Sign Language Technologies: Recognising Spatial Referencing Structures

Tom Pham

February 2025

Abstract

Sign languages rely heavily on spatial referencing to convey grammatical and contextual information through partially lexical signs (PLSs) such as depicting signs, pointing signs, and fragment buoys. However, current sign language recognition (SLR) systems predominantly focus on fully lexical signs (FLSs), leading to an underrepresentation of these spatial structures in popular datasets despite their prevalence in natural signing. This study investigates the capacity of existing SLR models to recognise spatially dependent PLSs across different signed languages, using Belissen et al.'s (2020) framework—a pose estimation and bidirectional LSTM model initially developed for French Sign Language (LSF)—applied to the British Sign Language (BSL) Corpus. The research evaluates the model's generalisability by analysing its performance in categorising FLSs and PLSs, with aims for improving systems targeting Australian Sign Language (Auslan). Results revealed that while the model achieved comparable accuracy for FLSs (F1: 0.52 vs. 0.54) and pointing signs (F1: 0.12 on BSL vs. 0.15 on LSF), depicting signs (F1: 0.14 vs. 0.30) and fragment buoys (F1: 0.00 vs. 0.10) showed markedly lower performance on BSL, correlating with their minimal representation in training data, and inconsistent pose estimation results. These disparities highlight the impact of data scarcity on underrepresented PLS categories and the need for linguistically diverse, balanced corpora to improve model robustness. Future work will focus on adapting the framework to Auslan-specific data to further evaluate the significance of spatial referencing in the production of accurate SLR translations, and an exploration of alternative pose estimation techniques to enhance tracking precision during rapid movements.

1 Introduction

The aim of this project was to investigate the role of spatial referencing in signed languages, and how effectively existing sign language recognition systems capture the nuanced spatial and referential elements of signing. This involved an exploration of the automatic SL structure recognition model built by Belissen et al. (2020) [1], which aims to validate whether spatial and referential elements of signing can be computationally detected through the recognition of partially lexical signs.

This research applies the above SL structure recognition model to segments of the BSL Corpus Project to evaluate whether different signing structures can be identified in British Sign Language, including: fully lexical signs, pointing signs, depicting signs, and floating buoys. This approach also allowed for the generalisability of the French LSF pretrained model to be tested against British Sign Language structures, offering a comparative baseline for the results. The scope of this project is limited to evaluating the existing LSF pre-trained model on BSL data. Due to time and resource constraints, retraining the model on BSL data was not undertaken.

This research contributes to the field by examining whether sign language recognition models, originally trained to detect spatial structures in one language, can be effectively adapted to another. BSL shares linguistic roots with Auslan (Australian Sign Language), and this relationship presents an opportunity to extend the research toward evaluating the model's applicability to Auslan, which could help improve SL recognition for Australian signers.

2 Literature Review

Sign language recognition (SLR) is the complex task of detecting and labelling signs from a video; it can be regarded a substep of sign language translation (SLT), which carries the end goal of translating sign language information into spoken language sentences. SLR systems typically involve assigning gloss labels to signs to represent their meaning (Liang et al., 2023) [2]. This can be a challenging problem, especially in the domain of continuous sign language recognition (CSLR). Not only do these systems need to segment and classify streams of multimodal signing information including handshape, hand orientation, movement and location, more nuanced features such as the signer's facial expressions and use of space also need to be considered for the lexical and grammatical information they convey.

Sign language translation, formalised by Camgoz et al. (2018) [3], goes one step further as to give coherence to the interpretation of these signs. In order to generate spoken language sentences, SLT systems face the challenge of translating unique linguistic and grammatical sign language structures, which often do not have a one-to-one mapping to their spoken language counterparts (Cihan Camgoz et al., 2020) [4].

Implementations of SLT pipelines have historically involved two key steps. First, a CSLR tokenisation system extracts meaningful features from sign language videos to generate a sequence of glosses. Then, a translation system translates the recognised sequence of glosses into spoken language (Yin & Read, 2020) [5]. However, as Yin et. al demonstrates, this Sign2Gloss2Text architecture faces several limitations. By transforming complex multi-channel signs into static labels, glosses strip away many non-manual features and spatial relationships essential for accurate translation. As a result, glosses often cause information bottlenecks when used as an intermediate representation for SLT systems.

