Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC 2022), pages 94–101
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

A Machine Learning-based Segmentation Approach for Measuring Similarity Between Sign Languages

Tonni Das Jui, Gissella Bejarano, Pablo Rivas
Baylor University
1311 S 5th St, Waco, TX 76706
{tonni_jui1, gissella_bejaranonic, pablo_rivas}@baylor.edu

Abstract

Due to the lack of more variate, native and continuous datasets, sign languages are low-resources languages that can benefit from multilingualism in machine translation. In order to analyze the benefits of approaches like multilingualism, finding the similarity between sign languages can guide better matches and contributions between languages. However, calculating the similarity between sign languages again implies a laborious work to measure how close or distant signs are and their respective contexts. For that reason, we propose to support the lexical similarity measurement between sign languages through a video-segmentation-based machine learning model that will quantify this match among signs of different sign languages. Using a machine learning approach, the similarity measurement process can run more smoothly than a more manual approach. We use a pre-trained temporal segmentation model for British Sign Language (or BSL). We test it on three datasets, an American Sign Language (ASL) dataset, an Indian Sign Language (ISL), and an Australian Sign Language (or Auslan) dataset. We hypothesize that the percentage of segmented and recognized signs by this machine learning model can represent the percentage of overlap or similarity between British and the other three sign languages. In our ongoing work, we evaluate three metrics considering Swadesh's and Woodward's list and their synonyms. We found that our intermediate-strict metric coincides with a more classical analysis of the similarity between British and American Sign Language, as well as with the classical low measurement between Indian and British sign languages. On the other hand, our similarity measurement between British and Australian Sign language holds for part of the Australian Sign Language and not the whole data sample.

Keywords: Sign language, Similarity, Machine learning, Segmentation model

1. Introduction

Measuring the similarity of sign languages can enhance research on genealogical, social, and other relations between different signed languages and regions. Besides, it can help understand Deaf culture, origins, and evolution. As reported in (Börstell et al., 2020), one of the largest sign language databases, Glottolog 4.1[1] (Hammarström et al., 2019) contains 194 sign languages datasets whose relations are not known or analyzed enough. Measuring similarity between specific sign languages can help reuse resources in a multilingualism approach, such as in machine translation (Bapna et al., 2019). We can bridge communication gaps between signers and speakers with properly annotated sign language datasets, scaled analysis, and machine learning technology.

Sign language similarity usually focuses on measuring lexical similarity across the signs, extracting features manually and under the subjectivity of the different experts. This approach can be very time-consuming due to the exhausting visual analysis that needs to be performed by a person. In that sense, more systematic approaches can support or complement this analysis by using machine learning methods. More specifically, sign languages similarity measurement is a process that can benefit from more computational approaches such as computer vision and natural language processing. Moreover, computer-vision approaches are preferred when working with sign language processing because they are less intrusive and less laborious. For example, most recent research is obtaining good results in sign language segmentation to find temporal boundaries of signs and recognition to identify a segment of a video with a corresponding sign (Renz et al., 2021a; Renz et al., 2021b; Bull et al., 2021; Varol et al., 2021; Camgoz et al., 2020).

Our work proposes to use a segmentation model to measure the sign languages similarity. For that goal, we use a pre-trained segmentation model in one sign language, such as BSL (Cormier and Fenlon, 2014; Fenlon et al., 2011), and measure how well it can segment and recognize signs in a second sign language. We evaluate different strict-level metrics, such as raw or exact match and match, considering synonyms. We use the vocabulary provided in Swadesh's list (Swadesh, 1971), and Woodward's list (Woodward, 2000) to compare to previous and future work. Our results show relative values to the previously-reported classical similarity-measure method comparing BSL to ASL and ISL. On the other hand, even when our similarity measure between BSL and AUSLAN categorizes them as the languages of the same family, the exact value is not close to the reported classical measurement when looking at the entire sample. When analyzed by the Australian region, our calculations are closer to the classical measure in the Melbourne sample. We have organized our paper as follows. In section 2, we review the background of simi-

[1] https://github.com/glottolog/glottolog

larity measures between languages and current similarity measurements between sign languages. In section 3, we describe our datasets. We provide more details of our proposed use of a video-based machine learning model to measure similarity in section 4.1 and the calculation and analysis of the metrics in section 4.2. Later in section 5, we present results and similarity analysis.

