Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 224–230
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

One Side of the Coin: Development of an ASL-English Parallel Corpus by Leveraging SRT Files

Rafael O. Treviño[1], Julie A. Hochgesang[2], Emily P. Shaw[1], Nic Willow[2]
[1]Department of Interpretation and Translation, [2]Department of Linguistics
Gallaudet University
800 Florida Ave NE, Washington, DC 20002, USA
{rafael.trevino, julie.hochgesang, emily.shaw, nic.willow}@gallaudet.edu

Abstract

We report on a method used to develop a parallel corpus of English and American Sign Language (ASL). The effort is part of the Gallaudet University Documentation of ASL (GUDA) project, which is currently coordinated by an interdisciplinary team from the Department of Linguistics and the Department of Interpretation and Translation at Gallaudet University. Creation of the parallel corpus makes use of the available SRT (SubRip Subtitle) files of ASL videos interpreted into or from English, or captioned into English. The corpus allows for one-way searches based on the English translation or interpretation, which is useful for translators, interpreters, and those conducting comparative analyses. We conclude with a discussion of important considerations for this method of constructing a parallel corpus, as well as next steps that will help to refine the development and utility of this type of corpus.

Keywords: corpus, parallel corpus, translation, interpreting, SRT

1. Introduction

The method of constructing a corpus presented in this paper addresses two issues. The first is in constructing a corpus. Even with written text, which is in a machine-readable format, constructing a corpus can be laborious and time-consuming. Sign language corpora are all the more difficult due to the need to adopt conventions for converting the data from video into a machine-readable format. In some sign language corpora, data is sometimes freely translated into a written language as a way to provide provisional access to the signed language. This translation task, however, still takes time that an annotator could use to work on annotation.

The second issue relates to the use of corpora by those from other disciplines. In looking toward the future in their text on corpus linguistics, McEnery and Hardie (2012) offer ideas on "the potential for corpus methods to extend beyond the field of linguistics into other areas of the humanities, sciences and social sciences" (p. 225). In our case, we exhibit an area for collaboration between sign language corpus linguistics and the field of sign language translation and interpreting. For example, when faced with challenges in how to translate a specialized term or phrase from English into ASL, students have had to rely on personal observations of deaf people and interpreters to build their vocabulary and repertoire of interpretation choices. However, their observations are fleeting (i.e., cannot be accessed later for review) and limited in number and kind.

Thus, the parallel corpus described in this paper presents a possible solution to the issues raised here. In the first case, our proposed method can be used to leverage SRT files to save annotators time and allow linguists (or others) to establish a provisional corpus expeditiously. In the second case, the resulting parallel corpus, while it cannot be exploited in its initial state by linguists, the parallel corpus can be used by others who can profit from the ability to investigate the signed language through another, written language. In other words, one linguist's provisional corpus is another field's treasure.

In addition, we must point out the two cases can work in tandem: a provisional corpus can be used by people other than linguists while the signed language side of the corpus is being annotated.

In any event, the parallel corpus described in this paper provides the means to conduct powerful analyses of larger interpreted datasets. We suspect, moreover, it may have applications beyond the suggestions and ideas presented herein.

2. Background

2.1 Gallaudet University Documentation of ASL (GUDA) Project

Gallaudet is an ASL-English bilingual university. The campus community consists of Deaf, DeafBlind, hard of hearing, and hearing people, all of whom have varying degrees of fluency in ASL (visual and tactile varieties), written and spoken English, not to mention other written, spoken, and signed languages. Because of its bilingual mission (Gallaudet University, 2007), the university commits to providing video content of lectures, announcements, and other communications in both ASL and English (written or spoken, or both).

The creation of the parallel corpus emerged from work on the GUDA project. The project aims to digitally organize the ASL video collections on campus so they may be accessed by scholars and the public (see Hochgesang, Willow, Treviño, and Shaw, 2019, for a more complete description of the project). Notably, the research team behind the project is currently composed of faculty and graduate students from both the Department of Linguistics and the Department of Interpretation and Translation. It is partly the intersecting interests of these two disciplines that helped uncover the benefits of combining the needs of translators and interpreters with the technology for building sign language corpora.

2.2 Definitions

Later in the paper, we present some important considerations regarding terminology in the face of multi-modal parallel corpora. For the moment, however, it may be useful to the reader for us to review a few preliminary terms.

