

# Evidence of Zipfian distributions in three sign languages

Inbal Kimchi<sup>1</sup>, Lucie Wolters<sup>1</sup>, Rose Stamp<sup>2</sup>, and Inbal Arnon<sup>1</sup>

<sup>1</sup>The Hebrew University of Jerusalem | <sup>2</sup>Bar Ilan University

One striking commonality between languages is their Zipfian distributions: A power-law distribution of word frequency. This distribution is found across languages, speech genres, and within different parts of speech. The recurrence of such distributions is thought to reflect cognitive and/or communicative pressures and to facilitate language learning. However, research on Zipfian distributions has mostly been limited to spoken languages. In this study, we ask whether Zipfian distributions are also found across signed languages, as expected if they reflect a universal property of human language. We find that sign frequencies and ranks in three sign language corpora (BSL, DGS and NGT) show a Zipfian relationship, similar to that found in spoken languages. These findings highlight the commonalities between spoken and signed languages, add to our understanding of the use of signs, and show the prevalence of Zipfian distributions across language modalities, supporting the idea that they facilitate language learning and communication.

**Keywords:** Zipfian distributions, sign language, universal properties of language

## Introduction

The languages of the world differ in many ways, but they also share commonalities. These can shed light on how our shared cognition impacts language structure. One striking commonality between languages is that word frequencies follow a Zipfian distribution (Zipf, 1949), in which there is a power-law relation between frequency and rank. This is reflected by a small number of very frequent words, forming a narrow peak in the distribution, a large number of infrequent

words, forming a long “tail”, and a non-linear decrease in frequency.<sup>1</sup> This Zipfian distribution has been found **across many languages** (Mehri & Jamaati, 2017; Piantadosi, 2014), within different parts of speech (Piantadosi, 2014), and even in child-directed speech (Lavi-Rotbain & Arnon, 2023). Importantly, most of what we know about Zipfian distributions results from research on spoken languages. In this paper, **we explore the distribution of lexical signs in three sign languages, to ask whether they are also Zipfian.** Doing so will further our understanding of the prevalence of such distributions in language and provide additional insights on the similarities between spoken and signed languages.<sup>2</sup>

Zipfian distributions are also found outside the linguistic domain – for example, in the population size of cities in the United States (Clauset et al., 2009) and the size of craters on the moon (Newman, 2005) – where they are thought to **reflect general mathematical principles** (e.g., **scale-invariance**, Chater & Brown, 1999). However, their recurrence in language, a product of the human mind, has **been argued to be driven, at least partially, by cognitive factors, reflecting properties of human communication and/or cognition**<sup>3</sup> (Christiansen & Chater, 2008; Ferrer-i-Cancho, 2016; Gibson et al., 2019; Semple et al., 2022). There are several hypotheses about the cognitive sources of Zipfian distributions in language. **The presence of Zipfian word frequency distribution has been claimed to minimize cognitive effort and facilitate fast communication (Ferrer-i-Cancho, 2016), to enable efficient communication by creating an optimal trade-off between listener and speaker effort (Ferrer-i-Cancho & Solé, 2003; Coupé et al., 2019), and to provide a form of optimal coding for the lexicon where maximal distinctions can be maintained (Ferrer-i-Cancho, 2016; Manin, 2008).**

Experimental support for the possible influence of cognitive factors on Zipfian distributions in language comes from recent studies showing that such distributions impact learnability (Bentz et al., 2017; Lavi-Rotbain & Arnon, 2022). Growing evidence suggests that exposure to Zipfian distributions facilitates learn-

1. In actual corpus data, natural languages follow a “near-Zipfian” distribution, and not a perfect Zipfian one, with **some deviation from the expected distribution** on both ends of the frequency scale. The most frequent words are not as frequent as they should be under a pure Zipfian distribution, and there is more variability than expected in the frequencies of the infrequent words (Piantadosi, 2014). However, since these prediction errors are not the focus of the current paper, and since our interest is in assessing the similarity between spoken and signed languages, we use the more common term Zipfian (and the corresponding formulae).

2. The term ‘signed language’ is used when contrasting signed languages with spoken languages. Otherwise, the term ‘sign language’ is used throughout the manuscript.

3. More broadly, the same phenomenon (power law distributions) can be driven by different pressures in different domains: the need for efficient communication is not relevant for crater size on the moon but is for the structure of the human lexicon.

ing across a range of linguistic and non-linguistic domains. Studies of word segmentation have found that children and adults show improved segmentation of an artificial language when exposed to Zipfian distributions compared to both uniform and less skewed distributions (Kurumada et al., 2013; Lavi-Rotbain & Arnon, 2019, 2020, 2022). Similar effects have been found in other domains: Grammatical categories were learned well in Zipfian distributions despite the lower frequency of some elements (Schuler et al., 2017), and both cross-situational word learning and the learning of novel argument structure was improved in Zipfian distributions compared to uniform ones (Hendrickson & Perfors, 2019; Goldbert et al. 2004). The learning of visual regularities was also facilitated in a Zipfian distribution compared to a uniform one (Lavi-Rotbain & Arnon, 2021), in line with the skewed distribution of objects that infants see in their environment (Clerkin et al. 2017; Lavi-Rotbain & Arnon, 2021).

While much work has examined the presence of Zipfian distributions in spoken languages, less attention has been devoted to investigating of similar distributions in signed languages. Signed and spoken languages share many foundational properties, including systematic structure, multiple levels of grammatical structure, and the provision of complete expressive and communicative power (Sandler & Lillo-Martin, 2001; Lillo-Martin & Gajewski, 2014; Emmorey, 2001), while also having modality-unique properties (see Brentari & Goldin-Meadow, 2017 for a discussion), such as the increased iconicity of signed languages (Perlman et al., 2018; Lillo-Martin & Gajewski, 2014; Talmy, 2001). With regard to the presence of Zipfian distributions, we hypothesize a resemblance between signed and spoken languages based on their shared communicative function and the shared cognitive architecture of speakers and signers: If the presence of Zipfian distributions in language is partially driven by their facilitative effect on learning and communication, as recent evidence would suggest, then we expect to find them in any communication system created by humans, whether spoken or signed.

While there is work on the lexical frequency of signs, only one study (Borstell, 2022, discussed in detail below) has examined the distribution of signs. The results of this study suggests that sign frequencies follow a Zipfian distribution, but leave several important questions unanswered, including whether the slope of the frequency distributions of signs is similar to that of words and whether there is a similarity in the fit of the rank/frequency distributions to Zipf's law across different signed languages. Looking at the distribution of signs will add to our understanding of the structure of sign languages and expand the empirical support for Zipfian distributions in language. In the next section we review the existing literature on lexical frequency in sign language studies.

## Lexical frequency in sign languages

Lexical frequency plays an important role in the acquisition and processing of spoken language: Frequency is a predictor of when a word is learned, and how it will be accessed and used in comprehension and production (see reviews in Diessel, 2007; Ellis, 2002 for spoken language). Consequently, lexical frequency has been estimated for many spoken languages and word frequencies are available for multiple languages via open access databases (e.g., UCREL for English, SUBTLEX-NL for Dutch and so on). **Frequency also impacts acquisition in signed languages, with earlier acquisition of more frequent signs** (Novogrodsky & Meir 2020; Caselli & Pyers, 2017; Sümer, et al., 2017). Early investigations into lexical frequency in signed languages primarily relied on subjective measures, often derived from self-reports, as opposed to the objective corpora-based frequency assessment prevalent in spoken language research. However, recent studies (including the current one) **utilize objective measures of frequency, aligning with the methods used in spoken language research, and ensuring a standardized and comparable measure across signed and spoken languages** (Smith & Hofmann, 2020; Fenlon et al., 2014a).

Despite the increasing acknowledgment of the importance of lexical frequency in sign language research, the availability of accessible datasets on the frequency of individual signs remains limited, for several reasons. First, signed language corpora are still in their infancy relative to spoken language corpora (Smith & Hofmann, 2020; Fenlon et al., 2015a; Fenlon et al., 2015b; Fenlon et al., 2014a). **There is a lack of large corpora in general**, and a lack of objective frequency data in particular (Fenlon et al., 2015a; Fenlon et al., 2014a; Smith & Hofmann, 2020). In addition, the relatively small size of the available corpora creates a challenge for estimating frequency in a reliable way (Smith & Hofmann, 2020). **Second, annotation conventions are not standardized across different sign language corpora, resulting in differences in glossing** (an issue which is discussed in-depth below). This means that in some corpora, multiple signs are lumped together, while in others they are separated, making it hard to compare frequencies across corpora (Johnston & De Beuzeville, 2016; McKee & Kennedy, 2006; Fenlon et al., 2014a; Schembri et al., 2017; Konrad et al. 2020b). For example, some corpora assign the same annotation to all depicting constructions<sup>4</sup> (classifier signs), while other corpora assign different glosses to different kinds of depict-

---

4. A number of different terms are used in the sign language literature: classifier constructions, verbs of motion and location, verbal predicates, lexical verbs, noun incorporation, classifier predicates, and depicting verbs (Liddell, 2003; Sandler & Lillo-Martin, 2006; Schick, 1987; Supalla, 1982; Zwitserlood, 2012). Here we use the term depicting construction, following Liddell (2003) and Cormier et al. (2012).

ing constructions (Konrad et al. 2020b; Schembri et al. 2017; Crasborn et al. 2015). Finally, assessing sign frequency requires a clear definition of the relevant sign categories, what constitutes a type, and what signs should be counted, a topic which is still under debate (see Johnston, 2010, 2012 for a discussion).

Despite these challenges, several studies examined lexical frequency in different sign language corpora (Morford & MacFarlane, 2003; Johnston, 2012; Fenlon et al., 2014a; McKee & Kennedy, 2006; Smith & Hofmann, 2020). These studies compared frequencies between signed and spoken language and/or between different sign languages and categories (e.g., depicting constructions, pointing signs, gestures, etc.). The studies report several findings relevant to the study of sign distributions. First, as in spoken languages, a small number of signs occur frequently and constitute a large proportion of the tokens, while a large number of signs occur rarely. For example, in the New Zealand Sign Language corpus, 11 signs account for 20% of the tokens in the corpus (McKee & Kennedy, 2006). Similarly, the top 10 most frequent signs in the British Sign Language corpus account for 28% of all tokens (Fenlon et al., 2014a), the top 100 most frequent signs in the Irish Sign Language corpus account for 32.6% of all tokens (Smith & Hofmann, 2020), and the top 100 most frequent signs in the Australian Sign Language (Auslan) corpus account for 53% of all tokens (Johnston, 2012). Second, the most frequent sign accounts for a large amount of the distribution and there is a sharp decrease in frequency between the most frequent sign and the next most frequent sign. For example, in Smith & Hofmann (2020), the two most frequent signs in the Irish Sign Language corpus, INDEX+1-person (a pointing sign expressed with extended index finger directed at oneself) and INDEX 2nd-and-3rd-person (a pointing sign directed away from the signer to mark second-person and third-person) each appear over 600 times, and the third most frequent sign (BUT) appears much less frequently, 106 times. These studies suggest that signed languages have a skewed frequency distribution, as in spoken languages.

