sense that there has been continuous collaboration between the respective corpus projects, striving towards cross-linguistic annotation conventions applied to both corpora, and also that they are available in the ELAN Annotation Format (EAF) (see section 3.1) in which annotation files are linked and time-aligned to a video file, unlike the ASL and NZSL studies. Whereas the Auslan study used predominately narrative data, the ASL study was based mostly on conversational data, and the BSL study focused entirely on conversational data (see Fenlon et al. 2014 for an overview of the different studies).

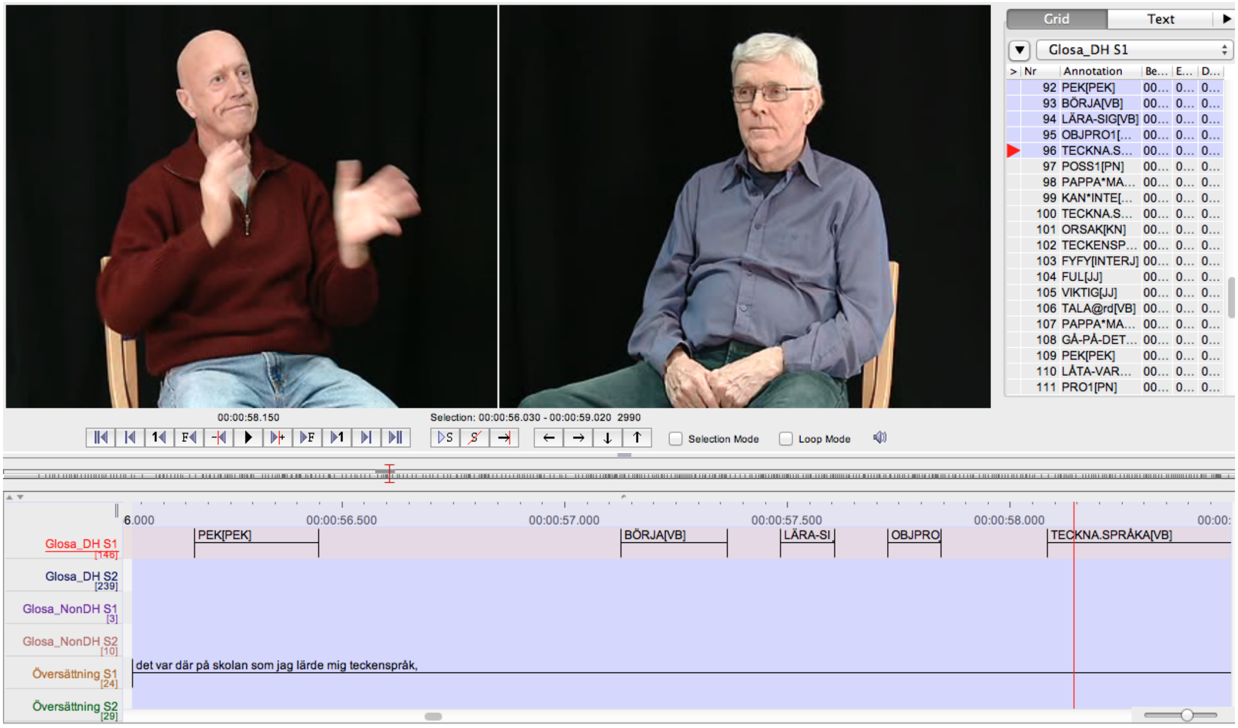
Looking at phonetic effects of frequency, there have been a couple of studies using sign language corpus data. Schembri et al. (2009) argued that frequency is a factor in sign-lowering, in the sense that high-frequency items are more likely to be subject to phonetic lowering than low-frequency items. Similarly, Fenlon, Schembri, Rentelis and Cormier (2013) looked at variation in the 1-handshape in BSL, finding that high-frequency items were more likely to exhibit variation in the handshape than low-frequency items. As mentioned earlier, frequency accounts for phonological reduction in spoken language (Browman & Goldstein 1992; Lahey & Ernestus 2014), and this is directly related to the idea of grammaticalization, which is a key property of language structure and evolution (Bybee & Hopper 2001b). Based on the few previous studies looking at frequency effects on signs, there are indications that frequency does affect the phonetics of signed language, as it does for spoken language, which is why we use duration as an indication of phonetic reduction and investigate whether it is correlated with frequency. Further study in this domain is important in order to investigate grammaticalization in the evolution of structure of signed language, and to see which aspects are modality-specific vs. modality-independent. One aspect of study with regard to grammaticalization would be the emergence of function signs from content signs in the signed modality. This will be important in the current study when looking at different categories of parts of speech and their respective durational properties. This study does not, however, delve deeper into the phonetic features of individual signs, but simply uses sign duration as a feature in its own right. Thus, the aim is to investigate frequency effects of sign duration (as indications of articulatory reduction) using pre-annotated corpus data, rather than analyze the exact phonetic realizations that constitute such reductions. Such phonetic analyses are left for future research using different methods.

3. Method and material

3.1 The Swedish Sign Language Corpus

The SSL Corpus is a still-expanding corpus of Swedish Sign Language (SSL), one of many sign languages of the world and the main sign language of the Deaf community in Sweden. The SSL Corpus project commenced in 2009 and was officially completed in 2011. Since then, the work on completing video editing, annotations and translations has continued. The data comprise approximately 25 hours of sign language video data from 42 signers of different age and geographical background signing in pairs, engaged in (semi-)spontaneous conversations or producing elicited narratives (cf. Mesch, Wallin & Björkstrand 2012; Mesch 2013; Mesch & Wallin 2015). The data of the SSL Corpus are being published open-access online continuously as transcribed and translated files are completed. The video files are uploaded along with annotation files in the ELAN Annotation Format (EAF) (see Wittenburg et al. 2006), a format that has been used for a number of other sign language corpora (Crasborn et al. 2007; Crasborn, Zwitserlood & Ros 2008; Crasborn & Zwitserlood 2008; Johnston 2008; Schembri et al. 2014). The version of the SSL Corpus used for this study contains 48,686 "raw sign token annotations", of which 44,786 tokens have been included here. The removed sign tokens are either uncertain glosses (tagged with "@z" in the SSL Corpus) and non-manual sign annotations (cf. Wallin & Mesch 2014).

The SSL Corpus data consist of the SSL video files, which are time-aligned to annotation tiers in the ELAN Annotation Format (EAF). The annotation tiers consist of four sign gloss tiers (one for each signer and hand), with sign glosses representing SSL signs as labels for signs (Wallin & Mesch 2014; Mesch & Wallin 2015), and a translation tier with idiomatic and stylistically approximate translations into Swedish for each signer (Mesch et al. 2012; Mesch, Rohdell & Wallin 2014).[1] Figure 1 illustrates the basic view of an SSL Corpus file as seen in the ELAN software interface.



Glosa_DH S1

>	Nr	Annotation	Be…	E…	D…
	92	PEK[PEK]	00…	0…	0…
	93	BÖRJA[VB]	00…	0…	0…
	94	LÄRA-SIG[VB]	00…	0…	0…
	95	OBJPRO1[…	00…	0…	0…
▶	96	TECKNA.S…	00…	0…	0…
	97	POSS1[PN]	00…	0…	0…
	98	PAPPA*MA…	00…	0…	0…
	99	KAN*INTE[…	00…	0…	0…
	100	TECKNA.S…	00…	0…	0…
	101	ORSAK[KN]	00…	0…	0…
	102	TECKENSP…	00…	0…	0…
	103	FYFY[INTERJ]	00…	0…	0…
	104	FUL[JJ]	00…	0…	0…
	105	VIKTIG[JJ]	00…	0…	0…
	106	TALA@rd[VB]	00…	0…	0…
	107	PAPPA*MA…	00…	0…	0…
	108	GÅ-PÅ-DET…	00…	0…	0…
	109	PEK[PEK]	00…	0…	0…
	110	LÅTA-VAR…	00…	0…	0…
	111	PRO1[PN]	00…	0…	0…

Selection: 00:00:56.030 - 00:00:59.020 2990

6.000 00:00:56.500 00:00:57.000 00:00:57.500 00:00:58.000 00:00:

Glosa_DH S1 [146]: PEK[PEK] BÖRJA[VB] LÄRA-SI OBJPRO TECKNA.SPRÅKA[VB]
Glosa_DH S2 [239]
Glosa_NonDH S1 [3]
Glosa_NonDH S2 [10]
Översättning S1 [24]: det var där på skolan som jag lärde mig teckenspråk,
Översättning S2 [29]

Figure 1. The user interface view of the SSL Corpus in ELAN. The two video windows show each of the two signers, the right-most window shows a grid with the signs produced by Signer 1, sitting on the left, with time-stamps for the signs' on- and offset. The bottom window shows the current viewing point on a timeline with each of the annotation tiers stacked vertically.

In the SSL Corpus, the only segmentation of the sign language data is done on the lexical level, with each sign corresponding to a single annotation cell in ELAN. Signs have been segmented based on a principle similar to that of the Corpus NGT (see Crasborn, Bank, et al. 2015)—a corpus of Sign Language of the Netherlands (NGT)—namely that signs are considered to start when the hands have assumed the handshape of the sign and/or commenced the movement of the sign, and end before the hands leave the final location and/or change the handshape of the sign (cf. Wallin & Mesch 2014; Wallin & Mesch 2015). Because the resolution of the video files in the SSL Corpus is 25 frames per second, the smallest meaningful unit in duration is 40 milliseconds. In the SSL Corpus data, we find instances of sign gloss annotations with durations that do not fall into 40 ms intervals, and because of this, all the durations in our data have been rounded off to the nearest 40 ms interval (with 40 ms as the minimum duration allowed) in order to remove meaningless duration intervals attributed to frames-to-annotation mismatches in the manual annotations.[2] In the SSL Corpus, holds of signs (usually by the non-dominant hand) spreading over succeeding signs (articulated by the dominant hand) are annotated accordingly, since each hand has its own annotation tier. However, instances of holds are not included in this study due to the nature of the annotation procedure, in which sign holds are annotated with a second annotation cell *after* the initial execution of the sign (i.e. its 'basic' articulation).

Besides the lexical (i.e. sign) segmentation, the SSL Corpus is segmented on the Swedish translation tiers. However, the translation tiers normally contain longer annotation cells segmenting the texts into larger chunks not necessarily corresponding to conventional linguistic units in SSL. Though these translation annotation cells are based on chunks adequate for representing the Swedish translations, they do not always correspond to consistent linguistic units in Swedish, such as clauses or sentences, but rather appropriate information-segments. As an illustration of this, in Figure 1, a single annotation cell on the translation tier for Signer 1 is selected, spanning eight consecutive signs on the sign gloss tier. A previous study found that the translation tier segmentations in the SSL Corpus normally stretch across more than a single 'clausal unit' in SSL, and the preliminary attempts at segmenting the SSL Corpus into clausal units based on visual cues (mainly prosodic markers) were found to be quite time-consuming (Börstell, Mesch & Wallin 2014). Instead, a coming release of the SSL Corpus is currently being prepared, in which the data are segmented into clausal units based on syntactic and prosodic information, and annotated for basic syntactic functions, but thus far only a small portion of the corpus has been segmented and annotated for this (Börstell et al. 2016), which is why these data are not available in the corpus version used for the present study.

The available release of the SSL Corpus also features part of speech tagged sign glosses. Expanding on a previous attempt to induce part of speech categories for SSL (Sjons 2013), and a semi-automatic method of part of speech transfer based on Bayesian word alignment between the Swedish translations and the sign glosses of the SSL signs (Östling, Börstell & Wallin 2015), the SSL Corpus is now part of speech tagged on the level of sign types—that is, each sign type is associated with a fixed part of speech. Type-level tagging could potentially be a problem if there are sign glosses associated with different functions, but this approach is reasonable for the SSL Corpus seeing as functionally distinct related sign pairs (e.g. related noun-verb pairs such as CHAIR vs. SIT) are differentiated in the sign glosses used (i.e. using different glosses for related nouns and verbs). Although the initial phase of the annotation procedure included a semi-automatic method, the complete list of sign types in the SSL Corpus has been manually corrected by two language experts, and subsequent expansions of the corpus are manually tagged for parts of speech directly by the annotators. The part of speech tags are attached directly to the sign gloss (for instance, "PRO1[PN]"), and any novel sign gloss used is immediately assigned a part of speech tag, attached to the sign gloss. In total, 18 parts of speech have been tagged for in the SSL Corpus, of which eight have previously been described in Ahlgren and Bergman (2006), and the other 10 have been added to make for a more fine-grained division and account for categories identified when working with the corpus data. Table 1 below lists the parts of speech annotated in the SSL Corpus, together with the tag labels used and a note whether or not it was included in Ahlgren and Bergman's description.

Table 1. Parts of speech and corresponding tags in the SSL Corpus.

Part of speech	**Tag**	**In Ahlgren & Bergman (2006)**
Adjective	JJ	yes
Adverb	AB	yes
Buoy	BOJ	–
Conjunction	KN	yes
Gesture	G	–
Interjection	INTERJ	–
Nominal classifier	NNKL	–
Noun	NN	yes
Numeral	RG	yes
Point	PEK	–
Preposition	PP	yes
Pronoun	PN	yes
Verb	VB	yes
Verb (CA)	VBCA	–
Verb (depicting)	VBAV	–
Verb (locative)	VBPP	–
Verb (stative)	VBS	–
Unlabeled/uncertain	?	–

As can be seen in Table 1, sign glosses for which a (single) part of speech cannot be assigned are tagged with the symbol "?".

The existence of part of speech tagged corpora of sign languages is rare (although see the label 'Grammatical category' in Johnston 2014), and this version of the SSL Corpus is, to the best of our knowledge, unique in the world by having an annotation procedure in which the initial phase is based on semi-automatic computational methods (albeit, as explained, having been manually corrected). This puts us in an exceptional position, since we have the opportunity of exploring some general phenomena in corpus linguistics based on a number of different linguistic features, by utilizing new tools.

Although the SSL Corpus is compiled with the intention of being balanced across signers (with regard to age, gender, and geographical background) and text types, the latest release is not balanced between signers or text types, since it only contains the material that has been edited and annotated to date. Table 2 shows the distribution of signer age and gender for the version used in this study. Similarly, Table 3 shows the distribution of files and sign tokens across text types in the SSL Corpus version used in the present study. The three text types available in the SSL Corpus are: 'Conversation', constituting semi-spontaneous dialogues (some general conversation topics were presented to the signers); 'Elicited narrative', consisting of re-tellings of picture-stories; and 'Presentation', consisting of the introductions signers made when first sitting down with their signing partner, telling each other about their background and current situation, etc.

Table 2. Distribution of each age group and gender of the informants in the SSL Corpus.

Age group	**Females**	**Males**
20–29	8	1
30–39	1	5
40–49	4	3
50–59	4	4
60–69	3	5
70–100	-	4
Total	*20*	*22*

Table 3. Distribution of text types in the SSL Corpus.