In this paper, *translation* refers to the act of rendering ASL in a video into written English after the recorded event has occurred. The written English may appear either as subtitles or as a tier in ELAN, or both. This activity can be carried out by an annotator, a professional, or, in the case of some videos, an unknown person.

By *interpreting* or *interpretation*, we refer to the act of rendering either ASL into English or English into ASL, most often in the simultaneous mode.

We typically use *transcription* to refer to a representation of spoken English in written form. Transcription can be used to represent spoken English, either as the source message or as the interpretation of an ASL source message. When a transcription is provided at the time of the recorded event, we refer to this as *real-time transcription*. When a transcription is produced after the event has already taken place, we refer to it simply as *transcription*, or *offline transcription*.

2.3 Sign Language Corpora as Parallel Corpora

Baker, Hardie, and McEnery (2006) define a *parallel corpus* as "a set of texts and their translations" (p. 126). They note parallel corpora are often used to compare terms and grammatical structures between languages, to look at the features of translations, and to assist with machine translation.

Sign language corpora often include translation into a written language as one of the steps in converting the data into a machine-readable format. Meurant, Cleve, and Crasborn (2016) observe that this work to translate the signed language into a written language effectively converts sign language corpora into bilingual (a.k.a., "parallel") corpora.

Indeed, Meurant, Cleve, and Crasborn (2016) also emphasize that the bilingual nature of signed language corpora has yet to be fully exploited for purposes such as the ones noted by Baker, Hardie, and McEnery (2006). They draw upon the Corpus LSFB (Meurant, 2015) and the Corpus NGT (Crasborn, Zwitserlood, & Ros, 2008) to describe how linguists, interpreters, translators, teachers, and language learners can all use parallel corpora of signed languages. At the time of the Meurant, Cleve, and Crasborn (2016) paper, the Corpus LSFB had 2.5 hours (2,400 sentences) of LSFB with translations into French, and the Corpus NGT had 15 hours (15,000 sentences) of NGT with translations into Dutch.

2.4 Corpora and Sign Language Interpreting

Parallel sign language corpora are not just for looking at questions pertaining to sign language interpreting, but the two do seem to go hand-in-hand. In Translation and Interpreting Studies (TIS), the benefits of corpora have been recognized since at least the 1990s, especially with regard to investigating theoretical issues and the process of interpreting (Baker, 1993; Shelsinger, 1998). With regard to sign language interpreting, the benefits of using corpora have also been recognized, mostly in the realm of training interpreters. Early on, for instance, Heßmann and Vaupel (2008) argued for the need to implement the use of sign language corpora in interpreter education and outlined some of the challenges in doing so. One of the challenges was taking into consideration spoken language, even though it may "seem odd to include vocal language texts in a sign language corpus" (p. 75). However, when creating a parallel corpus for comparative purposes, it seems handling the data of spoken language cannot be ignored. To this end, Heßmann and Vaupel (2008) also identified challenges related to the classification of data and the use of metadata in parallel corpora, which we also address in our paper in section 4.2.

As an invited speaker from outside linguistics at a workshop given for the Sign Linguistics Corpora Network in Berlin, Nancy Frishberg (2010) mentioned a few of the possibilities corpora hold for interpreter education. She posited corpora could be used "within mode" for non-native users of sign language to improve their linguistic ability and "across modes" for comparative analyses. The importance of corpora as a vehicle is that they provide the advantage of annotations, which enable the data to be searched (e.g., "I want to work on conversations, especially those with head-tilt" [slide 33]). Moreover, the benefit could be mutual in that the learners themselves could also provide input to the annotations (i.e., to crowdsource the annotation effort).

Since Heßmann and Vaupel (2008) and Frishberg (2010), some progress has been made in developing corpora for sign language interpreting, but there are only a few references in the literature. Wehrmeyer (2019) provides an in-depth account of her work in constructing the South African Sign Language Interpreting Corpus (SASLIC), a parallel corpus of English–South African Sign Language (SASL) based on interpreted news bulletins. The English source text (ST) of the corpus was created using re-speaking software, and the SASL target text (TT) was annotated using a novel convention. Wehrmeyer (2019) notes that, "a typical half-hour of ST could be transcribed in a day, whereas the TT transcription for that selection took at least 160 hours" (p. 73). In her paper, she reports a total of approximately 3.3 hours (200 minutes) of source text that had been transcribed. She concludes by observing sign language parallel corpora such as hers could be used to investigate "referencing techniques, non-manual features, discourse devices and interpreting strategies" (p. 81).