This conclusion is contradicted by a recent study examining the sign distribution in American Sign Language (ASL) (Sehry et al., 2021). This study reports subjective frequency for 2,723 ASL signs, which was obtained by asking signers to estimate the frequency of individual signs on a scale from 1–7. Sehry et al. (2021) created a sign frequency distribution, based on the data collected as part of the ASL-LEX 2.0 project – a large database of signs in American Sign Language rated for various properties (Sehry et al. 2021; Caselli et al. 2017). The frequency of these signs appears to be normally distributed (Sehry et al., 2021, Figure 2), and not skewed. However, this could be driven by the use of a Likert rating scale to assess frequency, which has limitations as an estimate of actual frequency (Woltz et al. 2012), and by the use of a particular subset of words, and not the entire lexicon.

The possibility that the normal distribution is not an accurate reflection of the distribution of signs, is supported by evidence reported in a recent chapter: Borstell (2022) analyzed sign frequency from three corpora (British Sign Language, Sign Language of the Netherlands, and Swedish Sign Language); sign frequencies and ranks were calculated and then transformed into a logarithmic scale. The resulting plots, presented in the chapter (Figure 13, pp. 116), displayed a negative linear slope, indicating a potentially good fit to a Zipfian distribution. However, since this was not the goal of the chapter, the analysis of sign distribution does not specify which sign categories were included or excluded in the analysis, making it difficult to conduct a valid comparison to spoken language, or to assess differences and similarities between the three examined corpora. Furthermore, the fit to a Zipfian distribution was not evaluated mathematically: The correlation between  $\log(\text{rank})$  and  $\log(\text{frequency})$  is not reported, nor is the slope, or the fit to a power law. In sum, studies of lexical frequency in signed languages show that some signs are much more frequent than others and suggest that the skew is similar to that of spoken languages (in the proportion of the frequent signs/sign categories). However, prior studies did not assess how Zipfian the distribution is and whether the slope is similar to that of spoken language.

In this study, we ask whether signed languages follow a Zipfian distribution, like spoken languages, as predicted if such distributions have a cognitive and/or communicative source. We do this by examining the sign frequency distributions of three annotated and machine-readable corpora: (1) the British Sign Language (BSL) corpus, (2) the German Sign Language (DGS) corpus, and (3) the Sign Language of the Netherlands (NGT) corpus. Unlike past studies, we investigate not only their objective sign frequencies, but also the fit of their frequency-rank distribution to a Zipfian distribution, using two mathematical estimates. Given the recurrence of Zipfian distributions in spoken languages, we expect to find a similar distribution in signed languages as well.

## Method

For this study, we used three different corpora, from the following three sign languages: British Sign Language (BSL), German Sign Language (DGS) and Sign Language of the Netherlands (NGT). We extracted the relevant videos and their annotations using ELAN, a video annotation software (Crasborn & Sloetjes, 2008, <https://archive.mpi.nl/tla/elan>, last access 12 March 2024). We then cleaned the data, leaving only relevant signs and sign categories (see details of inclusion/exclusion criteria for the different sign categories in Section 2.1 below). For the included sign types, we calculated sign frequency and created the frequency-

rank distribution. Finally, we used two methods to determine “how Zipfian” the distributions are. We describe each dataset and the analyses below. All relevant files (datasets, analysis scripts, etc.) can be found here: <https://osf.io/eh8sy/> (last access 12 March 2024).

## Corpora

### *The British Sign Language (BSL) corpus project*

The *BSL Corpus* is a collection of video clips of native, near native and fluent deaf signers of BSL signing in a range of semi-spontaneous language tasks. The BSL Corpus is hosted by the Deafness, Cognition, and Language Research (DCAL) Centre, based at University College London. The corpus is made up of 125 hours of videos produced by 249 deaf people from 8 cities across the United Kingdom. Participants were mixed and balanced for age, region, and gender and were recruited from a range of social backgrounds and ethnicities. Most people who participated reported that they learned BSL before the age of 7 and had lived in the same region for the last ten years (Schembri et al. 2017; Fenlon et al. 2014a).

The corpus contains data from four linguistic tasks: (1) Narratives: participants told short personal stories or anecdotes about their lives; (2) Lexical Elicitation: participants produced their sign variants for 102 different concepts, elicited using a picture/word combination; (3) Interviews: participants answered questions about their language attitudes and awareness; and (4) Conversations: Dyads of participants, matched for age and region, engaged in free conversation for 30 minutes (Schembri et al. 2017).

To assess sign frequency, we used data extracted from narratives, interviews, and conversations. We excluded data from the lexical elicitation task since we are interested in assessing lexical frequency in actual usage. In total, we used 197 narrative video files, 75 conversation video files, and 10 interview video files which were readily available online. The size of the corpus analyzed here is 34,909 tokens.

### *The DGS-korpus project*

The *DGS-Korpus* is an open access online corpus of dialogues between native users of German Sign Language (Deutsche Gebärdensprache, DGS). The corpus project was carried out at the Institute for German Sign Language and Communication of the Deaf at Hamburg University. The data consists of signed conversations, narrations, discussions, retellings, and other sign uses produced by 330 informants and was filmed between 2010 and 2012. Participants were mixed in age, gender, and regions in Germany. Nearly 560 hours of signing were recorded

(Langer et al. 2018; Konrad et al. 2020a; Konrad et al. 2020b). We analyzed all data available at the time, which consists of 408 video files and 353,227 tokens.

### *Corpus NGT*

The *Corpus NGT* is an open access online corpus of dialogues between native users of Sign Language of the Netherlands (NGT). The *Corpus NGT* was created by Onno Crasborn, Inge Zwitterlood and Johan Ros at Radboud University and filmed between 2008 and 2011. It consists of 72 hours of video data, including recorded conversations produced by 92 signers (Crasborn & Zwitterlood 2008; Crasborn et al. 2008). We analyzed all available data, which consists of 2,281 video files and 108,434 tokens.

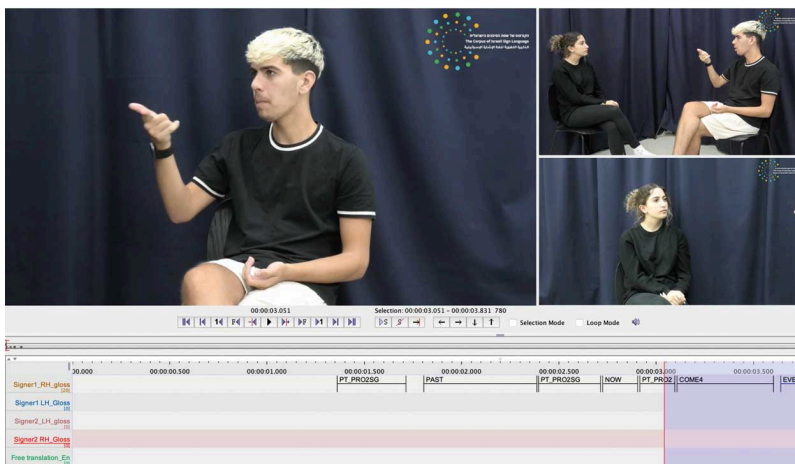
### Coding

#### *ELAN*

We used the ELAN annotation software (Crasborn & Sloetjes, 2008) to extract the annotations from the videos. ELAN is an annotation tool for audio and video recordings, which we used to extract the textual annotations of the video recordings for all three corpora (see Figure 1). The content of the annotations consists of Unicode text and annotation documents. The annotations are formed in layers called tiers. Each tier is hierarchically interconnected and consists of different kinds of annotations. For this project, as is done for spoken languages, we included only the tiers that contain glosses, rather than translations or annotations related to mouthing – the voiceless articulation of spoken words while simultaneously producing signs (Bank et al., 2016), or non-manual features.

A gloss, or an ID gloss, is an identifying label that is assigned to each unique lexical sign (Johnston, 2010; Fenlon et al. 2014a). As signs can be signed with one hand, usually the dominant one, or both hands, we extracted tiers for both left and right hands, and collapsed them to avoid duplicates as described below. For the BSL corpus, we included the right-hand (RH-ID gloss) and the left-hand (LH-ID gloss) tiers. For the DGS corpus, the tiers for both hands and both signers were included (“Lexeme\_Sign\_r\_A”, “Lexeme\_Sign\_l\_A”, “Lexeme\_Sign\_r\_B” and “Lexeme\_Sign\_l\_B”) because signers were filmed in pairs. The same tier inclusion was used for the NGT corpus (“GlossR S1”, “GlossL S1”, “GlossR S2”, and “GlossL S2”).





**Figure 1.** Example of ID-glossing in ELAN

### *Exclusion and inclusion criteria for sign categories*

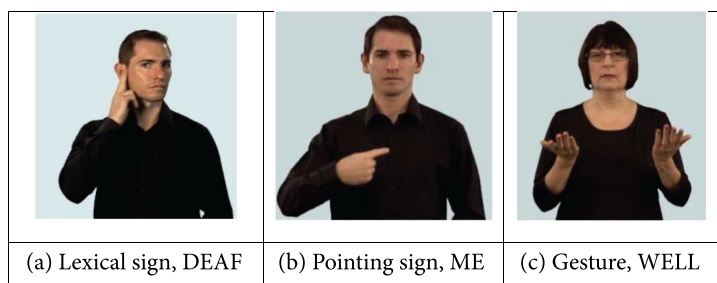
Before analyzing the lexical sign frequency of the three corpora, we had to decide how to categorize signs, and which signs to include in our analyses. This is not a trivial task, and one that has been the focus of debate within the sign language literature as it pertains to fundamental questions about the structure of the lexicon (Cormier et al., 2012; Brennan, 1982; see Johnston, 2012 for a discussion of this in relation to estimating sign frequency). Since such a debate is beyond the scope of the current paper, we decided to follow Fenlon et al. (2014a) in determining how to categorize signs and which categories to include. Next we will outline the lexicon we adopt and the specific categories we included or excluded (based on Fenlon et al. 2014a).