Text type	**Number of files**	**Number of tokens**
Conversation/narrative	63	38,232 (85.4%)
Elicited narrative	16	4,055 (9.1%)
Presentation	6	2,499 (5.6%)
Total	*85*	*44,786*

As shown in Table 3, the main portion of the data comes from conversational rather than elicited data. However, unlike Fenlon et al's (2014) study, our material is not composed exclusively of conversational data.

3.2 Sign categories

In order to compare the distribution of sign categories to previous frequency studies on sign languages as accurately as possible, we have aimed to label signs according to the category definitions described in Fenlon et al. (2014). An exact mapping to Fenlon et al's definitions is not possible to achieve due to differences in the annotation conventions for the different corpora, but the aim has been to match each category as closely as possible. The specific criteria used for our data from the SSL Corpus, and how they may differ from previous definitions, are listed below.

"Core" lexical signs are defined negatively in the SSL Corpus data. There are a number of designated tags used in the sign glosses in the SSL Corpus, and those tags define signs belonging to other categories (see below). Thus, all signs with an established gloss *not* containing any of the designated tags used in the SSL Corpus are defined as core lexical signs.

Fingerspellings carry the tag "@fs" (for fingerspelling, originally labeled '@b' for *bokstavering* 'fingerspelling') in the SSL Corpus. All signs for which the whole sign (i.e. an entire sign in isolation or all the elements of a compound sign) is fingerspelled is classified as a case of fingerspelling. In the SSL Corpus, all instances of fingerspelling are annotated as such, even for signs that have become lexicalized from a fingerspelled form, possibly having undergone certain phonological changes that are different from a novel or *ad hoc* fingerspelling.[3] One heavily reduced fingerspelled-derived sign is excluded from the category, namely the feedback sign for 'yes', YES@fb. This sign was excluded because its function as a feedback sign makes it highly frequent, thus skewing the statistics when comparing the relative distribution of fingerspelled items across sign languages (see section 4), and also patterning differently in terms of duration because of the repeated production (see section 5.2). Initialized signs are not annotated as fingerspellings in the SSL Corpus, and are also excluded from the sign category of fingerspellings.

Classifier signs are defined by the tag "@p" (for *polysyntetisk* 'polysynthetic', a term that has been used in Swedish sign language research to refer to depicting signs) in the sign gloss.

Gestures are defined by the tag "@g" (for *gest* 'gesture') in the sign gloss.

Pointing signs always contain either the string 'PRO' (for *pronomen* 'pronoun') or 'POINT' (originally labeled "PEK" for *pekning* 'point'). Sign glosses containing any of these substrings are labeled pointing signs. In the SSL Corpus, 'PRO' is only used for first person reference (singular or plural), whereas all other instances of index pointing are glosses as POINT. For POINT signs directed towards a physically present referent, there is a suffixed ">x" to the gloss, where "x" indicates the entity pointed to. This means that we can distinguish second person reference from other non-first pointing by the gloss "POINT>person", referring a point directed at a physically present person (which in the context of the SSL Corpus always refers to the addressee).

Buoys contain the string "BOJ" ('buoy') or "DELIMITER" (originally labeled "AVGRÄNS", for *avgräns(ning)* 'delimiter'), hence sign glosses with any of these (sub)strings are labeled as buoys (Wallin & Mesch 2015).

Sign names receive the designated tag "@pn" (for proper noun, originally '@en' for *egennamn* 'proper noun'), which means that all signs with this tag were automatically labeled as a sign name.

Other is a category of signs not falling into the categories above, and is defined based on different criteria. For instance, it covers instances of constructed action (containing the tag "@ca"). It also covers signs for which only a part of the sign, such as one element of a compound, is fingerspelled (following Fenlon et al. 2014). Unlike the Fenlon et al. study, uncertain/indecipherable signs are removed completely from the data (due to the difficulties of including these in a frequency–duration study), resulting in 3,900 tokens being stripped from the raw data of the SSL Corpus, giving us the 44,786 tokens investigated in this study. As in Fenlon et al., ambiguous or simultaneous signs that were not easily defined as belonging to a single category were also labeled as 'Other'.

3.3 Research questions

The research questions that this study set out to answer were the following:

a) What are the most frequent signs and sign categories in SSL and how do the sign and sign category distributions relate to previous corpus studies on other sign languages?
b) How does sign duration (sign length in time) relate to frequency of signs and their corresponding parts of speech?
c) Are there durational differences between the two types of compounds in the SSL Corpus: reduced vs. non-reduced?
d) How does target word length (in written characters) and frequency affect the likelihood of using fingerspelling?
e) Are there statistically significant differences in sign duration between older vs. younger, and female vs. male signers?

These questions will be addressed using the data from the SSL Corpus, making this the first study to address issues of frequency and duration in a sign language, based on extensive corpus data.

4. Distribution of signs and sign categories in SSL

4.1 Lexical frequency of sign types

Our first task when confronted with the part of speech tagged SSL Corpus data was to look at the distribution of sign types in the corpus. Table 4 below lists the top-20 sign types by frequency in the SSL Corpus, with corresponding part of speech tags, number of tokens, and mean duration.

Unsurprisingly, pronoun-like signs are frequent, with all singular forms of the pronominal points (PRO/POINT) being in the top-5. As the main part of the SSL Corpus consists of conversational data, the sign YES@fb (fingerspelled sign meaning 'yes') is a highly frequent item, since it is a common feedback signal in dialogues. Two items with a more pragmatic function are the PU@g and SO-TO-SPEAK items. The former is an acronym of "palms up" and refers to a gesture found in several different sign languages, the meaning of which is not entirely clear, but tends to represent a type of presentation or reference to a topic (Engberg-Pedersen 2002; McKee & Wallingford 2011; Ryttervik 2015). The latter, SO-TO-SPEAK, is a sign meaning approximately 'so to speak' or 'kind of', and happens to be a high-ranked item in our data mostly due to its frequent use by a specific signer whose production constitutes a fairly large proportion of the overall tokens in the current release of the SSL Corpus—the signer in question is responsible for roughly half of all occurrences of SO-TO-SPEAK. Apart from these, the lexical sign DEAF shows up in the top-20, quite unsurprisingly since it is a highly relevant concept for the Deaf community, and further such concepts are found as high-ranking items outside of the top-20 ranking.[4]

Table 4. The 20 most frequent sign types and mean duration in the SSL Corpus.

Rank	Gloss	Part of speech	Tokens	Mean duration (ms)
1	PRO1[abcde]	Pronoun	3361	104
2	POINT[abcde]	Point	2425	168
3	PU@g[abcde]	Gesture	1397	267
4	YES@fb[de]	Interjection	782	392
5	POINT>person[c]	Point	488	149
6	BUT[de]	Conjunction	363	161
7	POSS1[acd]	Pronoun	313	110
8	ONE[cde]	Numeral	304	134
9	TO-BE	Verb	299	126
10	DEAF[abd]	Noun	287	219
11	SO-TO-SPEAK	Adverb	282	270
12	PI[b] ('really')	Adverb	265	162
13	HAVE[abd]	Verb	260	133
14	HOW	Adverb	257	176
15	LATER	Adverb	241	196
16	THEN@fs[a]	Conjunction	235	149
17	PERF	Verb	221	146
18	NOT[ae]	Adverb	215	170
19	TO[a]	Preposition	214	123
20	GOOD[bcde]	Adjective	208	192

[a] A corresponding sign is also ranked top-20 in ASL.
[b] A corresponding sign is also ranked top-20 in Auslan.
[c] A corresponding sign is also ranked top-20 in BSL.
[d] A corresponding sign is also ranked top-20 in NZSL.
[e] A corresponding sign is also ranked top-20 in NGT.

In order to evaluate the similarity of sign frequency across sign languages, we used the frequency data from the four previous studies on lexical frequency in sign languages (Morford & MacFarlane 2003 for ASL; Johnston 2012 for Auslan; Fenlon et al. 2014 for BSL; McKee & Kennedy 2006 for NZSL), and then added frequency statistics based on the raw data on Sign Language of the Netherlands collected from English annotation files of the Corpus NGT (Crasborn & Zwitserlood 2008; Crasborn, Zwitserlood & Ros 2008; Crasborn, Zwitserlood, et al. 2015), as well as our SSL data from the SSL Corpus.[5] The superscript letters next to the signs in Table 4 show whether a corresponding sign in another sign language is found among the 20 most frequent signs in the corpus of that language. Though it is no easy task to directly match lexical items in one language to those in another, the meanings and functions of the signs in Table 4 are fairly straightforward. One problematic item is the SSL sign PI (meaning approximately 'genuine(ly), real(ly), very'), which could be linked to signs such as REAL, REALLY, and ACTUALLY in other corpora. The SSL sign PI has a range of uses, but is often used to emphasize something and/or express genuineness or asserting something as true. The sign PERF has a type of perfect aspect associated with it (cf. Bergman & Dahl 1994), such that its use may be similar to that of 'finish' type signs used in other sign languages (Kyle & Woll 1985:142; Johnston & Schembri 2007:153; Johnston et al. 2015), though no such sign is found in the top-20 most frequent signs of the other corpora.

Generally, we see a great deal of similarity across languages, with most of the high-frequency items in SSL being found as high-frequent items in at least one corpus of another sign language. This is true for both content signs (mostly concepts associated with Deaf culture), and function signs (e.g. pronouns/pointing). It is unsurprising to find function items such as pronouns among the most high-ranking lexical items, since this is a pattern found in many languages, spoken or signed.

Thus, it appears to be a modality-independent property. On the other hand, it is noteworthy that an item such as the palms up gesture is not only found among the most frequent items across these sign languages, but it is basically identical in form across languages. It is not remarkable to find pragmatic or discourse particles among high-ranking items, but the similarity in form and—to some extent—also function of this specific sign is interesting. It should, of course, be noted that three of the sign languages of the six compared here are related and sometimes considered dialects of the so-called 'BANZSL' family (cf. Johnston 2003), which arguably accounts for some of the similarities between those languages, but neither ASL or SSL are known to be related to each other or the BANZSL family. In order to investigate whether the palms up gesture is found and frequently occurring across sign languages, a larger language sample would be necessary, including sign languages with a non-European origin. Unfortunately, such data, especially in the sense of corpus data, is scarce.

4.2 Lexical frequency of sign categories

One interesting feature of the papers on sign frequency in ASL, Auslan and BSL is that they all include tables listing types of 'sign categories', apart from individual sign distributions. This is quite useful as it allows for a more general comparison of sign distribution across sign languages. Although the different studies used slightly different labels, Fenlon et al (2014) compiled a comparative table across ASL, Auslan and BSL, and Table 5 below is based on that comparison, with the addition of SSL Corpus data. Conveniently, most of these categories have designated tags in the SSL Corpus (as described in section 3.2 above), resulting in comparative data being fairly straightforward to obtain from the corpus (cf. Wallin & Mesch 2014; Mesch & Wallin 2015).

Table 5. Distribution of sign categories in SSL, ASL, Auslan and BSL.

Sign category	**SSL (*n* = 44,786)**	**ASL (*n* = 4,111)**	**Auslan (*n* = 63,436)**	**BSL (*n* = 24,823)**
'Core' lexical signs	68.3%	73.1%	65.0%	60.3%
Pointing signs	16.1%	13.8%	12.3%	23.0%
Gestures	3.4%	0.2%	6.5%	8.9%
Fingerspelling	4.0%	6.4%	5.0%	3.0%
Classifier signs	2.8%	4.2%	11.0%	2.3%
Buoys	1.6%	n/a	n/a	0.5%
Other	1.3%	n/a	n/a	1.9%
Sign names	2.5%	2.3%	0.2%	n/a

As Table 5 shows, the distribution of sign categories across the four sign languages is quite similar. For instance, lexical signs constitute the majority of tokens across languages, and pointing signs the second largest category of signs. The category 'Classifier signs' is noticeably higher in Auslan (11%) than in the other sign languages (2.3–4.2%), which is likely the result of the Auslan corpus data being heavily based on narrative texts. This difference in text types in the respective corpora would also account for the low number of sign names in the Auslan corpus compared to the other languages, since sign names are expected to be less prominent in narrative style signing.

Because the balance of text types used in the different corpora varies, Table 6 below shows the distribution of sign categories in the SSL Corpus based on the text type labels used within the corpus project, which is similar to what was done for the ASL study (Morford & MacFarlane 2003:220). The left-most genre column ('Conversation') is by far the largest one in terms of token and files (cf. Table 3 above), and is also the one most similar to the Fenlon et al. (2014) study based on conversational data from BSL. Table 6 shows interesting differences between the different text types, such as the 'Classifier signs', as expected, being over-represented in the elicited narratives, and 'Sign names' being over-represented in the 'Presentation' type texts, since that is when the signers mostly describe their background and heritage. Again, this would account for the Auslan distribution differing from the other sign languages in Table 5 above. As in the ASL study, it is only for narrative type texts that classifier signs constitute more than a couple of percent of the total tokens, but in these text types they are highly frequent (14.3% in the SSL Corpus; 17.7% in the ASL corpus data). Thus, it is clear that text type has an impact on the general sign distribution, and that similar tendencies in distributional patterns are found in different languages.

Table 6. Distribution of sign categories in the SSL Corpus based on text type.

Sign category	**Conversation** **(*n* = 38,232)**	**Elicited narrative** **(*n* = 4,055)**	**Presentation** **(*n* = 2,499)**	**All text types** **(*n* = 44,786)**
'Core' lexical signs	68.1%	70.4%	67.9%	68.3%
Pointing signs	17.2%	6.3%	15.2%	16.1%
Gestures	3.7%	1.7%	2.2%	3.4%
Fingerspelling	4.1%	3.1%	4.1%	4.0%
Classifier signs	1.8%	14.3%	0.4%	2.8%
Buoys	1.5%	1.8%	2.1%	1.6%
Other	1.2%	2.4%	0.8%	1.3%
Sign names	2.4%	0.0%	7.3%	2.5%

5. Sign frequency and duration

One of the claims made by Zipf was that word length and frequency are correlated for words in a language. In order to investigate if this pattern holds also for the signed modality, our task was thus to investigate whether high-frequency signs in the SSL Corpus have shorter duration than low-frequency signs. Figure 2 shows that this is, in fact, the case, with the y-axis showing the mean duration of sign types and the x-axis the logarithmic frequency of the sign types. Figure 2 shows a clear negative relationship between sign type mean duration and sign frequency. There is a decrease in sign duration as sign frequency increases. Two signs that diverge slightly from the general downward trend at the upper end of the x-axis (i.e. high-frequency items) are YES@fb (feedback sign meaning ‘yes’) and PU@g (a gestural ‘palms up’). The sign YES@fb is used as a feedback signal, and is a reduced and repeated derivation from a fingerspelled YES@fs. The difference from the fingerspelled YES@fb is that it is repeated, and often this is done as a long feedback sequence, which in the SSL Corpus has been annotated with a single annotation cell spanning over the entire sequence of repetitions, thus resulting in many long duration tokens of the sign. The sign PU@g has, as discussed in section 4, a pragmatic function. It has been shown that this sign has various functions in SSL, several of which are related to turn-taking, either as a way of keeping the turn while thinking, or as a way of handing the turn over to the addressee (Ryttervik 2015). Thus, the sign has functions associated with longer duration, either by filling a thinking pause in the signing, or by being in a sentence-final position, which is known as a position associated with lengthening of signs (cf. Grosjean 1979).

