Sign Languages and the Online World Online Dictionaries & Lexicostatistics

Shi Yu, Carlo Geraci, Natasha Abner

Sign Languages and the Online World
Online Dictionaries & Lexicostatistics

Shi Yu[1,2,3], **Carlo Geraci**[2,3], **Natasha Abner**[4]
[1]École des Hautes Études en Sciences Sociales, [2]Institut Jean Nicod CNRS, [3]École Normal Supérieure,
29, Rue d'Ulm, 75005 Paris, France
[4]Department of Linguistics, University of Michigan
450 Lorch Hall 611 Tappan Street Ann Arbor, MI 48109-1220, USA
shi@shiyu.fr, carlo.geraci76@gmail.com, nabner@gmail.com

Abstract

Several online dictionaries documenting the lexicon of a variety of sign languages (SLs) are now available. These are rich resources for comparative studies, but there are methodological issues that must be addressed regarding how these resources are used for research purposes. We created a web-based tool for annotating the articulatory features of signs (handshape, location, movement and orientation). Videos from online dictionaries may be embedded in the tool, providing a mechanism for large-scale theoretically-informed sign language annotation. Annotations are saved in a spreadsheet format ready for quantitative and qualitative analyses. Here, we provide proof of concept for the utility of this tool in linguistic analysis. We used the SL adaptation of the Swadesh list (Woodward, 2000) and applied lexicostatistic and phylogenetic methods to a sample of 23 SLs coded using the web-based tool; supplementary historic information was gathered from the Ethnologue of World Languages and other online sources. We report results from the comparison of all articulatory features for four Asian SLs (Chinese, Hong Kong, Taiwanese and Japanese SLs) and from the comparison of handshapes on the entire 23 language sample. Handshape analysis of the entire sample clusters all Asian SLs together, separated from the European, American, and Brazilian SLs in the sample, as historically expected. Within the Asian SL cluster, analyses also show, for example, marginal relatedness between Chinese and Hong Kong SLs.

Keywords: Web-based annotation tool, Sign Language, Lexicostatistics, Phylogenesis, Online dictionaries

1. Introduction

Lexicostatistics provides a means of determining the degree of similarity across languages by simply looking at portions of their vocabulary (Swadesh, 1971). Though such studies have been largely limited to spoken languages, promising results have been documented once similar methodologies are applied to sign languages (Woodward, 2000; McKee and Kennedy, 2000). In particular, Woodward (2000) proposes a sign language (SL) adaptation of the Swadesh list. Like the original Swadesh list, Woodward's list contains 100 items that are meant to identify basic/universal concepts which are supposed to reveal the degree to which pairs of SLs are related. However, this method has not been systematically tested or applied to SLs. A key reason for this is a lack of reliable data and the absence of software applications that allow for easy annotation of video data. In recent years, however, many SL dictionaries have appeared on the internet and can be freely consulted, solving the empirical problem of gathering the relevant data. The existing applications (ELAN, Ilex, SignStream, etc.) for annotating video data are stand-alone applications designed primarily to work with files stored on local machines. Moreover, these applications are designed with research flexibility in mind and do not come "pre-equipped" with theoretically-informed annotating codes or coding categories.

We have created a web-based tool that addresses these outstanding issues. The web-based application imports videos of signs from online dictionaries and provides a theoretically-informed annotation schema for the main articulatory properties of these signs. We show here how this tool facilitates theoretically-informed typological and historical analysis of sign languages, using the interface to systematically investigate the degree of similarities across 23 SLs. Thus, our methodology implements Woordward's original idea of comparing pairs of sign languages in such a way as to conduct an effective cross-linguistic comparison of a large sample of SL. The video data used for the present analyses come mainly from the online dictionary of the Spread The Sign Project (Domfors and Fredäng, 2008) and also from LSD Visual Sign Language Dictionary (Hong Kong Sign Language: `http://www.sign-aip.net/sign-aip/en/home/index.php`), Taiwan Sign Language Online Dictionary (Tsay et al. (2008): `http://lngproc.ccu.edu.tw/TSL/indexEN.html`) and NHK Sign Language CG (Japanese Sign Language: `https://www2.nhk.or.jp/signlanguage/index.cgi`). We present a case study of four Asian SLs (Japanese SL: *JSL*, Chinese SL: *CSL*, Taiwanese SL: *TSL*, and Hong Kong SL: *HKSL*) and explore relations within this historically and areally related group. We also apply this analytic approach to the handshapes of the 23 SLs in our database and use a cluster analysis to identify relationships across the sample.

