10 Documentary and Corpus Approaches to Sign Language Research

Jordan Fenlon, Adam Schembri, Trevor Johnston, and Kearsy Cormier

Chapter Overview

In this chapter we discuss some key aspects of the methodology associated with sign language documentation and corpus-based approaches to sign language research. We first introduce the field of sign language corpus linguistics, carefully defining the term "corpus" in this context and discussing the emergence of technology that has made this new approach to sign language research possible. We then discuss specific details of the methodology involved in corpus-building, such as the recruitment of participants, the selection of language activities for the corpus, and the set-up for filming. We move on to a discussion of annotation for corpora, with a focus on the use of ID glossing. We close with a brief discussion of online archiving and accessibility.

Research Methods in Sign Language Studies: A Practical Guide, First Edition.
Edited by Eleni Orfanidou, Bencie Woll, and Gary Morgan.

Introduction

A modern linguistic corpus is understood to refer to a large collection of spoken, written, or signed language data (with associated metadata) that is in machine-readable form, is (as far as possible) maximally representative of the language and its users, and can be consulted to study the type and frequency of constructions in that language. One well-known example of a modern linguistic corpus is the British National Corpus (BNC) of English. At 100 million words, this corpus consists of samples of spoken and written British English that have been carefully selected from a wide range of sources so as to be representative of British English during the late twentieth century and is often consulted by researchers for linguistic analysis (Rayson, Wilson, and Leech, 2002). Like most modern linguistic corpora, the texts within the BNC have, appended to them, linguistic annotations in the form of grammatical tags that provide another level of detail to the corpus (although other types of linguistic annotation are often available in modern corpora; see McEnery, Xiao, and Tono, 2006). Subsequently researchers can use the BNC to examine not only the frequency with which a given word occurs, but also the relative frequencies of each of its grammatical functions – in order to gain a better picture of language use in British English.

Sign language linguists (like their spoken language counterparts) have generally used the term "corpus" to refer to any kind of data set upon which a linguistic description or analysis has been based, or to any collection of video recordings (Lucas, Bayley, and Valli, 2001; Neidle and Vogler, 2012). These uses of the term are at odds with the description given above; and none of the data sets they refer to can be said to constitute a linguistic corpus in the modern sense. For instance, the data set subjected to linguistic analysis in these cases sometimes consists of elicited examples from a comparatively small number of native signers and, as such, cannot be said to be representative of everyday sign language. Other – larger, more spontaneous – collections of video recordings have often not been digitized or appropriately annotated and thus are not currently machine-readable in the sense explained above.

Although sign language corpora are now being developed, they and subsequent studies based on these corpora represent an emerging practice in sign language research; thus the field of corpus-based approaches to sign language linguistics is still very much in its infancy. Most sign language corpora are currently works in progress: they are primarily large data sets consisting of spontaneous and elicited signing that have been digitized and are currently undergoing linguistic annotation – the latter proving to be the most laborious stage in the creation of a sign language corpus. Depending on how much work has been completed for their respective sign language lexicons (i.e., whether a lexical database is available to support annotation), sign language corpus projects are at different stages in the annotation process, some having to take on the task of developing a lexical database concurrently with corpus creation.

The Need for Corpora

Why do we need sign language corpora? One important reason is that there is much work that needs to be done within the field of sign language linguistics to further our understanding of the structure and use of these languages. There is a pressing need

to test the claims made by many existing linguistic descriptions and analyses of sign languages, because they have often been based on limited data sets from a small number of signers. This reliance on small data sets is problematic when one considers that sign language use is commonly reported to be highly variable (Schembri and Johnston, 2012). The variability owes much to the fact that sign languages exist in unique sociolinguistic circumstances: they are young, minority languages, with few native signers and with an interrupted pattern of intergenerational transmission. As a consequence, it is often difficult even for native signers to be certain about what is and is not an acceptable construction in their language. The justification for corpora is supplied in this case by the assumption that the processing of large amounts of annotated texts can reveal patterns of language use and structure not available to everyday user intuitions, or even to expert detailed analysis.

There are also potential benefits to the deaf[1] community in the creation of sign language corpora. Further empirical research on sign language structure and on the documentation of signs used in the language (e.g., via a corpus-based dictionary or lexical database) will inform and improve sign language teaching materials – which will, in turn, lead to the improvements in the training of sign language teachers and interpreters and in the education of deaf children. In some cases sign language corpora have themselves been used as tools in educational settings, in the training of interpreters or teaching linguistics students, both groups being invited to explore and annotate texts from within corpora for specific sign language phenomena (e.g., Leeson, 2010; Mesch, Nilsson, Wallin, and Bäckström, 2010). These corpora also provide an important means of recording sign languages as they are used today for posterity, particularly since they are now increasingly recognized as endangered languages (Johnston, 2004; Nonaka, 2004; Schembri, 2010). Furthermore, if some texts are oriented toward specific topics relevant to the field of Deaf studies (e.g., if participants are encouraged to describe their educational experience), this will also enhance the status of a corpus as an important and valuable resource with wide-ranging applications.

The Emergence of Sign Language Corpus Linguistics

Modern sign language corpora have not been possible until relatively recently, due to the absence of a widely recognized and favorable transcription system for sign language data and the lack of suitable technology for sign language capture and secondary processing (see Crasborn, Efthimiou, Hanke, Thoutenhoofd, and Zwitserlood, 2008). In recent years advancements in technology have gone some way toward addressing these shortcomings. Coupled with a growing recognition, among the sign language research community, of the role that larger empirical data sets can play for testing hypotheses about sign language structure and use, this has led to the rapid emergence of a number of sign language corpus projects worldwide, as described below.