Broadly speaking, the sign language lexicon is comprised of several components including the native and non-native lexicon (Brentari & Padden, 2001). The key difference between them is that the native lexicon contains signs that have developed within sign languages while the non-native lexicon includes features which exist as a result of contact between spoken and signed languages, such as fingerspelling<sup>5</sup> and mouthing (Cormier et al., 2012). The native lexicon can be further divided into two components: core and non-core. The core native lexicon consists of lexical signs, often referred to as the permanent, frozen, or established part of the lexicon, which are highly stable, standardized in form and meaning,

---

5. Fingerspellings are handshapes which represent each individual letter of a spelled word in the ambient spoken language. For example, the name 'Yael' is fingerspelled in ISL with each individual letter in the Hebrew alphabet represented by a different handshape.

usually context-independent, and frequently used (Brennan, 1982; McDonald, 1985; Fenlon et al., 2014a). The non-core lexicon includes a set of elements whose use is more variable and context-dependent, that are only weakly lexicalized, and are less standardized, among them are depicting constructions, pointing signs, and gestures (Cormier et al. 2012; Fenlon et al., 2014a; Johnston, 2012). Depicting constructions are complex lexical units where individual elements, such as handshape, orientation, location, and movement, carry specific meanings, and contribute iconically to the overall meaning. Depicting constructions encode the location and movement of entities, and the handling, size and shape of entities (Johnston & Schembri, 2007; Liddell, 2003). An extended index finger, for instance, may represent any long upright object such as a person (Cormier et al., 2017). Another example of a feature in sign languages which are not fully lexicalized is pointing signs. Pointing signs, often articulated with an extended index finger, are used to refer to concrete and abstract locations in space, as shown in Figure 2b (Klima & Bellugi, 1979; Lillo-Martin & Klima, 1990; Meier, 1990). Pointing signs are argued to be partly lexicalized, with the handshape typically conventionalized, but the location and movement not (see Figure 2b). Gestures refer to communicative actions that are usually context dependent, are not conventionalized in form and meaning or are similar to some gestures that hearing non-signers produce (Johnston, 2012; Fenlon et al. 2014a; Johnston & De Beuzeville, 2016). Thus, the non-core lexicon defers from the core lexicon, that consists of fully lexicalized signs, like the sign “DEAF” in BSL (see Figure 2a), whose form is stable and consistent across signers. Elements of the non-core lexicon can become lexicalized over time (Cormier et al. 2012). An example of this is the ‘palm-up’ gesture, shown in Figure 2c, which is used frequently by both hearing and deaf individuals, meaning ‘well’ in BSL (Cooperrider et al. 2018).



**Figure 2.** Examples from the BSL Sign Bank (Fenlon et al., 2014b)

Following Fenlon et al. 2014a, we include all signs belonging to the core lexicon and a few categories from the non-core lexicon, in our frequency count, each analyzed corpus has different annotation conventions for signs from the non-core

lexicon. For example, while some corpora differentiate between different kinds of depicting constructions (e.g., specifying the handshape and the movement and differentiating between different handshapes), others collapse them all under one depicting construction category. For this reason, we created two datasets of each of the three corpora: One which contains all the categories annotated in the specific corpus (which we will refer to as the minimally excluded dataset), and another which only contains those categories whose annotation is shared across corpora (which we will refer to as the comparative dataset), the latter is a subset of the former. Few sign categories were excluded from the first dataset of each corpus, as detailed below, resulting in a comprehensive view of the different kinds of linguistic elements found within each sign language corpus. This ensures that we do not overlook potentially valuable sign categories that have different annotations between the corpora. The second dataset of each corpus includes only those sign categories with consistent and comparable annotations across corpora, allowing for a more valid and accurate comparison of sign distribution across the three sign languages. In what follows, we describe the different sign categories, and our decisions for inclusion or exclusion from our analysis. Because of differences in annotation across the three corpora, some sign categories (like depicting constructions) were included in the analysis of each corpus but not in the comparison between them (since their coding is not comparable across corpora, see discussion below). Appendix 1 details all the differences in annotation between the three corpora.

### *Sign categories included*

We included all signs which form part of the core lexicon, as well as several sign categories from the non-core lexicon that are frequent, usually context-independent across signers and corpora, and that were included in prior studies of lexicon frequency in sign languages (Fenlon et al. 2014a).

#### *Fully lexical signs (core lexicon)*

All signs belonging to the core lexicon were included, these are roughly equivalent to a word in spoken language.

#### *Depicting constructions*

Depicting constructions are complex lexical items in which each of the units of handshape, orientation, location, and movement may have their own meaning (Cormier et al., 2012; Zwitserlood, 2012; Sandler & Lillo-Martin, 2006). Depicting constructions are iconically motivated and have a general meaning to which each iconic value of its components contributes (Konrad et al. 2020b). There are big differences in how depicting constructions are coded across corpora (as shown in Appendix 1). For example, in the DGS corpus there is only one gloss category

(“\$PROD”) for all depicting constructions (Konrad et al. 2020a; Konrad et al. 2020b), while the BSL and NGT corpora have more specification (Schembri et al. 2017; Crasborn et al. 2015). Including the depicting constructions in the DGS dataset would impact the frequency distribution, because the lack of specification would make it one of the most frequent types, therefore, we decided to exclude it from the DGS dataset. We included depicting constructions in the datasets of the BSL and NGT corpora, but excluded them from the comparative dataset, to allow for a more comparable analysis.

### *Pointing signs*

Pointing (deictic) signs are indexical signs that typically use the index finger or other parts of the hand(s) for pointing. The coding of this category varies between the corpora (as shown in Appendix 1). In the DGS corpora, pointing signs include the signs “I” and “you”, which are lexicalized pointing signs, but the remaining 6 pointing signs are not specific, and encode variation in handshape only (for example “\$INDEX<sub>1</sub>” is a handshape with extended index finger, “\$INDEX<sub>4</sub>” is a thumb handshape) or location (for example “\$INDEX-TO-SCREEN<sub>1</sub>” for pointing towards the monitor, Konrad et al. 2020b). In contrast, the BSL corpus contains 21 more specific pointing signs, including a different gloss for 1st, 2nd and 3rd person, singular and plural, and so on (Schembri et al. 2017). The NGT corpus contains 17 pointing signs (Crasborn et al. 2015). Despite the variation, we decided to include pointing signs in all datasets of each corpus, due to their importance and frequency in sign languages.

### *Buoys*

Buoys are configurations that are used as a physical reference point for a referent. They are usually made with the non-dominant hand while the dominant hand continues to sign and they (Fenlon et al. 2014a). There are several types of buoys that can be expected in signed discourse. The four main kinds of buoys are list-, pointer-, fragment-, and theme-buoys (Liddell 2003; Schembri et al. 2017; Johnston & De Beuzeville, 2016). We included buoys in our frequency counts following Fenlon et al. 2014a. While the DGS and BSL corpora have specific glosses for buoys, the NGT corpus does not (as shown in Appendix 1). Thus, we excluded buoys from the three comparative datasets, and included them in the separate analysis of the DGS and BSL datasets.

### *Gestures*

Gestures refer to communicative actions that are non-lexical since they do not appear to be highly conventionalized in form and meaning (i.e., they rely on context to be properly understood), or are similar to some gestures that hearing non-signers produce (Johnston, 2012; Fenlon et al. 2014a; Johnston & De Beuzeville,

2016). The decision about what constitutes a gesture in this category differs across corpora, with some coding schemes including semi-lexicalised discourse markers (e.g., well, so, etc.) and others not (Johnston, 2012). For example, gestures like ‘G:Well’ in the BSL corpus (example shown in Figure 2c) are not classified as gestures in other corpora and vice versa. Therefore, we included gestures in the separate datasets, but not in the comparative ones, due to the variation between the corpora (as detailed in Appendix 1).

### *Excluded sign categories*

We excluded several kinds of categories, based on the annotation provided in the corpora: (1) Uncertain signs: signs that could not be recognized for various reasons (see details below), (2) Mouthing, (3) Extra-linguistic manual activity, (4) Fingerspelling, (4) Names, and (5) Cued Speech and Initializations.

### *Uncertain signs*

This category includes several kinds of signs that could not be clearly identified:

1. False starts, in which the signer started to sign something but then changed their mind, were excluded only in cases where the sign was not clear.
2. Unknown signs, in which the annotators did not recognize the sign or they were not sure of its meaning, were excluded.
3. “Invisible” signs are those that were not fully shown in the video or were poorly articulated or not completed. These signs were excluded only in cases where the sign was not clear
4. Signs not in Sign Bank. Signs that have not been added yet to the Sign Bank were excluded, as we do not know whether they are already lexicalized.

### *Mouthing*

These are instances in which there is no significant manual movement, and the meaning is expressed only via mouthing, which can be observed particularly in older informants (Konrad et al. 2020b). They are annotated in the corpora as \$ORAL^ (in the DGS corpus) and were excluded.

### *Extra-linguistic manual activity*

These tokens refer to extra-linguistic manual activity like rubbing one’s nose or brushing off one’s clothes. They are annotated in the corpora (coded as % or any lower-case letters in the NGT and \$\$EXTRA-LING-ACT^ in the DGS), but are not manual signs. And hence were excluded as well.

### *Fingerspelling*

Fingerspelled forms represent a sequence of hand configurations of the letters of the correspondence spoken language’s alphabet (Fenlon et al. 2014a). As some of

the fingerspelled signs are lexicalized and some are not, we decided to exclude all of them, as we cannot differentiate between the lexicalized and the non-lexicalized ones.

### *Names*

In the NGT and DGS corpora, personal names are glossed with a general gloss “NAME”, with no further information. The inclusion of this category would result in an artificial (and higher inflated) frequency count for the name category. Since names tend to appear in the tail of the distribution anyway (lower frequency), we decided to exclude them.

### *Cued speech and initializations*

In the 1970s, a cued speech system was developed in Germany for teaching the articulation of phonemes to deaf children. Some of these cued speech hand signs are used in DGS today, much like initialized signs (signs which incorporate the handshape of the first letter of the word, e.g., repeated “G” handshape in BSL to mean “Geography”), to express names where no conventional sign is at hand (Konrad et al. 2020b; McKee & Kennedy, 2006). Thus, as decided with names, these tokens were excluded as well.