2. Background for Sign Language Similarity

As mentioned in (Mathur and Napoli, 2011), many factors have an effect on similarities and dissimilarities across different sign languages. For this reason sign language similarity analysis often provides new information that helps linguists to study sign languages. For example, in spite of USA and UK sharing English as their spoken language, ASL is closer to French Sign Language (usually abbreviated as LSF) than to BSL (Cagle, 2010; Brooks, 2018; Mathur and Napoli, 2011). The factors that influence sign languages, can be geographic or historic ones (Cheek et al., 2002). Recent methods measure sign language similarity from a lexicostatistics perspective (Yu et al., 2018; Börstell et al., 2020). These four features are usually considered to measure similarity of signs: hand shape, location, movement, and palm orientation. Besides these features, it is worth to notice other cases of similarity. For example, signs may or may not encode the same meanings in different sign languages. For example, as reported by (Börstell et al., 2020), the NGT (Sign Language of the Netherlands) sign WAAR-A ('where') is identical to the ASL sign WHAT, while the sign WAT-A ('what') is identical to the ASL sign WHERE. This form overlap may produce cross-linguistic mismatch.

Language similarity is usually measured by the Swadesh method, which started being a list of 225 words (Swadesh, 1952) and ended up being a list of 100 universally used meanings (Swadesh, 1971). Initially, a similar process was followed to measure the similarity between a pair of sign languages. However, (Woodward, 2000) considered Swadesh's method an overestimation of the similarity measure. As mentioned by (Yu et al., 2018), Woodward highlights that the use of pointing for signs, such as pronouns and body parts, can be misleading. Woodward list was developed from swadesh list in (Woodward, 1978) but then modified to a list of 100 words (Woodward, 2000). Other work compares the similarity overlap obtained from a lexical database of 50 signs and the Swadesh list (Minton-Ryan et al., 2019). For instance, (McKee and Kennedy, 2000) reported similarity measures in three categories: identical (match in the four features), related but different (differ only in one feature), and completely different considering swadesh list. They reported 25% and 77% of identical similarity of between BSL-ASL and BSL-Auslan, respectively, and 31% and 87% including related-but-different. Similarity measures between 12% and 36% are considered families of a stock; between 36% and 81% make two sign languages of the same family, while the overlap of larger than 81% makes them dialects of the same sign language. These ranges were proposed in (Crowley and Bowern, 2010). Other previous work uses computational and more systematic approaches to measure similarity and intelligibility between and within sign languages. The work in (Hildebrandt and Corina, 2002) measures the similarity of different signs within the same sign language by asking native and hearing subjects. (Brentari et al., 2020) analyzes properties such as marking agency and number in four sign languages for their cross-linguistic similarities. (Sáfár et al., 2015) evaluates the mutual intelligibility through genre among three sign languages and the benefit of mouthing to measure the effect of the overlap between the spoken languages.

Some automatized methods include a comparison between finger-spelling only (Kishore et al., 2017) and automatic distance measures such as Dynamic Time Warping (DTW) on videos over the four previously mentioned features (Wang et al., 2014). Machine learning models for recognition, segmentation, and translation can contribute to analyzing larger corpora and with more detail. Moreover, we estimate that they would become a powerful tool to support similarity analysis of languages. More standardized and multi-sign language datasets are needed to approach these tasks.

3. Datasets and Preprocessing

In this section we describe our datasets and preprocessing methods. For ASL and Auslan, we found existing dataset. However, for ISL, we downloaded Youtube videos and match them with their transcripts. We use the python library moviepy to segment the videos according to their annotations per sentence. For testing the similarity with BSL (Schembri et al., 2013; Fenlon et al., 2011), we analyzed ASL, Auslan and ISL datasets. As we do not perform any preprocessing step for BSL, we provide details about it in section 4

3.1. ASL

We used How2Sign [2] (Duarte et al., 2021), a large-scale multi-modal and multi-view continuous American Sign Language dataset. It originally had significantly large training, testing, and validation datasets each consisting of video files and ground-truth annotation files. We work with its test set where the video files have multiple sentences in 24 fps.