In another parallel corpus, Roush (2016) compiled the translations produced by deaf translators from English into ASL of famous speeches in U.S. history. The corpus, known as the American Freedom Speeches (AFS) Translation Corpus, consists of 29 minutes of video. The English source text was fully transcribed and the ASL target text was fully annotated. The purpose of the AFS corpus was to explore its pedagogical utility in teaching sign language interpreters. Roush (2016) notes one of the advantages of the AFS corpus is that it shows learners how native users of ASL expressed particularly problematic constructions in the English source into ASL.

The foregoing are two examples of corpora developed for broad purposes, with the former having a pedagogical slant. Due to technological and other constraints, neither is publicly accessible on a website. It is also quite possible, if not certain, other small-scale parallel corpora exist, even if they have only been compiled on an *ad hoc* basis to answer specific research questions. A case in point is the well-known study conducted by Cokely (1986). In his study, Cokely (1986) recorded, transcribed, and annotated a total of approximately 32 minutes of an interpretation conducted from English into ASL. From this data, Cokely (1986) identified there was a negative correlation between lag time and the number of errors committed by the interpreter. Like the other corpora mentioned above, the data used by Cokely (1986) is also not publicly accessible.

2.5 Summary of Issues

Several issues related to parallel corpora for sign language interpreting have been raised in this brief section. In comparison to sign language corpora in general, parallel corpora are fewer in number and smaller in size, lack a common framework for classification and metadata, and their utility is still open to possibilities. Though often geared toward pedagogical purposes, parallel corpora can also be used to research theoretical issues and the process of interpreting. In sum, if sign language corpora are now starting to come of age, then sign language interpreting parallel corpora are the younger sibling toddling behind, tugging at the sleeves.

3. The ASL-English Parallel Corpus

As part of the GUDA project, our research team has access to an archive of over 2,000 videos held by Gallaudet University's library service and other departments on campus. Work is underway to annotate the ASL that appears in the videos in order to analyze the data from a linguistic perspective (see Hochgesang, Crasborn, and Lillo-Martin, 2018, for a review of our ASL annotation principles). One of the steps in annotating the videos includes the creation of a "free translation" tier in ELAN (Version 5.8) into English of the ASL that appears in the video. During our process of cataloging the videos available for annotation, we observed that certain collections of the ASL videos had English subtitles. We attempted to automatically create the free translation tier based on the subtitles, which were readily available. We outline this attempt in this section.

3.1 Materials

Materials for the corpus consisted of videos housed by Gallaudet University's library service with an available SRT (SubRip Subtitle) file. An SRT file is effectively an English representation of the ASL that appears in the video. Thus, its presence in the video eliminates the need for annotators to provide the translations and allows them to focus on (the more typically time-consuming act of) annotation of the ASL instead.

After eliminating all of the videos that did not have an SRT file available, we identified 590 videos as potential candidates for our parallel corpus. Of those, five were duplicates, leaving us with 585 videos and their respective SRT files. In total, the video data equal 107.48 GB and the SRT files equal 24.7 MB. Using a rough calculation based on 1 hour of video for every 500 MB (0.5 GB), we can estimate the size of the corpus to be approximately 215 hours (107.48/0.5 = 214.96).

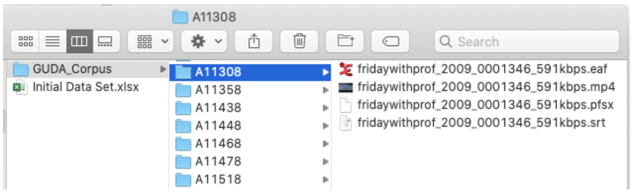
3.2 Construction

All scripts referred to in this section were written in Apple's Script Editor (Version 2.11) on a MacBook Pro (2017) running macOS Catalina (Version 10.15.3). We use ELAN (Version 5.8).

We first created EAF files for the video files using ELAN's batch-processing functionality (File > Multiple File Processing > Create Transcription Files for Multiple Media Files). We directed ELAN to the folder containing the 585 video files. We did not select a template for the new transcription files (but see section 4.2 for a discussion on the type of information we may want to include in a template in the future). We directed the location for the new transcription files to the same folder as the media files.