Prior to this, few sign language texts of any kind could be subjected to linguistic annotation (e.g., be tagged for parts of speech) for the purposes of studying the frequency of occurrence of specific signs or grammatical constructions. The absence of such texts reflected the fact that sign languages do not have a widely accepted writing system (although Sutton SignWriting is used by some for this purpose: see

www.signwriting.org), or even a standardized specialist notation system that can be used for transcription (a number of such systems exist, such as the Hamburg Notation System, but are time-consuming to use and require special font software; see Prillwitz, Leven, Zienert, Hanke, and Henning, 1989). Instead, researchers have often used contextually based glosses in place of such systems, particularly for studies of sign language grammatical structure. This practice is not without its disadvantages, since contextually based glosses do not provide information on a sign's form, do not have standardized ways to accurately represent some key formational aspects involved in articulation (e.g., the spatial and non-manual features of sign language utterances), and are likely to be inconsistent across studies, as one cannot be certain which specific lexical variant a particular gloss represents in each case (the disadvantages of relying on glossing in sign language research have long been recognized; see Johnston, 1991; Pizzuto and Pietrandrea, 2001; Frishberg, Hoiting, and Slobin, 2012). Furthermore, the use of glosses (or even a dedicated notation system) in these cases is also a major problem, since one sacrifices any connection with the primary data source, which, in itself, is far more likely to be informative for sign language research.

The development of time-aligned video annotation software, together with more widely available computer capacity for the capture and storage of large amounts of digital video, provided a solution for the "transcription problem" described above and set the scene for the emergence of corpus-based approaches to sign language linguistics. One multimedia annotation software program in widespread use among sign language researchers is ELAN (Wittenburg, Brugman, Russel, Klassmann, and Sloetjes, 2006). Using software like ELAN, transcriptions can be directly time-aligned with a media file representing a signed segment. Overlapping annotations can also be stored on separate tiers that represent a different level of analysis or a different articulator (e.g., tiers can be specifically devoted to grammatical tags denoting a sign's function or to a non-manual articulator such as the eyebrows). The use of such video annotation software also means that the source of the digital video material can remain the primary data rather than being replaced by a transcription that is less transparent (one needs only to look at the media file linked to an ELAN annotation file to see a specific sign's form). This innovation also meant that researchers could now systematically use glosses as annotations rather than transcriptions to which further linguistic annotations are appended. A systematic approach to glossing within sign language corpora, known as ID glossing (Johnston, 2010), involves identifying conventional linguistic units and types with the help of unique sign identifiers (i.e., ID glosses). We discuss ID glosses and their importance in more detail in the section on corpus methods below.

Further developments to ELAN over the years have enhanced capacity for working on sign language corpora (Crasborn and Sloetjes, 2008). In ELAN it is now possible to extract frequency statistics for any annotation on any tier and to examine the environments in which the frequencies occur. These searches can be modified further by specifying co-occurring annotations within tiers, which may include metadata values associated with each individual ELAN file (such as age, gender, or region). With these capabilities, the possibilities available to sign language researchers who use ELAN are numerous. For example, one can search multiple files that represent a wide cross-section of a given sign language community for all tokens of a specific sign (glossed on one tier), co-occurring with a particular grammatical function

(annotated on a grammatical category tier), and/or certain social factors (e.g., age) in order to gain a more thorough picture of patterns in sign language use.

Other video annotation software has also been developed for use with sign language corpora. In particular, iLex has been created by the team at the Centre for German Sign Language at the University of Hamburg, and, while it is similar to ELAN, it has the additional advantage of making it possible to link glosses in the annotation file to entries in a lexical database (Hanke and Storz, 2008). Work is currently underway to facilitate links between ELAN annotation files and lexical database tools such as LEXUS (Crasborn, Hulbosch, and Sloetjes, 2012).

Data Collection

The first sign language corpus project began in Australia in 2004 with a digital video archive of recordings of 100 deaf native or early learner/near-native signers of Auslan; the expectation was that it would later become a large machine-readable corpus in the modern linguistic sense. A similar project on a smaller scale began in Ireland in the same year to collect Irish Sign Language (ISL) data. Since then a number of other sign language corpus projects have begun (e.g., Netherlands Sign Language (NGT), British Sign Language (BSL), German Sign Language (DGS), Swedish Sign Language, and Polish Sign Language). For most of these projects, at the time of writing, the first stage (i.e., data collection and archiving) has been completed and, for NGT and BSL, the video data and some ELAN annotation files have been made openly available online (some of the Auslan corpus video data and ELAN annotation files are available online through the Endangered Languages Archive at the University of London, but they have restricted access). In this section we elaborate on the design criteria for these corpus projects, with a specific focus on the BSL and Auslan corpora. We also discuss specific issues unique to the design of sign language corpora, such as representativeness and the problem of the observer's paradox (Schembri, 2010; Schembri, Fenlon, Rentelis, Reynolds, and Cormier, 2013).