In the annotation file of large-scale ASL, sentence-wise time boundary was available for each video. We represent the duration distribution of sentences in Figure 1. We sample 100 sentences that last between 1 second and 6 seconds. Along with their respective annotation or English translation, this became our final ASL dataset. We converted the video files into 25fps as 25fps was the required rate of Renz's model. There

[2] https://how2sign.github.io/

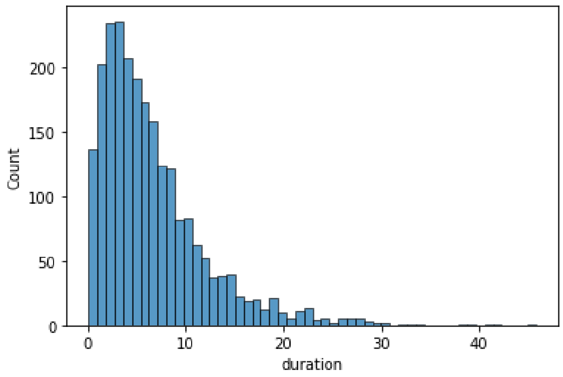

Figure 1: Duration of each sentences in seconds in x axis and Number of sentences in y axis. The majority of the sentences had no more than 10 seconds and there were no more than 100 sentences having length larger than 20 seconds.

are a total of 548 tokens (including repetition of some signs or tokens in different sentences) in the sample we took from ASL testing dataset for our experiment.

3.2. Auslan

We collected videos and their annotations in EAF files from Auslan corpus[3] mentioned in (Johnston and Schembri, 2006). The annotation files provided several tiers such as FreeTransl, LitTransl, Comments-Linzi, CLUwithinCLU etc. However, some of them were for isolated signs and some of them were for sentences. FreeTransl and LitTransl were for sentences and we took the datasets that had translations of LitTransl tier. We refined our collection criteria to be within the area of Melbourne (1 large video file containing 21 sentences), Sydney (1 large video file containing 31 sentences) and Northern Australia (3 large video files containing 66 sentences) from their collection of endangered Australian Sign Language. Similarly to ASL, we extracted a total of 118 sentences in new video files of 25 fps along with their corresponding English sentence translation in a ground-truth annotation file. There were total of 1186 tokens, including repetitions. For example, in these two sentences: “He ran. Then she ran”, 5 signs) tokens are counted.

3.3. ISL

We extracted 50 video files in 30 fps from a Indian Sign Language (ISL) tutorial along with their corresponding ground-truth annotation file (was available in English translation) from (CADREE, 2020). The tutorial was in English whereas the sign representation is in ISL. There were total of 112 tokens, including repetition of some tokens in different sentences. We converted the video files to 25 fps. The process that we followed to measure the classical similarity score of ISL with BSL is addressed in Section 5.

4. ML-based Segmentation Model for Similarity

We estimate the similarity by counting the number of recognized signs of one sign language that the segmentation model found even when pre-trained in another sign language. In other words, we interpret the test accuracy as the overlap between these two sign languages. This section explains the sign-segmentation model and our proposed evaluation metrics for sign language similarity analysis.

4.1. Segmentation Model

The temporal segmentation process of signs is a costly process that needs expert annotators to distinguish the boundaries or start and end of each sign in a semantic unit. Motivated by this, (Renz et al., 2021a) presents a 3D multi-stage temporal convolutional network trained as binary classification to determine if each frame is in a boundary or in a sign segment. To get the sign boundaries, they use a very well known action recognition model, I3D, to get spatio-temporal features. These spatio-temporal features are processed with a multi-stage temporal convolutional network. A classification layer on top of this feature vector generates the sign class probabilities.

Renz *et al.* propose a segmentation model trained on two datasets, BSLCorpus and PHOENIX14 (DGS) German Sign Language and tested in those 2 and BSL-1K. We take this model pretrained with BSLCorpus (Fenlon et al., 2011) learned weights. BSLCorpus is a BSL linguistic corpus that provides various types of manual annotations, and a portion of it carries individual signs with their sign categories and temporal boundaries. The BSLCorpus dataset consists of videos of 4.8 hours. The sign classification procedure followed numerous rules, including allocating lexical variations of the same word to the same class and selecting classes with less than 10 occurrences. Merging the categories for constructing a generalized training dataset focuses on priority for dominant hand. This work provides code and a pretrained model[4] in BSL that we use to test in ASL, Auslan and ISL.