Recent developments by Camgoz et. al. [4] has shifted from this two-step approach towards a transformer-based architecture utilising joint learning of CSLR and SLT. Vastly outperforming all comparable previous approaches at that time in both recognition and translation accuracy, their Sign Language Recognition Transformer (SLRT) exploits the supervision power of glosses without having an explicit gloss representation as an information bottleneck. The obtained state-of-the-art results on the PHOENIX2014T dataset have henceforth become the baseline for further research in the fields of SLT.

2.1 Challenges in CSLR and Impact on SLT

Despite these recent advancements in CSLR and SLT, several factors still contribute to the persistent challenge of achieving full end-to-end translations - one of these challenges being how these systems handle contextually nuanced linguistic features such as spatial referencing. Belissen et al. [1] attributes a key cause of this limitation to the fact many state-of-the-art CSLR systems almost exclusively focus on recognising fully lexical signs (FLSs) which contain [conventionalised form and meaning, but struggle with partially lexical signs (PLSs) that instead derive meaning from context and the use of space (Schonstrom & Holmstrom, 2022) [6]. Enabling important linguistic features such as spatial referencing, PLSs include depicting signs, pointing signs and fragment buoys, which are fundamental to natural signing but are underrepresented in popular training datasets like PHOENIX-2014.

2.1.1 Pointing Signs

Pointing signs are essential for maintaining coherence across sentences. They are used to introduce referents in discourse either by pointing to their actual locations in space, or by assigning to them a region of the signing space. This enables the signer to refer back to people, objects and places simply by pointing to the region previously assigned (Moryossef, 2024) [7].

2.1.2 Depicting Signs

Depicting Signs are typically one-handed signs that can be used to convey how a referent relates to other entities, its movement, and other characteristics like size and shape. For example, using a specific hand shape to represent a vehicle, a signer can demonstrate how a car swerves and crashes into another entity by moving their hand through space (Ferrara & Hodge, 2018) [8].

2.1.3 Fragment Buoys

Fragment Buoys are used to hold a fragment or represent the final posture of a two-handed sign, typically on the non-dominant hand. For example, in a sign involving the action of a building being constructed, the weak hand might hold a fragment buoy to maintain the spatial reference of the building's location while the dominant hand continues the action of describing the construction (Johnston, 2007) [9].

Depending on the type of discourse, Sallandre et al. [10] demonstrate that the ratio of PLSs to FLSs in dialogue ranges from 1:4 to as much as 4:1. As a result, many existing models trained using these datasets overlook the crucial role PLSs have in structuring discourse and grammar. Models trained on the linguistically restricted corpus PHOENIX-2014 in particular can thus be ill-equipped to pick up on these natural SL structures containing non-conventionalised meaning.

2.2 How is Poor Recognition of Spatial Referencing is Being Addressed

2.2.1 Datasets

Numerous considerations have been proposed by researchers to help improve recognition of spatial referencing in CSLR systems. Many in particular advocate for the use of more linguistically-diverse corpora for training. While datasets like PHOENIX-2014 have been widely used, they often are limited in their representation of natural signing, particularly in PLSs. Instead, researchers like Belissen et al. [1] highlight corpora developed by linguists such as the Dicta-Sign-LSF-v2 corpus. These corpora contain a more natural balance of FLSs and PLSs, consistent fine-grained annotation of different sign types, and dialogue-based content.

Recognising very few CSLR studies look at grammatical sentences, Mertz et al. [11] have also since introduced the LSF-SHELVES corpus, a Motion Capture corpus specifically aimed at providing material for studying iconicity and spatial referencing in LSF. It contains very few lexical signs, focusing more on grammatical parameters such as the location of entities relatively to spatial references. Both of these datasets have the goal of providing more comprehensive and natural sign language data for training CSLR systems that can better handle the spatial features of sign languages.