Sites and participants

One of the key criteria used to define a linguistic corpus is that it should be representative of a language's variety (McEnery and Wilson, 2001). This is important if the corpus is to be used as a basis for making generalizations about language. Sampling provides a way to ensure as wide a representation as possible and is based on explicit linguistic criteria. However, there are challenges facing those involved in sign language corpus design when representativeness is to be achieved. It is very likely that not enough is known about the size of the deaf community and its distribution in order to create a representative data set. For all sign languages, published estimates as to the size of the signing population are known to vary or are unreliable, if they exist at all (Johnston, 2004). Faced with this lack of information, sign language corpus projects have tended to favor the methodology employed in studies investigating sociolinguistic variation (e.g., Lucas et al., 2001), where participants are selected as

part of a quota sample, according to a set of demographic variables (e.g., gender, age, region, ethnicity, socioeconomic class, and age of sign language acquisition) that are considered relevant to deaf communities. Although the resulting data set may or may not be representative of the wider deaf community (considering that many deaf signers learn to sign in later childhood, or even in adulthood), recruiting participants via a quota sample with these demographic variables does take us some way toward capturing the full range of variability in the deaf community.

Sign language corpus projects have sought to recruit participants from a number of cities across their respective countries, partly as a way to achieve representativeness generally, but also because regional variation within sign language communities is known to be significant (Johnston and Schembri, 2007; Sutton-Spence and Woll, 1999). For the BSL corpus, eight cities in the United Kingdom were selected (Belfast, Birmingham, Bristol, Cardiff, Glasgow, London, Manchester, and Newcastle) that broadly represented the major countries within the United Kingdom (England, Scotland, Wales, Northern Ireland) and five of the most important regions in England (the country in the United Kingdom with the largest population). For Auslan, the five largest cities were selected (Sydney, Melbourne, Brisbane, Perth and Adelaide), because this is where both the general population and the deaf community are concentrated (Australia is a highly urbanized country, with half of its total population in these five urban centers). For both projects, the selection of these regions was motivated partly by the fact that they were all sites of centralized deaf schools (although in some cases these schools have now closed due to the trend to mainstream deaf children in schools with their hearing peers) and partly by the fact that they contain thriving adult deaf communities that have developed around these schools. The decision to recruit from major cities is also strategic, since it is easier to find a sufficient number of signers from a variety of backgrounds (e.g., of different ages, from deaf and hearing families, and so on) in these areas. In the BSL project, which drew on methodology employed in previous sociolinguistic studies of American Sign Language (ASL) and Auslan, care was taken to ensure that participants had lived or worked in a given region for a minimum amount of time (10 years) so that they could be considered representative of the signing used in that region.

It is well known that the age at which a child is exposed to sign language as a first language can affect language proficiency in adulthood (Cormier, Schembri, Vinson, and Orfanidou, 2012; Emmorey, 2002). With this in mind the Auslan and BSL corpus projects aimed to recruit only signers who reported having learned to sign by 7 years of age. Each project also aimed to recruit a large number of native signers (i.e., signers who learnt to sign from birth from deaf parent(s) or an elder sibling). Native signers represent nearly a third of the participants in the BSL corpus (31 percent or 76–249 signers) and more than three quarters in the Auslan corpus (79 percent or 79–100 signers). One could argue that these figures represent a disproportionately large sample of native signers, given estimates for the total number of native signers in the deaf community: the actual figure is widely thought to be around 5 percent (see Mitchell and Karchmer, 2004). This focus is motivated, however, by a desire to document the most proficient sign language users in the community, and this is standard practice in collecting language samples in linguistics.

Age-related variation is also given important consideration in corpus design. In many corpus projects, the selection of participants according to age is motivated partly by a need to document language change as the result of changes in language

policy in deaf education. For example, participant selection for the BSL corpus was balanced across four age groups that roughly reflect different periods in deaf education. In the oldest age group, participants are likely to have been educated in residential schools for deaf children that focused on the use of fingerspelling and/or the development of speech-reading skills in the classroom, while BSL may have been used among school children in dormitories or on the playground. In contrast, participants in the youngest group (18–30 years of age) are much more likely to have been educated in mainstream school settings with few (if any) deaf peers, although some may have been educated in schools that used BSL as the medium of instruction (some schools in the UK have introduced bilingual approaches, using BSL together with English since the 1980s).

Other variables that have been given consideration in sign language corpus design are gender, ethnicity, and social class. These variables have all been found to be relevant for spoken languages, and there is evidence to suggest that they are significant for sign languages too (Lucas et al., 2001; McKee, Schembri, McKee, and Johnston, 2011; Schembri et al., 2009).

Although each of these corpus projects has clearly defined criteria regarding participant selection, some flexibility is often required. First, many deaf community members may be understandably nervous about being filmed for an open-access archive and may decline invitations to participate; this concern can significantly reduce the pool of potential participants. Second, the research design criteria may not be realistic, because – in the absence of specific statistics such as census information – they are based on guesswork about the composition of deaf communities. For example, the BSL corpus aimed to recruit 10 percent non-white participants in each of its eight regions. However, this proved difficult, as few non-white signers (e.g., black and South Asian deaf people) could be recruited in the smaller urban centers. Additionally, when one is faced with an increasingly mobile deaf community (and a small one too), fulfilling some quotas can prove a challenging task. In some of the smaller cities in the BSL corpus it was often difficult to find early learners in the youngest age group who had remained within their region for at least 10 years and who were willing to participate. Lastly, some demographics traditionally applied to the wider community are much more difficult to define within the deaf community. For example, in the BSL project, participants were identified according to two broad social classes – working class or middle class – on the basis of their educational background and occupation. However, the emergence of a professional class is a relatively recent phenomenon for most western sign language communities, following recent improvements in access to university education, and this means that it is difficult to achieve a balance of social classes across age groups. To overcome these problems in future, it is advised that projects either attempt to recruit more signers in the more populous regions known to contain a higher concentration of a specific group (e.g., non-white or middle-class signers) or adjust their design criteria slightly in some cases (e.g., by allowing participants who report learning to sign later or living in the area for a shorter period of time to be recruited), so that other key criteria that are underrepresented may be fulfilled. A sensitive awareness of these issues and of how they may affect the recruitment process is therefore required of the team in the planning stages of corpus-building, so that concessions may be made and alternatives quickly sought (e.g., less insistence on the recruitment of middle-class signers may be required in the older age groups).