They explain their results with two metrics mF1B and mF1S, which are calculated based on the correct identification of boundary positions and extent of the sign segments, respectively. They defined *boundary* as a series of the frames labeled with value of 1. Consequently frames of a sign segment are labeled with value of 0. To measure correct segment identification they work with two thresholds. One establishes the maximum distance between the middle of the ground-truth boundary and the middle of the predicted boundary. On the other hand, they count as a correct identified sign segment the value of IoU (intersection over union) of the ground-truth and predicted sign segments. Although they reach values of 68.68 and 47.71 in mF1B

[3] http://hdl.handle.net/2196/00-0000-0000-0000-D7CF-8

[4] https://github.com/RenzKa/sign-segmentation

and mF1S, these metrics mainly focus on lengths and positions and not in the recognition of the class or sign. Moreover, they mention that semantic class labels were not fundamental to achieve good segmentation performance. From our understanding, they also used a pre-trained model on sign language recognition. Up to date this paper is written, we were not able to access details on the accuracy of the sign language recognition model. However, we hypothesise they rely on some of the most advanced sign language recognition models looking at their collaborators an their previous work.

4.2. Evaluation Metrics

To measure the similarity of two sign languages, dataset-A and dataset-B, we train or use a pre-trained model on dataset-A. The input for this model is a video of a sequence of signs and the prediction is the written or annotated signs, in a sequence as well. Then, we test this model in dataset-B and compare the prediction of sequence signs to the ground-thruth annotation. In our case, the segmentation model is trained on BSL and we will test the model for ASL, Auslan and ISL to measure the similarity between them and BSL. We represent the similarity between ground-truth annotation (part of our dataset) and the predicted-annotation (model's output) using the 3-metric measurement system with different level of strictness: EXACT_MATCH, MATCH_SYNG, MATCH_SYNGP.

```
00 : 00 : 00, 312 −− > 00 : 00 : 00, 815
let

00 : 00 : 01, 135 −− > 00 : 00 : 01, 492
forget

00 : 00 : 01, 535 − − > 00 : 00 : 01, 998
emotion
```

Figure 2: **The ground truth-annotation files:** These files have signs of a sentence with their corresponding boundaries.

In Figure 2 and 3, we show some examples of how the ground-truth and the prediction annotation files look like. In Figure 2, the single file (of a sentence) contains a total of three signs (after filtering stop words such as " 's", "an", "the") and the line before the sign contains the time boundaries of that single sign. The first and last boundaries of the file represent a single sentence's time boundary. We represent the corresponding prediction-annotation file for that ground-truth annotation file in Figure 3. The "WEBVTT" writing on top of the file represents that these files are in .vtt extensions. We get these files by testing the input video files (containing one sentence each) on the pre-trained sign segmentation model in BSL.

```
WEBVTT

00 : 00 : 00.440 − − > 00 : 00 : 00.920
EMPTY,DISCUSS,DIRTY,TIDY,LET,WANT

00 : 00 : 01.120 − − > 00 : 00 : 01.440
SLIP,LOW,NOT,GROW-UP,SURPRISE,SAY

00 : 00 : 01.525 − − > 00 : 00 : 01.980
LOW,SAY,ENOUGH,CLOTHES,PANIC,ANGRY
```

Figure 3: **The predicted-annotation files:** These files have multiple predicted signs of a sentence with their corresponding boundaries.

We work with continuous sign language and not isolated sign language. We look for any match throughout the sentence boundary instead of the sign boundary. According to EXACT_MATCH, there is one sign ("let") common in both files. For the sign "let" from Figure 2, "EMPTY,DISCUSS,DIRTY,TIDY,LET,WANT" are the predicted signs in Figure 3.

4.2.1. Exact Match

The ground truth annotation files have signs of a single sentence in them. We get the corresponding predicted signs in individual predicted-annotation files. Regardless of lexical ordering differences, the sentence boundaries of each prediction-annotation file should contain the matched sign if there were any matched sign between the ground-truth annotation file and prediction annotation file. For EXACT_MATCH, we first calculate the total signs throughout all the ground truth-annotation files that had any match in its corresponding prediction-annotation file. We then divide this number by the number of signs from the ground truth-annotation file. We finally represent the percentage of the ratio. Equation 1 below represents the formula to calculate metric1.

$$\text{EXACT\_MATCH} = \frac{n}{N} * 100, \tag{1}$$

where n is the number of groundtruth signs that matched with a sign from it's corresponding predicted-annotation sentence, and, N is the total number of groundtruth annotation signs.