The rest of the paper is organized as follows: Section 2 introduces some basic principles of lexicostatistics applied to SL, Section 3 describes the web application we created for phonological annotation of signs, Section 4 describes the comparative approach applied to the data, Section 5 reports the results of a case study on four Asian SLs, and Section 6 provides the analysis of handshapes of the sample of 23 SLs. Finally, Section 7 concludes the paper.

2. Lexicostatistics & Sign Language

Lexicostatistics is a method used in historical and comparative linguistics to determine the relationship between pairs of languages based on the degree of shared lexicon (Dobson, 1969; Rea, 1990). List(s) of concepts/meanings which are assumed to be universally instantiated in the world's languages (e.g., blood, many, leaf, etc.) are used to compare the lexica (Swadesh, 1971). Large scale comparison may then be made by means of distance matrices and cluster analysis. Although the lists and the methodology have been criticized (Hoijer, 1956; Gudschinsky, 1956), lexicostatistics has proven to be a good method to work with underdescribed and unwritten languages (Crowley and Bowern, 2010; Lehmann, 2013). The percentage of overlapping properties across items from the list determines the linguistic distance between two sign languages. For spoken languages, languages that share more than 81% of signs are treated as dialects of the same language; if the percentage is between 36% and 81%, they are treated as different languages from the same family; while if the percentage is below 36%, the two languages then belong to distinct families (Crowley and Bowern, 2010).

Woodward (1993) adapted the original Swadesh list for the purpose of sign language comparison (Figure 1). In particular, he removed body parts and pronouns because they are often represented in SLs by pointing to the referent; thus, they may lead to an overestimation of the relationship between SLs. In his works Woodward compared pairs of languages like American and French SL (Woodward, 1978) and several South Asian and East Asian sign languages (Woodward, 1993). McKee and Kennedy (2000) used Woodward's list to compare British (BSL), Australian (Auslan) and New Zealand SLs (all closely historically related) to the historically unrelated American SL. For each item in the list, pairs of languages may be evaluated on the similarity of the articulatory properties used in the languages' signs for that item. The articulatory properties themselves may be drawn from the four major phonemic classes of SLs: handshape, location, movement, and palm orientation. Such a comparison produces results like those shown in Table 1 for Auslan and BSL (adapted from McKee and Kennedy (2000)).

Auslan &BSL	Hs	Loc	Mov	Ori	Notes
egg			x		
grass			x		Different weak hand
look for	x				Two handed in BSL

Table 1: Example of lexical comparison

These approaches are based on pairwise comparisons of SLs and they show that the lexicostatistics method can be successfully applied to languages in the visual modality (but see Section 4 for commentary on some of the problems of this method of comparison). However, previous research has not attempted a systematic comparison of a large sample of SLs.

3. An Annotation Tool For Online Dictionaries

In this section we describe the front-end of the web-based tool that we created for annotation videos from online dictionaries. The annotation tool has been created using JavaScript, and a JavaScript plugin (Video.js) was used to display the video files fetched from online SL dictionaries, the video is displayed continuously with repetition. The workspace is accessible by standard web browsers and is divided in three major areas: 1) on the top-left side, the video-streaming for annotation, 2) in the central part, the main annotation area, and 3) one the right, the list of words to be annotated (Figure 2). Languages are chosen by using a dropdown menu on top of the list. The results of annotations for a specific sign are summarized below the video.

The data set is first imported by using the English version of Woodword's list, a script is used to fetch the corresponding words and videos in other languages on the Spread The Sign online dictionary, where the same word is grouped together across languages. Our data set thus include the word in English and the corresponding word in the original language. All words are checked during the annotation whenever ambiguity arises (e.g., two entries for the word "dust", as noun and as verb)

We included 55 handshapes in our annotation tool. These 55 handshapes are supposed to be representative of handshapes used in sign languages and have been proven to be able to capture most handshape configuration in our data set. Several categories of handshape include multiple handshape images that are allophonic variations in SLs. For annotation, this step requires only a click on the correspondant handshape. Also in this section, the hand part feature (i.e., Orientation (Brentari, 1998)) can be selected using the dropdown list, the two-handed option is used to annotate signs with identical articulation of both hands.