Using a master spreadsheet that contained all of the video file names, we assigned each video an ID. We wrote a script to automatically create a folder for each Video ID. We then wrote a script to automatically move the video files (almost all in MP4 format), the newly created EAF files, and the SRT files into their respective folder.

ELAN (Version 5.8) provides the functionality to import an SRT file and automatically create a tier, which it names "Subtitle-Tier," based on the text and timestamps contained in the SRT file. See the user manual for a review of this functionality (Hellwig et al., 2019, p. 91). However, ELAN will not run this functionality on a batch of files; it will only import an SRT file and attach it to the EAF that is currently open. Therefore, we wrote another script to (a) open each folder, (b) open the EAF file inside the folder, (c) import the associated SRT file, (d) close the EAF, and (e) open the next folder, and so on until the process was completed. See Figure 1 for the resulting file structure.



GUDA_Corpus
Initial Data Set.xlsx

A11308
A11358
A11438
A11448
A11468
A11478
A11518

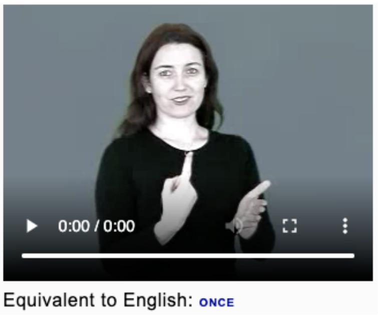
fridaywithprof_2009_0001346_591kbps.eaf
fridaywithprof_2009_0001346_591kbps.mp4
fridaywithprof_2009_0001346_591kbps.pfsx
fridaywithprof_2009_0001346_591kbps.srt

Figure 1: File structure of the ASL-English parallel corpus.

3.3 Application

At this early stage, the ASL-English parallel corpus has not been used to investigate any research questions and has only recently been introduced to interpreting and translation students. Leveraging the SRT files has served to automatically add what is effectively the free translation tier in ELAN, saving GUDA annotators a precious amount of time. Nonetheless, we would like to demonstrate one example of the utility of a parallel corpus constructed by leveraging SRT files: terminological searches.

The website HandSpeak (www.handspeak.com) is an online ASL Dictionary. A search for the English term "once" returns the ASL equivalent shown in Figure 2. The site provides two other ASL equivalents based on the usage of the word in the phrases "once a week" and "once in a while," but the initial forms are much the same as the one that appears in Figure 2.



Equivalent to English: ONCE

Figure 2: Word of the day (www.handspeak.com, retrieved on February 15, 2020).

By comparison, a search for the English term "once" in the ASL-English parallel corpus returns 118 occurrences, all of them in context. The results include an instance of an English interpretation — "*Once* someone transcribes video footage through [system]"—for which the ASL source text contained the sign shown in Figure 3. Note that, to ensure signs can be clearly seen and to make them accessible to the reader, we will use images from ASL Signbank (Hochgesang, Crasborn, & Lillo Martin, 2020) and include the ID gloss.

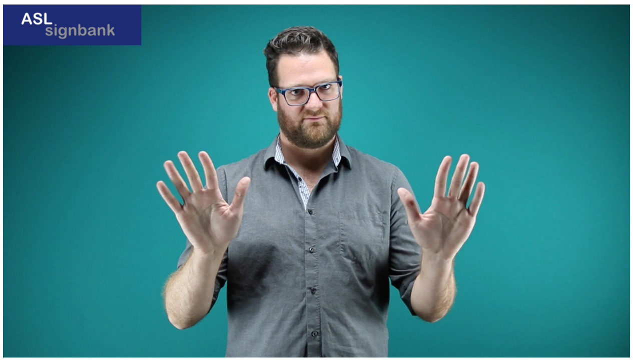


Figure 3: FINISH (ASL Signbank, 2020).

Another occurrence is from a segment of spoken English interpreted into ASL. The English segment was "because they would see an entire line of text all at *once,*" and the ASL sign corresponding to the concept of *all at once* in that context is the sign shown in Figure 4.