### *Collapsing over specific tokens within a sign type*

To ensure accurate counts for the different sign types (unique signs), we had to collapse different tokens within certain sign types:

1. Two handed signs were collapsed into a single sign if both hands signed the same sign at the same time (or up to 5 milliseconds apart), so that two handed signs were counted only once.
2. Phonological variants were collapsed together. In spoken languages, phonological variants refer to differences between accents, and are included in the analysis as one type, with no reference to the variants (e.g., two different pronunciations of ‘park’ will not be counted as two types). In sign languages, phonological variation refers also to variation in phonological form, in terms of changes in movement, location, handshape or orientation, however the variations do not signal differences in meaning. As in spoken language, we counted different phonological variants under the same type, as they all refer to one specific sign. For example, the signs “WOMAN<sub>3a</sub>” and “WOMAN<sub>3b</sub>” in the DGS corpus reflect phonological variants (indicated by the addition of ‘a’ and ‘b’) and were both glossed as “WOMAN<sub>3</sub>”.
3. In the coding for the NGT corpus, we glossed both types “PT-Bhand:B” (point to self with B handshape) and “PT-1hand:1” (point to self with extended index finger) as one type which we called “I”, as they both refer to this meaning.

### *Creating a frequency distribution and assessing the fit to a “Zipfian” one*

The next step after applying our exclusion criteria as explained above, was to estimate the distribution by counting the frequency of each unique sign. We created a list of unique signs and their frequencies and calculated the rank of each sign (the most frequent sign was given rank 1, the 2nd most frequent sign was given rank 2, and so on). Signs with the same frequency were assigned ranks randomly, one after the other, as is commonly done in investigations of Zipfian distributions in spoken languages (Lavi-Rotbain & Arnon, 2023). After obtaining the frequency distributions, we used two methods to evaluate “how Zipfian” they are. The first method examines the correlation between frequency and rank on a log-log basis to assess how linear this relation is; under a Zipfian distribution, we expect a negative correlation close to  $-1$ . However, this method is an imperfect one, as other distributions besides Zipfian ones are also linear on a log-log scale. Consequently, the result of a Pearson correlation close to  $-1$  is insufficient for concluding that the original distribution follows a power law (Clauset et al. 2009).

Therefore, we employed an additional method, used in previous studies of Zipfian distributions (Piantadosi, 2014, Lavi-Rotbain & Arnon, 2023), that estimates the parameters of the distribution (as described below, in Equation 1) and then calculates the correlation between the observed frequencies and the expected frequencies under a Zipfian distribution with the estimated parameters. There are two parameters: (1)  $\alpha$ , the exponent of the power law, which determines the slope of the distribution, which is close to 1 in pure Zipfian distributions; and (2)  $\beta$ , a correction added to the original Zipf’s law by Mandelbrot to create a better fit to actual language data (Mandelbrot 1953). The parameters are found by using the maximum likelihood estimator (MLE), which is a commonly used algorithm to solve parameter estimation problems (Linders & Louwerse, 2020; Piantadosi, 2014).

$$f(r) \propto \frac{1}{(r + \beta)^\alpha} \quad \text{Equation 1.}$$

We then use these parameters to find the expected frequency of the signs, by using the probability mass function of a Zipfian distribution (see Equation 2) and evaluate the goodness of the fit between the observed frequency distribution (as calculated) and the expected one under a Zipfian distribution. If the distribution is Zipfian, the correlation between the two should be linear and positive, with a correlation close to 1.

$$p(r) = \frac{1}{(r + \beta)^\alpha} * \sum_{r=1}^N \frac{1}{(r + \beta)^\alpha} \quad \text{Equation 2.}$$

## Results

We first outline the results of the correlation between frequency and rank on a log-log basis to see how linear it is, and then we present the results using the parameter estimates to see the correlation between the observed frequency and the expected frequency under a Zipfian distribution.

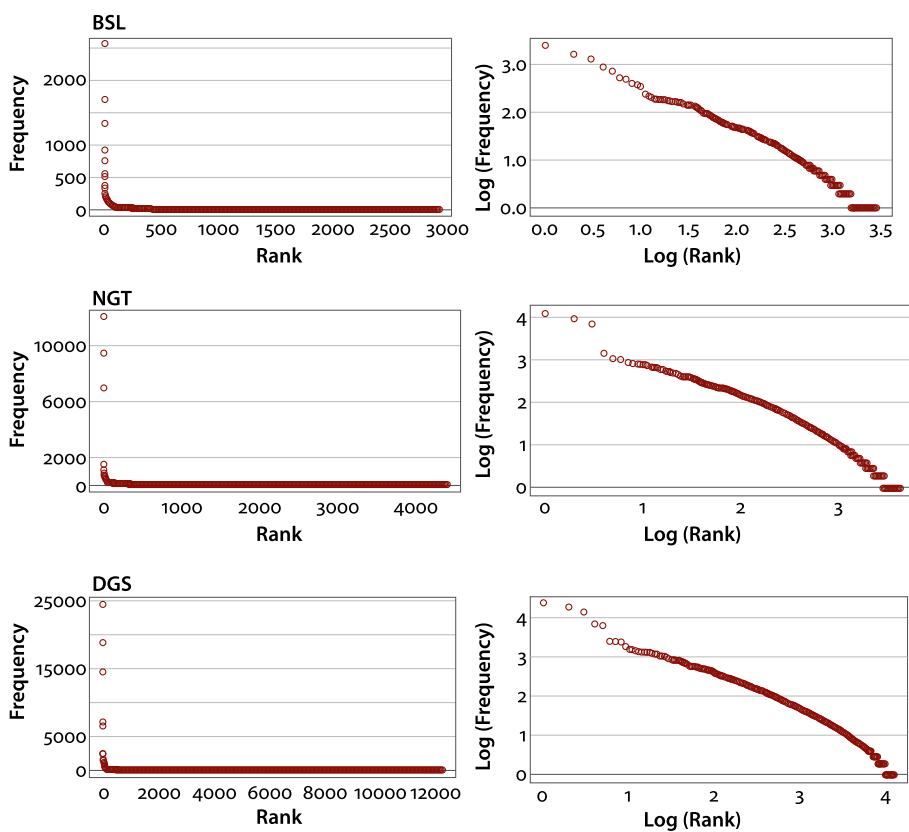
We started by applying our exclusion and inclusion criteria and creating the two datasets for each corpus. Importantly, the minimally excluded datasets of each corpus retained a very high proportion of the complete corpus (93%–98% percent). That is, only 2–7% of the data was excluded before analysis. These numbers indicate that the samples we conducted our analysis on provide a good estimate of the use of signs in the complete corpora. The minimally excluded BSL dataset retained 93% of the tokens from the BSL corpus, after excluding uncertain signs (3%), fingerspelling (3%), and names (1%). For the comparative BSL dataset, 83% of the tokens from the complete BSL corpus were included, with additional exclusions for depicting constructions (2%), gestures (8%), and buoys (0.5%). Similarly, the minimally excluded DGS dataset preserved 96.38% of the tokens from the complete DGS corpus, with exclusions for uncertain signs (0.0006%), fingerspelling (0.7%), names (0.3%), mouthing (0.5%), extra linguistic manual activity (0.1%), initializations (0.03%), and cued speech (0.07%). For the comparative DGS dataset, 88% of the tokens from the DGS corpus were included, with further exclusions for depicting constructions (2%), gestures (9%), and buoys (0.5%). Finally, the minimally excluded NGT dataset retained 98% of the tokens from the NGT corpus, excluding uncertain signs (0.4%), fingerspelling (1%), names (0.04%), and extra linguistic manual activity (0.3%). For the comparative NGT dataset, 83% of the tokens from the NGT corpus were included, with additional exclusions for depicting constructions (5%) and gestures (10%). Appendix 2 provides a detailed breakdown of the percentage of each sign category in each corpus. Table 1 shows the total count of signs and the number of unique signs in each corpus, across the entire dataset, the minimally excluded dataset, and the comparative dataset (always smaller, since sign types that were not coded in a comparable way across corpora were excluded).

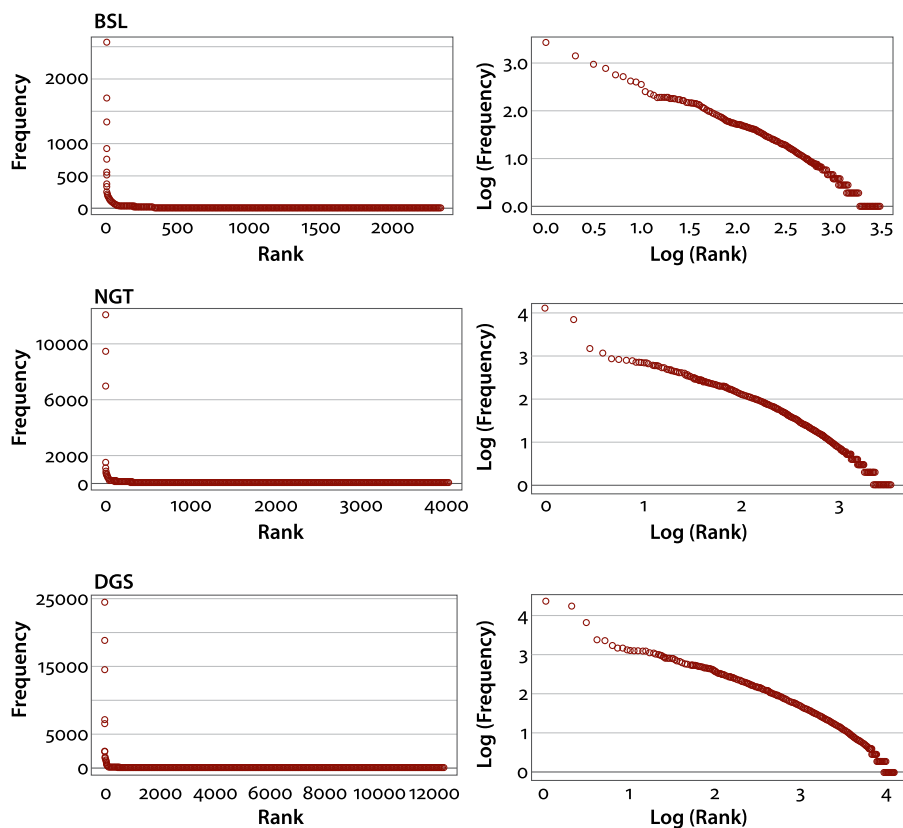
We started by looking at the Pearson correlation between frequency and rank in log space in the three corpora over the minimally excluded datasets. As predicted, the correlations were close to  $-1$ , indicating a good fit to a Zipfian distribution (BSL:  $R_2 = -0.98$ , DGS:  $R_2 = -0.97$ , NGT:  $R_2 = -0.98$ ). Figure 3 shows the distributions in regular and log space. We then conducted the same analysis on the comparative datasets. Again, the correlations were close to  $-1$ , indicating a good fit to a Zipfian distribution (BSL:  $R_2 = -0.98$ , DGS:  $R_2 = -0.97$ , NGT:  $R_2 = -0.98$ ). Figure 4 shows the distributions in regular and log space for the comparative datasets for each corpus.