The second section contains features of place of articulation, based on Brentari (1998) model, we included neutral space and four major regions. For signs produced in neutral space, the choice is between horizontal, vertical, or lateral. For signs produced on a major body region (*head*, *torso*, *arm*, *hand*), the annotator may use the dropdown list to specify one among eight micro regions each.

- Head: *top*, *forehead*, *eye*, *cheek*, *nose*, *lip*, *mouth*, *chin*, and *below-chin*
- Arm: *upper*, *elbow-front*, *elbow-back*, *forearm-front*, *forearm-back*, *forearm-ulnar*, *wrist-front*, and *wrist-back*
- Hand: *palm*, *finger-fronts*, *back of palm*, *back of fingers*, *radial-side*, *ulnar-side*, *tip*, and *heel*
- Torso: *neck*, *shoulder*, *clavicle*, *torso-top*, *torso-mid*, *torso-bottom*, *waist*, and *hips*

1. all	26. grass	51. other	76. warm
2. animal	27. green	52. person	77. water
3. bad	28. heavy	53. play	78. wet
4. because	29. how	54. rain	79. what
5. bird	30. hunt	55. red	80. when
6. black	31. husband	56. right	81. where
7. blood	32. ice	57. river	82. white
8. child	33. if	58. rope	83. who
9. count	34. kill	59. salt	84. wide
10. day	35. laugh	60. sea	85. wife
11. die	36. leaf	61. sharp	86. wind
12. dirty	37. lie	62. short	87. with
13. dog	38. live	63. sing	88. woman
14. dry	39. long	64. sit	89. wood
15. dull	40. louse	65. smooth	90. worm
16. dust	41. man	66. snake	91. year
17. earth	42. meat	67. snow	92. yellow
18. egg	43. mother	69. star	93. full
19. fat	44. mountain	70. stone	94. moon
20. father	45. name	68. stand	95. brother
21. feather	46. narrow	71. sun	96. cat
22. fire	47. new	72. tail	97. dance
23. fish	48. night	73. thin	98. pig
24. flower	49. not	74. tree	99. sister
25. good	50. old	75. vomit	100. work

Figure 1: Woodward's vocabulary list for sign language comparison.