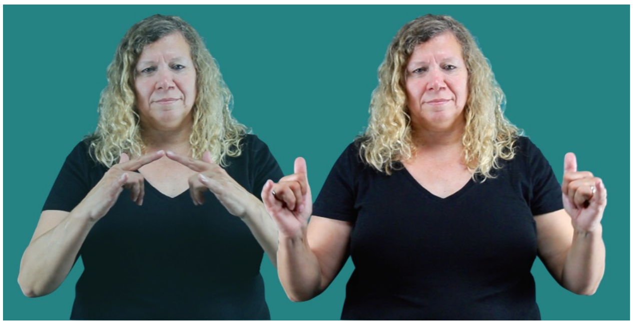


Figure 4: SAME-TIME (ASL Signbank, 2020).

In another result, a segment of ASL was interpreted into English as, "At school I can frequently interact as much as possible, but *once* I left school [...]." The ASL rendition did not contain a signed equivalent for "once." Rather, a shift in the signer's body posture from right to left signaled the contrast in time: before and after leaving school.

The foregoing is an abbreviated example of the potential of a parallel corpus. A few of the uses in teaching and researching sign language interpreting have already been mentioned. However, there are many more potential uses, and an exhaustive review is beyond the scope of this paper. For the time being, we ask our readers to let their imagination soar.

The creation of the ASL-English parallel corpus was a fruitful endeavor; however, it is not complete, and there are opportunities for us and others to improve upon the process. In the following section, we report on issues we encountered that we deem must be taken into consideration in the construction of a sign language parallel corpus.

4. Issues and Next Steps

4.1 One Side of the Coin

The most salient feature of this type of parallel corpus is that it is machine-readable in only one of the languages (specifically here, English). Sophisticated algorithms could probably mine the images and video that are aligned with the English text; however, most users interested in conducting a text-based search derived from the signed language, such as an ID gloss, will be disappointed. We must therefore emphasize this is a parallel corpus that can only be accessed through one of the languages. In other words, searches and analyses can only be initiated from the English representations of the ASL.

Nonetheless, those who wish to create bilingual corpora can still use the methods described herein to prepare a provisional corpus. In fact, leveraging SRT files to produce a free translation tier in ELAN does not preclude any other annotation work or linguistic analyses. All the videos used for this parallel corpus, for instance, are still in the queue for ASL annotation for the linguistic research that still needs to be done. In fact, it would be a useful time-saving strategy considering the lengthy effort of ASL annotation.

4.2 Classification

We use this section to address broad issues of classification that seem to be particular to sign language parallel corpora. We feel clear standards regarding the classification, and thus organization, of the data that feeds into parallel corpora will enable all stakeholders to take the fullest advantage of them.

4.2.1 Classification of Event Types

In section 3.1, we reported on the materials used to create this ASL-English parallel corpus, which were videos with SRT files. However, we feel more discussion is merited regarding what the SRT files represent. For the purposes of this discussion, we will use the term *online* to refer to communicative events that occurred at the time the video was recorded and *offline* to refer to events that occurred afterward.

In general, there are three types of online events that are of interest for our parallel corpus: interactions that occurred in (a) ASL only, (b) ASL interpreted into spoken English, and (c) spoken English interpreted into ASL. The English representation of the ASL that appears in the SRT can come from a number of different sources.

In Event Type A, the ASL in the video is translated offline into English, and there is time to disambiguate any utterances in the source. In the case of signed language corpora, this is a common scenario: the signed language is often translated by the annotator using the free translation tiers. In our corpus, however, there are many instances of ASL videos being translated offline into English by an unknown translator (although often carried out by a professional). The difference is not without theoretical and practical implications, as the intended audience of a translation has a significant effect on textual choices made by the translator (or annotator). For instance, a translation produced for subtitles for a public audience may synthesize information in order not to overload the screen with text. Annotators, on the other hand, do not concern themselves with how the translation will display on screen and, therefore, may approach translation differently.

Event Type B (ASL>English) represents the most complex group of the three. In straightforward cases, a verbatim transcript is produced offline of the English interpretation. However, there are many cases in which the ASL is either re-translated offline into English, or the verbatim English transcript of the interpretation is edited. In another scenario, a real-time captioner transcribes the English interpretation on-site. The online nature of real-time transcription means there is another opportunity for infelicities between the spoken English interpretation and its real-time transcription. Offline, this transcript may be used as-is to produce the subtitles or it may be edited. In sum, SRT files for Event Type B events may come from any one of three sources: 1) a re-translation of the ASL into

English, 2) a verbatim transcript of the English interpretation, or (if the service was provided) 3) a transcript of the real-time captioning based on the English interpretation, all with the possible intervention of editing.