**Table 1.** Lexical properties of the three corpora

Language	Dataset	Number of tokens	Number of types	Frequency range
BSL	Complete	34,909	4,275	1-2542
	Minimal Exclusion	32,436	2,903	1-2542
	Comparative	28,848	2,343	1-2542
DGS	Complete	353,227	13,120	1-24,408
	Minimal Exclusion	340,429	12,276	1-24,408
	Comparative	310,001	12,232	1-24,408
NGT	Complete	108,434	4,796	1-12,101
	Minimal Exclusion	106,501	4,383	1-12,101
	Comparative	89,609	4,027	1-12,101

**Figure 3.** The distribution of raw frequency (Left) and log frequency (Right) for the three minimally excluded datasets



**Figure 4.** The distribution of raw frequency (Left) and log frequency (Right) for the three comparative datasets

We then examined the fit to a Zipfian distribution using the second method, in which we estimate the parameters of the distribution and test how well the observed frequency fits the expected frequency. For both the minimally excluded datasets, and for the comparative ones, alpha (the slope) was very similar across corpora (ranging from 0.93 to 1.05), and similar to the expected alpha of 1. The range of beta was between  $-0.22$  and  $1.14$ , smaller than what has been reported for spoken languages (i.e., in Lavi-Rotbain & Arnon's CDS study, the range was 6.27–19.48 (Lavi-Rotbain & Arnon, 2023)). Importantly, the value of beta does not determine whether a distribution is Zipfian and is not expected to be stable across corpora. In addition, the Pearson correlation between the observed and expected frequencies of all corpora is close to 1 (see Table 2), indicating a very good fit to a Zipfian distribution. These alpha values seem similar to those found for one hundred translations of the Bible, where the range of the Zipf's exponent varied from 0.77–1.44 (Mehri & Jamaati 2017), and somewhat lower than the alpha in child-

directed speech for 15 languages, where alpha ranged from 1.16–1.57 (Lavi-Rotbain & Arnon, 2023). That is, the slope found for sign distributions is very similar to that of word distributions. One note of caution: Our estimates for BSL may be less reliable since their calculations are based on a smaller corpus and alpha estimates are only reliable from around 50,000 tokens, see Lavi-Rotbain & Arnon, 2023).

**Table 2.** Correlations and parameter estimates for each corpus

	No. of tokens	No. of types	Log*log correlation	Frequency range	$\alpha$	$\beta$	Pearson's r (observed* expected)
BSL	32,436	2,903	-0.98	1-2,542	1.05	1.14	0.99
BSL (comparative)	28,848	2,343	-0.98	1-2,542	1.03	1.00	0.99
DGS	340,429	12,276	-0.97	1-24,408	0.98	0.26	0.97
DGS (comparative)	310,001	12,232	-0.97	1-24,408	0.93	-0.08	0.96
NGT	106,501	4,383	-0.98	1-12,101	1.04	0.09	0.96
NGT (comparative)	89,609	4,027	-0.98	1-12,101	0.99	-0.22	0.97

What do the most frequent signs look like across the three corpora?

Next, we wanted to make a qualitative comparison of the frequent signs in each corpus. To do this, we extracted the 20 most frequent signs from the minimally excluded dataset of each corpus, and calculated the proportion of tokens they represent (Table 3). Some of the frequent signs are shared across corpora: The sign “I” appears in the top three most frequent signs across all three corpora, the signs “good” and “no/not” appear in the top 20 most frequent signs in all three corpora, and the signs “but”, “deaf”, “one”, “right”, “can”, “same”, “self”, “my” and “look” appear in the top 20 most frequent signs in two of the three corpora. Moreover, pointing signs and different kinds of gestures also appear as frequent signs across the three corpora. The distribution of sign categories among the frequent signs is also quite similar: In the BSL corpus, 7 of the most frequent signs are pointing signs, 1 is a gesture, and 12 are content signs. In the DGS, 6 are pointing signs, 4 are gestures, and 10 are content signs. In the NGT corpus, 2 of the most frequent signs are pointing signs, 2 are gestures, and 16 are content signs.

**Table 3.** 20 most frequent words in each corpus (exact percentages given for signs that had a frequency of over 1%)

Rank	BSL sign	Frequency	DGS sign	Frequency	NGT sign	Frequency
1	PT:PRO <sub>1</sub> SG (I)	2542 (7.8%)	\$INDEX <sub>1</sub>	24408 (7.2%)	PT-1hand	12101 (11.4%)
2	G:WELL	1722 (5.3%)	I <sub>1</sub>	18736 (5.5%)	PO	9469 (8.9%)
3	PT:PRO <sub>3</sub> SG	1317 (4%)	\$GEST-OFF <sup>^</sup>	14516 (4.3%)	I	6951 (6.5%)
4	PT:	924 (2.8%)	I <sub>2</sub>	7079 (2.1%)	GEBAREN-A (GESTURES)	1440 (1.4%)
5	GOOD	757 (2.3%)	\$GEST <sup>^</sup>	6524 (1.9%)	JA-A (YES)	1134 (1%)
6	PT:LOC	552 (1.7%)	\$GEST- DECLINE <sub>1</sub> <sup>^</sup>	2607	PO+PT	1048
7	PT:PRO <sub>2</sub> SG	519 (1.6%)	YOU <sub>1</sub>	2552	ATTENTIE (ATTENTION)	864
8	PT:DET	409 (1.3%)	DEAF <sub>1</sub> A	2427	GOED-A (GOOD)	844
9	SAME	391 (1.2%)	GOOD <sub>1</sub>	1912	KUNNEN-A (CAN/ BE ABLE TO)	818
10	WHAT	350 (1%)	BUT <sub>1</sub>	1595	WETEN-A (KNOW)	787
11	PT:POSS <sub>1</sub> SG	248	ALSO <sub>1</sub> A	1582	HOREN-A (TO BELONG)	770
12	RIGHT	225	MUST <sub>1</sub>	1426	HEE (HEY)	675
13	LOOK	212	NOT <sub>3</sub> A	1418	ZIEN-A (SEE)	673
14	BAD	191	RIGHT-OR- AGREED <sub>1</sub> A	1374	ZELFDE-A (SAME)	672
15	NO	190	CAN <sub>1</sub>	1350	ZEGGEN (SAY)	608
16	NOW	190	MY <sub>1</sub>	1347	NIET-A (NOT)	598
17	ONE	189	\$GEST-NM- NOD-HEAD <sub>1</sub> <sup>^</sup>	1336	ZELF-A (SELF)	562
18	THINK	187	SELF <sub>1</sub> A	1336	1-A	545
19	BUT	178	\$NUM-ONE- TO-TEN <sub>1</sub> A:1d	1291	DOOF-A (DEAF)	531
20	HAVE	177	PRESENT-OR- HERE <sub>1</sub>	1243	KIJKEN-A (LOOK)	520

## Discussion

One of the striking commonalities between languages is their Zipfian distribution of word frequencies (Zipf, 1949). The recurrence of such distributions, across languages, has been the topic of much research and debate, with different views as to whether and how the distribution reflects foundational properties of human cognition and/or communication (e.g., Ferrer-i-Cancho & Sole, 2003; Bentz et al. 2017). Recent work has shown that Zipfian distributions provide a facilitative environment for learning in a range of linguistic tasks, including word segmentation, word learning, and grammatical category learning, e.g., Lavi-Rotbain & Arnon, 2022, supporting the possibility that such distributions have cognitive sources in language. To date, research on Zipfian distributions has been mostly limited to the word frequencies of spoken languages. If Zipfian distributions reflect properties of human cognition/communication, as suggested by previous research, they should also be found in signed languages. In this study, we examined whether the distribution of signs in three sign languages follow a Zipfian distribution. While previous research has looked at lexical frequency in sign languages, studies to date have not mathematically assessed the fit of sign distributions to Zipf's law.

In this paper, we examined the distribution of signs in corpora of three languages: BSL, DGS, and NGT, using two methods to assess how well the distribution fits Zipf's law. First, we examined the correlation between frequency and rank on a log-log basis to see how linear it is. We then estimated the parameters of the distribution (alpha and beta) and examined the correlation between the observed frequency and the expected frequency under a Zipfian distribution. As hypothesized, the distribution of sign frequency showed a close fit to a Zipfian distribution in all three corpora: The log-log correlations of the analyzed corpora were close to  $-1$ , reflecting an almost perfect negative linear correlation, and the correlation between the expected frequency and the observed one was close to  $1$ , reflecting a close fit to a Zipfian distribution. The results show that the use of signs is highly skewed and follows a similar distribution to that of words in spoken language. With that, they reveal another dimension of similarity between signed and spoken languages: The frequency with which linguistic building blocks (words or signs) are used. Moreover, these findings are consistent with the idea that Zipfian distributions reflect pressures and needs shared by all human languages.

Our analyses revealed several similarities between the three signed languages. Despite having somewhat different annotation systems and differently sized samples (ranging from 34,909–108,434), the slope of the distribution (the alpha), reflecting the decrease in sign frequency, was similar across the three corpora (alpha 0.93 to 1.05). This range of alpha is similar to the range found in spoken

languages when looking at Bible translations (Mehri & Jamaati, 2017, this is the only paper that has a large cross-language comparison of alpha values). Interestingly, the alpha values of sign languages and bible translations – both adult produced registers – seem smaller than those found in child-directed speech (Lavi-Rotbain & Arnon, 2023): In Mehri and Jamaati’s study the range was 0.77–1.44, in the current study it was 0.93–1.05, while in the child-directed speech from 15 languages, the range was 1.16–1.57. These differences may reflect the unique properties of child-directed speech, and, in particular, their smaller lexicon, which may lead to a steeper slope (Mehri & Jamaati, 2017). In other words, the most frequent words in child-directed speech may take up a larger part of the distribution compared to the most frequent words in adult-to-adult conversation, or in written text. Importantly, the relevant property in determining alpha seems to be register (e.g., child-directed vs. adult-to-adult) rather than modality. Another point of similarity can be seen when we look at the 20 most frequent signs in the three languages. More than 50% of the frequent signs are shared across the three languages (e.g., “I”, “good”, “no”, “but”, “deaf”, “one”, “can”, “same”, “you”, “my”, etc.), and the distribution of sign categories within the most frequent signs is similar across the languages.