4.2.2. Ground Truth Synonyms

MATCH_SYNG is similar to EXACT_MATCH, except we first get a set of synonyms for each word sign of ground truth-annotation files. Then, We look for the sign or any sign synonyms of that sign in the corresponding prediction-annotation files. For example, if there is a word "small" in the ground-truth annotation file, and we get a set of synonyms for that word as "{little, slight, tiny, minor}" and the prediction-annotation file contains "tiny", we calculate it as one match. Finally, we calculate the number of matches, and we divide this number by the total number of signs in the ground truth-annotation files and present its percentage. Equation 2 below represents the formula to calculate MATCH_SYNG. It is the procedure for MATCH_SYNG.

$$\text{MATCH\_SYNG} = \frac{n}{N} * 100, \qquad (2)$$

where n is the number of groundtruth signs or any synonym of that sign that matched with a sign from it's corresponding predicted-annotation sentence, and, N is the total number of groundtruth annotation signs.

4.2.3. Ground Truth and Prediction Synonyms

For MATCH_SYNGP, along with considering the synonyms of ground-truth words, we also consider the synonyms of predicted words for matching. It is similar to MATCH_SYNG except that we also collect a set of synonyms for each word signs of prediction-annotation files and the synonyms of ground-truth annotation files' words. So, we look for the sign or any synonyms of a sign from ground-truth annotation files in the corresponding prediction-annotation files' words or any synonyms of that word. Finally, we calculate the number of matched signs considering the synonyms of both files. We divide this number by the number of original signs in the ground truth-annotation files and present its percentage. Equation 3 below represents the formula to calculate MATCH_SYNGP. It is the procedure for MATCH_SYNGP.

$$\text{MATCH\_SYNGP} = \frac{n}{N} * 100, \qquad (3)$$

where, n is the number of groundtruth signs or any synonyms of a sign from ground-truth annotation files in the corresponding prediction-annotation files' words or any synonyms of that word that matched with a sign from it's corresponding predicted-annotation sentence, and, N is the total number of groundtruth annotation signs.

5. Experiments and Result Analysis

In this section we present the overlap of our datasets and the Woodward' and Swadesh's lists to have a better perspective and interpretation of our results. Although the two lists are traditionally identical (as woodward list was mainly developed from swadesh list), we included results for both the lists. The reason is that we compare our results to classical measurements that use swadesh list such as (McKee and Kennedy, 2000) for BSL-ASL and BSL-Auslan, and our manual calculation for ISL. However, more recent works use woodward list and there is more probability to compare our analysis with others future work. We present the values obtained for our 3-metric system and analyze which one gets closer results to classical similarity measurements. Finally we analyze Australian results by each region.

5.1. Signs from Swadesh list and Woodward list in our dataset

We have described in Section 1 that 100 signs of Swadesh list and 100 signs of Woodward list have the possibility of lexical similarity of any two sign language all over the world. The occurrences of signs from Swadesh list and Woodward list are around $1/10$ times of the total signs in our datasets (represented in Table 1).

SL	Sign entries	w_s	% Overlap	s_s	% Overlap
ASL	548	21	3%	43	7.85%
Auslan	1186	78	6.58%	156	13.15%
ISL	112	11	9.8%	10	8.93%

Table 1: Occurrences of words of sign language dataset in Woodward and Swadesh lists. Here, w_s = Occurrences of words from Woodward list # of times (including repetition in different sentences) and s_s = Occurrences of words from Swadesh list # of times (including repetition in different sentences.

In this table, total sign entries for ASL is 548. As we are working with continuous signs instead of isolated signs, it includes repetition of signs. Also, Renz's model tries to predict each sign in a sentence. So, we calculated total of how many signs we are putting as input to the Renz's model that it is trying to predict (excluding stop words such as 'a', 'the'), instead of total of how many unique words are there in the dataset. For example, from "The person is picking a pen from the other person's hand", total signs are 'person', 'is', 'pick', 'pen', 'from', 'other', 'person', 'hand' and the number is 8 (including the repetition of the word 'person', because the model is trying to that word twice). Sign entries column represent this count for all the datasets. Also, the overlap percent 3% means that from 548 signs altogether, 3% times a Woodward word appeared. The overlap column of Table 1 represents this count for all of our datasets.