There are numerous other continuous SL RGB datasets that also include annotations other than lexical, including the BSL Corpus Project (Schembri, 2008) [12], Corpus NGT (Zwitserlood et al., 2008) [13], DGS Korpus (Prillwitz et al., 2008) [14], Auslan Corpus (Johnston, 2009) [15], NCSLGR (Neidle and Vogler, 2012) [16] and Corpus LSFB (Meurant et al., 2016). These datasets also benefit from being conversational or narrative-like in nature, providing a variety of natural sign language sentences and PLSs. However, as noted by Belissen et al. [1], these datasets are usually ignored in the field of automatic SLR and their annotations are often incomplete.

2.2.2 PLS Classification Framework

Aiming to validate whether a CSLR system could effectively recognise FLSs and PLSs in continuous signing, Belissen et al. [1] built a two-part CSLR framework that outputs binary predictions for four distinct categories - FLSs, depicting signs, pointing signs, and fragment buoys. Performing a series of experiments on the Dicta-Sign–LSF–v2 corpus, their work first involved using pose estimation to extract signer representations combining body pose, hand shape, and facial features, then feeding these through a bidirectional LSTM network to output predictions. The results showed promising capabilities in recognising PLSs, with depicting and pointing signs achieving frame-wise F1-scores

of 60%, comparable to FLSs, demonstrating that spatial and referential elements of signing can be computationally detected. However, fragment buoys were more challenging with only a 14% F1-score. This discrepancy in performance reflects a strong correlation between recognition accuracy and data availability. FLSs, which accounted for 75% of the recorded manual units, outperformed depicting signs (11%) and fragment buoys (2%), which reaffirms the need for more high-quality annotated SL data to improve CSLR model performance on PLSs.

2.2.3 Addressing Gloss & Contextual Understanding Bottlenecks

To overcome the limitations of traditional gloss annotations which oversimplify linguistic nuances and require labour-intensive labelling, recent advancements in SLT include explorations into gloss-free representation learning and context-aware modelling. Gloss-free learning methods, including the Universal Gloss-level Representation (UniGloR) framework proposed by Hwang et al. [17], aim to eliminate gloss dependency and preserve subtle articulations often lost in gloss-based approaches. The UniGlor framework achieves this by using dense spatio-temporal representations of sign keypoint sequences in place of glosses. Unlike glosses, these multi-dimensional latent representations more effectively capture subtle nuances of signing motions, providing a richer representation of sign language when used as an intermediate for SLT tasks.

To address the challenge of poor contextual understanding in CSLR systems during dialogue, Zhou et al. [18] introduced the use of Cross-Sentence Context Integration, exemplified by their SCOPE framework. Recognising that most existing methods, both in SLR and SLT, focus on translating just one sentence at a time, the SCOPE framework instead processes the contextual information of the last three dialogue sentences alongside the current motion. By processing both contextual embeddings and motion data through a transformer encoder, this approach not only improves translation coherence in multi-turn interactions but also achieves a 4.15/5 user experience rating in real-world testing, demonstrating the critical role of contextual awareness in natural signing scenarios (Liu et al., 2024) [19].

3 Methodology

3.1 Research Approach

This project evaluates the sign categorisation model developed by Belissen et al. (2020) [1] by applying it to the British Sign Language (BSL) Corpus Project. The goal is to determine whether the model can computationally detect spatial and referential elements of signing—specifically, fully lexical signs (FLS), pointing signs (PTs), depicting signs (DSs), and floating buoys (FBs)—and to assess how well a model trained on French Sign Language (LSF) performs on BSL data without retraining. The Belissen et al. model utilises a Convolutional Recurrent Neural Network (CRNN) architecture. It processes input features (body pose, hand shapes, and facial dynamics) through a 1D convolutional layer for temporal pattern extraction, followed by 1-4 bidirectional LSTM layers with dropout regularisation. A final fully connected layer with softmax activation outputs frame-wise probabilities for linguistic descriptors (e.g., lexical signs, depicting signs). Training employs RMSProp optimisation with categorical cross-entropy loss.