Our comparison of the most frequent signs across languages also revealed differences in annotation, and how those may impact the estimation of sign frequency. For example, “PO” (palm-up), ranks as the second most frequent sign in the NGT corpus, constituting almost 9% of the sign tokens in the corpus. However, palm-up can have several meanings, including ‘see what I mean?’, ‘I agree with you’, and ‘it’s your turn’ (Crasborn et al., 2015). Annotating each meaning separately may result in lower frequencies for the individual meanings, consequently affecting their ranks. More broadly, this highlights the need to generate consistent annotations and coding schemes across different corpora (Johnston & De Beuzeville, 2016; Fenlon et al., 2014a; Johnston, 2010). As mentioned above, the corpora differ in the level of specification for various sign categories, as well as in how certain sign categories are annotated. For example, the DGS corpus includes the lexicalized pointing signs “I” and “you”, while the remaining six pointing signs lack specificity and only encode variation in handshape or location (Konrad et al. 2020b). In contrast, the BSL corpus contains 21 pointing signs, differentiating between 1st, 2nd and 3rd person, singular and plural, and so on (Schembri et al. 2017). These differences could lead to the creation of sign categories which appear more frequent than their actual use, since multiple signs are collapsed together. Another limitation of the usage of these corpora for our analysis, that is also an advantage, is that the texts originate from different kinds of linguistic interactions and are not parallel in content. On the one hand, this means that we are comparing different linguistic content across the three languages. The fact that we,

nevertheless, see a similar distribution despite differences in content attests to its robustness.

Before analyzing the sign distributions, we excluded several sign categories (e.g., mouthing, fingerspelling, names, etc.), as well as cases where multiple signs were collapsed into one category (e.g., all depicting constructions were coded as “\$PROD” in the DGS corpus). We created two datasets of each corpus: a minimally excluded dataset, and a subset of that that only contained sign categories coded similarly across the three languages. Importantly, the minimally excluded dataset included a high proportion of the complete corpus, with only 2%–7% of the data excluded. Nevertheless, we wanted to see what the distribution looks like without exclusions. To do this, we extended our analysis to the complete datasets, applying the same two methods applied above to assess how closely the distributions fit Zipf’s law (see analyses in Appendix 3 and 4). The complete datasets (without exclusions) also showed a very good fit to the Zipfian distribution, suggesting that the presence of this distribution is a robust phenomenon found in different language samples, and is not an artifact of a specific exclusion criteria.

For spoken language, it has been suggested that Zipfian distributions can facilitate word segmentation by having the highly frequent words serve as anchors for segmenting less frequent ones (e.g., Kurumada et al. 2013), and by creating a more predictable learning environment where there is more room for making predictions and learning from the violation of those predictions (Lavi-Rotbain & Arnon, 2022). Both factors may also facilitate sign language acquisition. The task of segmentation, though different from the one facing speakers of spoken languages, is also present in sign languages. Infants learning to sign need to segment continuous, transient input into discrete lexical units, as well as learn which movement transitions are likely to belong to a sign, and which indicate a boundary between two signs (this difference is labeled as “horizontal” vs. “vertical” information by Brentari, 1998). Indeed, child-directed signing, like child-directed speech, contains modifications that could assist segmentation such as slower signing, larger sign sizes, and increased repetition (Holzrichter & Meier, 2000; Masataka et al., 2000; Erting et al., 1990). Several recent studies find parallels in the factors impacting segmentation of spoken and signed languages, despite the different cues for segmentation in the two modalities (Brentari 2006; Orfanidou et al. 2010, 2015). For example, deaf signers of BSL were better at detecting real BSL signs appearing in a stream of real and made-up signs when the made-up signs were possible BSL signs (ones made up of handshapes or movements that are used in BSL but not in the presented combinations), a preference similar to that shown by listeners when avoiding segmentations that include impossible words (Orfanidou et al. 2010). If signers need to segment signs from a continuous input and are impacted by (some of) the same factors that impact speech segmentation, then sign acquisition may

also be facilitated in Zipfian distributions. This prediction can be tested experimentally by exposing signers to novel signs in different distributions. If the facilitative effect of Zipfian distributions is a general one (Lavi-Rotbain & Arnon, 2021, 2022; Shufaniya & Arnon, 2022), it should also be found for signers, and for sign language acquisition.

The main goal of the current study was to learn more about the distribution of signs and ask whether it is Zipfian. Furthermore, the study of sign language distributions can contribute to our understanding of the possible sources of Zipfian distributions. One open question is how such distributions emerge and whether they are present from the initial stages of language. The data we have from spoken language is insufficient to answer this question as we do not have data on the structure of spoken languages at the time of their emergence. Pidgin and creole languages, which emerge from the need of speakers of two mutually unintelligible languages to communicate, can inform us of how structure emerges (Bickerton, 1983; Siegel, 2008). However, the structure of these languages is heavily influenced by existing languages, making them less suitable for questions of language emergence (Blasi et al., 2017). In contrast, there are multiple sign languages emerging at this time in communities of language users who did not previously have a shared language. Some emerged when deaf people were brought together for educational purposes, like Nicaraguan Sign Language (Senghas & Coppola, 2001) or Israeli Sign Language (Meir & Sandler, 2007). Others, like Al Sayyid Bedouin Sign Language in Israel (Meir et al. 2010) and Kata Kolok in Bali (De Vos, 2012) emerged in communities with a relatively high incidence of deafness. These emerging languages grow and develop over time, providing a window into how linguistic structure initially emerges and how it changes through transmission and learning (Sandler, 2016). For some of these languages, like Israeli Sign Language, there is a corpus of the language over the first three generations of signers (Stamp et al., 2022). Using such corpora, we can ask whether the distribution of signs changes during emergence, and how it is affected by changing lexicon size. In particular, we may expect the distribution to start out skewed, but not Zipfian, when the initial lexicon is both smaller and more variable (i.e., when different signs are used by different signers for the same meaning) and become more Zipfian over time, as it is learned and transmitted by multiple signers.

## Conclusions

Despite being regarded as a hallmark of language, the presence of Zipfian distributions in language has only been shown in spoken languages. In this study, we used three sign language corpora to show that the distribution of signs is Zipfian,



and similar to that found in spoken languages. These findings add novel insights to our understanding of sign language use and to the generality of Zipfian distributions in language.

## Funding

The funding for this project was provided by Israeli Science Foundation grant 445/20 awarded to I. Arnon.

## Acknowledgements






The data in this article was collected from three corpora, readily available online: (1) the British Sign Language (BSL) corpus, (2) the German Sign Language (DGS) corpus, and (3) the Sign Language of the Netherlands (NGT) corpus. We thank the contributors of these corpora.

The BSL data in this article was collected from the British Sign Language Corpus Project (BSLCP) at University College London, funded by the Economic and Social Research Council UK (RES-620-28-6001), and supplied by the CAVA repository. The data are copyright. The DGS data was collected as part of the DGS-Korpus project hosted by the Academy of Sciences in Hamburg at the Institute for German Sign Language and Communication of the Deaf, Hamburg University. The NGT data is taken from the *Corpus NGT* that was created by Onno Crasborn, Inge Zwitserlood and Johan Ros at Radboud University between 2008 and 2011.













## Data and code availability

The full scripts are available via the Open Science Framework (OSF) website, at: <https://osf.io/eh8sy/> (last access 12 March 2024).

## References

-  Bank, R., Crasborn, O., & van Hout, R. (2016). The prominence of spoken language elements in a sign language. *Linguistics*, 54(6), 1281–1305.
-  Bentz, C., Alikaniotis, D., Samardžić, T., & Buttery, P. (2017). Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. *Journal of Quantitative Linguistics*, 24(2–3), 128–162.
-  Bickerton, D. (1983). Creole languages. *Scientific American*, 249(1), 116–123.
-  Blasi, D. E., Michaelis, S. M., & Haspelmath, M. (2017). Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour*, 1(10), 723–729.
-  Borstell, C. (2022). Searching and utilizing corpora [Review of Searching and utilizing corpora]. In J. Fenlon & J. A. Hochgesang (Eds.), *Signed Language Corpora*, pp. 115–118. Gallaudet University Press.

- Brennan, M. (1982). An introduction to the visual world of BSL. In D. Brien (Ed.), *Dictionary of British Sign Language/English*, pp. 1–133. Faber & Faber.
- Brentari, D. (1998). *A prosodic model of sign language phonology*. MIT Press.
- [doi](#) Brentari, D. (2006). Effects of language modality on word segmentation: An experimental study of phonological factors in a sign language. In S. Anderson, L. Goldstein, & C. Best (Eds.), *Papers in laboratory phonology* (Vol. 8), pp. 155–164. De Gruyter Mouton.
- [doi](#) Brentari, D., & Goldin-Meadow, S. (2017). Language emergence. *Annual review of linguistics*, 3, 363–388.
- [doi](#) Brentari, D., & Padden, C.A. (2001). Native and foreign vocabulary in American Sign Language: A lexicon with multiple origins. In D. Brentari (Ed.), *Foreign vocabulary in sign languages: A cross-linguistic investigation of word formation*, pp. 87–119. Lawrence Erlbaum.
- [doi](#) Caselli, N.K., & Pyers, J.E. (2017). The road to language learning is not entirely iconic: Iconicity, neighborhood density, and frequency facilitate acquisition of sign language. *Psychological Science*, 28(7), 979–987.
- [doi](#) Caselli, N., Sevcikova Sehyr, Z., Cohen-Goldberg, A.M., & Emmorey, K. (2017). ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods*, 49(2), 784–801.
- [doi](#) Chater, N., & Brown, G.D. (1999). Scale-invariance as a unifying psychological principle. *Cognition*, 69(3), B17–B24.
- [doi](#) Christiansen, M.H., & Chater, N. (2008). Language as shaped by the brain. *The Behavioral and Brain Sciences*, 31(5), 489–508; discussion 509–558.
- [doi](#) Clauset, A., Shalizi, C.R., & Newman, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661–703.
- [doi](#) Clerkin, E.M., Hart, E., Rehg, J.M., Yu, C., & Smith, L.B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372 (1711), 1–8.
- [doi](#) Cooperrider, K., Abner, N., & Goldin-Meadow, S. (2018). The palm-up puzzle: Meanings and origins of a widespread form in gesture and sign. *Frontiers in Communication*, 3, 1–14.
- Cormier, K., Fenlon, J., Gulamani, S., & Smith, S. (2017). BSL corpus annotation conventions. *Annotation Convention*, Vol. 3, 2–15.
- [doi](#) Cormier, K., Quinto-Pozos, D., Sevcikova, Z., & Schembri, A. (2012). Lexicalisation and de-lexicalisation processes in sign languages: Comparing depicting constructions and viewpoint gestures. *Language & Communication*, 32(4), 329–348.
- [doi](#) Coupé, C., Oh, Y., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9), eaaw2594.
- Crasborn, O. & Zwitserlood, I. (2008). The Corpus NGT: An online corpus for professionals and laymen, In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd (eds.), *Construction and exploitation of Sign Language corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, pp. 44–49. ELDA.
- Crasborn, O., Bank, R., Zwitserlood, I., Van Der Kooij, E., De Meijer, A., Sáfár, A., & Ormel, E. (2015). *Annotation conventions for the Corpus NGT, version 3*. Centre for Language Studies & Department of Linguistics, Radboud University Nijmegen.