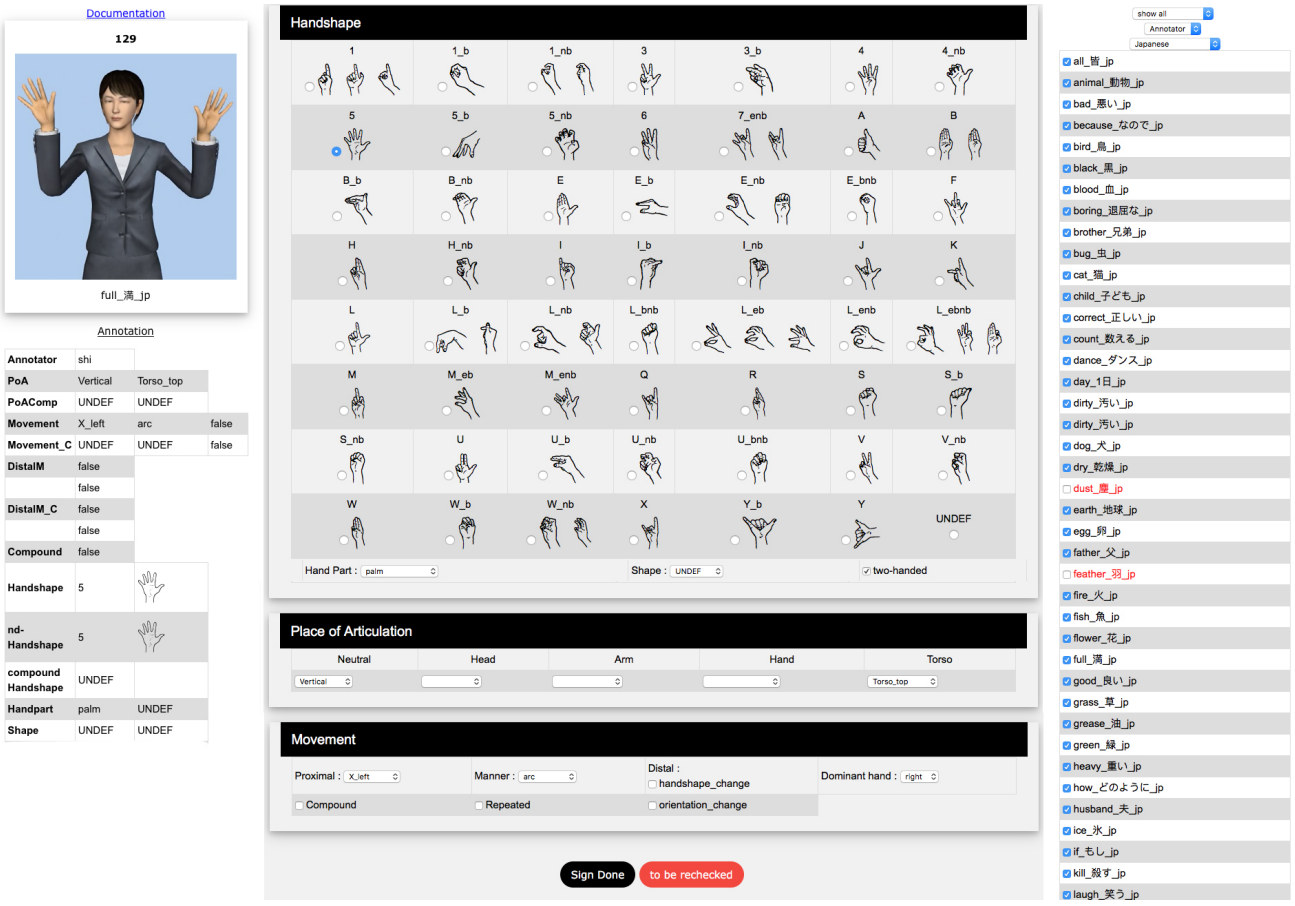

Figure 2: Workspace of the web-based annotation tool.

For movement in the third section, both dropdown list and check button are used for annotation. For proximal movements, we annotate the axis on which the movement is performed and its direction (forward, backward, down, up, left, and right). For distal movements the non-exclusive options are handshape change and orientation change. We also annotated the manner of movement (straight, circular, arch, etc.) and presence/absence of repetitions.

Each of the previous sections are duplicated in the case of compound sign, when the option "Compound" is selected, additional sections will display on the screen for the annotation of the second part of the sign. For all the options of annotation, we reserved an "undefined" option for empty value and also for the review of ambiguous signs for annotation. Annotation results are sent to the server and saved in the JSON format. This file is then transformed to a *.csv* file for the purpose of linguistic analysis.

4. Methodology

Previous studies compared signs by looking at the global similarities of the four main classes of phonemes (Handshape, Location, Movement and Orientation). However, none of them has been explicit on how similarity is measured. In particular, for each class of phonemes it was never specified the set of contrastive features that would determine a significant difference between any two phonemes. For instance, consider the following handshapes:

They all have four selected fingers, some of them also have a selected thumb. Some of them have spread or stacked fingers, some have flexed non-base joints, others have flexed base joints. Under a holistic analysis all these handshapes could be considered similar. However, a feature-based analysis would distinguish the handshapes not just on the number of selected fingers, but also based on thumb selection, whether the selected fingers are spread or not, the base and/or non-base joints are flexed, etc.

Similar considerations extend to the other classes of phonemes. For instance, it is unclear whether the neutral space was treated as a single entity or whether different planes have been distinguished (horizontal, vertical, lateral). Even more problematic is the case of orientation where the definition itself may lead to different interpretations of what counts as similar/identical. Indeed, orientation can be defined either in absolute terms with respect to signer's body or relative to the plane of articulation (Quer et al., 2017).