Unlike Event Types A and B, in which the SRT file represents English as a target language, the SRT files for Event Type C (English>ASL) videos represent English as the source language. For these events, the ASL interpretation is not typically back-translated into English, although that is possible. The SRT file may either be based on a transcript prepared by a real-time captioner on-site or it could be based on a verbatim transcript prepared offline.

We emphasize that the event types outlined above refer to communicative *events* and not entire videos. Some videos can include instances of each event type, such as a video of an interpreted panel discussion with both deaf and hearing members. In this case, the source language alternates between ASL and English.

4.2.2 Classification of the Corpus

A full discussion on the classifications of parallel corpora is beyond the scope of this paper. There are several. However, we will throw a proverbial wrench into the mix. In broad strokes, according to Fantinuoli and Zanetti (2014), as cited in Wehrmeyer (2019), a corpus is classified with regard to the number of languages it represents (1 = monolingual, 2 = bilingual, 3 or more = multilingual), architecture (comparable or parallel), purpose (general or specialized), modality, and directionality.

In terms of modality, we must consider the visual nature of signed languages and the need for conventions to annotate them so they are machine-readable. We are no strangers to multimodal corpora. In our case, echoing the advice of Heßmann and Vaupel (2008), at some point we will have to decide what consideration we want to give the spoken English lurking within our data. Is a rough transcript sufficient? Is a phonetic transcription merited? What degree of "verbatim" is needed? Moreover, multimodality should also take into account co-speech gestures produced by hearing people who are visible in the videos.

However, the issue of modality is not the wrench. The wrench was alluded to earlier in our discussion in section 4.1, on the ability to access the ASL only by searching through English (at least until we are able to annotate the ASL). This is not to be confused with the directionality of a corpus (unidirectional or bidirectional). In a unidirectional corpus, the translations occur from a source language to a target language. In a bidirectional corpus, the translations occur in both directions. Searches, however, can still be conducted in either language. What do we call a bilingual, multimodal corpus that is only machine-readable in one of its languages? Undoubtedly, in this case, it is an asymmetrical one favoring the majority language with a written system.

4.2.3 Classification of Metadata

Because the GUDA project utilizes Gallaudet's diverse video collection that was originally recorded for numerous reasons, measures are being taken to gather metadata that is useful to corpus research, such as the IMDI initiative and elaborative considerations presented during the ECHO workshop (e.g., Crasborn & Hanke, 2003). Additionally, because the videos were not originally collected for the purpose of being included in a corpus, GUDA researchers are also engaging in re-consent measures, as considered by others (e.g., Chen Pichler, Hochgesang, Simons, & Lillo-Martin, 2016).

The parallel corpus has additional considerations as well. In traditional corpora, a "free translation" is created by an annotator who is not considered a primary participant. However, the interpretations and translations in the parallel corpus were not created by researchers but by participants in the original communicative event. For this reason, participants may or may not be visible in the recordings. That is, in the case of a presentation, the presenter as well as the interpreter, and, if present, the real-time captioner, are all participants and are all providing analyzable utterances in various modalities.

As discussed, annotators are typically not considered participants. However, the work of an annotator has potential as another data point, in which case the "free translation" tier provided by an annotator would have non-traditional benefit within a parallel corpus. Basic identifying information is collected on annotators, but a parallel corpus seaking to analyze the resulting translations may need to consider metadata on par with that collected for any other participant.

4.3 Data Quality of SRT Files

A closer look at the data showed some SRT files were either scarce, partially incomplete, or significantly misaligned. To resolve SRT files with scarce, unuseable information, we recommend including a step to eliminate files below a certain size from the process. Partially incomplete SRT files may be difficult to detect. At this time, we do not have a recommendation for how to eliminate or correct them, other than manual inspection.

Some subtitles in SRT files are significantly misaligned, usually due to errors in timestamping. One possible solution is to add code to check whether the beginning and ending timestamps of an SRT file fall within the duration of its respective video.

4.4 Alignment

Alignment is a significant issue in creating parallel corpora, and it is especially difficult with sign language corpora. Both Meurant, Cleve, and Crasborn (2016) and Wehrmeyer (2019) report on the difficulties in aligning translations with their signed segments.