- Crasborn, O., Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In: *6th International Conference on Language Resources and Evaluation (LREC 2008)/3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pp. 39–43.
- Crasborn, O., Zwitterlood, I. & Ros, J. (2008). *The Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands*. Centre for Language Studies, Radboud University Nijmegen. Available at [https://archive.mpi.nl/tla/islandora/object/tla:1839\\_00\\_0000\\_0000\\_0004\\_DF8E\\_6?asOfDateTime=2018-03-02T11:00:00.000Z](https://archive.mpi.nl/tla/islandora/object/tla:1839_00_0000_0000_0004_DF8E_6?asOfDateTime=2018-03-02T11:00:00.000Z) (last access 12 March 2024). ISLRN: <https://www.islrn.org/resources/175-346-174-413-3/> (last access 13 March 2024).
- De Vos, C. (2012). Sign-spatiality in Kata Kolok: How a village sign language in Bali inscribes its signing space [Doctoral dissertation, Radboud University Nijmegen].
-  Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25(2), 108–127.
-  Ellis, N.C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, 24(2), 143–188.
-  Emmorey, K. (2001). *Language, cognition, and the brain: Insights from sign language research*. Psychology Press.
-  Erting, C.J., Prezioso, C., & O’Grady Hynes, M. (1990). The interactional context of deaf mother-infant communication. In *From gesture to language in hearing and deaf children*, pp. 97–106. Springer Verlag.
-  Fenlon, J., Cormier, K., & Schembri, A. (2015a). Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography*, 28(2), 169–206.
-  Fenlon, J., Schembri, A., Johnston, T., & Cormier, K. (2015b). Documentary and corpus approaches to sign language research. *Research methods in sign language studies: A practical guide*, pp. 156–172. Wiley-Blackwell.
-  Fenlon, J., Schembri, A., Rentelis, R., Vinson, D., & Cormier, K. (2014a). Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua*, 143, 187–202.
- Fenlon, Jordan, Kearsy Cormier, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam, & Bencie Woll. (2014b). *BSL SignBank: A lexical database of British Sign Language* (1st edn). London: Deafness, Cognition and Language Research Centre, University College London.
-  Ferrer-i-Cancho, R. & Solé, R.V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3), 788–791.
-  Ferrer-i-Cancho, R. (2016). Compression and the origins of Zipf’s law for word frequencies. *Complexity*, 21(S2), 409–411.
-  Gibson, E., Futrell, R., Piantadosi, S.T., Dautriche, I., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5), 389–407.
-  Goldberg, A.E., Casenhiser, D.M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, 15(3), 289–316.
-  Hendrickson, A.T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition* 189, 11–22.

- Holzrichter, A. S., & Meier, R. P. (2000). Child-directed signing in American sign language. In C. Chamberlain, J. P. Morford, & R. I. Mayberry (Eds.), *Language acquisition by eye*, pp. 25–40. Lawrence Erlbaum.
- [doi](#) Johnston, T. (2012). Lexical frequency in sign languages. *Journal Of Deaf Studies And Deaf Education*, 17(2), 163–193.
- Johnston, T., & De Beuzeville, L. (2016). Auslan corpus annotation guidelines. *Auslan Corpus*.
- [doi](#) Johnston, T., & Schembri, A. (2007). *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press.
- [doi](#) Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15, 10–131.
- Klima, E. S., & Bellugi, U. (1979). *The Signs of Language*. Harvard University Press.
- Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., Böse, O., Jahn, E., Schulder, M. (2020a). *MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release [Dataset]*. Hamburg University.
- Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. (2020b). *Öffentliches DGS-Korpus: Annotationskonventionen / Public DGS Corpus: Annotation conventions*. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University.
- [doi](#) Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, 127(3), 439–453.
- Langer, G., Müller, A., & Wähl, S. (2018). Queries and Views in iLex to Support Corpus-based Lexicographic Work on German Sign Language (DGS). In M. Bono, E. Efthimiou, S. E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, J. Mesch & Y. Osugi (eds.) *Involving the Language Community. Proceedings of the 8th Workshop on the Representation and Processing of Sign Language. 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*, pp. 107–114. ELRA.
- Lavi-Rotbain, O., & Arnon, I. (2019). Children learn words better in low entropy. *Proceedings of the 41th Annual Conference of the Cognitive Science Society*, pp. 631–637. Cognitive Science Society.
- [doi](#) Lavi-Rotbain, O., & Arnon, I. (2020). The learnability consequences of Zipfian distributions: Word segmentation is facilitated in more predictable distributions. *PsyArXiv*. [preprint MS, pp. 1–17]
- [doi](#) Lavi-Rotbain, O., & Arnon, I. (2021). Visual statistical learning is facilitated in Zipfian distributions. *Cognition*, 206, 104492, 1–8.
- [doi](#) Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of Zipfian distributions in language. *Cognition*, 223, 105038, 1–14.
- [doi](#) Lavi-Rotbain, O., & Arnon, I. (2023). Zipfian distributions in child-directed speech. *Open Mind*, 7, 1–30.
- [doi](#) Liddell, S. K. (2003). *Grammar, gesture and meaning in American Sign Language*. Cambridge University Press, Cambridge.
- [doi](#) Lillo-Martin, D. C., & Gajewski, J. (2014). One grammar or two? Sign Languages and the nature of human language. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(4), 387–401.

- Lillo-Martin, D., & Klima, E. S. (1990). Pointing out differences: ASL pronouns in syntactic theory. *Theoretical Issues in Sign Language Research*, 1, 191–210
- [doi](#) Linders, G. M., & Louwerse, M. M. (2020). Zipf's law in human-machine dialog. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pp. 1–8. Association for Computing Machinery.
- Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication Theory*, 2, 486–502.
- [doi](#) Manin, D. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, 32(7), 1075–1098.
- Masataka, N., Morford, J., & Mayberry, R. (2000). The role of modality and input in the earliest stage of language acquisition: Studies of Japanese Sign Language. In Chamberlain, C., Morford, J. P., & Mayberry, R. (Eds.), *Language acquisition by eye*, pp. 3–24. Lawrence Erlbaum.
- McDonald, B. H. (1985). Productive and frozen lexicon in ASL: An old problem revisited. In W. Stokoe & V. Volterra (Eds.), *SLR '83: Proceedings of the 3rd International Symposium on Sign Language Research*, pp. 254–259. CNR & Linstok Press.
- [doi](#) McKee, D., & Kennedy, G. D. (2006). The distribution of signs in New Zealand Sign Language. *Sign Language Studies*, 6(4), 372–390.
- [doi](#) Mehri, A., & Jamaati, M. (2017). Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. *Physics Letters, Section A: General, Atomic and Solid State Physics*, 381(31), 2470–2477.
- Meier, R. (1990). Person deixis in ASL. In S. Fischer & P. Siple (Eds.), *Theoretical issues in sign language research*, Vol. 1, pp. 175–190. University of Chicago Press.
- [doi](#) Meir, I., & Sandler, W. (2007). *A language in space: the story of israeli sign language*. Psychology Press.
- [doi](#) Meir, I., Sandler, W., Padden, C., & Aronoff, M. (2010). Emerging sign languages. In M. Marschark & P. Spencer (Eds.), *Oxford handbook of deaf studies, language, and education*, Vol. 2, pp. 267–280. Oxford University Press.
- [doi](#) Morford, J. P., & MacFarlane, J. (2003). Frequency Characteristics of American Sign Language. *Sign Language Studies*, 3(2), 213–225.
- [doi](#) Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5), 323–351.
- [doi](#) Novogrodsky, R., & Meir, N. (2020). Age, frequency, and iconicity in early sign language acquisition: Evidence from the Israeli Sign Language MacArthur–Bates Communicative Developmental Inventory. *Applied Psycholinguistics*, 41(4), 817–845.
- [doi](#) Orfanidou, E., Adam, R., Morgan, G., & McQueen, J. M. (2010). Recognition of signed and spoken language: Different sensory inputs, the same segmentation procedure. *Journal of Memory and Language*, 62(3), 272–283.
- [doi](#) Orfanidou, E., McQueen, J. M., Adam, R., & Morgan, G. (2015). Segmentation of British Sign Language (BSL): Mind the gap! *Quarterly Journal of Experimental Psychology*, 68(4), 641–663.
- [doi](#) Perlman, M., Little, H., Thompson, B., & Thompson, R. L. (2018). Iconicity in signed and spoken vocabulary: a comparison between American Sign Language, British Sign Language, English, and Spanish. *Frontiers in psychology*, 9, 1433, pp. 2–14.

- doi Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5), 1112–1130.
- Sandler, W., & Lillo-Martin, D. (2001). Natural sign languages. In M. Aronoff and J. Rees-Miller (Eds.), *Handbook of linguistics*, pp. 533–562. Blackwell.
- doi Sandler, W., & Lillo-Martin, D. (2006). *Sign Language and linguistic universals*. Cambridge University Press.
- Sandler, Wendy. (2016). What comes first in language emergence? In N. Enfield (Ed.) *Dependency in language: On the causal ontology of language systems* (Studies in Diversity in Linguistics 99), pp. 67–86. *Language Science Press*.
- Schembri, Adam, Jordan Fenlon, Ramas Rentelis, & Kearsy Cormier. (2017). *British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008–2017* (3rd edn). University College London. Available at <https://www.bslcorpusproject.org> (last access 12 March 2024).
- Schick, B. S. (1987). The acquisition of classifier predicates in American Sign Language. [Doctoral Dissertation, Purdue University Indiana].
- doi Schuler, K. D., Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). The Effect of Zipfian Frequency Variations on Category Formation in Adult Artificial Language Learning. *Language Learning and Development*, 13(4), 357–374.
- doi Sehry, Z. S., Caselli, N., Cohen-Goldberg, A. M., & Emmorey, K. (2021). The ASL-LEX 2.0 Project: A database of lexical and phonological properties for 2,723 Signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education*, 26(2), 263–277.
- doi Semple, S., Ferrer-i-Cancho, R., & Gustison, M. L. (2022). Linguistic laws in biology. *Trends in Ecology and Evolution*, 37(1), 53–66.
- doi Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological science*, 12(4), 323–328.
- doi Shufaniya, A., & Arnon, I. (2022). A cognitive bias for Zipfian distributions? Uniform distributions become more skewed via cultural transmission. *Journal of Language Evolution*, 7(1), 59–80.
- doi Siegel, J. (2008). *The emergence of pidgin and creole languages*. Oxford University Press.
- doi Smith, R. G., & Hofmann, M. (2020). Lexical frequency analysis of Irish Sign Language. *TEANGA, the Journal of the Irish Association for Applied Linguistics*, 11, 18–47.
- Stamp, R., Ohanin, O. & Lanesman, S. (2022). The Corpus of Israeli Sign Language. *Conference Proceedings (LREC): Language Resources (LRs) and Evaluation for Human Language Technologies (HLT)*, pp. 192–197. ELRA.
- Sümer, B., Grabitz, C., & Küntay, A. (2017). Early produced signs are iconic: Evidence from Turkish Sign Language. In *The 39th Annual Conference of the Cognitive Science Society (CogSci 2017)*, pp. 3273–3278. Cognitive Science Society.
- Supalla, T. (1982). Structure and acquisition of verbs of motion and location in American Sign Language. [Ph.D. dissertation, University of California at San Diego].
- doi Talmy, L. (2001, June). Spatial structuring in spoken and signed language. *Annual Meeting of the Berkeley Linguistics Society*, 27(1), pp. 271–300.
- doi Woltz, D. J., Gardner, M. K., Kircher, J. C., & Burrow-Sanchez, J. J. (2012). Relationship between perceived and actual frequency represented by common rating scale labels. *Psychological Assessment*, 24(4), 995–1007.

