BSL-Hansard: A parallel, multimodal corpus of English and interpreted British Sign Language data from parliamentary proceedings

Euan McGill
Universitat Pompeu Fabra
Barcelona, Spain
euan.mcgill@upf.edu

Horacio Saggion
Universitat Pompeu Fabra
Barcelona, Spain
horacio.saggion@upf.edu

Abstract

BSL-Hansard is a novel open source and multimodal resource composed by combining Sign Language video data in BSL and English text from the official transcription of British parliamentary sessions. This paper describes the method followed to compile BSL-Hansard including time alignment of text using the MAUS (Schiel, 2015) segmentation system, gives some statistics about this dataset, and suggests experiments. These primarily include end-to-end Sign Language-to-text translation, but is also relevant for broader machine translation, and speech and language processing tasks.

1 Introduction

In the United Kingdom (UK), there are an estimated 151,000 British Sign Language (BSL) signers according to the British Deaf Association[1] many of whom constitute the d/Deaf and Hard-of-Hearing (DHH) community in that country. BSL is a flourishing language, and has seen a 40% increase in the number of people who identify their main language as BSL in the ten years between the 2011 and 2021 Census in England and Wales[2].

d/Deaf signers prefer to access information and use technology in their native language (Yin et al., 2021) which is, in many cases, a sign language (SL). However, technologies such as machine translation (MT) for sign languages (SLT) are much less well-established compared to their spoken language counterparts (Bragg et al., 2019; Núñez-Marcos et al., 2023). This means that many DHH individuals must opt for resources in their non-primary language, often the ambient spoken language in the territory - for example English where BSL is used.

In recent years, there has been marked progress in the provision of information and services for BSL signers. For example, a growing proportion of public service television broadcasting is available with BSL interpretation and members of the DHH community are becoming more prominent in the national media[3]. The recent British Sign Language Act 2022 has also enshrined in law BSL's status as an official language of England, Scotland, and Wales. However, there remains a comparatively small amount of data available to develop language technology resources for BSL. The BSL-Hansard dataset intends to make a large amount of parallel English-BSL data available to researchers.

Section 2 explores resources available for BSL, before Sections 3 and 4 introduce and describe the parallel BSL-Hansard dataset of English-BSL parliamentary utterances. Section 5 then discusses possible uses and experiments with the dataset, and offers concluding remarks.

2 BSL resources

There is already a body of extant resources available for SLT research using BSL. Perhaps the most prominent is the BSL Corpus (Schembri et al., 2013) which is the first digitised corpus of continuous BSL. It contains an impressive amount of vari-

[1] https://bda.org.uk/help-resources/

[2] https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/bulletins/languageenglandandwales/census2021#main-languages-varied-across-england-and-wales

[3] https://www.theguardian.com/society/2021/dec/25/rose-ayling-elliss-strictly-come-dancing-win-gives-deaf-children-huge-confidence-boost

ation from elicited and natural conversation, with 249 signers across eight British cities and is intended for a broad range of research tasks. Annotation is currently incomplete, so it is important to pursue other data collection projects. There exist other resources for BSL including the ECHO corpus (Brugman et al., 2004) with sign video and extensive linguistic annotation, and Dicta-Sign[4] which contains isolated sign videos.

Another resource is the BOBSL (Albanie et al., 2021) parallel English-BSL dataset. Similar to BSL-Hansard, BOBSL was created by collating $1,400$h of a broad range of BBC television programmes and their companion BSL interpretation. It is a valuable resource which is large enough to conduct machine learning research, shown by the researchers' experiments on SLT, sign language recognition (SLR), and sentence alignment. However it seems that although the corpus is free to use, each researcher must request access individually which makes it impractical to leverage its data in large, commercial projects (De Sisto et al., 2022).

Other resources may be generated through data augmentation, transfer learning, and bootstrapping techniques from better-resourced SLs such as American Sign Language or from spoken languages (Moryossef et al., 2021; Zhou et al., 2021). This type of data may be suboptimal (Yin and Read, 2020) as they are not a genuine representation of a SL, and the same may be said about data from SL interpretation (Bragg et al., 2021). However, these are currently frequently-utilised ways of obtaining sufficient quantities of data for data-hungry machine learning approaches.

2.1 BSL in Parliament

There has been BSL interpretation for every edition of Prime Minister's Questions (PMQs) and Budget statements in the British House of Commons since early 2020[5]. More recently (since January 2023), the session immediately before PMQs is interpreted. In addition, there are plans to interpret a greater number and wider range of parliamentary from summer 2023 which will provide an even larger amount of parallel data available.