In Event Type A (ASL) events, SRT files typically segment translations into what can fit and be comfortably read on a screen at the time of its corresponding ASL utterance. While segmentation in traditional sign language corpora is concerned with linguistic boundaries (e.g., Ormel and Crasborn, 2011), segmentation in SRT files is bound more by technological and pragmatic constraints.

In Event Type B (ASL>English) and C (English>ASL) the issue of segmentation is compounded by latency effects. In Event Type B events, the onset of the English translation will typically occur some brief amount of time after the ASL utterance; similarly, in Event Type C events, the onset of the ASL translation will occur some brief amount of time after the English utterance. If real-time transcription services are provided, another level of latency may be introduced between the source and its representation in written English.

Depending on the purpose of the corpus, the latency may be informative in and of itself. For example, in section 2.4 we reported on the study by Cokely (1986), who identified a negative correlation between the duration of the latency and the number of errors committed by the interpreter.

For our initial application of the corpus, which was to find ASL matches used in context for English terms, matches do not have to be perfectly aligned. Users will likely scroll some time before and after the match to understand the context. Nonetheless, matches must appear within a reasonable window of the segment returned by ELAN.

Other purposes may require a more exact alignment than that provided by SRT files. For instance, data used to train a machine-learning algorithm would have to be well aligned so as not to train the algorithm on the wrong data.

4.5 User Engagement

Many uses of parallel sign language corpora beyond research have been proposed (e.g., Meurant, Cleve, & Crasborn, 2016; Roush, 2016; Wehrmeyer, 2019), yet we do not fully understand the average user's needs, wants, or expectations. Indeed, there is arguably no average user at this point.

We offer the idea that if parallel sign language corpora are to extend beyond the researcher's laboratory, we must investigate users' engagement with the corpora. Therefore, our next steps include gathering reactions from students and professionals to our corpus. Questions to investigate include, for example, how useful is a parallel corpus that can only be searched in one of its languages (e.g., based on English searches)? What features would users want added? What groups of users (translators, interpreters, language learners, educators, etc.) find the corpus most beneficial? Can users think of other implementations of the corpus the researchers have yet to identify?

4.6 Open Access and Sustainability

Currently, the corpus is stored on a hard drive and on three computers in a lab housed by the Department of Interpretation and Translation at Gallaudet University. The next step would be to make this corpus available online so that it is citable and others can benefit from it (e.g., Berez et al, 2018). But this is an issue that other interpreting corpora have. Once the GUDA project identifies a suitable digital location, the data (including the parallel corpus described here) will be made available. We are exploring the possibility of maintaining a dynamic site which we would continuously update with our ongoing work. In addition, we would periodically archive our data (most likely, on a triennial basis) with reputable language archives such as The Language Archive. Stable archived data will ensure access to the corpus beyond the researchers' time at Gallaudet.

4.7 Ethics of Mining Existing Videos

Along with the issue of citability and open access of our parallel corpus is the issue of ethics. By that, we are referring to the need to respect the privacy of the people in the data (their images, voices, and any other identifiable information). Our data comes from existing online videos. Currently there is little consensus regarding treatment of such data—usually because most of those have been based on wholly written texts. The situation is different when it comes to signed languages because we cannot avoid picturing people when we represent their language use. Because our data contains English data, we are including voices of people as well. Management of this kind of potentially identifying data needs to be considered. We will attempt to re-consent all videos by contacting people who are included in the videos (much like outlined in Chen Pichler et al., 2016). Given the immense logistics of such an endeavour, however, one other ongoing solution we are test-driving is an "opt out" mechanism to be provided with each video. We will defer to the preferences of those who appear in the data.

4.8 Growth and Refinement

As stated earlier, the quality of the SRT files was sometimes less than desirable, and it skewed our corpus data. Scripts were written to handle the data at hand and were therefore useful to construct only this corpus. Moreover, since the construction of this corpus in November 2019, we have identified other video sources and collaborators.

Therefore, a future iteration of this parallel corpus would include writing more general scripts designed to work with any incoming data and that can be shared with the wider research community. In addition, procedures would be put in place to handle and flag the issues identified in this paper (e.g., removing or repairing corrupt SRT files). With new video sources and a more streamlined process for categorizing the videos, we could also investigate the feasibility of creating specialized corpora.