The data were analyzed with linear regression in the statistical language R (R Core Team 2014), using the built-in stats package. Linear regression analyzes a linear relationship between one or several predictor variables and an outcome variable. A linear regression model of the relationship between sign duration and sign log frequency, with sign data for YES@fb and PU@g excluded, found a significant negative correlation ($\beta = -64.2$, $t(4606) = -18.39$, $p < .0001$). In other words, sign duration decreases as a function of sign frequency, which is exactly what we would expect assuming that signed language (represented by SSL) patterns like spoken language. Thus, the Zipfian correlation between frequency and word length is not a modality-specific property of spoken language, but rather a universal feature of language.

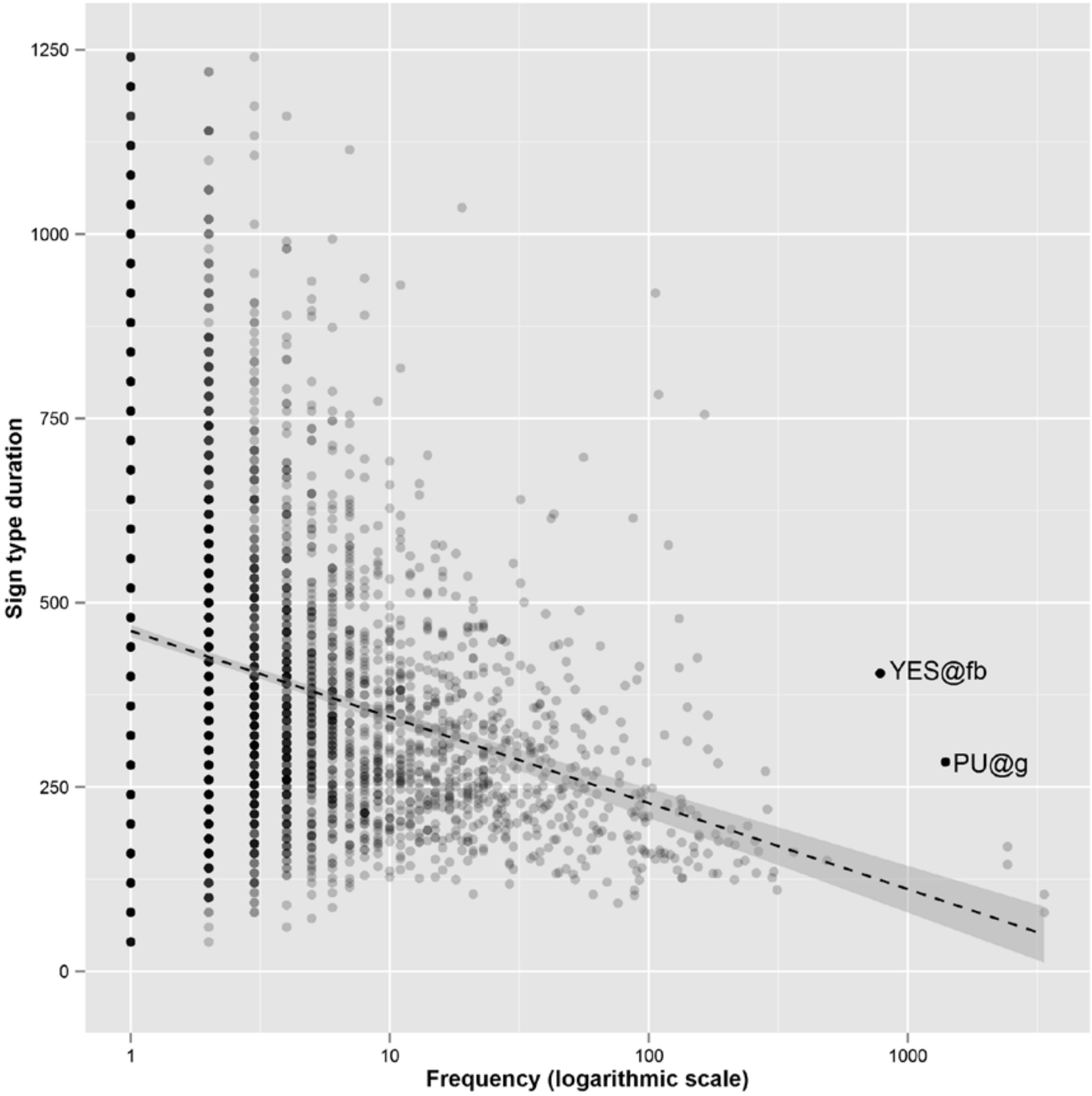


Figure 2. Sign type mean duration as a function of log frequency in the SSL corpus. The regression line shows the fit of a linear regression model of the relationship between log sign frequency and sign duration. The shaded area illustrates the 95% confidence interval of the fitted values. Frequency and duration for sign types YES@fb and PU@g are also shown in the figure.

These findings adhere to the general idea of grammaticalization, which has been accounted for in both spoken (e.g. Bybee & Hopper 2001a; Bybee 2007) and signed (e.g. Pfau & Steinbach 2006; Janzen 2012) language, supporting the idea that frequency plays an important role in the structure of language: higher frequency means shorter duration, which in turn should be attributed to phonetic (and over time possibly phonological) reductions in sign forms. The fact that many of the most frequent lexical items, not only in SSL, but also in other sign languages, are function type signs (e.g. pointing) is unexpected, as is the fact that the most frequent items are shorter in terms of duration, as reduction correlates with frequency. For spoken language, this has been shown to be true also for content words, such that the English word *time* is reduced in production compared to its homophone *thyme* (Gahl 2008). In SSL, we find that there is a correlation between duration and frequency across all items of the SSL Corpus, and a future venture could be to investigate whether durational differences are found between homophonous lexical items. This is not yet possible on the basis of corpus data, due to the limited size of the SSL Corpus. However, the following sections investigate the interaction between lexicalization and duration by looking at specific subgroups of signs: first, the characteristics of two types of compound signs (reduced vs. non-reduced) is explored (section 5.1); second, we look at fingerspelled signs in the SSL Corpus, looking at the interplay between target word (the written words being re-coded as fingerspellings) length and frequency and the duration and frequency of fingerspelled types and tokens in SSL (section 5.2).

5.1 Compounds

One specific type of signs that is interesting from the point of view of lexicalization is compound signs (see Lepic 2015a; Lepic 2015b for an overview). As has been shown in studies on ASL and SSL, signs are articulated faster when they act as a part of a compound rather than being articulated in isolation, in particular the initial element of the compound (Wallin 1982), and compounds tend to assimilate internally into a more monosyllabic form over time (Frishberg 1975; Liddell & Johnson 1986; Wallin 1982; Sandler 2012). However, in many of the studies on sign language compounds, it is not entirely clear whether the signs labeled as compounds are mainly those that are reduced into a monosyllabic form, those in which the elements are clearly separable from each other, or both. In the SSL Corpus, compounds are annotated with designated symbols, differentiating two types of compound signs: compounds with two clearly identifiable forms (usually calqued from Swedish) are separated by a "^" in the gloss (e.g. HEARING^INJURE 'hard of hearing'); and high-frequent collocation compounds that have merged into a co-articulated/monosyllabic form separated by a "*" in the gloss (e.g. BELIEVE*NOT).[6] We decided to investigate the distribution and duration of these types of compounds in the SSL Corpus in order to identify if the two categories exhibit differences that mirror the characteristics that would be expected, namely that the reduced types have significantly shorter duration than the non-reduced, and also that the reduced are more frequent than the non-reduced. Considering that compounding is a type of word formation, we would also expect the non-reduced category to be larger in terms of sign types, since it is, in a way, the open category, whereas the reduced category is expected to contain compounds that are lexicalized rather than novel/productive, and therefore should be a smaller category but characterized by more tokens per type. For the purpose of this study, we excluded any compound containing one or more fingerspelled elements, and excluded numerals. We also restricted our analysis to compounds with exactly two elements. This was done to avoid differences in duration based on differences in the number of elements, as the non-reduced category contains many items with three or more elements as part of the compound.

As shown in Table 7, the mean duration for each compound type confirms that there is a difference between the two categories, with the non-reduced compounds having a mean duration that is significantly (275 ms) longer than that of the reduced compounds ($t(391) = 9.4, p < .0001$). These findings support the idea that high-frequency items tend to be more reduced, and also suggests that the division into separate categories in the annotation is valid. Tables 8 and 9 show all sign types with more than six tokens in the SSL Corpus. Table 7 also shows that the reduced forms are more frequent per type (6.6 tokens per type on average), whereas the non-reduced compounds are represented by more types but occur on average less times per type (2.5 tokens per type). Unsurprisingly, the distinct compounds are more frequent as hapax (i.e. single occurrence) items, with 194 hapax distinct compounds, but only 53 hapaxes for the reduced compounds. This is to be expected since the reduced compounds are likely to have been reduced as a result of frequency, hence high-frequency items tend to be reduced, and vice versa.

Table 7. Types, tokens, and mean duration of the two types of compound signs in the SSL Corpus.

Type	Types	Tokens	Mean duration (ms)
Non-reduced (marked "^")	302	769	654
Reduced (marked "*")	91	600	379

Tables 8 and 9 show the most frequent sign types in each of the two compound categories (non-reduced vs. reduced, respectively), covering signs that occur more than six times in the SSL Corpus.

Table 8. Non-reduced compound signs with >6 tokens in the SSL Corpus.

Gloss	Tokens	Mean duration (ms)
SELF^CLEAR ('of course')	52	277
SNOW^OLD-MAN ('snowman')	27	450
DEAF(L)^ASSOCIATION	23	471
PICTURE^TELEPHONE	23	466
MANILLA^SCHOOL	20	536
VOCATION^SCHOOL	16	577
NOT^THING ('nothing')	16	350
SICK^HOUSE ('hospital')	13	412
KRISTINA^SCHOOL	12	400
HEARING^INJURE ('hard of hearing')	11	574
VÄNERSBORG^SCHOOL	10	692
NEWS^SIGN (news show in SSL)	9	568
DEAF(L)^SCHOOL	8	600
WORLD^CONGRESS	8	545
PLAY^SAND (the town of Leksand)	8	445
WORK(J)^FRIEND ('colleague')	8	510
ASSOCIATION^LIFE	8	390
SIGN^SQUARE (online platform for SSL)	7	462
GUARD^MASTER ('janitor, handyman')	7	565
HIGH^STADIUM ('(junior) high school')	7	462
TRAVEL^LEADER ('travel guide')	7	514
PRE-K^SCHOOL ('preschool')	7	617
WORK(J)^PLACE	7	628

Table 9. Reduced compound signs with >6 tokens in the SSL Corpus.

Gloss	Tokens	Mean duration (ms)
TO-BE*POINT	95	211
EXIST*NOT	74	299
HAVE*NOT	53	241
KNOW*PERF	49	197
KNOW*NOT	38	303
SEE*SHOW	24	258
MOTHER*FATHER	23	293
SAID*POINT	21	251
CAN*NOT	20	232
REMEMBER*NOT	16	300
ONE*DAY	10	380
FATHER*MOTHER	9	222
BELIEVE*NOT	8	185
WANT*HAVE	8	215
UNDERSTAND*NOT	7	377
NEED(B)*NOT	7	268

A noteworthy difference between the items in Tables 8 and 9 is which types of signs constitute the parts of the compounds. For the non-reduced compounds, the items in Table 8 are mainly noun-noun compounds, which reflects a common compounding pattern of Swedish. For the reduced compounds, these high-frequency items are to a large extent verb-negation collocations, or other collocations in which at least one element is a function rather than content type sign (such as KNOW*PERF 'know already'). Negation collocations have been investigated in several sign languages, finding that verb-negation chunks tend to form a close phonological unit, to the extent that the items may become fused together, and that frequency plays a part in forms becoming reduced, over time giving rise to irregular negation (Zeshan 2006; Wilkinson 2016). Again, we see that the interaction between grammaticalization, reduction and frequency are closely intertwined, and the two categories of compounds annotated for in the SSL Corpus do exhibit statistically significant durational differences.

5.2 Fingerspelling

One rather unique aspect of many sign languages in the world is the use of a manual alphabet to represent the words of a spoken language (in its written form) by different handshapes produced in a sequential manner. Fingerspelling is basically lexical borrowing from a spoken language to a sign language, and while this strategy is sometimes an *ad hoc* solution in order to express concepts for which there are no established signs, it may result in certain concepts being expressed by a fingerspelled form by default (i.e. lexicalized fingerspellings). Due to the sequential nature of using a manual alphabet (cf. Wilcox 1992), fingerspelling tends to break normal phonological structure of signed language (Liddell & Johnson 1989; Sandler 1989; Brentari 1998). In ASL, it has been shown that fingerspellings that become lexicalized tend to be reduced over time, with handshapes, orientation and movement changing from their citation forms (Battison 1978). Although we know of many heavily reduced lexicalized signs originating as fingerspellings in SSL, no study has investigated fingerspelling based on a phonological or phonetic analysis, let alone in terms of reduction and lexicalization.

In the SSL Corpus, all fingerspelled items are marked by the tag "@fs" (originally "@b", for the Swedish word *bokstavering* 'fingerspelling'). We extracted all single element signs (i.e. no compounds) for which the entire sign was fingerspelled from the SSL Corpus, and looked at the distribution and duration of these items.

In the first step, we investigated the mean duration of fingerspelled signs based on the number of written characters they have in their target word form. Unlike the Auslan Corpus (Johnston 2014:41), the SSL Corpus does not gloss fingerspelled items based on the realized articulation (i.e. which letters are actually spelled out), but only the target form, hence only the target form was considered in our study. This is in most cases straightforward as the target word is understood from the context even when heavily reduced, but a potential problem would be if a production were inaccurate in terms of the characters articulated (i.e. if an item is misspelled). However, as a corpus-based study, we believe the data quantity will ensure any general patterns to emerge regardless.

First of all, we looked at the general duration of fingerspellings in relation to the target word length in number of characters. As shown in Figure 3, there is a general trend towards fingerspellings of longer words to have longer duration, which is to be expected since the target words have more sequential information to articulate, i.e. should take longer to articulate. This positive relationship between fingerspelled sign type duration and word length in characters is highly significant, as shown by linear regression analysis ($\beta = 91.81$, $t(260) = 11.19$, $p < .0001$, see Figure 3). Note that since the number of sign types for the longer words (e.g. those with more than 6 characters in their target words) are counted in single digit numbers (cf. Figure 5 below), the data in the higher end on the x-axis in Figure 3 should be considered with caution.