Every session in the UK Parliament is transcribed in English in a "substantially verbatim"[6] manner, and is kept as public record in Hansard. Every session is also publicly available in video and audio on the Parliamentlive.tv web service. As such, this allows for alignment in parallel between BSL video and English text and audio. The following sections first describe the amount of data that is available and used for the purpose of compiling this parallel resource, followed by the method used to compile it, and then a discussion of its use and place in the wider literature.

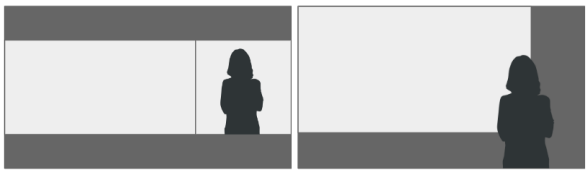

Figure 1: Signer framing type in SL videos

3 Dataset statistics

BSL-Hansard contains 86h40m of SL video in *.mp4* format from 19 individual signers. There is no additional demographic information, as there is no extant source of the interpreters self-identifying. Appendix 1 shows the amount of sessions interpreted by each signer by alias, as well as a suggested split into train, development, and test splits whereby no individual signer appears in more than one of these sets. The exact split is 62% for training, 18% for development, and 20% for test. These sets are slightly uneven due to the fact that some signers co-appear in some videos.

The videos frame the signer in two distinct ways, shown in Figure 1. The first separates the signer into a box with a plain background (left), which takes up approximately one third of the video frame. The second superimposes the signer in the bottom right-hand corner of the screen over a mixture of footage from partially the parliamentary chamber and partially a plain background (right). There are 34 instances of the former type in the corpus, and 78 instances of the latter.

The accompanying transcripts total 871k words in English, which are aligned on timestamped sentences to the appropriate video. There are 18.9k unique words where 4.6k overlap with the large SignBSL[7] dictionary resource. The most frequently-occurring non-stopword in the dataset is "prime" which appears 8.7k times. There are 112 individual sessions, and the nine session types

[4]https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/

[5]https://www.parliament.uk/business/news

[6]As well as text on parliamentary procedure, "members' words are recorded, and then edited to remove repetitions and obvious mistakes, albeit without taking away from the meaning of what is said" (https://hansard.parliament.uk/)

[7]https://www.signbsl.com/about

are distinguished by video and transcript titles in the dataset. Appendix 1 provides information about the types of parliamentary session which make up the dataset and define how the files are labelled.

4 Dataset compilation

Videos in *.mp4* format are manually downloaded from the Parliamentlive.tv web service, and the official transcripts are manually downloaded from the Hansard web page.

The videos and texts are then processed predominantly using the functionality of the Munich Automatic Segmentation System (MAUS) (Schiel, 2015). MAUS is a Hidden Markov Model-based statistical forced aligner which first predicts the phonetic label based on an input transcript, and then aligns the predicted phones with an input audio signal. This service is available on the web (Kisler et al., 2017), and can be used with other functionalities such as pre-processing, grapheme-to-phoneme conversion, and subtitle generation.

Figure 2 provides an overview of the processing tasks and file types involved. A given input video with a maximum duration of no more than nine minutes is matched with the appropriate Hansard transcript, and converted to *.wav* format using the *ffmpeg* library. The input text is pre-processed by removing all content inside parentheses and square brackets, as well as the first three lines of procedure in each Hansard document.

The resulting *.txt* and *.wav* files are input into the 'WebMAUS Basic'[8] web service where the British English language model is chosen, and output format is set to *.bpf* - a file type which allows for time alignment between the phonetic transcription and the audio signal.

The text file containing the original transcription and the aligned *.bpf* file are subsequently input into the BAS 'Subtitle'[9] web tool which maps the alignment with the original transcript in order to generate sentence-type utterances. In order to preserve full phrases as well as possible, the parameters are set to split subtitles on punctuation marks, or otherwise at a maximum length of 20 words - the result is output to *.vtt* file format. These files may be converted to a researcher-friendly *.csv* or *.json* formats by means of straightforward, freely-available conversion scripts[10].

4.1 Dataset storage, usage and reproducibility

This dataset is stored as open access in a Zenodo[11] repository. The processing scripts and tools, as well as tools to isolate the signer in both framing types, are stored in a Github repository[12].

BSL-Hansard is stored in this way due to the terms of use[13] of the UK Parliament's web services. It is possible to store excerpts of parliamentary sessions in a manner available to everyone, but in context and without editing or manipulating the video or audio feeds in any way. It also allows this resource to be available on a platform which is robust and secure.

5 Uses, discussion and future steps

It is possible for researchers, particularly those on machine translation between signed and spoken languages, to use BSL-Hansard in many ways. This section describes some experiments that are possible to conduct with this data, and experiments that will improve the data inside the dataset. It also describes some of the limitations of the dataset and this type of dataset in general. Finally, there is a brief note on the extensibility of this dataset and the methods used to compile it before some concluding remarks.