5. Conclusion

This is a first attempt at creating a sizable parallel corpus, albeit only searchable through one of its languages, by leveraging SRT files. Lessons learned and considerations outlined in this paper may serve as a blueprint for future endeavors and other scholars.

6. Bibliographical References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honor of John Sinclair* (pp. 233–252). John Benjamins.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics.* Edinburgh University Press.

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., & Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1), 1–18.

Chen Pichler, D., Hochgesang, J., Simons, D., & Lillo-Martin, D. (2016). Community Input on Re-consenting for Data Sharing. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), Workshop Proceedings: 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining / Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (29-34). Paris: European Language Resources Association (ELRA).

Cokely, D. (1986). The effects of lag time on interpreter errors. *Sign Language Studies, 53*, 341–375. https://doi.org/10.1353/sls.1986.0025

Crasborn, O., & T. Hanke. (2003, May 8–9). (version Jan 2010). Metadata for sign language corpora. Background document for an ECHO workshop, Radboud University Nijmegen.

Fantinuoli, C. & Zanettin, F. (2014). Creating and using multilingual corpora in translation studies. In C. Fantinuoli & F. Zanettin (Eds.), *New directions in corpus-based translation studies* (pp. 1–10). Language Science Press.

Frishberg, N. (2010, December 3). *Repurposing corpus materials for interpreter education* [Workshop presentation]. Sign Linguistics Corpora Network, Berlin, Germany. https://www.ru.nl/publish/pages/607111/slcn4_frishberg.pdf

Gallaudet University. (2007, November). *Mission and Goals. https://www.gallaudet.edu/about/planning-for-the-future/mission-and-goals*

Hellwig, B., Van Uytvanck, D., … & Geerts, J. (2019). *ELAN - Linguistic annotator* (Ver. 5.8). The Language Archive. https://www.mpi.nl/corpus/manuals/manual-elan.pdf

Heßmann, J., & Vaupel, M. (2008). Building up digital video resources for sign language interpreter training. In O. Crasborn et al. (Eds.), *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora* (pp. 74–77). European Language Resources Association.

Hochgesang, J. A., Crasborn, O., & Lillo-Martin, D. (2018). Building the ASL Signbank: Lemmatization principles for ASL. In M. Bono et al. (Eds.), *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community* (pp. 69–74). European Language Resources Association.

Hochgesang, J. A., Willow, J., Treviño, R., & Shaw, E. (2019, September 26–28). *Gallaudet University Documentation of ASL (GUDA) - Whither a corpus for ASL?* [Poster presentation]. 13th Conference of Theoretical Issues in Sign Language Research (TISLR13), Hamburg, Germany. https://doi.org/10.6084/m9.figshare.9842696.v1

McEnery, T., & Hardie, A. (2012). *Corpus linguistics.* Cambridge University Press.

Meurant, L., Cleve, A., & Crasborn, O. (2016). Using sign language corpora as bilingual corpora for data mining: Contrastive linguistics and computer-assisted annotation. In E. Efthimiou et al. (Eds.), *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining* (pp. 159–166). European Language Resources Association.

Ormel, E., & Crasborn, O. (2011). Prosodic Correlates of Sentences in Signed Languages. A Literature Review and Suggestions for New Types of Studies. *Sign Language Studies*, 12(2), 279–315.

Shlesinger, M. (1998). Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta, XLIII* (4).

Roush, D. (2016, October 26–29). *Learning benefits of a translation corpus for novice ASL-English interpreters* [Poster presentation]. 2016 Biennial Conference of the Conference of Interpreter Trainers, Lexington, KY.

Wehrmeyer, E. (2019). A corpus for signed language interpreting research. *International Journal of Research and Practice in Interpreting, 21*(1) 62–90. https://doi.org/10.1075/intp.00020.weh

7. Language Resource References

Crasborn, O., Zwitserlood, 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, ISLRN 175-346-174-413-3.

Hochgesang, J.A., Crasborn, O., & Lillo-Martin, D. (2020) ASL Signbank. New Haven, CT: Haskins Lab, Yale University. https://aslsignbank.haskins.yale.edu/

Meurant, L. (2015). Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB). Laboratoire de langue des signes de Belgique francophone (LSFB-Lab). FRS-F.N.R.S. et Université de Namur.