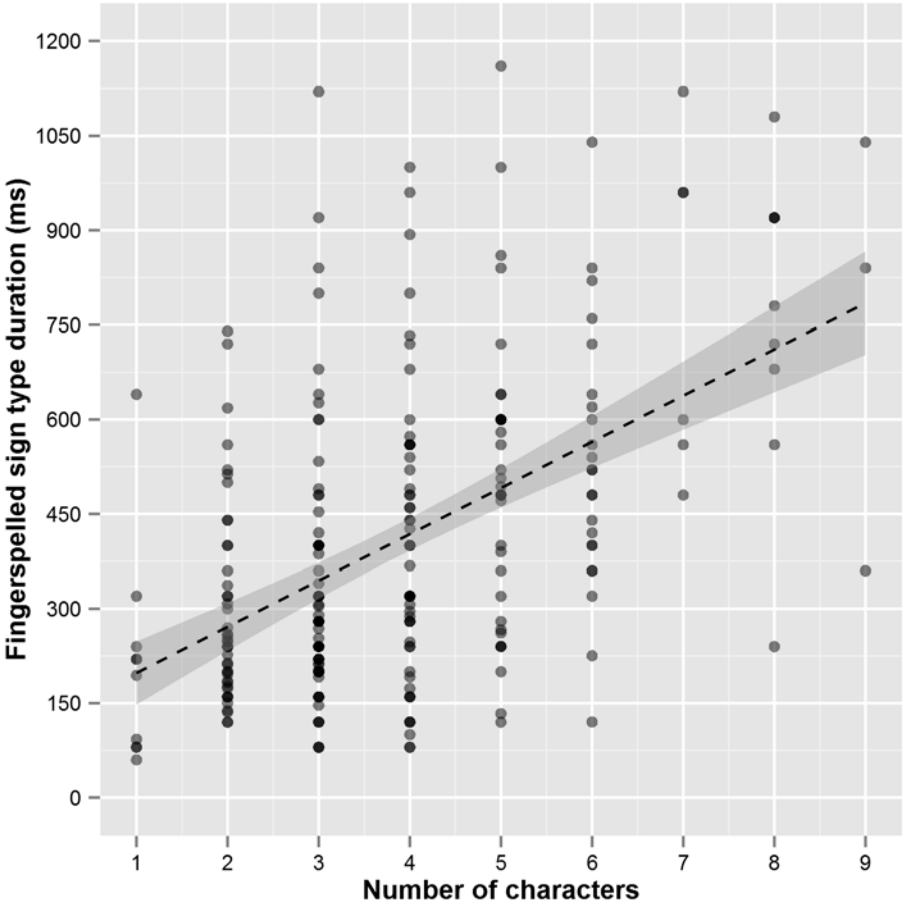


Figure 3. Mean duration of fingerspelled words by number of characters in the written target word. The regression line illustrates the fit of a linear regression model regressing mean duration against number of characters. The shaded area illustrates the 95% confidence interval of the fitted values.

In Figure 2 above, we saw that all signs (fingerspellings or not) in the SSL Corpus have predicted durations of about 100–500 ms, regardless of frequency. Comparing this with the data in Figure 3 means that fingerspellings of target words with >5 characters exceed this interval, thus making these fingerspelled signs longer than what signs in general tend to be. This would mean that fingerspelling more than five characters generally takes longer than the average sign, which is an indication of longer fingerspellings being a less good fit in SSL phonology based on duration. Unsurprisingly then, we find in Figure 4, in which the number of sign types grouped by character lengths that SSL favors shorter written words as fingerspellings, that the most fingerspelled sign types are found among 3- and 4-letter words. Besides the distribution of fingerspelled sign types, Figure 4 also shows the distribution of Swedish word types by word length (in written characters), as they appear in the Swedish translation tiers in the SSL Corpus.[7] As can be seen by comparing the patterns of each language, the distribution of Swedish words peaks in frequency between 5 and 7 letters, whereas the distribution of fingerspelled signs peaks between 2 and 4 letters. This is yet another indication of SSL preferentially utilizing fingerspelling for shorter words, which could, in turn, be an indication of lexicalization of fingerspelled items being easier for shorter words, as they require less reduction to fit into a preferred phonological structure.

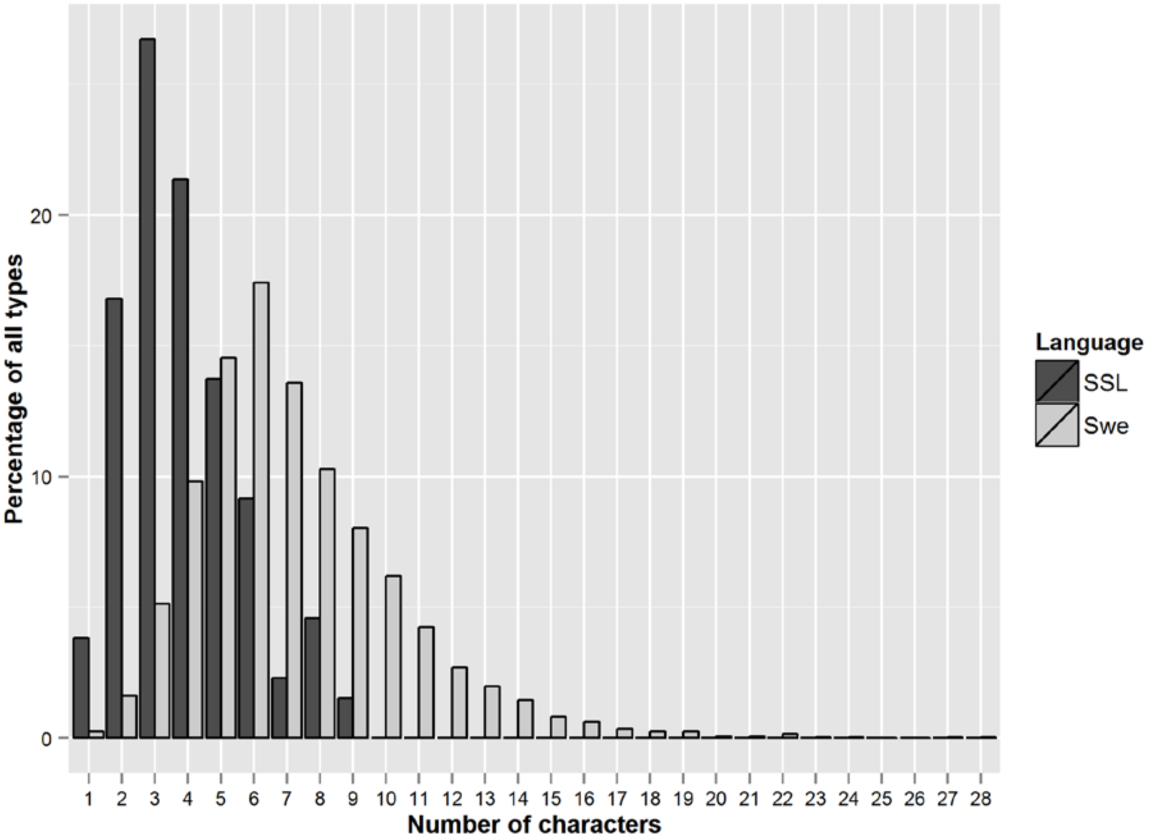


Figure 4. The relative distribution of fingerspelled sign types by number of characters in the target written word compared to the distribution of Swedish words in the SSL Corpus translations by number of characters in the word (written form).

The preference for shorter words to be fingerspelled in SSL may also be influenced by the fact that for Swedish—as for any language—shorter words are more frequent in terms of tokens, and as high-frequency items should tend to be more susceptible to borrowing.[8] Figure 5 shows that this may, in fact, be the case. The graph in Figure 5 plots the number of fingerspelled sign types by the number of characters in the target written words, on the one hand, and the logarithmic mean word token frequency in Swedish as estimated from the Swedish Blog Sentences corpus (Östling & Wirén 2013), on the other.[9] Whereas the number of fingerspelled sign types and number of characters show a strong negative Pearson correlation (r = –0.54), the number of fingerspelled sign types is conversely positively correlated with logarithmic mean word token frequency in Swedish (r = 0.69). Unsurprisingly, Swedish word token frequency is also negatively correlated with number of characters (r = –0.91). In sum, all three variables are highly correlated and it is thus hard to tease apart whether fingerspellings are preferred for shorter words or simply for high-frequency items.

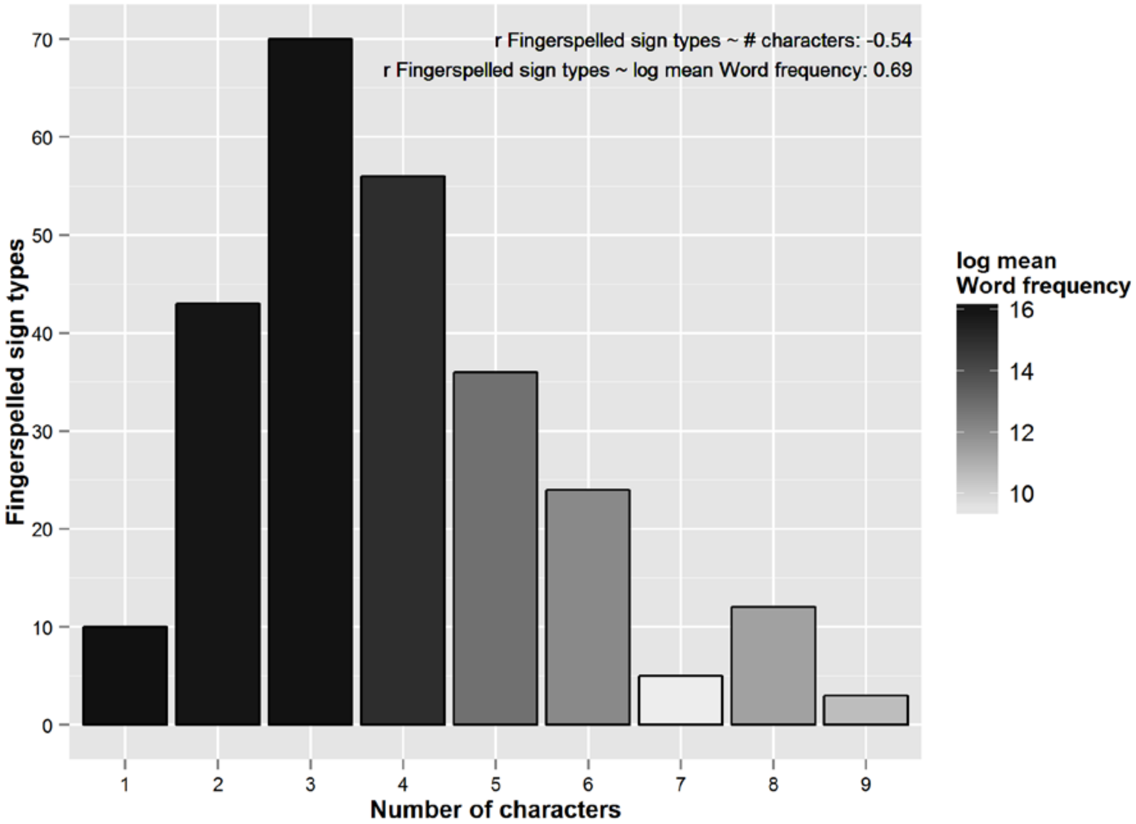


Figure 5. The number of fingerspelled sign types by number of characters in the target written word and logarithmic mean word token frequency in Swedish.

Another comparison with the data in Figure 2 can be made when looking at the frequency and duration of fingerspelled signs in the SSL Corpus. As Figure 6 shows, there is again a negative relationship between fingerspelled mean sign duration and sign frequency, according to which fingerspelled signs tend to have shorter duration if they occur frequently. A linear regression model of the negative relationship between fingerspelled sign frequency (logarithmic) and duration found that the relationship is significant ($\beta = -74.76$, $t(260) = -4.87$, $p < .0001$), showing that also for fingerspelled signs, frequency and duration are correlated.

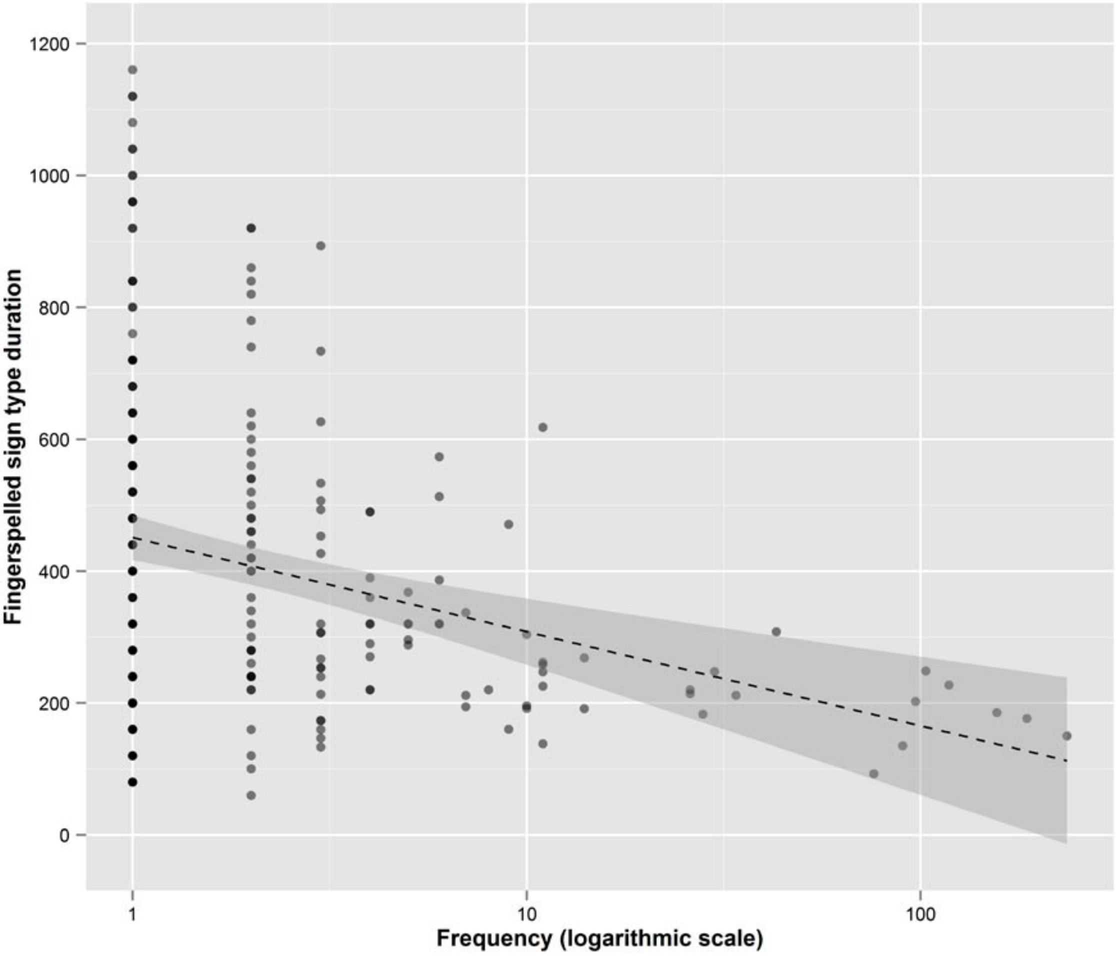


Figure 6. Fingerspelled sign type duration as a function of log frequency in the SSL corpus. The regression line shows the fit of a linear regression model of the relationship between log sign frequency and sign duration. The shaded area illustrates the 95% confidence interval of the fitted values.