5.1 Sign Language translation

The first is end-to-end (E2E) sign language translation, in other words going from sign language video directly to text. These methods are based on Transformer encoder-decoder architecture (e.g. (Liu et al., 2020)). A system introduced in Camgöz et al. (2020) can jointly learn SLR and translation, and negates the need to go through an intermediate step of SL gloss-to-text transformation. They achieved state-of-the-art performance at the time on the PHOENIX-Weather (Camgöz et al., 2018) German Sign Language corpus. An interesting next step would be to implement an E2E method using BSL-Hansard videos. The BOBSL

[8] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic

[9] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Subtitle

[10] e.g. https://github.com/iTrauco/vtt-to-csv-python-script

[11] https://zenodo.org/record/7974945

[12] https://github.com/LaSTUS-TALN-UPF/BSL-Hansard-tools

[13] https://www.parliament.uk/site-information/copyright-parliament/pru-licence-agreements/downloading–sharing-terms–conditions/

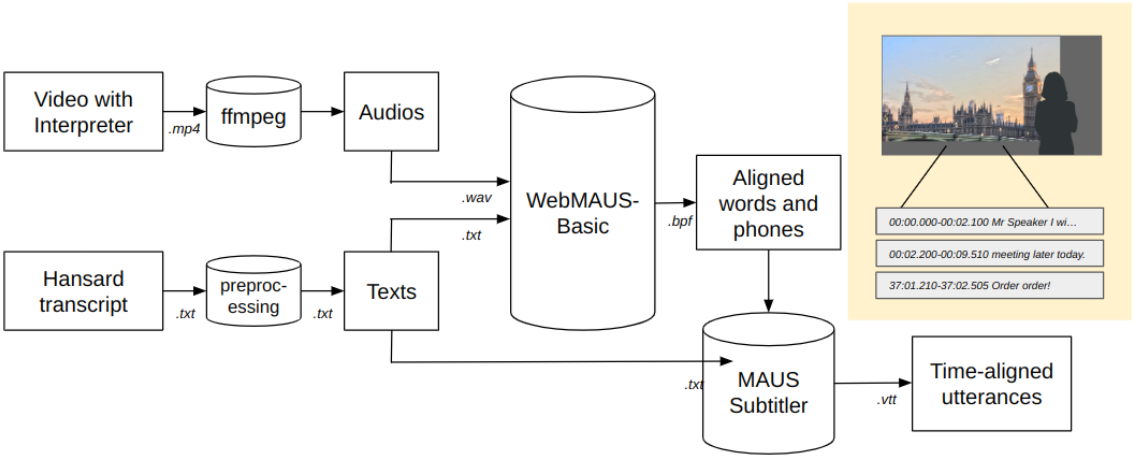

Figure 2: Processing pipeline between input video and text into time-aligned labelled text, including the file input and output types involved. The dataset is made up of the video files and time aligned text captions (pictured, top right).

authors (Albanie et al., 2021) did implement an E2E transformer-based methodology with limited success (1.00 BLEU-4), citing the unconstrained settings, large vocabulary and wide domain of their dataset. It is possible that better results may be achieved on this dataset despite its much smaller size as it is much more domain-specific by only containing parliamentary exchanges.

It may also be beneficial to implement other E2E methods, such as the recent SLTUNET model (Zhang et al., 2023). Also, the STMC transformer (Yin and Read, 2020) has been shown to outperform Camgöz et al. (2020)'s results and to be generalisable to other datasets.

5.2 Annotation, alignment and recognition

As E2E translation is the only translation type possible, due to the SL video not having annotations, it may be beneficial to label this dataset with SL glosses. Fortunately, BSL has rich dictionary resources to draw from with relatively large vocabulary sizes.

Outwith the joint approach in E2E SLT, labelling can be achieved through continuous SLR (Wadhawan and Kumar, 2021). The two different video framing settings may be challenging, but signers are consistently directly facing the camera and dressed in grey formal clothing.

While the dataset is labelled with aligned English text, it is also possible to perform alignment as an automated annotation strategy known as 'sign spotting'. As proposed in Albanie et al. (2020), keywords may be spotted through mouthings (a frequently-used articulator in SL inventories) using computer vision techniques. In addition, signs can be spotted through comparing them to SL lexicons - as previously mentioned, these are well resourced for BSL. Sign spotting may be a fruitful technique for this dataset specifically, as parliamentary procedure makes terms such as 'Mr. Speaker' or 'prime minister' occur very frequently which means these can be used as temporal keypoints for further annotation.

These methods may be considerably less accurate than manual transcription, but are far less human resource-intensive.