Using the methodology outlined above, the Auslan corpus project recruited 100 deaf signers from Sydney, Melbourne, Brisbane, Perth and Adelaide, namely 20 participants at each site; this data set has been combined with other data from previous projects to create a total collection of data from 256 deaf Australians. The BSL corpus consists of a data set collected from 249 signers; 30 signers were filmed in Belfast, Birmingham, Cardiff, Glasgow, Manchester and Newcastle, 32 in Bristol, and 37 in London. The DGS corpus consists of an even larger data set, collected from 330 deaf signers in 13 German regions (Langer, 2012). The ISL corpus is based on data collected from 40 deaf signers (Leeson and Saeed, 2012), and the Swedish Sign Language corpus is based on data collected from 42 deaf signers (Mesch and Wallin, 2012). When these corpora are compared to one another, the difference in participant numbers between some of them may reflect the relative size of the general population in these countries and, consequently, the estimated size of the signing community. For example, the general population of the Republic of Ireland is much smaller than that of the United Kingdom and, consequently, a smaller number of deaf signers are said to use ISL (Leeson and Saeed, 2012).

Recruitment and filming

In order to ensure the recruitment of large numbers of participants from many regions, many of the sign language corpus projects mentioned so far have employed local members of the deaf community in the role of "fieldworkers," to locate suitable participants – a methodology first used by Ceil Lucas and colleagues (see Lucas et al., 2001). For both the BSL and the Auslan corpus projects, fieldworkers were deaf signers who were well known and respected members of their local deaf community, and nearly all were native signers from deaf families. Fieldworkers were also present on the day of filming, to assist with data collection and to lead on some tasks given to the participants. An important reason for having deaf fieldworkers present on the day of filming to perform these functions, instead of a hearing researcher, is that they would assist in keeping language contact influences to a minimum (as reported by Lucas and Valli, 1992). Furthermore, it is important for *local* deaf community fieldworkers to lead on lexical elicitation, particularly if elicitation of regional variants is expected. Deaf fieldworkers also provide a key link to the local community that can be maintained long after the data collection stage has been completed; thus deaf fieldworkers on the BSL corpus were also involved in a series of local community presentations delivered in each region involved in data collection.

The BSL corpus fieldworkers were also on hand to explain the project's aims to participants before filming commenced and to assist with the collection of background information about the participants. This was obtained through a questionnaire consisting of 39 questions that aimed to elicit comprehensive background information on each participant's language experience. These questions followed metadata standards for sign language corpora proposed in Crasborn and Hanke (2003), which also serve as the basis for metadata categories in the NGT and DGS corpora. Questions covered a range of topics that included the participant's general language preference (e.g., signing, sign with speech) with different members of their family, the language used at school in and out of the classroom, the language used during childhood, where else the participant may have lived in the UK, and the extent to which (s)he interacted with the deaf community.

Figure 10.1 Screenshots from BSL Corpus Project video data (pair view and individual view). Schembri, Fenlon, Rentelis, and Cormier, 2011.

It is also necessary to give serious consideration to the layout and location of the filming studio and to the pairing of participants in order to maximize the quality and range of the data collected. For the BSL corpus, participants were filmed in pairs, with two high-definition cameras focused respectively on each member of the pair and a third one focused on both, as shown in Figure 10.1. In the DGS corpus project five cameras were used, two additional ones being positioned above the signers, to give a bird's eye view of their hands as they moved in the signing space around their body (Hanke, König, Wagner, and Matthes, 2010). BSL corpus participants were seated in chairs without arms (so they could not rest their elbows during signing, which may have interfered with sign language production), in front of a blue background screen with appropriate studio lighting, and wearing plain colored clothing on the upper body in order to ensure the best possible capture of the sign language data. Participants were always filmed in same-sex or mixed pairs made up of people of similar ages. They were also paired with someone familiar to them – friends or acquaintances – as one of the tasks required them to engage in spontaneous conversation for 30 minutes. Pairs of participants in a long-term relationship (particularly retired married couples) were not filmed together, because in these cases the conversational data obtained was often not natural due to a high level of familiarity. Filming sessions were always located in settings familiar to the participants, such as deaf social clubs or the offices of deaf organizations, to ensure that participants felt comfortable and that it was appropriate to use a relatively informal variety of BSL.