We argue that the data in Figures 3, 4, and 6 should be seen as different aspects of a similar idea, and a possible relationship between these aspects could be rendered as: a) shorter words are fingerspelled faster as they require fewer elements to be articulated; b) shorter (which also tend to be more frequent) words are favored as lexicalized fingerspellings as they fit the phonological structure of simplex signs better; c) since shorter words are favored as lexicalized fingerspellings, they occur more frequently as fingerspellings than do longer words, and are therefore more likely to be reduced (i.e. have shorter duration). As shown in Figure 5, it is however difficult to distinguish word length from frequency in the source language for fingerspellings (i.e. normally Swedish), and thus both word length and frequency may play a part in which items are borrowed as fingerspellings. To further investigate the interaction between fingerspelling, word length, phonological structure, and articulatory economy, it would be necessary to gloss all fingerspellings in the SSL Corpus with the realized output of each occurrence (i.e. which letters are actually articulated), and possibly to distinguish phonetic forms that are the result of general assimilation rules from those that are the result of phonological changes connected to lexicalization (cf. Battison 1978).

6. Parts of speech and duration

A quite unique feature of the SSL Corpus is the fact that it contains part of speech tags on the level of sign types (see Östling, Börstell & Wallin 2015). By using these tags, we can generalize on our distribution and duration study by grouping signs by their parts of speech. The part of speech categories used are initially based on the categories given for SSL by Ahlgren and Bergman (2006), though extended by the addition of some categories deemed necessary for the sake of acknowledging the specifics of SSL grammar without assuming a 1-to-1 correspondence to Swedish (cf. Haspelmath 2007; Schwager & Zeshan 2008).

As frequency-based accounts on language in general state that high-frequency items are more reduced than are low-frequency items (Zipf 1935; Zipf 1949), and functional (i.e. grammatical) type words are reduced as high-frequency items having been subject to grammaticalization and hence form reduction (e.g. Bybee 2003), we expect function type parts of speech to have shorter durations than content parts of speech. We take the most prototypical content type parts of speech (nouns, verbs, and adjectives) and compare them to the most prototypical function type parts of speech (pronouns/points, prepositions, and conjunctions). Note that the category 'points' subsumes non-first person index points (previously labeled as pronouns), which is why we have included it into function type category. The parts of speech not clearly assigned to either content or function types were excluded from this statistical comparison, but are still provided in Table 10 below, in which each part of speech is accounted for in terms of type/token distribution and mean duration (mean of sign type means).[10]

Table 10 shows that the expected pattern is, in fact, visible, with content parts of speech

Table 10. PoS types and mean duration

PoS type	PoS	Types	Tokens	Mean duration (ms)
Content	Noun	1,670	7,833	466
	Verb	969	10,358	371
	Adjective	283	2,182	338
	Total	**2,922**	**20,373**	**422**
Function	Pronoun	70	5,401	260
	Pointing	35	861	331
	Preposition	34	1,368	215
	Conjunction	23	3,458	184
	Total	**162**	**11,088**	**255**
Other	Gesture	32	1,730	412
	Verb (depicting)	405	1,493	534
	Verb (stative)	90	406	351
	Verb (CA)	102	215	659
	Verb (locative)	2	182	213
	Adverb	232	4,971	308
	Numeral	162	1,204	534
	Interjection	32	1,384	360
	Nominal classifier	25	171	385
	Buoy	24	708	729
	Unlabeled	420	861	519
	Total	**1,526**	**13,325**	**489**

averaging at 422 ms in duration, while function parts of speech average at 255 ms. This difference is unsurprisingly highly significant ($t(3082) = 8.08, p < .0001$).

As several previous studies on part of speech distinctions in sign languages have considered the formal differences between nouns and verbs, one can note that nouns in our data have longer duration on average than do verbs. In Hunger's (2006) investigation of related noun-verb pairs (i.e. pairs of nouns and verbs that share aspects of form and meaning) in Austrian Sign Language, she found that verbs have substantially longer durations than nouns, but notes that it is unclear if this formal difference can be generalized to unrelated nouns and verbs. Judging by our results, such a formal difference is not found for SSL—rather the opposite—but it should be noted that our analysis does not take morphology or syntax into consideration. This means that we do not compensate for complex forms such as compounding, or syntactic position (e.g. sentence-final position, known to be associated with longer duration). Going back to Tables 8 and 9 in section 5.1, we saw that many of the non-reduced compounds were nouns, whereas verbs were found mainly in the reduced category of compounds. Looking at all compounds (regardless of the type or number of elements), we find that compounding is more common among nouns (n=374) than verbs (n=79). This could be one possible explanation for why verbs have seemingly shorter duration on average than nouns. In order to test this idea, we compared the difference in sign duration between nouns and verbs using linear regression. The regression model used included predictors for the part of speech (verb vs. noun) and for whether the sign type at hand was a compound as well as a parameter for the part of speech × compound interaction. This analysis showed that the sign duration of compounded nouns is significantly longer than that of compounded verbs ($\beta = 212.19$, $t(2635) = 6.59, p < .0001$). However, it is also found that nouns in general have longer duration than verbs, even when controlling for the effect of compounding ($\beta = 41.63$, $t(2635) = 3.9, p < .0001$). In other words, the difference in sign duration between nouns and verbs cannot be attributed to the fact that compounding is more prevalent among nouns than verbs.

Although we attempted a sub-analysis of related noun-verb pairs (e.g. CHAIR vs. SIT) in the SSL Corpus, the size of our dataset was too small to find suitable pairs for which we could investigate this further. Also, since the clausal segmentation of the SSL Corpus is only in its initial stage, we cannot use clause-position as a factor in order to control for sentence-final lengthening, which is needed in order to do even a qualitative analysis of such sign pairs. Thus, we leave these questions for future studies to investigate in more detail.

7. Age, gender, and duration

The final aim of our study was to investigate the interaction between sign duration and the age and gender of the signers in our data. This was done in order to compare the SSL data to previous studies on spoken language, for which it has been established that older speakers have lower articulation rate than younger speakers (cf. e.g. Ramig 1983; Jacewicz et al. 2009; Horton, Spieler & Shriberg 2010). Thus, we separated the sign duration data for each signer, and used the age and gender labels from the SSL Corpus metadata in order to group the data into age groups (six groups, as illustrated previously in Table 2) and gender groups (female vs. male). The purpose of this was to see if there is any variation between these groups when looking at sign duration as a variable.

The relationship between mean sign duration per signer on the one hand, and the age and gender of that signer on the other, is illustrated in Figure 7. Sign durations that were higher than 3 standard deviations from the grand mean sign duration (>1231 ms) were excluded before the means were calculated.

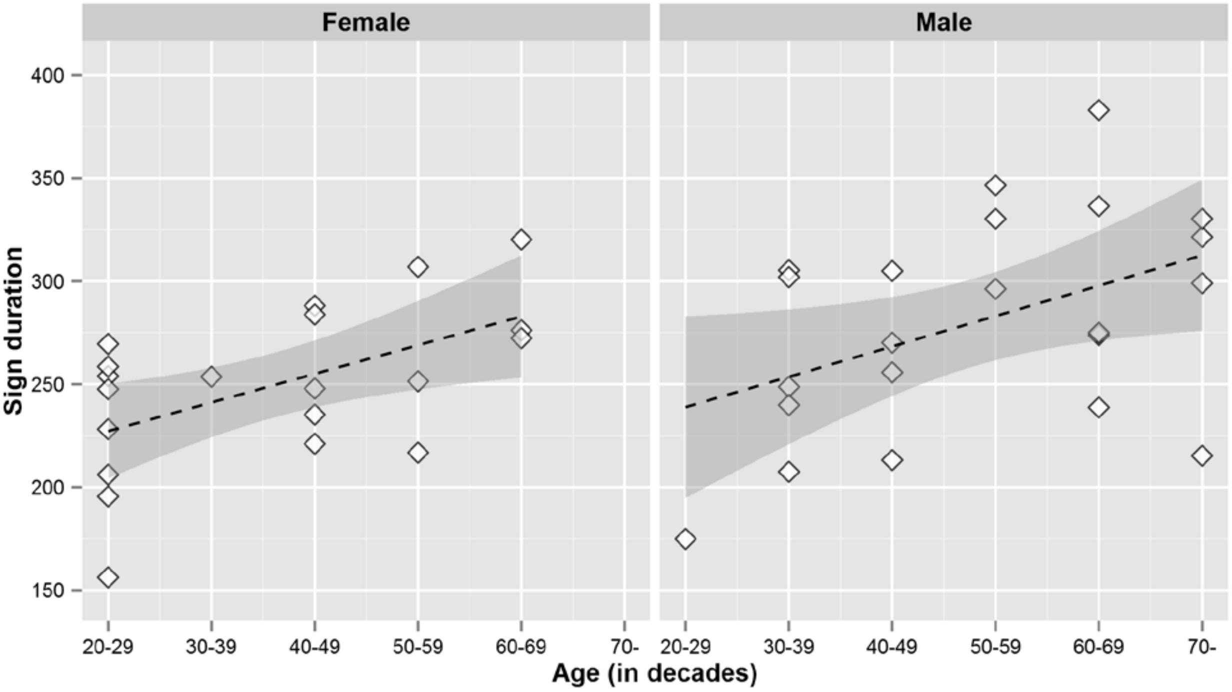


Figure 7. Mean sign duration per signer as a function of the age and gender of the signer. Sign durations above 3 SDs from the grand mean sign duration (>1231 ms) were excluded from the data. The shaded areas illustrate the 95% confidence intervals of the fitted values.

The slopes of the regression lines in Figure 7 indicate that sign durations tend to increase by age for both men and women. Men also appear to sign somewhat slower than women. In order to determine whether these effects are significant, we conducted linear mixed effects modeling to the data set. Mixed effects modeling allows for testing these effects while at the same time controlling for differences in duration between signers and signs (cf. e.g. Gelman & Hill 2006). It is a special type of the general linear model that allows for the inclusion of random effects, such as, for instance, random differences between individual speakers or individual signs. Mixed effects models can therefore control for differences in sign duration between individual signers and individual signs. By doing so, it is possible to rule out that any observed effect is not an artifact of a bias in the distribution of signs or signers across age groups or genders—for example, the somewhat longer sign durations of males in comparison to females might be due to the male signers in the corpus being more prone to use signs with longer durations than females, rather than a gender difference *per se*.

Analyses were conducted in the statistical language R. Degrees of freedom for the calculation of *p*-values was estimated using Welch-Satterthwaite approximation, as implemented in the lmerTest() package (Kuznetsova, Brockhoff & Christensen 2014). For the analysis we excluded data points for signs that were produced by less than four signers. Sign durations that deviated by more than 3 standard deviations from the overall mean duration were also excluded. Due to the rather low number of signers (n=42) included in the data set, the six age categories shown in Table 2 were used. The model includes a fixed effect for age, gender, as well as age × gender interaction. It also includes random intercepts for signer and sign, respectively, thereby controlling for duration differences between signers and signs. The model shows that the general increase in sign duration by age across both males and females is significant ($\beta = 12.8$, $t(38.36) = 3.85$, $p < .001$). When controlling for differences in sign durations between signs and signers, however, sign durations as well as the increase in sign durations by age do not differ significantly between males and females. In other words, the analysis indicates that signers generally get somewhat slower by age, but that male and female signers on average sign equally fast and are equally affected by age. An age difference of 10 years is generally associated with a difference in sign duration of 12.8 ms.

Our findings are the first to show that articulation rate in a sign language is correlated with signer age based on corpus data. It is perhaps unsurprising that age and mean sign duration correlate, seeing as this pattern has been found for spoken language, and it would be reasonable to think that physiological factors associated with aging (e.g. joint flexibility, motor control) could be causes of the decrease in articulation rate, as has been argued for spoken language (cf. Ramig 1983). However, there are at least two sociolinguistic factors that could be at play here. First, signers of SSL (as with most sign languages) differ with regard to their age of onset for acquiring the language. For some signers of SSL, the first exposure to the language came relatively late, for instance from when they entered the deaf school. Thus, it is possible that the older signers in the SSL Corpus are late(r) learners compared to the younger signers, which could account for many differences in language production, articulation rate being one. Unfortunately, the metadata concerning age of onset in language acquisition for the individual signers are incomplete, and thus we cannot investigate this aspect further. Second, there might be differences in articulation rate that stem from the age (or other attributes) of the addressee, such that the signing production is adapted to the receiver. However, the pairs of signers in the SSL Corpus knew each other beforehand, and the recruitment process was even based on getting one signer and asking them to bring a signing partner themselves, meaning that they were familiar with each other already (i.e. no first meetings between signers at the recording sessions). Nonetheless, seeing as our findings mirror patterns found among spoken languages, it is unsurprising that the age-duration correlation is present, although age is probably not the only factor correlating with articulation rate.

8. Conclusions

Our study of the distribution and duration of signs in the SSL Corpus is the first study to investigate sign frequency in SSL, and only the fifth study to do so in any sign language on the basis of corpus data. Our study is also unique in the sense that it makes use of a part of speech tagged corpus, thus also including parts of speech into the investigation of frequency, distribution, and duration, which has not been done for any sign language previously. Frequency phenomena have proved to be integral to the concept of grammaticalization and the emergence and development of linguistic structure on many levels. In this study, we have investigated several aspects of grammaticalization (content vs. function signs) and lexicalization (compounding; patterns and preferences of fingerspelling) from the perspective of lexical frequency through various sub-studies, with a focus throughout the paper on duration as a correlate of frequency. We conclude this paper here by returning to our initial research questions to summarize our findings and give some suggestions about future research.

a) What are the most frequent signs and sign categories in SSL and how do the sign and sign category distributions relate to previous corpus studies on other sign languages?