In our study, we decided to use a theoretically-informed annotation procedure and implement a feature based analysis directly in the annotation tool (see Section 3). Rather then establishing identity/similarity based on the global assessment of pairs of (video) signs, we used Brentari (1998) model to generate the set of features upon which difference is then measured. The signs of each language are independently annotated by selecting the relevant feature values. Pairwise comparison is made post-hoc by counting the number of identically specified features. Pairs of signs sharing all features are considered identical. Pairs of signs where only one feature value is different feature are treated as similar. Pairs of signs that are different for more than one feature are treated as different. This procedure of assessing the articulatory properties of signs is in many respects stricter than those used in previous studies and it has the risk of biasing the data by maximizing differences. It also treats as equally relevant features that generates macroscopic differences (like selected fingers) and features that creates less perceivable differences (like flexed non-base joints). However, these biases can be mitigated by neutralizing some differences (e.g., collapsing [$\pm$ spread] handshapes in one single group, grouping locations by major regions, etc.) or by weighting features. In this study, we decided to consider all features and not to apply any weight correction. However, we show the effect of collapsing some feature values for handshape and place of articulation.

In previous studies, Annotators' subjective perception could affect data evaluation in two steps of the procedure. First when s/he tries to identify the individual phonemes for each sign, and then when s/he has to establish whether pairs of phonemes are identical, similar or different. Our procedure is based on the annotation of the articulatory properties of individual signs. It does not mitigate the subjective evaluation occurring when identifying the correct phoneme, but it removes any subjectivity from the evaluation of similarities between two signs.

5. Comparing Asian Sign Languages

We applied our annotation procedure to investigate potential relations between pairs of languages. Our data set is annotated by one sign language expert to keep the homogeneity and the correctness of the annotation. The kinds of comparisons and analyses reported here are similar to those reported in previous studies (a.o., Woodward (2000) and McKee and Kennedy (2000)). However, the fact that we adopted an extremely rich set of features allows us to perform a more effective comparison of the articulatory properties of the signs.

As a case study, we conducted an analysis on four Asian SLs: JSL, CSL, TSL and HKSL. Unfortunately, very little is known about historical relations among these SLs. We cross-checked information available to us such as the Ethnologue of World's Languages and Wikipedia and we found that JSL is related to TSL (and Korean SL), while CSL (variety of Shanghai) is related to although not mutually intelligible with HKSL.

In the following tables we provide the results. On the first column we indicate the pairs of languages; on the second column we report the percentage of signs that are identical in the two languages; on the third column we report signs that are similar (i.e. only one feature/phoneme is different); while the last column reports the percentage of signs that are different (i.e. two or more features/phonemes are different).

Table 2 reports the results of the comparison made with the full set of phonological features; Table 3 reports the results after handshapes with the same selected fingers but different joint flection have been collapsed into one