Data types

While there are some differences between projects in the type of data collected, there is a clear consensus among projects that different genre types should be sampled in order to maximize representativeness. The type of texts contained in sign language corpora is often linked to a project's overall aims and to the initial studies intended/proposed. For the BSL corpus, data collection was limited to four situational varieties which are considered the bare minimum for studies on sociolinguistic variation and language change: a personal experience narrative, free conversation, an interview, and a word list. The personal experience narratives (primarily intended as a warm-up, to acquaint participants with the filming studio) lasted no more than five minutes and featured the retelling of an amusing, poignant, or significant event in the participants'

lives. This was followed by a 30-minute conversation where participants were left to themselves and were free to talk about anything they wanted. A short 15-minute interview on language attitudes and awareness, led by the deaf fieldworker, followed the conversation session. Lastly, participants took part in a lexical elicitation task in which they were asked to produce signs that they used for 102 concepts chosen for their known or suspected high level of sociolinguistic variation.[2]

Other corpus projects have used a variety of elicitation tasks in order to elicit language that included specific grammatical constructions. The Auslan corpus project used the video stimuli from two tests in the test battery for ASL morphology (Supalla et al., n.d.): the verbs of motion production (VMP) and the noun–verb production (NVP) tasks. The VMP involves 40 stop-animation movies that are intended to elicit representations of a subset of motion events with a selected number of referent types; this is a very useful way to elicit classifier/depicting verbs of motion. The NVP was selected because the video material not only elicited short sentences that involved descriptions of various types of transitive events, but also often resulted in the production of noun–verb pairs (i.e., related lexical items in which the nominal and the verbal forms are distinguished by movement and/or non-manual features).

A number of narratives were also elicited using methods widely employed in the sign language literature. The "Canary Row" video stimulus from the Warner Brothers' *Tweety and Sylvester* cartoon series was employed as a way to elicit a story from the participants. This methodology was originally borrowed from gesture studies (see McNeill, 1992), but it is also used in sign language research (e.g., Emmorey, Borinstein, Thompson, and Gollan, 2008). Written stimuli to elicit narratives were also used: written English versions of two Aesop's fables – "The Boy Who Cried Wolf" and "The Hare and the Tortoise" – were given to the participants, who were then asked to retell the stories in Auslan (see Crasborn et al., 2007). An additional narrative was also elicited by using the children's picture story book *Frog, Where Are You?* (see Engberg-Pederson, 2003). The DGS corpus project collected deaf community jokes in addition to elicited narratives.

Discussions, exchanges of different viewpoints, and argumentation were also elicited in the Auslan and NGT corpora through the use of an interview that focused on issues of concern or controversy in the deaf community. Discussion in the DGS corpus was stimulated by informants looking at warning signs collected from different places in the world and talking about what they may or may not mean. Also in the DGS corpus data collection, negotiation was elicited through a calendar task in which participants agreed on dates for separate meetings at times that did not clash with other timetabled tasks. A barrier game task in which participants had to identify a number of differences in two related pictures was also used with similar results in the Auslan corpus project. Explanatory discourse was involved in the Auslan and DGS projects, where participants had to explain the meaning and origin of their name signs.

Observer's paradox, audience design, and consent

Data collected for sign language corpora are at risk of the observer's paradox – i.e., the fact that the vernacular form of a language, used by speakers when they are not being observed, can only be studied by observation. In order to lessen the effects of

this paradox (whereby signers may adjust their signing style because they are being observed/filmed), participants in the BSL corpus were reassured during the process of obtaining informed consent that the conversational data would form part of a restricted-access corpus: it would not be made publicly available on the Internet but would only be shared with other university researchers who have declared an academic interest in the data. It was hoped that participants would relax and converse freely as a result and that the data collected would be as close to the vernacular variety as possible. Additional measures, such as pairing participants with someone well known to them and filming in settings conducive to spontaneous conversation, also ensured that the effects of the observer's paradox were lessened. The remaining three situational varieties – the personal narratives, the interviews, and the lexical elicitation data – all formed part of the open access archive, although access to the interview data was later restricted, as issues arose over the content and its appropriateness within the public domain (for details, see Schembri et al., 2013). The results of a subsequent phonological variation study (Fenlon, Schembri, Rentelis, and Cormier, 2013) suggests that phonological variation in the conversational data is not too dissimilar to phonological variation in the data collected by Bayley, Lucas, and Rose (2002), which was not intended to be part of an online open access collection.

Annotation and Translation

Once the data have been collected and converted into a digital video archive, the next stage of corpus-building can begin. It is after this stage, where annotation work is undertaken, that the digital video archive becomes a modern linguistic corpus. Here it is important to prioritize the type of annotation most appropriate for allowing one to search the corpus effectively – particularly since the process of annotation for sign language corpora is a time-consuming one. Johnston (2010) stresses that two types of annotation are essential for all sign language corpora: ID glossing and a written translation. It is only these two types of annotation at a minimum that will allow the video data set to become a searchable and machine-readable resource. By contrast, transcription – i.e., a notation system adopted in order to describe a sign's form – is likely to take much longer and will not necessarily result in a corpus that can be readily searched by others. In this section we discuss in more detail ID glossing, written translations, and their rationale – with reference to the BSL and Auslan corpora.

ID glossing refers to the practice of using one unique identifying gloss or "ID gloss" for each sign. This ID gloss is used to represent the sign in its citation form along with all its phonological and morphological variants. The procedure and principles of grouping sign variants together under a single lemma form what is called "lemmatization" and are outlined, for signed languages, in Johnston and Schembri (1999), Cormier, Fenlon, et al. (2012), and Fenlon, Cormier, and Schembri (under review). One such principle (which applies equally to spoken languages) is that all inflectional morphological variants of a sign are identified as a single lemma (e.g., the directional verb GIVE is always assigned the same ID gloss regardless of how it has been modified spatially for person or number). Both the

BSL and the Auslan corpora use English words as ID glosses, the choice of the English word being often motivated by its strong association with a particular sign form (e.g., the English word "like" may be one of the primary meanings that signers associate with the sign LIKE). Therefore the choice of an English word as a surrogate marker for a sign's form aids in consistent glossing and is strongly preferred over notation systems for corpus-based approaches to sign language research.