Comparing our results from SSL to previous lexical frequency studies on other sign languages shows that sign distribution across sign languages is fairly similar with regard to which items tend to occur with high frequency. However, there are differences when comparing the sign language results to spoken languages. For instance, as noted by previous lexical frequency studies on sign languages, grammatical items tend to be less common among the most common items when compared to spoken languages: Fenlon et al. (2014) report 22 function words in the top-100 frequent items (compared to 46 in the top-100 in the British National Corpus, for English), whereas we identify a similar number of 25 function words among the top-100 items in the SSL Corpus. The reason for this difference between spoken and signed language is explained in the previous studies

(Morford & MacFarlane 2003; McKee & Kennedy 2006; Johnston 2012; Fenlon et al. 2014) as being a result of modality effects on the grammar, which has consequences for the lexicon: since signed language is produced in space, the modality allows for spatial configurations to be expressed directly on a (classifier) verb or noun, without the need for an preposition (as would often be required in e.g. Swedish or English, and syntactic functions like subordination (Pfau, Steinbach & Herrmann 2016) and negation (Zeshan 2006) may be expressed solely by simultaneous non-manuals and thus fall outside a lexical frequency count. We observe both prepositions and conjunctions among the most frequent lexical items in the SSL Corpus, but note that these are few: three prepositions and six conjunctions among the top-100 items (see Appendix). This points to a high degree of similarity across sign languages (i.e. intra-modally), but some differences between the modalities (i.e. spoken vs. signed language) in terms of lexical frequency and parts of speech.

We find that the distribution of signs across categories of sign types depend on text type, which has been shown also in previous research (see Morford & MacFarlane 2003). Notably the category of classifier signs shows a great deal of variation according to text type, such that it constitutes a substantial part of the total number of sign tokens in narrative texts (14.3%), but only about 1–2% in conversational texts. This is to be expected since classifier predicates (or, depicting verbs) are typically associated with a narrative structure. The distribution of sign categories is overall very similar across sign languages, but a specification of the balance between text types is crucial when reporting any such distributions, which is also important for the compilation of any corpus (cf. McEnery, Xiao & Tono 2006).

b) How does sign duration (sign length in time) relate to frequency of signs and their corresponding parts of speech?

One of the most well-known frequency phenomena in language is that frequency correlates with word length (Zipf 1935; Zipf 1949). This has direct consequences in the theories concerning grammaticalization, that grammatical items and structures arise from lexical through a process of frequent use and form reduction (e.g. Bybee & Scheibman 1999; Bybee & Hopper 2001a; Bybee 2002; Bybee 2007), which in turn is a cornerstone in the emergence and development of language structure. Because of this, we wanted to look at the interaction between duration and frequency of signs and their parts of speech in the SSL Corpus. We find several correlations that support previous claims regarding word frequency vs. form reduction and duration, corroborating findings from spoken language research. For instance, we find, as accounted for among spoken languages, that the higher the frequency of an item, the shorter duration it has. For lexical items, we find two high-frequency items that deviate from this tendency in the SSL Corpus: YES@fb, which is a feedback sign often repeated many times in quick succession but annotated a single occurrences; and PU@g, which is a discourse-oriented sign often functioning as either a signal for keeping the turn (filled pause) or as a marker of a possible change in turn-taking. The sign PU@g is quite interesting as it shows up in several sign languages (and as a high-frequency item in all previous lexical frequency studies). With regard to parts of speech, we find that the pattern predicted by frequency and grammaticalization research, namely that function type parts of speech should be shorter than content type parts of speech, is indeed present in the SSL Corpus, with function parts of speech having significantly shorter parts of speech than content parts of speech. We also observe the expected pattern that content parts of speech are represented by more sign types than function parts of speech, but that the latter types occur with more tokens per type on average than do the former. Thus, we observe modality-independent properties of language in terms of distribution and duration of lexical items, but also some possible modality-specific properties, such that sign languages have fewer function type items among the most high-frequency items compared to spoken languages (as identified in previous lexical frequency studies on sign languages).

c) Are there durational differences between the two types of compounds in the SSL Corpus: reduced vs. non-reduced?

In the SSL Corpus, there are two types of compounds that have been distinguished by specific annotation tags: reduced vs. non-reduced. We investigated whether there is a statistical difference between these types of compounds, and looked at some distributional differences between them. For instance, we observe that there is a statistically significant difference between the two types in terms of mean sign duration, which reflects the labels—that is, reduced compounds have shorter duration than non-reduced compounds. We also see that the non-reduced class contains more types, but fewer tokens per type, which is to be expected from frequency research and properties of lexicalization and grammaticalization: when compounding is used as a word formation strategy, it is done so by juxtaposing two pre-existing lexical items, and as the compound is used more frequently, its form is reduced, moving towards a monosyllabic form (cf. Frishberg 1979; Wallin 1982; Lepic 2015b). Seeing as novel compounds are the result of productive word formation, they pass through an "open class" of compounds in the shape of non-reduced compounds, but only enter the more exclusive class of reduced compounds after extensive usage (which is only available to a smaller number of items, according to lexical frequency research).

d) How does target word length (in written characters) and frequency affect the likelihood of using fingerspelling?

Another subtype of signs in SSL is fingerspellings, which is associated with lexical borrowing (Battison 1978), and we looked in more depth at this category of signs to see interactions between lexical frequency in both source and target language and lexicalization preferences in SSL. Fingerspelling normally disobeys the regular structure of signs by having a complex sequential structure (Wilcox 1992), and as such we predicted that shorter words tend to be favored as lexicalized fingerspellings over longer words, because shorter words fit the regular phonological structure of signed language better, and are therefore not too costly in terms of articulatory economy. We find that Swedish words with few letters are indeed preferred as fingerspelled borrowings in SSL: most fingerspelled sign types are found among short words (2–4 characters), which differs from the distribution of Swedish words, which peaks at slightly longer words in number of characters. We find that there is a statistical correlation between Swedish word length and fingerspelling, but that frequency of the Swedish words is also correlated here. This is, of course, unsurprising seeing as frequency and word length are highly correlated to begin with. However, we argue that it is reasonable to interpret our results such that word length is associated with the likelihood of not employing, but retaining, fingerspelling as a lexicalization process. That is, while any written word can be encoded as a fingerspelled sign through the use of the manual alphabet in SSL, shorter words are less costly in terms of articulatory economy and thus less likely to be substituted for a non-fingerspelled form. This idea could be investigated further in future studies by observing new concepts entering in SSL as fingerspelled borrowings, and how often these fingerspellings are replaced by other non-fingerspelled sign forms as the concept is established and occurring more frequently.

e) Are there statistically significant differences in sign duration between older vs. younger, and female vs. male signers?

The general focus of this paper has been on the duration of sign across all signers, but we decided to look specifically at possible differences between subcategorizations of the signers in the SSL Corpus. Based on the available information in the metadata files, we had the opportunity to extract age groups (n=6) and gender (female vs. male). Though age- and gender-related differences have been found in spoken languages, for instance that older speakers have lower articulation rate than younger speakers, and that there may be differences between genders (e.g. Ramig 1983; Jacewicz et al. 2009; Horton, Spieler & Shriberg 2010), no such investigation has previously been done on sign language corpus data. Since we had this type of signer metadata available, we decided to look at age group and gender categorizations of signers in the SSL Corpus with regard to duration. Here, we found a significant difference between age groups in the duration of signs, with the positive correlation that an increase in age is associated with a longer duration of signs. This could possibly be attributed to physiological effects of aging, as has been argued for spoken languages (for signed language this could be associated with muscle capacity, limb movement, and join flexibility), but we are missing important metadata information about the age of onset, which could also be a factor influencing the articulation rate of signing. The age–duration correlation is found for both the female and male group, but there is, however, no significant difference in mean sign duration between the female and male groups. A topic that could be pursued in future studies using sign language corpora is in which ways age, gender, age of onset for language acquisition, relationship with addressee, and the text type affect articulation rate. At least the latter could possibly be done by using the SSL Corpus data when the basic annotation work is complete and there is a better balance across signers and text types.

This paper has investigated several aspects of sign frequency and duration from perspectives related to core properties of language structure, mainly grammaticalization and lexicalization. Many of our findings corroborate basic frequency phenomena, either for language in general (modality-independent) or specifically for signed language (modality-specific). However, the investigation of duration in relation to frequency has not been done previously for any other sign language, and our sub-studies lexicalization preferences and the correlation between age and articulation rate are also novel within sign language research. Thus, there is much to gain from using time-aligned, annotated corpus data when investigating basic linguistic structure, and the SSL Corpus has been a fruitful resource for investigating both sign distribution and variations within and across signers with regard to the frequency and duration of signs. Further expansions of the SSL Corpus, which is already today the most elaborate collection of SSL data, will entail more data, and thus ensure better a foundation for future studies.

9. References

Ahlgren, Inger & Brita Bergman. 2006. Det svenska teckenspråket. *Teckenspråk och teckenspråkiga: Kunskaps- och forskningsöversikt (SOU 2006:29)*, 11–70. Statens offentliga utredningar.

Battison, Robbin. 1978. *Lexical borrowing in American Sign Language*. Silver Spring, MD: Linstok Press.

Bellugi, Ursula & Susan D. Fischer. 1972. A comparison of sign language and spoken language. *Cognition* 1(2-3). 173–200.

Bergman, Brita & Östen Dahl. 1994. Ideophones in Sign Language? The place of reduplication in the tense-aspect system of Swedish Sign Language. In Carl Bache, Hans Basbøll & Carl-Erik Lindberg (eds.), *Tense, Aspect and Action. Empirical and Theoretical Contributions to Language Typology*, 397–422. Mouton de Gruyter.

Brentari, Diane. 1998. *A prosodic model of sign language phonology*. Cambridge, MA: MIT Press.

Brentari, Diane & Carol A. Padden. 2001. Native and foreign vocabulary in American Sign Language: A lexicon with multiple origins. *Foreign vocabulary in sign languages: A cross-*

linguistic investigation of word formation, 87–120.

Browman, Catherine P. & Louis Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49(3-4). 155–180. doi:10.1159/000261913.

Bybee, Joan L. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14(3). 261–290. doi:10.1017/S0954394502143018.

Bybee, Joan L. 2003. Mechanisms of change in grammaticalization: The role of frequency. In Brian D. Joseph & Richard D. Janda (eds.), *The Handbook of Historical Linguistics*, 602–623. Oxford: Blackwell.

Bybee, Joan L. 2007. *Frequency of use and the organization of language*. New York, NY: Oxford University Press.

Bybee, Joan L. & Paul Hopper (eds.). 2001a. *Frequency and the emergence of linguistic structure*. Amsterdam/Philadelphia, PA: John Benjamins.

Bybee, Joan L. & Paul Hopper. 2001b. Introduction to frequency and the emergence of linguistic structure. *Frequency and the Emergence of Linguistic Structure*. 1–24. doi:10.1075/tsl.45.

Bybee, Joan L., Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago, IL: University of Chicago Press.

Bybee, Joan L. & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics* 37(4). 575–596. doi:10.1515/ling.37.4.575.

Byrd, Dani. 1994. Relations of sex and dialect to reduction. *Speech Communication* 15(1-2). 39–54. doi:10.1016/0167-6393(94)90039-6.

Börstell, Carl, Johanna Mesch & Lars Wallin. 2014. Segmenting the Swedish Sign Language Corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In Onno Crasborn, Eleni Efthimiou, Evita Fotinea, Thomas Hanke, Jette Kristoffersen & Johanna Mesch (eds.), *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, 7–10. Paris: European Language Resources Association (ELRA).

Börstell, Carl, Mats Wirén, Johanna Mesch & Moa Gärdenfors. 2016. Towards an annotation of syntactic structure in Swedish Sign Language. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen & Johanna Mesch (eds.), *Proceedings of the 7th workshop on the Representation and Processing of Sign Languages: Corpus Mining*, 19–24. Paris: European Language Resources Association (ELRA).

Cormier, Kearsy, Jordan Fenlon, Ramas Rentelis & Adam Schembri. 2011. Lexical frequency in British Sign Language conversation: A corpus-based approach. In Peter K. Austin, Oliver Bond, Lutz Marten & David Nathan (eds.), *Proceedings of the Conference on Language Documentation and Linguistic Theory 3*, 81–91. London: School of Oriental and African Studies.

Cormier, Kearsy, Adam Schembri & Martha E. Tyrone. 2008. One hand or two?: Nativisation of fingerspelling in ASL and BANZSL. *Sign Language and Linguistics* 11(1). 3–44. doi:10.1075/sl.

Crasborn, Onno. 2015. Transcription and Notation Methods. In Eleni Orfanidou, Bencie Woll & Gary Morgan (eds.), *Research Methods in Sign Language Studies*, 74–88. Chichester: John Wiley & Sons, Ltd. doi:10.1002/9781118346013.ch5.

Crasborn, Onno, Richard Bank, Inge Zwitserlood, Els van der Kooij, Anne de Meijer & Anna Sáfár. 2015. *Annotation conventions for the Corpus NGT*. Nijmegen.

Crasborn, Onno, Johanna Mesch, Dafydd Waters, Els van der Kooij, Bencie Woll & Brita Bergman. 2007. Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics* 12(4). 535–562.

Crasborn, Onno & Inge Zwitserlood. 2008. The Corpus NGT: An online corpus for professionals and laymen. In Onno Crasborn, Thomas Hanke, Eleni Efthimiou, Inge Zwitserlood & Ernst Thoutenhoofd (eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, 44–49. Paris: ELDA.

Crasborn, Onno, Inge Zwitserlood, Els van der Kooij, Anna Sáfár, Johan Ros, Ellen Nauta, Merel van Zuilen, Lianne van Dijken, Frouke van Winsum & Anne de Meijer. 2015. Corpus NGT gloss annotations (3rd release). Centre for Language Studies, Radboud Universiteit, Nijmegen. doi:10.13140/RG.2.1.2303.7525.

Crasborn, Onno, Inge Zwitserlood & Johan Ros. 2008. Het Corpus NGT: A digital open access corpus of movies and annotations of Sign Language of the Netherlands. Nijmegen: Centre for Language Studies, Radboud Universiteit.

Diessel, Holger. 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25(2). 104–123. doi:10.1016/j.newideapsych.2007.02.002.

Ellis, Nick 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. 143–188.

Emmorey, Karen (ed.). 2003. *Perspectives on classifier constructions in sign languages*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engberg-Pedersen, Elisabeth. 2002. Gestures in signing: The presentation gesture in Danish Sign Language. In Rolf Schulmeister & Heimo Reinitzer (eds.), *Progress in sign language research: In honor of Siegmund Prillwitz*, 143–162. Hamburg: Signum.

Fenlon, Jordan, Adam Schembri, Ramas Rentelis & Kearsy Cormier. 2013. Variation in handshape and orientation in British Sign Language: The case of the "1" hand configuration. *Language and Communication* 33(1). 69–91. doi:10.1016/j.langcom.2012.09.001.

Fenlon, Jordan, Adam Schembri, Ramas Rentelis, David Vinson & Kearsy Cormier. 2014. Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua* 143. 187–202. doi:10.1016/j.lingua.2014.02.003.

Fischer, Susan D. & Wynne Janis. 1990. Verb sandwiches in American Sign Language. In Siegmund Prillwitz & Tomas Vollhaber (eds.), *Current trends in European sign language research*, 279–293. Hamburg: Signum Verlag.

Frishberg, Nancy. 1975. Arbitrariness and iconicity: Historical change in American Sign Language. *Language* 51(3). 696–719.