5.2. Evaluation of metrics

We process three datasets, ASL, Auslan, ISL datasets for testing them on a pre-trained segmentation model in BSL. For each of these datasets, we have video files with their corresponding ground truth-annotation files and obtain the prediction-annotation files after testing. We provide a repository[5] for reproducible experiments. We analyze the similarities between ground truth-annotation files and predicted-annotations files with respect to EXACT_MATCH, MATCH_SYNG, MATCH_SYNGP. As the EXACT_MATCH does not consider any synonym sign matching, rather matches directly, we address it as stricter metric. On the other hand, MATCH_SYNGP considers synonyms of both ground-truth signs and predicted-signs which increases its possibility of getting a match per pair. Nevertheless,

[5] https://github.com/tonnidas/sign_similarity

	Woordward Similarity (in %)			Swadesh Similarity (in %)			Classic Similarity (in %)
Sign Language	EXACT_ MATCH	MATCH_ SYNG	MATCH_ SYNGP	EXACT_ MATCH	MATCH_ SYNG	MATCH_ SYNGP	
ASL	28.57	**33.33**	47.62	13.95	**23.26**	39.53	25
Auslan	23.72	**50**	57.69	34.62	**46.15**	48.72	77
ISL	0	**9.09**	9.09	0	**0**	0	7

Table 2: All numbers represent the percent for that column in that particular row. The first row after the header is for ASL dataset which matches with classical similarity measurement. There are focused two rows. First one is "Woodward similarity" that represents how many Woodward words occurrences found a match among all the Woodward words occurrences in datasets and the second one is "Swadesh similarity" which represents how many Swadesh occurrences found a match among all the Swadesh occurrences in datasets.

MATCH_SYNG is a semi-strict metric as we consider the synonyms of only ground-truth signs. Our results show that this metrics MATCH_SYNG is the more reasonable and correlated sign language similarity measurement compared to the classical method of similarity score. We compare our results with the scores of identical categorized from (McKee and Kennedy, 2000).

From Table 1, we can see that among the signs in ASL dataset, 3% of the time Woodward appeared and 7.85% of the time Swadesh words appeared. It indicates that great part of the dataset is out of Swadesh and Woodward's lists, and this also holds for both Auslan and ISL. In our Auslan dataset, the times of occurrences of Woodwards and Swadesh words were 78 and 156 respectively and in our ISL dataset, the times of occurrences of Woodwards and Swadesh words were only 11 and 10 respectively. In Table 2, we presented our results for the two datasets with respect to two lists of words: *Woodward similarity*' and *Swadesh similarity*'.

Our results show that, in general, metrics considering '*Swadesh Similarity*' are closer to the '*Classical Similarity*'. For BSL and ASL, we see in Table 2, our similarity metric, MATCH_SYNG, is 23.26% that supports the classical similarity score is 25%, which is close. It is a common assumption that ASL and BSL are similar as both American and British speak English. Nevertheless, ASL and BSL are independent sign languages, fully unique and distinct, and cannot be understood by each other's users.

We could not find a reported similarity score for ISL and BSL considering Swadesh list and Woodward list. Thus, we calculated the classical measurement value for ISL manually considering only the appeared Woodward words and Swadesh words following (McKee and Kennedy, 2000)'s method for the category of identical. This process considers four features: location, handshape, orientation and movement for each single isolated sign. If any all of the four features match with another sign of same meaning, that is considered identical. As our ISL dataset has low number of swadesh and woodward words (10 and 11), the similarity percentage according to the category of identical signs may not represent the similarity score for all 100 swadesh or woodward listed words. Thus, we considered calculating similarity score for all the features(location, handshape, orientation and movement) individually and put a score of 1 for all these features for a individual sign. If a feature is matched in both sign representations from different sign languages that has the same meaning, a score of 1 is calculated. If all four of the features are matched for a sign representation that a score of 4 is achieved. We calculate the percentage of scores by dividing the scores that is achieved with the scores that we would achieve if all of them were identical according to (McKee and Kennedy, 2000) and then we calculate the percentage.

We considered only the isolated words (manually picked) that appeared combined in both standardized Swadesh and Woodward words lists and count 10 and 11, respectively. Our calculations indicate that BSL-ISL has a 7% of classical similarity. We see in Table 2 that the similarity score between BSL and ISL, is around 9% for our MATCH_SYNG metric, which is close to the classical measurement.

Comparing BSL and Auslan, according to MATCH_SYNG, Woodward and Swadesh similarity is 50% and 46.15%, respectively. we can see that Auslan results for '*Woodward Similarity*' do not fully support its classical similarity measurement with value 77%. In spite of this result, some specific dialects of Auslan correlates better with the classical measurement as we will describe in Subsection 5.2.1.