3.2 Experiments

The research was conducted in two phases. The first phase consisted of a literature review. The second phase focused on adapting the sign categorisation model to recognise fully lexical signs (FLS) and partially lexical signs (PLS) in BSL. This involved identifying BSL content featuring PLSs, and then extracting keypoint features using the feature extraction framework from Belissen et al. (2020) [1], configured according to their performance assessment for different signer representations. Specifically, 2D features were extracted using OpenPose for body and face keypoints, and Koller's DeepHand model [20] was used for handshape predictions. This 2D feature configuration was selected based on its strong performance across different sign types, with FLS and

pointing signs (PTs) showing slight preference for 3D features, and depicting signs (DSs) achieving optimal performance with 2D features. The extracted features were then input through the sign categorisation model, and the resulting sign predictions were evaluated against ground truth annotations. Performance was assessed using frame-wise accuracy and the F1-score, providing a balanced measure of precision and recall. Finally, the overall recognition performance on the BSL dataset was compared to the baseline results obtained using the DictaSignV2 dataset.

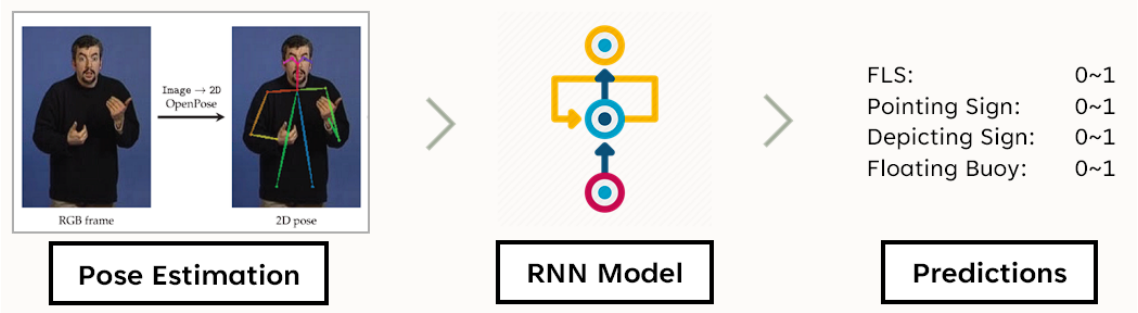

Figure 1: Sign categorisation framework created by Belissen et al. (2020) [1]

3.3 Data

Five video extracts were obtained from the BSL Corpus Project to evaluate the generalisability of Belissen et al.'s Sign Categorisation Model trained originally on French LSF (DictaSignV2 dataset). This BSL Corpus Project was chosen specifically for the similar roots shared between BSL and Auslan, the comprehensive level of annotations provided for partially lexical sign types, and the availability of narrative style dialogue which contains a natural mixture of spatial structures.

The five extracts used–BF1n, BF5n, BF18n, BF24n and G1n–are each between 0-4 minutes in length and feature a single signer seated in front of a camera. These videos were filmed at 25 fps and are stored in a .mp4 format. Annotation files are provided in ELAN Annotation Format (.eaf) format, and provide both sentence-level and gloss-level annotations (Schembri, 2017) [21].

4 Results & Discussion

Sign Type	BSL Corpus (F1-score)	DictaSignV2 (F1-score)
Pointing Signs (PTs)	0.12	0.15
Fully Lexical Signs (FLSs)	0.52	0.54
Depicting Signs (DSs)	0.14	0.30
Fragment Buoys (FBs)	0.00	0.10

Table 1: Signer-independent recognition performance comparison between the BSL corpus and DictaSignV2.

When comparing signer-independent performance between the BSL corpus and the original DictaSignV2 dataset, the model showed similar recognition accuracy for pointing signs (PTs) and fully lexical signs (FLSs). The F1-score for pointing signs was 0.12 on BSL compared to 0.15 on DictaSignV2, while FLSs achieved 0.52 on BSL versus 0.54 on DictaSignV2. These results suggest that common structural patterns exist between the two sign languages, allowing the model to generalise effectively for these categories. However, performance was significantly lower for depicting signs (DSs) and fragment buoys (FBs). The F1-score for depicting signs dropped from 0.30 in DictaSignV2 to 0.14 in BSL, while fragment buoys saw a more severe decline, from 0.10 to 0.00. Several factors may have contributed to this disparity:

- Linguistic Differences: The structure of DSs and FBs may vary significantly between French Sign Language (LSF) and BSL, leading to poor transferability of the trained model.
- Lack of Data: In the DictaSignV2 training dataset, FLSs represent 75% of manual signs, whereas DSs and PTs only account for 11%, and FBs are even more scarce, with just 589 annotated instances (2% of the dataset). The lack of training data for these categories likely hindered recognition performance.
- Absence of Eye Gaze Information: The feature extraction techniques used in this study did not capture eye gaze information as part of signer representation, despite its important role in depicting signs. This omission may have impacted the model's ability to correctly classify these signs.
- Motion Blur & Pose Estimation Errors: The OpenPose estimation framework struggled with fast hand movements, leading to motion blur that degraded recognition accuracy, particularly for signs involving rapid transitions.

4.1 Limitations

One key limitation observed was that the categorisation model rarely assigned probabilities above 0.5 (shown in figure 2), which was the positive classification threshold used in Belissen et al (2020) [1]. This lower confidence is likely the result of dissimilarities in linguistic structures between the two languages, as well as in filming conditions. To compensate, this study lowered the threshold to 0.25, which increased detection sensitivity but also raised the likelihood of false positives. Additionally, while the DeepHand model (Koller et al., 2016) [20] provided valuable handshape classification, it ignored hand orientation, which is essential for distinguishing certain signs in BSL. The model was originally trained on over one million frames from three sign language corpora (Danish Sign Language, New Zealand Sign Language, and German Sign Language). However, BSL was not among them, which may have affected its effectiveness for this specific application.

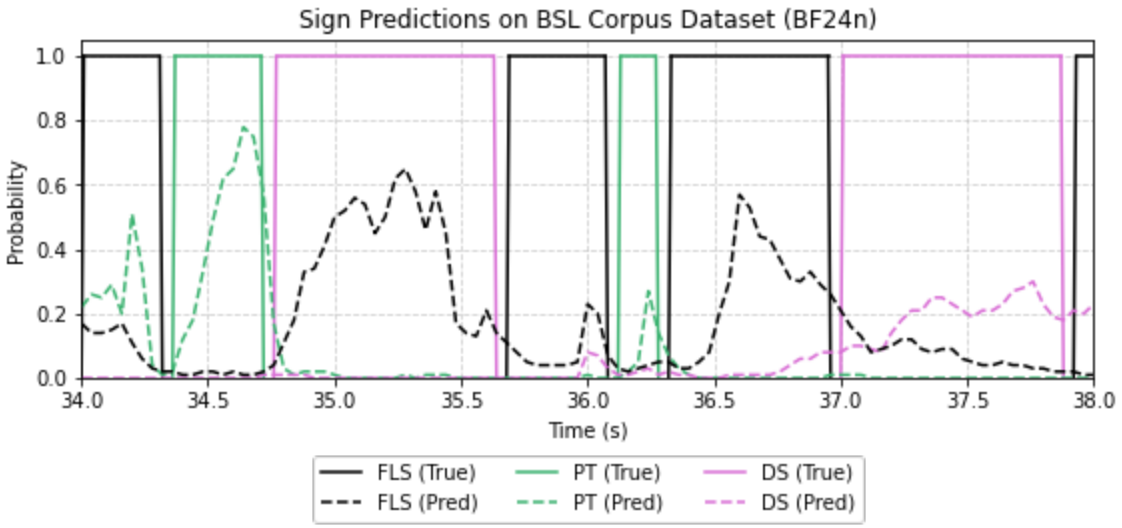

Figure 2: Prediction sequence from a four-second extract of a signed narrative video in BSL (video reference: BF24n)

5 Conclusion

This study investigated the performance of a French Sign Language (LSF) trained sign language recognition model to British Sign Language (BSL), focusing on the recognition of partially lexical signs (PLSs) crucial for conveying spatial information. The findings highlight both the potential and limitations of adapting sign language recognition models across languages. The high performance on FLSs and PTs suggests that some linguistic structures are transferable, yet the challenges with DSs and FBs emphasise the need for more diverse training data and better feature extraction methods—such as those that incorporate eye gaze tracking and more robust motion detection. Future work should focus on expanding the training dataset with additional PLS annotations and exploring alternative pose estimation techniques that can better handle fast hand movements and spatial referencing. Given that BSL shares linguistic roots with Auslan, there is also potential to extend this research to Australian Sign Language, and its application to sign language translation.