5.3 Limitations and opportunities

The main limitation of interpreted SL data, which makes up all of BSL-Hansard, is that it lacks the naturalness and regional (Sutton-Spence and Woll, 1999) and sociolinguistic (Lucas and Bayley, 2016; Schembri et al., 2018) variation of native and conversational BSL. It is also important to note that interpreted SL may not convey the entire message of the spoken language data due to brevity restrictions and errors which naturally occur during live interpretation. That being said, this data is still valuable as the sheer amount of parallel sentences in one domain allow the implementation of machine learning techniques.

This dataset is also readily extensible, as there is a constant and increasing stream of BSL-interpreted parliamentary sessions becoming available. It may also be possible to extend this methodology of dataset compilation into, for example, the Scottish and Catalan Parliaments which both have signed video and official transcripts available to download.

Acknowledgements: This work has been conducted within the SignON project. SignON is a Horizon 2020 project, funded under the Horizon 2020 program ICT-57-2020 - "An empowering, inclusive, Next Generation Internet" with Grant Agreement number 101017255.

We also acknowledge support from the Spanish State Research Agency under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M).

References

Albanie, Samuel, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*.

Albanie, Samuel, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset.

Bragg, Danielle, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

Bragg, Danielle, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. 2021. The fate landscape of sign language ai datasets: An interdisciplinary perspective. 14(2):1936–7228.

Brugman, Hennie, Onno Crasborn, and Albert Russel. 2004. Collaborative annotation of sign language data with peer-to-peer technology. In Lino, Maria Teresa, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 213–216, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Camgöz, Necati, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *CVPR 2018*, pages 7784–7793, 03.

Camgöz, Necati Cihan, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR 2020*, pages 10020–10030.

De Sisto, Mirella, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France, June. European Language Resources Association.

Kisler, Thomas, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Liu et al. 2020. Multilingual denoising pre-training for neural machine translation. *CoRR*, pages 1–17.

Lucas, Ceil and Robert Bayley. 2016. Quantative sociolinguistics and sign languages: Implications for sociolinguistic theory. In Coupland, Nikolas, editor, *Sociolinguistics: Theoretical Debates*, chapter 16, pages 349–366. Cambridge University Press, Cambridge, United Kingdom.

Moryossef, Amit, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation.

Núñez-Marcos, Adrián, Olatz Perez de Viñaspre, and Gorka Labaka. 2023. A survey on sign language machine translation. *Expert Systems with Applications*, 213:118993.

Schembri, Adam, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. 2013. Building the british sign language corpus. *Language Documentation and Conservation*, 7:136–154.

Schembri, Adam, Rose Stamp, Jordan Fenlon, and Kearsy Cormier. 2018. Variation and change in varieties of british sign language in england. In Braber, Natalie and Sandra Jansen, editors, *Sociolinguistics in England*, chapter 7, pages 165–188. Palgrave Macmillan, London, United Kingdom.

Schiel, Florian. 2015. A statistical model for predicting pronunciation. In *International Congress of Phonetic Sciences*.

Sutton-Spence, R. and B. Woll. 1999. *The Linguistics of British Sign Language: An Introduction*. The Linguistics of British Sign Language: An Introduction. Cambridge University Press.

Wadhawan, Ankita and Parteek Kumar. 2021. Sign language recognition systems: A decade systematic literature review. 28:785–813.

Yin, Kayo and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *COLING 2020*, pages 5975–5989, Online. ICCL.

Yin, Kayo, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing.

Zhang, Biao, Mathias Müller, and Rico Sennrich. 2023. SLTUNET: A simple unified model for sign language translation. In *The Eleventh International Conference on Learning Representations*.

Zhou, Hao, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation.

Appendix 1: Dataset statistics

Type	Description	Total
PMQs	Prime Minister's Questions	110
Budget	Financial statements/budgets from the Chancellor of the Exchequer	7
Covid	Statements about the Coronavirus pandemic	4
NIQs	Questions to the Secretary of State for Northern Ireland	2
SCQs	Questions to the Secretary of State for Scotland	2
WEQs	Questions to the Minister for Women and Equality	2
AGQs	Questions to the Attorney General	1
CYQs	Questions to the Secretary of State for Wales	1
SITQs	Questions to the Minister for Science Technology and Innovation	1
Afghanistan	Updates on the conflict in Afghanistan	1

Table 1: Session types and frequency in the dataset

Signer	# sessions	Partition
S101	1	Test
S102	16	Train
S103	2	Test
S104	2	Train
S105	1	Dev
S106	1	Test
S107	6	Train
S108	7	Train
S109	3	Train
S201	5	Test
S202	4	Dev
S203	10	Test
S204	15	Dev
S205	24	Train
S207	16	Train
S208	1	Test
S209	1	Train
S210	6	Test
S211	2	Dev

Table 2: Individual signer IDs used in the corpus, number of occurrences in sessions, and place in the dataset partition