Frishberg, Nancy. 1979. Historical change: From iconic to arbitrary. In Edward S. Klima & Ursula Bellugi (eds.), *The signs of language*, 67–87. Cambridge, MA: Harvard University Press.

Frishberg, Nancy, Nini Hoiting & Dan I. Slobin. 2012. Transcription. In Roland Pfau, Markus Steinbach & Bencie Woll (eds.), *Sign language: An international handbook*, 1045–1075. Berlin/Boston, MA: De Gruyter Mouton.

Gahl, Susanne. 2008. "Time" and "Thyme" are not homophones: The effect of lemma frequency on word durations in spontaneous. *Language* 84(3). 474–496. doi:10.1371/journal.pone.0005772.

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press. doi:10.2277/0521867061.

Grosjean, François. 1979. A study of timing in a manual and a spoken language: American sign language and English. *Journal of psycholinguistic research* 8(4). 379–405.

Haspelmath, Martin. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology* 11(1). 119–132.

Horton, William S., Daniel H. Spieler & Elizabeth Shriberg. 2010. A corpus analysis of patterns of age-related change in conversational speech. *Psychology & Aging* 25(3). 708–713. doi:10.1037/a0019424.A.

Hunger, Barbara. 2006. Noun/Verb Pairs in Austrian Sign Language (ÖGS). *Sign Language & Linguistics* 9(1/2). 71–94. doi:10.1075/sll.9.1.06hun.

Jacewicz, Ewa, Robert A. Fox, Caitlin O'Neill & Joseph Salmons. 2009. Articulation rate across dialect, age, and gender. *Language Variation and Change* 21(2). 233–256. doi:10.1017/S0954394509990093.

Janzen, Terry. 2012. Lexicalization and grammaticalization. In Roland Pfau, Markus Steinbach & Bencie Woll (eds.), *Sign language: An international handbook*, 816–841. Berlin/Boston, MA: De Gruyter Mouton.

Jerde, Thomas E., John F. Soechting & Martha Flanders. 2003. Coarticulation in fluent fingerspelling. *The Journal of Neuroscience*.

Johnston, Trevor. 2003. BSL, Auslan and NZSL: Three signed languages or one? In Anne Baker, Beppie Bogaerde van de & Onno Crasborn (eds.), *Cross-linguistic perspectives in sign language research: Selected papers from TISLR 2000*, 47–69. Hamburg: Signum.

Johnston, Trevor. 2008. The Auslan Archive and Corpus. In David Nathan (ed.), *The endangered languages archive*. London: Hans Rausing Endangered Languages Documentation Project, School of Oriental and African Studies, University of London.

Johnston, Trevor. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15(1). 106–131. doi:10.1075/ijcl.15.1.05joh.

Johnston, Trevor. 2012. Lexical frequency in sign languages. *Journal of Deaf Studies and Deaf Education* 17(2). 163–193. doi:10.1093/deafed/enr036.

Johnston, Trevor. 2014. *Auslan Corpus Annotation Guidelines*. Sydney.

Johnston, Trevor, Donovan Cresdee, Adam Schembri & Bencie Woll. 2015. FINISH variation and grammaticalization in a signed language: How far down this well-trodden pathway is Auslan (Australian Sign Language)? *Language Variation and Change* 27. 117–155. doi:10.1017/S0954394514000209.

Johnston, Trevor & Adam Schembri. 1999. On defining lexeme in a signed language. *Sign language & linguistics* 2(2). 115–185. doi:10.1075/sll.2.2.03joh.

Johnston, Trevor & Adam Schembri. 2007. *Australian Sign Language: An introduction to sign language linguistics*. Cambridge: Cambridge University Press.

Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 229–254. Amsterdam/Philadelphia, PA: Benjamins.

Kuznetsova, Alexandra, Per Bruun Brockhoff & Rune Haubo Bojesen Christensen. 2014. lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). *R package version 2.0-6*.

Kyle, Jim G. & Bencie Woll. 1985. *Sign language: The study of deaf people and their language*. Cambridge: Cambridge University Press.

Lahey, Mybeth & Mirjam Ernestus. 2014. Pronunciation variation in infant-directed speech: Phonetic reduction of two highly frequent words. *Language Learning and Development* 10(4). 308–327. doi:10.1080/15475441.2013.860813.

Lepic, Ryan. 2015a. Motivation in Morphology: Lexical patterns in ASL and English. University of California, San Diego (PhD dissertation), Retrieved from http://escholarship.org/uc/item/5c38w519.

Lepic, Ryan. 2015b. The Great ASL Compound Hoax. In Aubrey Healey, Ricardo Napoleão de Souza, Pavlína Pešková & Moses Allen (eds.), *Proceedings of the 11th High Desert Linguistics Society Conference*, vol. 11, 227–250. Albuquerque, NM: University of New Mexico.

Liddell, Scott K & Robert E. Johnson. 1989. American Sign Language: The phonological base. *Sign Language Studies* 64. 195–278.

Liddell, Scott K. 1978. Nonmanual signals and relative clauses in American Sign Language. In Patricia Siple (ed.), *Understanding language through sign language research*, 59–90. New York, NY: Academic Press.

Liddell, Scott K. 2003. *Grammar, gesture, and meaning in American Sign Language*. Cambridge: Cambridge University Press.

Liddell, Scott K. & Robert E. Johnson. 1986. American Sign Language compound formation processes, lexicalization, and phonological remnants. *Natural Language and Linguistic Theory* 4. 445–513. doi:10.1007/BF00134470.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. Abingdon: Routledge.

McKee, David & Graeme Kennedy. 2006. The distribution of signs in New Zealand Sign Language. *Sign Language Studies* 6(4). 372–391.

McKee, Rachel Locker & Sophia Wallingford. 2011. "So, well, whatever": Discourse functions of palm-up in New Zealand Sign Language. *Sign Language & Linguistics* 14(2). 213–247. doi:10.1075/sll.14.2.01mck.

Mesch, Johanna. 2013. *Korpus för det svenska teckenspråket: Att bygga en (långsiktig) korpusdatabas*. Stockholm: Department of Linguistics, Stockholm University.

Mesch, Johanna, Maya Rohdell & Lars Wallin. 2014. Annotated files for the Swedish Sign Language Corpus. Version 2. Sign Language Section, Department of Linguistics, Stockholm University.

Mesch, Johanna & Lars Wallin. 2015. Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics* 20(1). 103–121. doi:10.1075/ijcl.20.1.05mes.

Mesch, Johanna, Lars Wallin & Thomas Björkstrand. 2012. Sign Language Resources in Sweden: Dictionary and Corpus. In Onno Crasborn, Eleni Efthimiou, Evita Fotinea, Thomas Hanke, Jette Kristoffersen & Johanna Mesch (eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]*, 127–130. Paris: European Language Resources Association (ELRA).

Mesch, Johanna, Lars Wallin, Anna-Lena Nilsson & Brita Bergman. 2012. Dataset. Swedish Sign Language Corpus project 2009--2011 (version 1). Sign Language Section, Department of Linguistics, Stockholm University.

Miller, Christopher. 2006. Sign language: Transcription, notation, and writing. In Keith Brown (ed.), *Encyclopedia of Language & Linguistics*, 353–354. Oxford: Elsevier.

Morford, Jill P. & James MacFarlane. 2003. Frequency characteristics of American Sign Language. *Sign Language Studies* 3(2). 213–226.

Pfau, Roland & Markus Steinbach. 2006. Grammaticalization in sign languages. In Bernd Heine & Heiko Narrog (eds.), *Handbook of Grammaticalization*, 683–695. Oxford: Oxford University

Press.

Pfau, Roland, Markus Steinbach & Annika Herrmann (eds.). 2016. *A matter of complexity: Subordination in sign languages*. Boston, MA/Berlin & Preston: De Gruyter Mouton & Ishara Press.

Pierrehumbert, Janet B. 2002. Word-specific phonetics. In Aditi Lahiri (ed.), *Laboratory Phonology VII*, 101–140. Berlin/New York, NY: Walter de Gruyter.

Quinto-Pozos, David. 2010. Rates of fingerspelling in American Sign Language. Poster presented at Theoretical Issues in Sign Language Linguistics 10 (TISLR10), Purdue University, West Lafayette, Indiana.

R Core Team. 2014. R: A language and environment for statistical computing.

Ramig, Lorraine A. 1983. Effects of physiological aging on speaking and reading rates. *Journal of Communication Disorders* 16(3). 217–226. doi:10.1016/0021-9924(83)90035-7.

Ryttervik, Magnus. 2015. Gesten PU@g i svenskt teckenspråk – en studie i dess form och funktion. Department of Linguistics, Stockholm University (MA thesis).

Sandler, Wendy. 1989. *Phonological representation of the sign: Linearity and nonlinearity in sign language phonology*. Dordrecht: Foris.

Sandler, Wendy. 2012. The phonological organization of sign languages. *Language and Linguistics Compass* 6(3). 162–182. doi:10.1002/lnc3.326.

Schembri, Adam & Onno Crasborn. 2010. Issues in creating annotation standards for sign language description. In Phillipe Dreuw, Eleni Efthimiou, Thomas Hanke, Trevor Johnston, Gregorio Martinéz Ruiz & Adam Schembri (eds.), *Proceedings of the 4th Workshop of the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 212–216. Paris: European Language Resources Association (ELRA).

Schembri, Adam, Jordan Fenlon, Ramas Rentelis & Kearsy Cormier. 2014. British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2014 (Second edition). London: University College London.

Schembri, Adam, David McKee, Rachel McKee, Sara Pivac, Trevor Johnston & Della Goswell. 2009. Phonological variation and change in Australian and New Zealand Sign Languages: The location variable. *Language Variation and Change* 21(02). 193. doi:10.1017/S0954394509990081.

Schwager, Waldemar & Ulrike Zeshan. 2008. Word classes in sign languages: Criteria and classifications. *Studies in Language* 32(3). 509–545. doi:10.1075/sl.32.3.03sch.

Sigurd, Bengt, Mats Eeg-Olofsson & Joost van de Weijer. 2004. Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica* 58(1). 37–52. doi:10.1111/j.0039-3193.2004.00109.x.

Sjons, Johan. 2013. Automatic induction of word classes in Swedish Sign Language. Stockholm

University (MA thesis).

Tkachman, Oksana & Wendy Sandler. 2013. The noun–verb distinction in two young sign languages. *Gesture* 13(3). 253–286. doi:10.1075/gest.13.3.02tka.

Tyrone, Martha E. & Claude E. Mauk. 2010. Sign Lowering and Phonetic Reduction in American Sign Language. *Journal of Phonetics* 38(2). 317–328. doi:10.1016/j.wocn.2010.02.003.

Wallin, Lars. 1982. Sammansatta tecken i svenska teckenspråket. Forskning om teckenspråk VIII. (Forskning Om Teckenspråk). Stockholm: Department of Linguistics, Stockholm University.

Wallin, Lars & Johanna Mesch. 2014. *Annoteringskonventioner för teckenspråkstexter. Version 5. [Annotation guidelines for sign language texts]. Projektet Korpus för det svenska teckenspråket 2009-2011 (version 1).*

Wallin, Lars & Johanna Mesch. 2015. Annoteringskonventioner för teckenspråkstexter. Forskning om teckenspråk (FoT) XXIV, Department of Linguistics, Stockholm University.

Vermeerbergen, Myriam, Lorraine Leeson & Onno Crasborn (eds.). 2007. *Simultaneity in signed languages: Form and function*. Amsterdam/Philadelphia, PA: John Benjamins.

Wilbur, Ronnie B. 1999. Stress in ASL: Empirical evidence and linguistic issues. *Language and Speech* 42(2-3). 229–250. doi:10.1177/00238309990420020501.

Wilbur, Ronnie B. & Susan Bobbitt Nolen. 1986. The duration of syllables in American Sign Language. *Language and Speech* 29(3). 263–280.

Wilcox, Sherman. 1992. *The phonetics of fingerspelling*. Amsterdam/Philadelphia, PA: John Benjamins.

Wilkinson, Erin. 2016. Finding frequency effects in the usage of NOT collocations in American Sign Language. *Sign Language & Linguistics* 19(1). 82–123.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.

Woodward, Jr., James C. 1976. Signs of change: Historical variation in American Sign Language. *Sign Language Studies* 10. 81–94.

Wright, Charles E. 1979. Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition* 7(6). 411–419. doi:10.3758/BF03198257.

Zakia, Richard D. & Ralph Norman Haber. 1971. Sequential letter and word recognition in deaf and hearing subjects. *Perception & Psychophysics* 9(1). 110–114. doi:10.3758/BF03213041.

Zeshan, Ulrike (ed.). 2006. *Interrogative and negative constructions in sign languages*. Nijmegen: Ishara Press.

Zipf, George K. 1935. *The psycho-biology of language: An introduction to dynamic philology*. New York, NY: Houghton Mifflin.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.

Östling, Robert, Carl Börstell & Lars Wallin. 2015. Enriching the Swedish Sign Language Corpus with part of speech tags using joint Bayesian word alignment and annotation transfer. In Beáta Megyesi (ed.), *Proceedings of the 20th Nordic Conference on Computational Linguistics (NODALIDA 2015), NEALT Proceedings Series 23*, 263–268. Vilnius: ACL Anthology.

Östling, Robert & Mats Wirén. 2013. Compounding in a Swedish Blog Corpus. In Laura Álvarez López, Charlotta Seiler Brylla & Philip Shaw (eds.), *Computer mediated discourse across languages*, 45–63. Stockholm: Acta Universitatis Stockholmiensis.