The practice of using ID glossing is made considerably easier if this lemmatization work has already been completed for each sign language and if these groupings have been recorded in a lexical database that can be said to be representative of all the signs that make up a sign language's core lexicon. Annotators can then consult this database to find a form's corresponding ID gloss and possible meanings (via English translation equivalents) linked to this ID gloss. For the Auslan corpus, a lexical database that follows such principles was already in place prior to annotation work (see Johnston, 2001). Such a resource has given the Auslan corpus an advantage that has resulted in its becoming the largest annotated sign language corpus that is based on a lemmatized lexical database (c. 105,000 sign token annotations at the time of writing). In contrast, no such resource existed for BSL prior to the BSL corpus project. Available dictionaries for BSL (e.g., Brien, 1992) had not followed lemmatization practices: many homonyms are grouped together as a single entry, and many phonological variants are often presented as separate, unrelated entries. As a result, lemmatization work had to take place concurrently with annotation work, as part of a lexical frequency study of approximately 25,000 tokens. This has resulted in a database of approximately 1,800 entries, each entry showing, at a minimum, the sign's form, its ID gloss, and English keywords associated with its meaning. At the time of writing, this lexical database is being converted into BSL SignBank, an online dictionary for BSL (see Cormier, Fenlon, et al., 2012 and Fenlon et al., under review).

Johnston (2010) also advocates a written translation as a second type of annotation that should be prioritized in order to enhance usability and to maximize access to the corpus. This is because the act of ID glossing is unlike that of context-based glossing, which is often seen as a type of translation in itself. Since the same ID gloss is consistently applied to a sign form regardless of any variation in grammatical use (i.e., GIVE labels one sign regardless of whether it functions as a verb "give" or as a noun "gift") or in its particular meaning/sense (e.g., EXCITED labels a particular sign even where the English words "motivate," "interest," "interesting," or "eager" would appear to be a better translation), it is difficult to get a sense of the intended meaning of a signed utterance from ID glosses. The task of translation is instead assigned to a separate tier in the ELAN annotation viewer, which is also time-aligned with the ID glosses and the corresponding media file. Therefore anyone accessing the corpus can use both the ID gloss tier and the English translation to gain a full understanding of what is being signed via which lexical items. Although both ID glossing and translation are needed for a corpus, a written translation is much faster to achieve than ID glossing and can render a larger proportion of the corpus accessible in a shorter space of time than ID glossing alone.

The annotation stage is the most time-consuming and laborious stage in the creation of sign language corpora. For most sign language corpus projects, this is because the task of annotation must be conducted concurrently with the creation of

a lexical database. Furthermore, the process of annotation cannot be automated, as the technology required to do so is still in its formative years. Thus it may be some time before any sign language corpus is minimally complete – in the sense that ID glossing and translation have been completed for the entire corpus. For example, work on ID glossing and translation of the Auslan corpus began in 2006 and is expected to take at least five more years before a basic reference corpus of the entire video archive is achieved. Detailed multi-tier annotation beyond ID glossing and translation (e.g., to the level of detail described in Johnston, 2011) would take even longer. For the BSL corpus, after the first two years of lexical annotation work carried as part of a lexical frequency study, only 5 percent of the conversational data had been glossed. Although this seems very little, the creation of these sign language corpora signals in fact the beginning of a period of corpus-based research in sign language linguistics, in which questions that are difficult or impossible to answer in any other way can finally be addressed. This is certain to lead to a much better understanding of the structure and use of sign languages.

Archiving, Interfaces, and Access

Today it is possible to access some sign language corpus data over the Internet. However, this access is often restricted to the video data alone; annotation data are limited or not available at all (depending on the amount of annotation work completed). The first stage of the NGT corpus was completed in 2008, in the sense that the archived video recordings had been edited and catalogued and were made openly accessible through a digital video archive on the Internet. However, at the time of writing, only a small percentage of the NGT video files have been annotated and/or translated into a parallel written text. Similarly, video files for the Swedish Sign Language corpus are now openly available online but only a small percentage have been annotated and translated (see Mesch, Wallin, Nilsson, and Bergman, 2012). From 2012 the Auslan corpus videos have been available on the SOAS Endangered Languages Documentation Archive (University of London); most recordings are openly accessible, though some have limited or restricted access. Registered users may apply for increased access if they agree to SOAS and deposit-specific conditions. ELAN annotation files are also available for a limited set (only about half of the Auslan corpus has been annotated with ID glosses and translations). The Auslan deposit at SOAS is updated yearly, as additional annotations become available. Since 2011 the BSL corpus video data have been made available online via CAVA (human Communication Audio-Visual Archive), a secure system that allows anyone to view and download the open access corpus data – that is, narratives and lexical elicitation data – and allows researchers access to the restricted corpus data – conversations and interviews – via a user license that includes a confidentiality agreement. An initial set of ELAN annotation files (containing lexical annotations from conversation data from four regions) has been available online since September 2014. For the NGT, Auslan, and BSL corpora, it is possible to search the video data by using the participant metadata collected.

Conclusion

Corpus methods can (and, we argue, should) be used in any kind of language documentation. Sampling sign language data from a variety of signers with different backgrounds and different text types helps ensure that the data set is as representative as possible, even if it is not large for whatever reason (e.g., due to lack of resources). The use of ID glossing via a lexical database in particular is crucial for any kind of language documentation. Without ID glossing, no claims can be made about the nature of the lexicon (e.g., lexical frequency), and ultimately the study of any other level of language structure (phonology, syntax) is also compromised. The use of ID glossing, coupled with translation into one or more written languages, ensures maximum searchability and accessibility, in addition to machine readability.