6 Acknowledgments

Data in this report were collected for the British Sign Language Corpus Project (BSLCP) at University College London, funded by the Economic and Social Research Council UK (RES-620-28-6001), and supplied by the CAVA repository. The data are copyright.

I would also like to express my gratitude to Dr. Jessica Korte for her generous support, kindness, time and for giving me the opportunity to work on such an engaging project. Your mentorship has made this experience incredibly enjoyable and rewarding!

References

[1] V. Belissen, A. Braffort, and M. Gouiffès, "Experimenting the automatic recognition of non-conventionalized units in sign language," *Algorithms*, vol. 13, no. 12, 2020.

[2] Z. Liang, H. Li, and J. Chai, "Sign language translation: A survey of approaches and techniques," *Electronics*, vol. 12, no. 12, p. 2690, 2023.

[3] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sub-word level sign language translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 702–718, 2018.

[4] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10352–10362, 2020.

[5] K. Yin and J. Read, "Better sign language translation with stmc-transformer," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5543–5555, 2020.

[6] K. Schonstrom and I. Holmstrom, "L2m1 and l2m2 acquisition of sign lexicon: The impact of multimodality on the sign second language acquisition," *Frontiers in Psychology*, vol. 13, p. 896254, 2022.

[7] A. Moryossef, "sign.mt: Real-time multilingual sign language translation application," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2024.

[8] L. Ferrara and G. Hodge, "Language as description, indication, and depiction," *Frontiers in Psychology*, vol. 9, p. 716, 2018.

[9] T. Johnston and R. Adam, *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press, 2007.

[10] M.-A. Sallandre, B. Antonio, and G. B. B. e. Garcia, "Étude exploratoire de la fréquence des catégories linguistiques dans quatre genres discursifs en lsf," *Lidil*, no. 71, pp. 165–190, 2019.

[11] C. Mertz, V. Barreaud, T. Le Naour, D. Lolive, and S. Gibet, "A low-cost motion capture corpus in french sign language for interpreting iconicity and spatial referencing mechanisms," in *Language Resources and Evaluation*, pp. 1–27.

[12] A. Schembri, "British sign language corpus project," 2008–2017.

[13] I. Zwitserlood, O. Crasborn, D. van Gessel, E. Kooijman, C. Koster, M. de Weerd, G. Baker, and T. Hanke, "Corpus ngt: A video corpus of sign language of the netherlands," 2008. Technical report, Radboud University Nijmegen.

[14] S. Prillwitz, B. Zienert, and M. Hennig, "Dgs-korpus: A video corpus of german sign language," 2008. Technical report, University of Hamburg.

[15] T. Johnston, "The auslan corpus: A resource for sign language and linguistic research," 2009. Technical report, Macquarie University.

[16] C. Neidle and C. Vogler, "A new web interface to facilitate access to corpora: Development of the asllrp data access interface," 2012.

[17] E. J. Hwang, S. Lee, H. Lee, Y. Yoon, and J. C. Park, "A spatio-temporal representation learning as an alternative to traditional glosses in sign language translation and production," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, July 2024.

[18] Y. Zhou, Y. Liu, S. Ren, C. Huang, J. Yu, and L. Xu, "Sign language contextual processing with embedding from large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Association for Computational Linguistics, October 2024.

[19] Y. Liu, W. Zhang, S. Ren, C. Huang, J. Yu, and L. Xu, "Scope: Sign language contextual processing with embedding from llms," 2024. Preprint.

[20] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3792–3801, 2016.

[21] A. Schembri, J. Fenlon, R. Rentelis, and K. Cormier, "British sign language corpus project: A corpus of digital video data and annotations of british sign language 2008-2017," 2017. `https://www.bslcorpusproject.org`.