Zipf, G.K. (1949). *Human behavior and the principle of least effort. Human behavior and the principle of least effort*. Addison-Wesley Press.

Zwitserslood, I. (2012). Chapter 8 Classifiers. In R. Pfau, M. Steinbach, & B. Woll (Eds.), *Sign Language: An international handbook*, pp. 158–181. De Gruyter Mouton.

## Appendix 1

### Annotation differences between the three corpora

	DGS	BSL	NGT
Depicting constructions	\$PROD	prefix DS (depicting constructions) Shapes: DSEW (Depicting sign: entity (whole)), DSEP (Depicting sign: entity (part)), DSH (Depicting sign: Handling), DSS (Depicting sign: showing size and shape) Movements: MOVE, PIVOT, AT, BE Type-like depicting signs: DSEW(1-VERT), DSEW(1-HORI), DSEW(2-DOWN), DSEW(2-HORI), DSEW(BENT2-HORI), DSEW(5-HORI), DSEW(FLAT-LATERAL), DSEW(FLAT-HORI)	Shapes: 1, 1_curved, V, 3, 4, 5, B, B_curved, O, C, C_spread, Beak, Beak_open, Baby_O, Baby_beak, T, Baby_C, Baby_beak_open, S, money, Y Movements: MOVE, PIVOT, AT, BE
Pointing signs	\$INDEX \$INDEX2 \$INDEX4 \$INDEX-TO-SCREEN1. \$INDEX-ORAL1 \$INDEX-AREA1 I1 YOU1	PT:PRO1SG PT:PRO2SG PT:PRO3SG PT:PRO1PL PT:PRO2PL PT:PRO3PL PT:DET PT:DETPL PT:LOC PT:LOCPL PT:POSS1SG	PT-1hand PT-Bhand PT-1hand:1 PT-Bhand:B PT:BL PT:up PT:down PT:thumb PT:index PT:mid PT:ring

	DGS	BSL	NGT
		PT:POSS <sub>2</sub> SG	PT:pinky
		PT:POSS <sub>3</sub> SG	PT:index-mid-ring
		PT:POSS <sub>1</sub> PL	PT-Vhand:index-
		PT:POSS <sub>2</sub> PL	mid
		PT:POSS <sub>3</sub> PL	PT-3hand:index-
		PT:BODY	mid-ring
		PT:LBUOY	PT:arc
		PT:FBUOY	PT:alt
		PT:BUOY	
		PT:	
Buoys	\$LIST <sub>1</sub>	LBUOY	–
	\$LIST <sub>2</sub>	PBUOY	
	\$LIST-TO-LIST <sub>1</sub>	FBUOY	
	\$LIST-TO- REMOVE <sub>1</sub> A	TBUOY	
	\$LIST- TOGETHER <sub>1</sub> C		
Gestures	\$GEST^	G: (all gestures)	PO (palm up)
	\$GEST-NM^	G:CA: (Tokens of	PB (palm down)
	\$GEST-OFF^	constructed action are also	PV (palm forwards)
	Beside these	recognized as instances of	POS (the word
	collective types, there	gestural activity)	category or
	are several gesture		categories of the
	type entries specified		particular token in
	by form and meaning		that particular
	much like lexical		context)
	signs: e.g. \$GEST- TO-PONDER <sub>1</sub> ^		



## Appendix 2

Proportion of sign categories in the minimally excluded corpora

	BSL	DGS	NGT
Fully lexical signs	60.67%	73.80%	63.92%
Depicting constructions	2.35%	1.96%	5.18%
Pointing signs	21.96%	13.51%	18.72%
Buoys	0.47%	0.45%	–
Gestures	7.93%	8.61%	10.39%
Uncertain signs	3.42%	0.00%	0.35%
Mouthing	–	0.52%	–
Extra linguistic manual activity	–	0.11%	0.29%
Fingerspelling	2.57%	0.67%	1.11%
Names	1.08%	0.26%	0.04%
Initialization	–	0.03%	–
Cued speech	–	0.07%	–

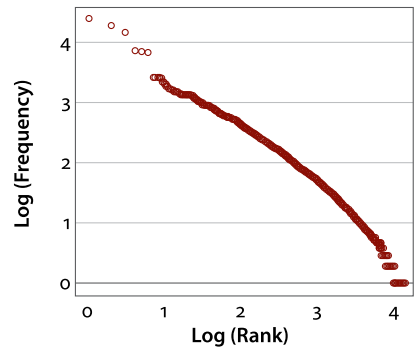
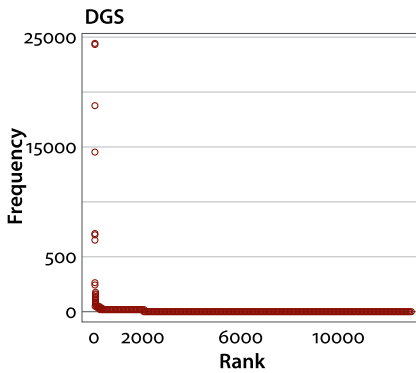
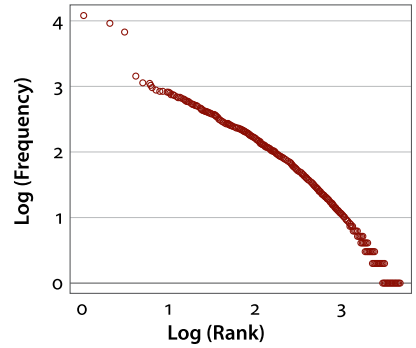
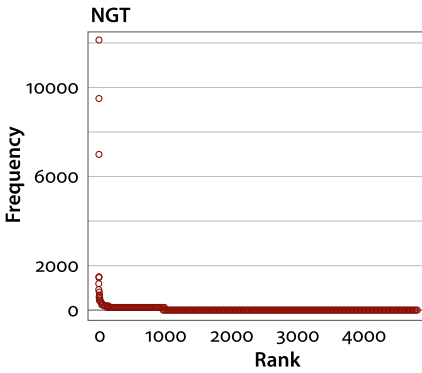
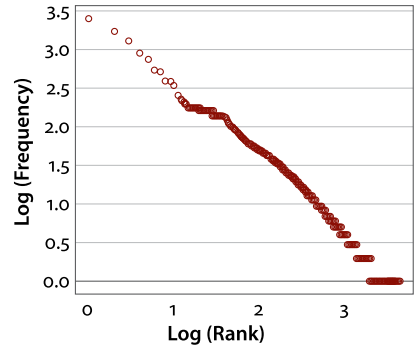
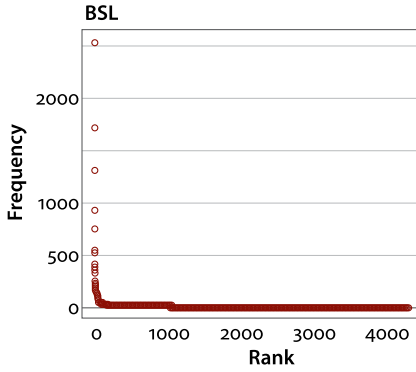
## Appendix 3

Assessing how well the distributions of the signs in the complete datasets fit zipf's Law

	No. of tokens	No. of types	Log*log correlation	Frequency range	$\alpha$	$\beta$	Pearson's r (observed* expected)
BSL	34,909	4,275	–0.98	1–2,542	1.03	1.02	0.99
DGS	353,227	13,120	–0.97	1–24,408	0.99	0.43	0.98
NGT	108,434	4,796	–0.98	1–12,101	1.04	0.10	0.96

As shown in the table above, the Pearson correlations between frequency and rank in log space in the three complete corpora were close to  $-1$ , indicating a good fit to a Zipfian distribution (BSL:  $R_2 = -0.98$ , DGS:  $R_2 = -0.97$ , NGT:  $R_2 = -0.98$ ). In addition, the Pearson correlation between the observed and expected frequencies of all corpora is close to  $1$ , indicating a very good fit to a Zipfian distribution.

## Appendix 4





The distribution of raw frequency (Left) and log frequency (Right) for the three corpora


## Address for correspondence


Inbal Kimchi  
Department of Cognitive Sciences  
Jack, Joseph and Morton Mandel School for Advanced Studies in the Humanities  
The Hebrew University of Jerusalem  
Mount Scopus  
Jerusalem 9190501  
Israel  
inbal.kimchi@gmail.com  
 <https://orcid.org/0000-0002-2553-7117>

## Co-author information

Lucie Wolters  
Department of Cognitive Sciences  
The Hebrew University of Jerusalem

 [lucia.wolters@mail.huji.ac.il](mailto:lucia.wolters@mail.huji.ac.il)  
 <https://orcid.org/0009-0003-7346-9037>

Rose Stamp  
Department of English Literature &  
Linguistics  
Bar Ilan University  
[rose\\_stamp@hotmail.com](mailto:rose_stamp@hotmail.com)  
 <https://orcid.org/0000-0002-1993-559X>

Inbal Arnon  
Department of Psychology  
The Hebrew University of Jerusalem  
[inbal.arnon@gmail.com](mailto:inbal.arnon@gmail.com)  
 <https://orcid.org/0000-0001-8934-718X>

## Publication history

Date received: 11 July 2023  
Date accepted: 13 March 2024  
Published online: 2 April 2024

Copyright of Gesture is the property of John Benjamins Publishing Co. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.