5.2.1. Analysis of dialects in Auslan

The total proposed values of similarity measurement between BSL and Auslan in Swadesh's and Woodward's lists are distant from the classical measurement of 77%. In this section we provide a desegregated analysis on Auslan Sign Language variations and how our proposed metric calculated individually by dialect might reflect a closer relation with the classical measurement. According to (Wikipedia contributors, 2022), the reason behind this is that Sydney and Melbourne dialects of Auslan is more inclined to BSL where Northern Auslan dialect is more prone to be different than BSL. In our dataset, we collected 5 video files for a total of 3 groups dialects; 3 files from Northern (total of 654 signs), 1 file from Sydney (total of 230 signs) and another 1 file from Melbourne (total of 302 signs).

	MATCH_SYNG (in %)		
Sign Language	Swadesh Similarity	Woordward Similarity	Classic Similarity
Auslan (Northern1)	39.29	25	77
Auslan (Northern2)	45	28.57	
Auslan (Northern3)	27.59	23.53	
Total Of this 3 files	36.36	25	
Auslan (Melbourne)	55.17	75	77
Auslan (Sydney)	68	61.11	
Total Of this 2 files	63.29	68.42	

Table 3: Dis-aggregated analysis of Australian dialects

Table 3 represents the similarity rate according to MATCH_SYNG aggregated into two groups, the first one is of northern dialect and the second one is of Sydney and Melbourne (combined). We combined these two in one group as we also mentioned earlier that these two dialects may have roots in older dialectal differences from the United Kingdom. From this table, we can see that the northern group has a similarity of only 36.36% for Swadesh words and only 25% for Woodward words which is far away from the classical measurement of 77%. On the other hand, the combined Sydney and Melbourne group has a similarity of around 64% for Swadesh words and around 70% for Woodward words which is very close the classical measurement of 77%. This different results on regions and dialects of the Auslan dataset explains our results in Table 2. The sign language from northern dialects are not much similar to BSL while the Melbourne and Sydney dialects are similar which is why we can see that the overall combined Auslan results are not close to classical measurement in Table 2.

6. Conclusions and Future Work

This work proposes the similarity measurement of three pairs of sign languages: BSL vs. ASL, BSL vs. Auslan, and BSL vs. ISL. This measurement consists of interpreting the accuracy of a model trained in one sign language and tested in another as the overlap or similarity measurement. Our work emphasises on cross-linguistic matching where forms of the signs and also the assigned English gloss for the signs match. The ground truth-annotations are provided by the signers (according to ASL, Auslan, ISL dataset repositories). The segmentation model identify the temporal boundaries of each signs and then predicts the sign. As the model is pre-trained in BSL, it can only predict a sign from the testing set (ASL, Auslan, ISL) successfully when similar sign is present in BSL. So, our similarity percentage reflects what percent of signs in ASL or Auslan or ISL would a BSL signer recognize based on their lexical forms.

We introduce three accuracy metrics of different strict levels using exact matches and considering synonyms for only the ground truth and for both ground truth and predictions. We found that the intermediate-strictness metric using woodward and swadesh lists are the closer measurements to the classical one for ASL and Auslan; and woodward list for ISL.

This approach could help provide a more systematic way to measure the similarity between two sign languages. Our approach can measure the similarity of any pair of sign languages once we compare our findings with previous manually reported similarities. However, we compare our similarity metrics to previous classical measurements reported. We cannot guarantee that the same calculations were followed in all the sign languages on those classical calculations. On the other hand, we do not report more information about the BSL dataset and its overlap with Woodward's and Swadesh's lists.

Naturally, another suitable model to measure similarity can be a sign language recognition, which directly focuses on the sign. In reality, isolated signs may be influenced by other signs when used inside a sentence, and continuous signs make up the English word related to the sign.

7. Bibliographical References

Bapna, A., Cherry, C. A., Lepikhin, D. D., Foster, G., Krikun, M., Johnson, M., Chen, M., Ari, N., Firat, O., Macherey, W., Wu, Y., Cao, Y., and Chen, Z. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges, July.

Börstell, C., Crasborn, O., and Whynot, L. (2020). Measuring lexical similarity across sign languages in Global signbank. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 21–26, Marseille, France, May. European Language Resources Association (ELRA).

Brentari, D., Horton, L., and Goldin-Meadow, S. (2020). Crosslinguistic similarity and variation in the simultaneous morphology of sign languages. *The Linguistic Review*, 37(4):571–608.