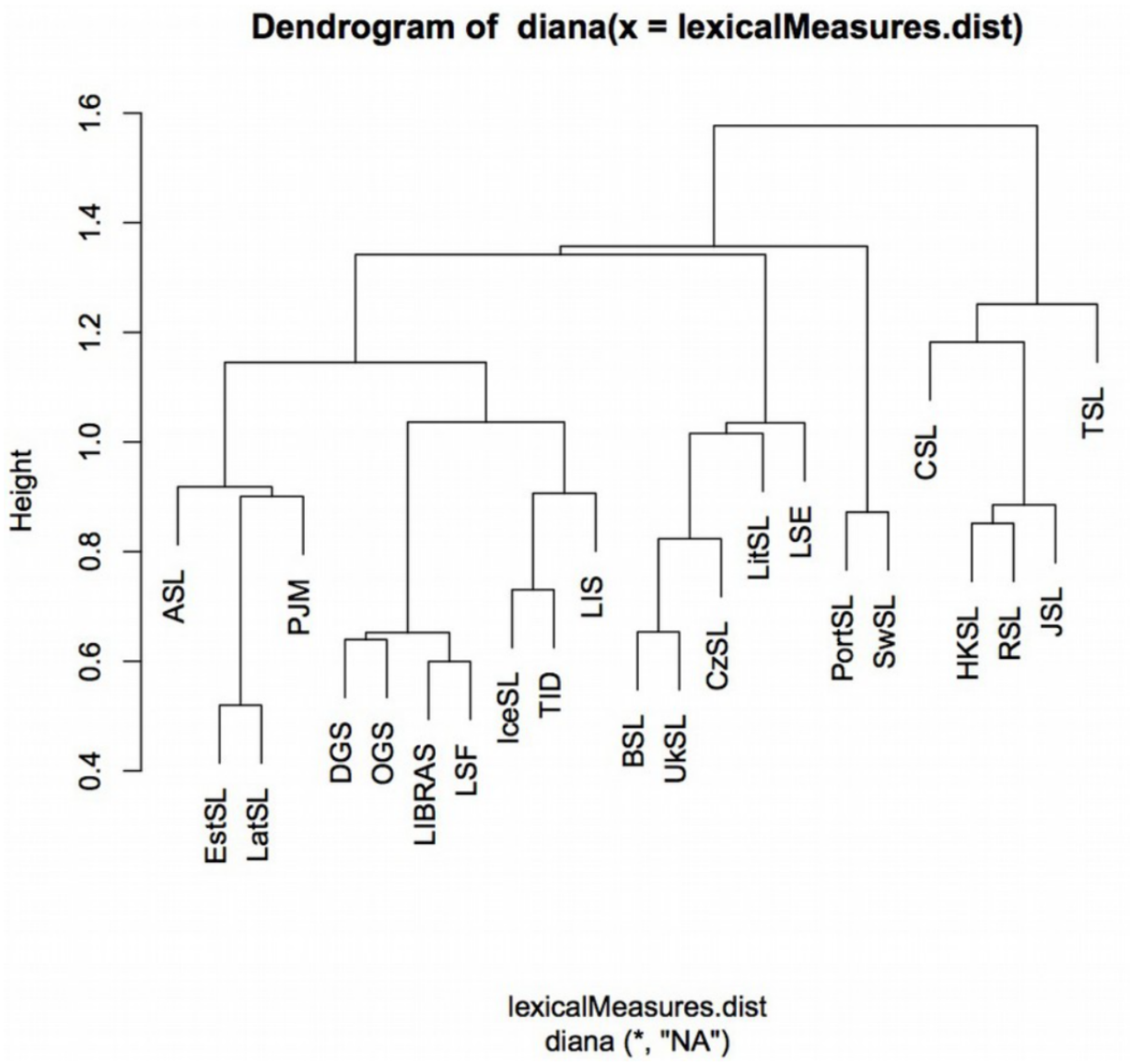

Figure 3: Divisive hierarchical cluster analysis of 23 SLs based on handshapes. ASL: *American SL*, BSL: *British SL*, CSL: *Chinese SL*, CzSL: *Czech SL*, EstSL: *Estonian SL*, LSF: *French SL*, DGS: *German SL*, OGS: *Austrian SL*, HKSL: *HongKong SL*, IceSL: *Icelandic SL*, LIS: *Italian SL*, JSL: *Japanese SL*, LatSL: *Latvian SL*, LitSL: *Lithuanian SL*, PJM: *Polish SL*, PortSL: *Portuguese SL*, LIBRAS: *Brazilian SL*, RSL: *Russian SL*, LSE: *Spanish SL*, SwSL: *Swedish SL*, ASL: *Taiwan SL*, TID: *Turkish SL*, UkSL: *Ukrainian SL*.

level; Table 4 reports results after place of articulation has been collapsed into five major regions (neutral space, head, torso, arm, hand), while table 5 reports results after handshape and place of articulation have been collapsed.

Languages	Identical	Similar	Different
JSL&CLS	0.00%	9.28%	90.72%
JSL&TSL	4.30%	13.98%	81.72%
JSL&HKSL	3.09%	9.28%	87.62%
CSL&TSL	4.26%	13.83%	81.91%
CSL&HKSL	9%	16%	74%
TSL&HKSL	3.19%	12.77%	84.04%

Table 2: Comparison made with the full set of features

Languages	Identical	Similar	Different
JSL&CLS	1.03%	11.34%	87.63%
JSL&TSL	5.37%	15.05%	79.57%
JSL&HKSL	4.12%	11.34%	84.84%
CSL&TSL	5.32%	17.02%	77.66%
CSL&HKSL	10.10%	22.22%	67.68%
TSL&HKSL	4.26%	14.89%	80.85%

Table 3: After collapsing handshapes with the same selected fingers

What emerges by looking at the percentages of the different tables is that a comparison based on pure articulatory features does not let emerge any cross-linguistic similarity. However, when the effect of some features is neutralized,

Languages	Identical	Similar	Different
JSL&CLS	2.06%	10.31%	87.63%
JSL&TSL	5.38%	19.35%	75.27%
JSL&HKSL	4.12%	12.37%	83.51%
CSL&TSL	4.26%	18.09%	77.66%
CSL&HKSL	14.14%	23.23%	62.63%
TSL&HKSL	6.38%	12.77%	80.85%