Notes

1 In this chapter we do not make a distinction between the spellings *deaf* and *Deaf*: we are using *deaf* throughout, because we wish not to make assumptions about individual deaf people's identity.
2 For a list of the 102 concepts that were included in the lexical elicitation task and the list of questions asked during the interview, see http://www.bslcorpusproject.org/cava/activities

Keywords

annotation; corpora; ELAN; ID glossing; lexical database; machine readability; observer's paradox; representativeness; sampling; translation

See Also

See Chapter 4 for more information on technical requirements and the use of annotation software when collecting sign language data.

Suggested Readings

For more on the background that led to the beginnings of sign language corpora, see Johnston and Schembri (2013). For more about the observer's paradox, audience design, and issues surrounding consent and confidentiality in sign language corpus studies, and about multiple interfaces and access levels for different audiences in the BSL corpus, see Schembri et al. (2013).

References

Bayley, R., Lucas, C., and Rose, M. (2002). Phonological variation in American Sign Language: The case of the 1 handshape. *Language Variation and Change* 14(1), 19–53.
Brien, D. (Ed.) (1992). *Dictionary of British Sign Language/English*. London: Faber & Faber.

Cormier, K., Schembri, A., Vinson, D., and Orfanidou, E. (2012). First language acquisition differs from second language acquisition in prelingually deaf signers: Evidence from grammatical processing of British Sign Language. *Cognition* 124, 50–65.

Cormier, K., Fenlon, J., Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R., Woll, B. (2012). From corpus to lexical database to online dictionary: Issues in annotation of the BSL corpus and the development of BSL signbank. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, and J. Mesch (Eds.), *Proceedings of the fifth workshop on the representation and processing of sign languages: Interactions between corpus and lexicon*. Paris: European Language Resources Association, pp. 7–12.

Crasborn, O., and Hanke, T. (2003). *Additions to the IMDI metadata set for sign language corpora*. Paper presented at the ECHO workshop, Nijmegen University.

Crasborn, O., and Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In O. Crasborn, E. Efthimiou, T. Hanke, E. D. Thoutenhoofd, and I. Zwitserlood (Eds.), *Proceedings of the third workshop on the representation and processing of sign languages: Construction and exploitation of sign language corpora*. Paris: European Language Resources Association, pp. 39–43.

Crasborn, O., Hulbosch, M., and Sloetjes, H. (2012). Linking Corpus NGT annotations to a lexical database using open source tools ELAN and LEXUS. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, and J. Mesch (Eds.), *Proceedings of the fifth workshop on the representation and processing of sign languages: Interactions between corpus and lexicon.* Paris: European Language Resources Association, pp. 19–22.

Crasborn, O., Efthimiou, E., Hanke, T., Thoutenhoofd, E. D., and Zwitserlood, I. (Eds.) (2008). *Proceedings of the third workshop on the representation and processing of sign languages: Construction and exploitation of sign language corpora*. Paris: European Language Resources Association.

Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., Van Der Kooij, E., Woll, B., and Bergman, B. (2007). Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics* 12(4), 535–562.

Emmorey, K. (2002). *Language, cognition, and the brain: Insights from sign language research*. Mahwah, NJ: Lawrence Erlbaum.

Emmorey, K., Borinstein, H. B., Thompson, R., and Gollan, T. H. (2008). Bimodal bilingualism. *Bilingualism: Language and Cognition* 11(1), 43–61.

Engberg-Pederson, E. (2003). How composite is a fall? Adults' and children's descriptions of different types of falls in Danish Sign Language. In K. Emmorey (Ed.), *Perspective on classifier constructions in sign languages*. Mahwah, NJ: Lawrence Erlbaum, pp. 311–322.

Fenlon, J., Cormier, K., and Schembri, A. (under review). Building BSL SignBank: The lemma dilemma revisited.

Fenlon, J., Schembri, A., Rentelis, R., and Cormier, K. (2013). Variation in handshape and orientation in British Sign Language: The case of the "1" hand configuration. *Language and Communication* 33, 69–91.

Frishberg, N., Hoiting, N., and Slobin, D. I. (2012). Transcription. In R. Pfau, M. Steinbach, and B. Woll (Eds.), *Sign language: An international handbook*. Berlin: Mouton de Gruyter, pp. 1045–1075.

Hanke, T., and Storz, J. (2008). iLex: A database tool for integrating sign language corpus linguistics and sign language lexicography. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, and E. D. Thoutenhoofd (Eds.), *Proceedings of the third workshop on the representation and processing of sign languages: Construction and exploitation of sign language corpora*. Paris: European Language Resources Association, pp. 64–67.

Hanke, T., König, L., Wagner, S., and Matthes, S. (2010). DGS corpus & Dicta-Sign: The Hamburg studio set-up. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martinez-Ruiz, and A. Schembri (Eds.), *Proceedings of the fourth workshop on the representation*

and processing of sign languages: Construction and exploitation of sign language corpora. Paris: European Language Resources Association, pp. 106–109.

Johnston, T. (1991). Transcription and glossing of sign language texts: Examples from Auslan (Australian Sign Language). *International Journal of Sign Linguistics* 2(1), 3–28.

Johnston, T. (2001). The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics* 1(2), 145–169.

Johnston, T. (2004). W(h)ither the deaf community? Population, genetics and the future of Australian Sign Language. *American Annals of the Deaf* 148(5), 358–375.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15(1), 106–131.