Brooks, R. (2018). A Guide to the Different Types of Sign Language Around the World. https://tinyurl.com/wex97vu6, May. Accessed: 2022-05-20.

Bull, H., Afouras, T., Varol, G., Albanie, S., Momeni, L., and Zisserman, A. (2021). Aligning subtitles in sign language videos. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV), pages 11552–11561, October.

CADREE. (2020). Simple Sentences in Indian Sign Language part 1| Simple Sentences of Everyday Usage in ISL. `https://tinyurl.com/429hzc85`, October.

Cagle, K. M. (2010). *Exploring the Ancestral Roots of American Sign Language: Lexical Borrowing from Cistercian Sign Language and French Sign Language*. ProQuest LLC.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020). 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 (CVPR)*, June.

Cheek, C., Cheek, A., Knapp, H., and Rathmann, C. (2002). *Modality and structure in signed and spoken languages*. Cambridge University Press.

Cormier, K. and Fenlon, J. B. (2014). Bsl corpus annotation guidelines.

Crowley, T. and Bowern, C. (2010). *An Introduction to Historical Linguistics*. Oxford University Press.

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., and Giro-i Nieto, X. (2021). How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744.

Fenlon, J., Stamp, R., Cooke, M., Rentelis, R., Hepner, A., Denmark, C., Parker, J., Wilkins, J., and Nelson, M. (2011). Bsl corpus project.

Hammarström, H., Forkel, R., and Haspelmath, M. (2019). glottolog/glottolog: Glottolog database 4.1, November. Type: dataset.

Hildebrandt, U. and Corina, D. (2002). Phonological similarity in american sign language. *Language and Cognitive Processes*, 17(6):593–612.

Johnston, T. and Schembri, A. (2006). Issues in the creation of a digital archive of a signed language. In *Sustainable data from digital fieldwork: Proceedings of the conference held at the University of Sydney*, pages 7–16.

Kishore, P., Kumar, D. A., Prasad, M., Sastry, A., and Kumar, E. K. (2017). similarity assessment of 30 world sign languages and exploring scope for a sign–to–sign translator. *International journal of control theory and applications*, 10:315–335.

Mathur, G. and Napoli, D. J. (2011). *Deaf around the world: The impact of language*. Oxford University Press.

McKee, D. and Kennedy, G. (2000). Lexical comparison of signs from american, australian, british and new zealand sign languages. *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima*, pages 49–76.

Minton-Ryan, C. A., Sorola, M., Brown, J., and Perez, P. R. (2019). A lexicostatistical study: Phonological similarity between american and malawi sign languages.

Renz, K., Stache, N. C., Albanie, S., and Varol, G. (2021a). Sign language segmentation with temporal convolutional networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139. IEEE.

Renz, K., Stache, N. C., Fox, N., Varol, G., and Albanie, S. (2021b). Sign segmentation with changepoint-modulated pseudo-labelling. *CoRR*, abs/2104.13817.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., and Cormier, K. (2013). Building the British Sign Language Corpus. *undefined*.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: With special reference to north american indians and eskimos. *Proceedings of the American Philosophical Society*, 96(4):452–463.

Swadesh, M. (1971). *The origin and diversification of language*. Edited *post mortem* by Joel Sherzer. Chicago: Aldine.

Sáfár, A., Meurant, L., Haesenne, T., Nauta, E., Weerdt, D. D., and Ormel, E. (2015). Mutual intelligibility among the sign languages of belgium and the netherlands. *Linguistics*, 53(2):353–374.

Varol, G., Momeni, L., Albanie, S., Afouras, T., and Zisserman, A. (2021). Read and attend: Temporal localisation in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16857–16866.

Wang, L.-C., Wang, R., Kong, D.-H., and Yin, B.-C. (2014). Similarity assessment model for chinese sign language videos. *IEEE Transactions on Multimedia*, 16(3):751–761.

Wikipedia contributors. (2022). Auslan — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Auslan&oldid=1081044982`. [Online; accessed 25-April-2022].

Woodward, J. (1978). Historical bases of american sign language. *Understanding language through sign language research*, pages 333–348.

Woodward, J. (2000). Sign Languages and Sign Language Families in Thailand and Viet Nam. In Karen Emmorey et al., editors, *The Signs of Language Revisited*, pages 30–52. Psychology Press.

Yu, S., Geraci, C., and Abner, N. (2018). Sign languages and the online world online dictionaries & lexicostatistics. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.