Appendix: The 300 most frequent types in the SSL Corpus (67% of all tokens)

Rank	Gloss	Part of speech	Tokens	Mean duration (ms)
1	PRO1	Pronoun	3361	104
2	POINT	Point	2425	168
3	PU@g	Gesture	1397	267
4	YES@fs@fb	Interjection	782	392
5	POINT>person	Point	488	149
6	BUT	Conjunction	363	161
7	POSS1	Pronoun	313	110
8	ONE	Numeral	304	134
9	TO-BE	Verb	299	126
10	DEAF	Noun	287	219
11	SO-TO-SPEAK	Adverb	282	270
12	PI ('very, really')	Adverb	265	162
13	HAVE	Verb	260	133
14	HOW	Adverb	257	176
15	LATER	Adverb	241	196
16	THEN@fs	Conjunction	235	149
17	PERF	Verb	221	146
18	NOT	Adverb	215	170
19	TO	Preposition	214	123
20	GOOD	Adjective	208	192
21	MUCH	Adverb	203	175
22	NOW@fs	Adverb	186	176
23	NEVERMIND	Gesture	185	282
24	WHAT	Pronoun	184	132
25	BE-INSIDE	Verb (locative)	181	185
26	POINT>person/POINT	Point	177	132
27	POINT.PL	Point	169	299
28	FUN	Adjective	169	347
29	LIST-BUOY.TWO	Buoy	164	755
30	TO-MEAN	Verb	157	182
31	YES@fs	Interjection	156	185

32	SEE	Verb	155	140
33	TO-SIGN	Verb	154	425
34	SAME	Pronoun	153	210
35	REASON	Conjunction	150	159
36	REMEMBER	Verb	149	168
37	PRO1.PL	Pronoun	145	218
38	ONLY	Adverb	142	149
39	LOOK-AT	Verb	141	357
40	COME-THERE	Verb	141	311
41	BELIEVE	Verb	135	125
42	ALSO	Adverb	135	221
43	CAN	Verb	134	126
44	HEARING	Noun	134	227
45	MOTHER	Noun	134	174
46	JUST	Adverb	133	233
47	SIGN-LANGUAGE	Noun	131	411
48	SIGN.FLUENT	Verb	131	478
49	SOMEBODY	Pronoun	131	215
50	WITH	Preposition	129	136
51	A-LITTLE	Adverb	126	152
52	MORE	Adverb	122	187
53	MUST	Verb	121	158
54	LIST-BUOY.ONE	Buoy	119	577
55	YEAR@fs	Noun	118	225
56	WANT	Verb	115	152
57	TO-WORK	Verb	115	320
58	FEEL	Verb	115	160
59	OR	Conjunction	111	160
60	EXIST	Verb	110	163
61	DELIMITER	Buoy	109	781
62	PLUS	Conjunction	106	151
63	POINT-BUOY	Buoy	106	919
64	OK@fs	Interjection	103	248
65	THAT/WHICH	Conjunction	102	159
66	KNOW-NOT	Verb	100	209
67	EVERYBODY	Pronoun	98	282
68	KINDA@fs	Adverb	97	201
69	HAVE-OPINION	Verb	97	159
70	IN@fs^DAY ('today')	Adverb	97	246
71	CHILDREN	Noun	96	208
72	WILL	Verb	96	163
73	TO-BE*POINT	Verb	95	211
74	TIME	Noun	94	234
75	POSS	Pronoun	94	140
76	TWO	Numeral	94	165
77	TYPEWRITE	Verb	93	241
78	CONVERSE	Verb	92	413
79	SAID	Verb	92	122
80	EVEN-SO	Adverb	92	224
81	IF@fs	Conjunction	90	134
82	BEFORE	Adverb	90	183
83	GROW-UP	Noun	90	394
84	TO-MEET	Verb	89	240

85	FATHER	Noun	89	216
86	MANY	Pronoun	88	209
87	ON	Preposition	88	108
88	THINK	Verb	88	101
89	BOY	Noun	87	145
90	PATH-EXTENT	Verb (depicting)	87	614
91	BEGIN	Verb	86	215
92	ALWAYS	Adverb	86	164
93	DIFFERENT	Adverb	85	285
94	OBJPRO1	Pronoun	84	178
95	OTHER	Pronoun	83	185
96	BE-CORRECT	Verb	81	387
97	PARENTS	Noun	80	317
98	SCHOOL	Noun	78	337
99	FINISHED	Adjective	77	265
100	NO-WAY	Interjection	77	255
101	IN	Preposition	76	92
102	EXIST*NOT	Verb	74	298
103	DO	Verb	72	183
104	THERE	Adverb	70	221
105	AND	Conjunction	69	120
106	GO-HOME	Verb	67	353
107	IMPOSSIBLE	Adverb	67	208
108	WITH.BE ('join')	Unlabeled	66	190
109	TIME-PASSES	Verb	65	440
110	TALK	Verb	64	186
111	HUG	Verb	64	300
112	DOG	Noun	63	239
113	THREE	Numeral	63	270
114	LEARN	Verb	63	191
115	MAYBE	Adverb	63	165
116	SELF	Pronoun	61	197
117	LIVE	Verb	61	103
118	START	Verb	61	263
119	BACK	Adverb	60	342
120	RECEIVE1	Verb	60	163
121	PERF-NEG	Verb	60	180
122	LIST-BUOY.THREE	Buoy	59	1338
123	FIRST	Adverb	59	228
124	INTERPRETER	Noun	58	347
125	HAND(AA)+HANDLE@p	Verb (depicting)	57	349
126	FROM	Preposition	57	142
127	MOVE-OUT	Verb	57	246
128	WALK	Verb	57	270
129	PRO1.DUAL	Pronoun	57	293
130	POINT.BUOY	Buoy	56	697
131	CONNECT	Verb	56	214
132	OFTEN	Adverb	56	247
133	SEEMS	Verb	55	152
134	EXERCISE	Verb	55	321
135	HOUSE	Noun	55	301
136	FROG	Noun	54	254
137	HAND(GG)+HANDLE@p	Verb (depicting)	54	488

138	SIT	Verb	53	189
139	HAVE*NOT	Verb	53	240
140	SELF^CLEAR ('of course')	Interjection	52	278
141	COME-HERE	Verb	52	330
142	WRITE	Verb	52	382
143	TOO	Adverb	52	124
144	LATER(L)	Adverb	51	226
145	USE-TO	Verb	51	130
146	WHY	Adverb	51	252
147	ALL-GOOD	Adjective	50	209
148	GO-TOGETHER	Verb	50	274
149	CREATURE(Vb)+MOVE@p	Verb (depicting)	49	446
150	HERE	Adverb	49	163
151	TIME-FUTURE	Adverb	49	398
152	KNOW*PERF	Verb	49	195
153	BE-CALLED	Verb	49	220
154	YEAR	Noun	48	420
155	WHERE	Adverb	48	258
156	SAY	Verb	48	138
157	SO	Pronoun	47	194
158	RIGHT	Verb (stative)	47	249
159	AFTER	Preposition	46	238
160	SOME	Pronoun	46	247
161	MAN	Noun	45	133
162	STOP	Verb	45	246
163	AREA@cl	Nominal classifier	45	442
164	REMAIN	Adverb	45	208
165	POSS>person	Pronoun	44	132
166	LEAD-TO	Conjunction	44	367
167	OLD	Adjective	44	178
168	ÖREBRO@pn	Noun	43	333
169	ENTER(L)	Verb	43	326
170	LOOK-AT.AROUND	Verb	43	620
171	TO-SIGN	Verb	43	441
172	EXIT@fs	Verb	43	306
173	PERSON@cl	Nominal classifier	43	177
174	DIFFICULT	Adjective	42	234
175	FOUR	Numeral	42	198
176	FINE ('pretty')	Adjective	42	281
177	ZERO	Numeral	42	347
178	HAND(G)+HANDLE@p	Verb (depicting)	42	611
179	HEAR	Verb	41	167
180	DECIDE	Verb	41	225
181	NAH	Interjection	41	314
182	ENTER(N)	Verb	40	288
183	CALL-ATTENTION@g	Gesture	40	223
184	TIME-FUTURE.FROM	Adverb	40	484
185	DEAF-CLUB	Noun	40	380
186	OH-REALLY	Interjection	40	346
187	BAD	Adjective	40	270
188	TELL-STORY	Verb	39	302
189	BROTHER	Noun	39	251
190	ORDER+ONE	Numeral	39	262

191	CLASS	Noun	39	219
192	SIT(VVb)	Verb (depicting)	38	228
193	AT	Preposition	38	149
194	COMMUNICATE	Verb	38	362
195	WHO	Pronoun	38	160
196	AT-HOME	Adverb	38	362
197	FAMILY	Noun	38	387
198	KNOW*NOT	Verb	38	302
199	VÄNERSBORG@pn	Noun	38	285
200	INCREDIBLE	Adverb	37	334
201	UNDER	Preposition	37	183
202	TRY	Verb	37	244
203	IMPORTANT	Adjective	37	241
204	UNDERSTAND-NOT	Verb	37	357
205	ONLY-ON	Numeral	37	193
206	RECEIVE	Verb	36	202
207	NEVER	Adverb	36	356
208	SOMETIMES	Adverb	36	221
209	GO	Verb	36	155
210	MEMBER	Noun	36	296
211	BOTH	Pronoun	35	303
212	LANGUAGE	Noun	35	297
213	SWEDISH	Noun	35	313
214	OWN	Adjective	35	276
215	TALK	Verb	35	410
216	POINT.DUAL	Point	35	300
217	KNOW	Verb	35	153
218	GROUP	Noun	34	308
219	LAST	Adverb	34	248
220	FOR-EXAMPLE@fs	Conjunction	34	211
221	STOCKHOLM@pn	Noun	34	183
222	BORN	Verb	34	239
223	TAKE	Verb	34	206
224	SOCIALIZE	Verb	33	499
225	UNDERSTAND(L)	Verb	33	209
226	COME-HERE(L)	Verb (depicting)	33	317
227	POINT.MULTI	Point	32	529
228	WEEK	Noun	32	226
229	MOVIE	Noun	32	415
230	SIT(Vb)	Verb	31	275
231	SLEEP	Verb	31	304
232	BIG	Adjective	31	294
233	BRING	Verb	31	201
234	LEAVE-HERE	Verb	31	321
235	TO-GESTURE	Verb	30	553
236	TO-MEAN	Verb	30	240
237	PLACE	Noun	30	170
238	INTO@fs	Verb	30	247
239	TIME-OCCURRENCE	Noun	30	137
240	FUT-NEG	Verb	30	278
241	BUILD	Verb	29	312
242	WRONG	Adverb	29	200
243	LET'S-SAY	Adverb	29	118

244	SEARCH	Verb	29	391
245	POINT.REL	Point	29	427
246	SMALL-PERSON	Adjective	29	337
247	BECOME	Verb	29	144
248	FORGET	Verb	29	145
249	EXPLAIN	Verb	28	407
250	PURPOSE	Noun	28	311
251	SWEDEN@pn	Noun	28	329
252	EIGHT	Numeral	28	199
253	JOB	Noun	28	263
254	NEW@fs	Adjective	28	183
255	OUTSIDE	Adverb	28	210
256	TEN	Numeral	28	227
257	GIVE.LIFT	Verb	28	446
258	USE	Verb	28	180
259	HAPPEN	Verb	28	237
260	CALL-ATTENTION@g>person	Gesture	28	313
261	POINT(B)	Point	28	189
262	TIME-FUTURE.TO	Adverb	27	351
263	HOME	Noun	27	274
264	DAY	Noun	27	233
265	PROBLEM	Noun	27	217
266	FRIEND	Noun	27	227
267	SNOW^OLD-MAN	Noun	27	449
268	HAND(SS)+HANDLE@p	Verb (depicting)	27	445
269	EASY	Adjective	27	301
270	COME-HERE	Verb	27	292
271	HAND(SS)+HANDLE@p	Verb (depicting)	26	378
272	ENTHUSIASTIC	Verb (stative)	26	344
273	CELL-PHONE	Noun	26	192
274	CONTINUE	Verb	26	412
275	FUNNY	Adjective	26	208
276	WOMAN	Noun	26	238
277	OBJPRO	Pronoun	26	251
278	PERFECT	Adjective	26	196
279	TV@fs	Noun	26	213
280	READ-LIPS	Verb	26	370
281	LUCK@fs	Noun	26	219
282	WANT-NOT	Verb	25	320
283	ACTIVE.DO	Unlabeled	25	366
284	WIFE	Noun	25	283
285	FUNCTION	Verb	25	274
286	EVENING	Noun	25	299
287	ONE-MORE	Numeral	25	240
288	WALK(N)	Verb	25	446
289	OUTSIDE	Preposition	25	337
290	PY ('also')	Adverb	25	193
291	NEXT-IN-TURN	Verb (stative)	25	329
292	CALL-ATTENTION(Jv)	Verb	25	180
293	TEACH	Verb	25	348
294	NEXT	Adjective	25	325
295	RELATIVES	Noun	24	255
296	SIGN	Noun	24	303

297	SEE*SHOW	Verb	24	257
298	APT	Adjective	24	253
299	CLUB	Noun	24	230
300	TENNIS	Noun	23	335

Contact information

Carl Börstell*
Dept. of Linguistics
Stockholm University
SE 106 91 Stockholm
Sweden

calle@ling.su.se

Thomas Hörberg
Dept. of Linguistics
Stockholm University
SE 106 91 Stockholm
Sweden

thomas_h@ling.su.se

Robert Östling
Dept. of Linguistics/Dept of Modern Languages
Stockholm University/University of Helsinki
SE 106 91 Stockholm/FI 00014 Helsinki
Sweden/Finland

robert@ling.su.se/robert.ostling@helsinki.fi

*Corresponding

Acknowledgments

We wish to thank Johanna Mesch, Kristina Nilsson Björkenstam, Johan Sjons, Lars Wallin, Mats Wirén, and the participants at the *Work in Progress* seminar in Stockholm, 3 September, 2016, and the *12th Conference on Theoretical Issues in Sign Language Research (TISLR 12)* in Melbourne, 4–7 January, 2016, for discussions during the earlier stages of this project. We are also very grateful to Ryan Lepic and Marcin Włodarczak as well as two anonymous reviewers for their comments and suggestions on drafts of this paper.

Copyright

The final version of this manuscript is published in

Börstell, C., Hörberg, T. & Östling, R. 2016. Distribution and duration of signs and parts of speech in Swedish Sign Language. *Sign Language and Linguistics*, 16(2): 143-196. DOI: 10.1075/sll.19.2

Notes

[1] The sign gloss annotations in the SSL Corpus are only available in Swedish in the official release but have been translated into English in this paper.

[2] We thank an anonymous reviewer for pointing out the issue of fine-grained annotations in relation to video resolution.

[3] This is a consequence of the annotation conventions of the SSL Corpus, not differentiating between lexicalized and non-lexicalized fingerspelled items. There is, however, no easy way of differentiating these types of fingerspellings, as they are rather scalar than complementary, and any attempt at making a distinction between different types of fingerspelling in SSL should be made on the basis of (adapted) definitions from other sign languages (cf. Brentari & Padden 2001; Cormier, Schembri & Tyrone 2008).

[4] A list of the 300 most frequent signs in the SSL Corpus is found in the Appendix.

[5] The English sign glosses in the Corpus NGT seem to lack translated glosses for classifier constructions, which have all been labeled with Xs instead of a unique identifying gloss. However, this should not affect the results looking at the

top-20 frequent items, seeing as these types of signs rarely constitute the most frequent signs.

[6] It should be noted that historical compounds that have merged into a completely monosyllabic form with indivisible internal elements are glossed as monosyllabic and monomorphemic signs in the SSL Corpus.

[7] We use the Swedish translations of the SSL Corpus as parallel data here in order to control for lexical content. Since the sign glosses and the written translations to some extent constitute a parallel corpus, we avoid differences in lexical distribution and frequency caused by differences in textual content.

[8] We thank Ryan Lepic for bringing this point to our attention.

[9] The Swedish Blog Sentences (SBS) corpus can be found at: http://www.ling.su.se/english/nlp/corpora-and-resources/sbs. NB: The size of the SBS corpus is currently about twice the size of that given in Östling & Wirén (2013).

[10] We have deliberately excluded parts of speech like adverbs and numerals from these counts, since they contain signs that would fall into different part of speech types (i.e. content vs. function).