Johnston, T., and Schembri, A. (1999). On defining lexeme in a signed language. *Sign Language & Linguistics* 2(2), 115–185.

Johnston, T., and Schembri, A. (2007). *Australian Sign Language: An introduction to sign language linguistics*. Cambridge: Cambridge University Press.

Johnston, T., and Schembri, A. (2013). Corpus analysis of sign languages. In C. A. Chapelle (Ed.), *Encyclopedia of applied linguistics*. Oxford: Wiley Blackwell, pp. 1312–1319.

Langer, G. (2012). A colorful first glance at data on regional variation from the DGS-corpus: With a focus on procedures. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, and J. Mesch (Eds.), *Proceedings of the fifth workshop on the representation and processing of sign languages: Interactions between corpus and lexicon*. Paris: European Language Resources Association, pp. 101–108.

Leeson, L. (2010). *From theory to practice: Sign language corpora in teaching and learning*. Paper presented at the sign linguistics corpora network: Workshop 4 (Exploitation), Berlin, Germany.

Leeson, L., and Saeed, J. I. (2012). *Irish Sign Language*. Edinburgh: Edinburgh University Press.

Lucas, C., and Valli, C. (1992). *Language contact in the American deaf community*. San Diego: Academic Press.

Lucas, C., Bayley, R., and Valli, C. (2001). *Sociolinguistic variation in American Sign Language* (vol. 7). Washington, DC: Gallaudet University Press.

McEnery, T., and Wilson, A. (2001). *Corpus linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

McKee, R., Schembri, A., McKee, D., and Johnston, T. (2011). Variable subject expression in Australian Sign Language and New Zealand Sign Language. *Language Variation and Change* 23(3), 375–398.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Cambridge: Cambridge University Press.

Mesch, J., and Wallin, L. (2012). From meaning to signs and back: Lexicography and the Swedish Sign Language corpus. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, and J. Mesch (Eds.), *Proceedings of the fifth workshop on the representation and processing of sign languages: Interactions between corpus and lexicon*. Paris: European Language Resources Association, pp. 123–126.

Mesch, J., Nilsson, A.-L., Wallin, L., and Bäckström, J. (2010). *Using corpus data for teaching purposes*. Paper presented at the sign linguistics corpora network: Workshop 4 (Exploitation), Berlin, Germany.

Mesch, J., Wallin, L., Nilsson, A.-L., and Bergman, B. (2012). Datamängd: Projektet Korpus för det svenska teckenspråket 2009–2011 (version 1). Avdelningen för teckenspråk, Institutionen för lingvistik, Stockholms universitet. Accessed September 16, 2014. http://www.ling.su.se/teckenspraskorpus

Mitchell, R. E., and Karchmer, M. A. (2004). Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign Language Studies* 4(2), 138–163.

Neidle, C., and Vogler, C. (2012). A new web interface to facilitate access to corpora: Development of the ASLLRP Data Access Interface (DAI). In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, and J. Mesch (Eds.), *Proceedings of the fifth workshop on the representation and processing of sign languages: Interactions between corpus and lexicon*. Paris: European Language Resources Association, pp. 137–142.

Nonaka, A. M. (2004). Sign languages: The forgotten endangered languages: Lessons on the importance of remembering. *Language in Society* 33(5), 737–767.

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

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., and Henning, J. (1989). *HamNoSys Version 2.0. Hamburg notation system for sign languages: An introductory guide*. Hamburg: Signum.

Rayson, P., Wilson, A., and Leech, G. (2002). Grammatical word class variation within the British National Corpus sampler. In P. Peters, P. Collins, and A. Smith (Eds.), *New frontiers of corpus research: Papers from the twenty-first international conference on English language research on computerized corpora, Sydney 2000*. Amsterdam: Rodopi, pp. 295–306.

Schembri, A. (2010). Documenting sign languages. In P. Austin (Ed.), *Language documentation and description* (vol. 7). London: School of African and Oriental Studies, pp. 105–143.

Schembri, A., and Johnston, T. (2012). Sociolinguistic aspects of variation and change. In R. Pfau, M. Steinbach, and B. Woll (Eds.), *Sign language: An international handbook*. Berlin: Mouton de Gruyter, pp. 788–816.

Schembri, A., Fenlon, J., Rentelis, R., and Cormier, K. (2011). British Sign Language Corpus Project: A corpus of digital video data of British Sign Language 2008–2011 (1st ed.). London: University College London. Accessed September 17, 2014. http://www.bslcorpusproject.org.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., and Cormier, K. (2013). Building the British Sign Language corpus. *Language Documentation and Conservation* 7, 136–154.

Schembri, A., Mckee, D., Mckee, R., Pivac, S., Johnston, T., and Goswell, D. (2009). Phonological variation and change in Australian and New Zealand Sign Languages: The location variable. *Language Variation and Change* 21(02), 193–231.

Supalla, T., Newport, E. L., Singleton, J., Supalla, S., Metlay, D., and Coulter, G. (n.d.). *The test battery for American Sign Language morphology and syntax*. Unpublished manuscript, University of Rochester, New York.

Sutton-Spence, R., and Woll, B. (1999). *The linguistics of British Sign Language: An introduction*. Cambridge: Cambridge University Press.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the fifth international conference on language resources and evaluation, LREC 2006*. Nijmegen, Netherlands: Max Planck Institute for Psycholinguistics, The Language Archive. Accessed September 17, 2014. http://tla.mpi.nl/tools/tla-tools/elan