Table 4: After collapsing locations in major areas

Languages	Identical	Similar	Different
JSL&CLS	3.09%	13.40%	83.51%
JSL&TSL	6.45%	23.66%	69.89%
JSL&HKSL	5.15%	15.46%	79.38%
CSL&TSL	5.32%	21.28%	73.40%
CSL&HKSL	15.15%	24.24%	60.60%
TSL&HKSL	7.45%	14.89%	77.66%

Table 5: After collapsing locations and handshape

some similarities emerge. In particular, Table 5 shows that CSL and HKSL share around 40% of the signs in the Woodward list and should be treated as two different languages of the same family. JSL and TSL share almost 30%. While this is not enough to consider them as languages of the same family, it somehow makes justice of the fact that the two are not mutually intelligible.
Traditional lexicostatistics methodologies leave open the question whether at a higher level detailed analysis these languages belong to the same linguistic group or not. We address this question in the next section.

6. On Handshape Features

In this section we report the analysis of handshape similarities conducted on the 23 SLs available in our annotated data set (see Figure 3)
We conducted a divisive cluster analysis (Baayen, 2008). What emerges somewhat clearly is that the Asian languages (plus RSL) are clustered together, while all European languages plus American and Brazilian SL are split in secondary clusters. Based only on handshape, we can readily distinguish two large sign language families.

7. Conclusions

Documentation of individual SL history is quite fragmented and often unreliable, especially when it comes to describe contact with other SLs. In this paper we used lexicostatistics and phylogenetic methods to investigate the degree of similarity across 23 SLs. This has been made possible thanks to the use of online resources and a new web-base annotation tool that we created specifically for this purpose. Results showed that lexicostatistics methods are reliable as long as the degree of analysis remains at a superficial level. The variability and degree of freedom introduced by more fine-grained annotations of the articulatory properties of signs make methods based on holistic assessment of similarity less reliable. However, once more sophisticated analysis are used, cross-linguistic similarities emerge even once looking at a relatively large sample of languages.

8. Acknowledgements

This research has been funded by Fyssen Fundation Research Grant (PI: Carlo Geraci).

9. Bibliographical References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Brentari, D. (1998). *A prosodic model of sign language phonology*. MIT Press.

Crowley, T. and Bowern, C. (2010). *An introduction to historical linguistics*. Oxford University Press.

Dobson, A. J. (1969). Lexicostatistical grouping. *Anthropological Linguistics*, pages 216–221.

Domfors, L.-Å. and Fredäng, P. (2008). Spread the sign. `http://www.spreadthesign.com/`.

Gudschinsky, S. C. (1956). The ABC's of lexicostatistics (glottochronology). *Word*, 12(2):175–210.

Hoijer, H. (1956). Lexicostatistics: A critique. *Language*, 32(1):49–60.

Lehmann, W. P. (2013). *Historical linguistics: An introduction*. Routledge.

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.

Quer, J., Cecchetto, C., Donati, C., Kelepir, M., Pfau, R., and Steinbach, M. (2017). *The SignGram Blueprint: A Guide to the Preparation of Comprehensive Reference Grammars for Sign Languages*. De Gruyter Mouton, Berlin/Boston.

Rea, J. A. (1990). Lexicostatistics. *Research Guide on Language Change, Trends in Linguistics Studies and Monographs*, 48:217–222.

Swadesh, M. (1971). *The origin and diversification of language*. Transaction Publishers.

Tsay, J., Tai, J. H.-Y., Lee, H.-H., Chen, Y., and Liu, C.-H. (2008). Taiwan sign language online dictionary. 3rd edition. Institute of Linguistics, National Chung Cheng University, Taiwan. `http://lngproc.ccu.edu.tw/TSL/indexEN.html`.

Woodward, J. (1978). Historical bases of American Sign Language. In Patricia Siple, editor, *Understanding Language through Sign Language Research*. Academic Press.

Woodward, J. (1993). Lexical evidencefor the existence of South Asian and East Asian sign language families. *Journal of Asian Pacific Communication*, 4:91–106.

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: An anthology to honor Ursula Bellugi and Edward Klima*. Lawrence Erlbaum Associates.