Improving phonological distance measures for signs: the CatFormCompare tool

Hope E. Morgan[1], **Amy Isard**[2], **Anh Dang**[3]
[1]Institute of German Sign Language and Communication of the Deaf
[2]Hub of Computing and Data Science
University of Hamburg, Germany
[3]Utrecht University
{hope.morgan, amy.isard}@uni-hamburg.de, t.t.a.dang@uu.nl

Abstract

This paper describes the CatFormCompare tool, designed to enable the comparison of phonological content between pairs of signs, especially in larger datasets. With this tool and a schema for coding categorical form (the SL CatForm coding schema), a pipeline is created that allows a feedback mechanism for advancing research—specifically by directly addressing one of the hard problems in sign language phonology: how to extract true minimal pairs from datasets coded for categorical form? Solving this problem would simultaneously improve phonological distance measurements for sign languages because it would mean that the units for measuring distance are grounded in the linguistic structure of the language and not simply a by-product of the coding system. Here we report on the tool and the first evaluation of its functioning.

Keywords: phonological distance, minimal pairs, string edit distance

1. Introduction

A robust research program for sign language phonology would reflect close alignments between these domains: signer perception of phonological differences, the encoding of categorical form units in a notation or orthographic system, and theoretical models that represent this internal structure. One indication of the current status of the field is that there is no system for representing units of form in sign languages that is calibrated well enough to automatically find minimal pairs—that is, two signs in the same language with different lexical meanings that differ by only one form value while the rest of the form values are identical.

Minimal pairs are important objects of study for many reasons. Such pairs reflect linguistic categories in each language and yield inventories of contrastive forms that can be compared cross-linguistically. Minimal pairs also reveal the cognitive organisation of words in the mind and reflect usage (Wedel et al., 2013; Sun and Poeppel, 2023), as well as provide evidence that sign languages have combinatorial structure, a prerequisite for *duality of patterning*, a design feature of human language (de Boer et al., 2012).

Currently, notation or orthographic systems for phonology are not clearly aligned with mental categories of sign form—phonemes—in the same way that International Phonetic Alphabet (IPA) characters can be used to represent phonemes spoken languages. In an effort to move the field closer to this alignment and reveal where gaps lie, we introduce the CatFormCompare tool as part of a pipeline to compare the phonological content of sign pairs. This pipeline has two components: (i) a string of characters representing categorical forms in a sign, the SL CatForm string, and (ii) the CatFormCompare tool, a simple interface housing a tool written with the Python programming language.[1] The tool and the coding schema, along with a small sample of annotated Kenyan Sign Language (KSL) data, are available from the Hamburg University Research Data Repository.[2]

This paper focuses specifically on the CatFormCompare tool and how it evaluates the content of two strings to find minimal pairs. As a test of the pipeline, we use a dataset of 1,880 signs in KSL coded in SL CatForm coding schema and a list of 449 minimal pairs that were found in this dataset. Only minimal pairs that came from the same signer were included in the list to make sure they were from a single grammar. These minimal pairs function as a ‘ground truth’ for evaluating the output of the tool and thus the conditionalities described in this paper. Our goal is to locate pairs that differ by one phonological variable and from this derive a measure of phonological distance. In summary, we find that the pipeline functions, but still requires adjustments and improvements at different points.

2. Background: ‘Edit distance’ in different language modalities

In spoken languages, in order to quantify the phonological differences between words, a string edit distance such as Levenshtein distance can be used

[1] https://docs.python.org
[2] https://doi.org/10.25592/uhhfdm.18534

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 341–350
16 May 2026.

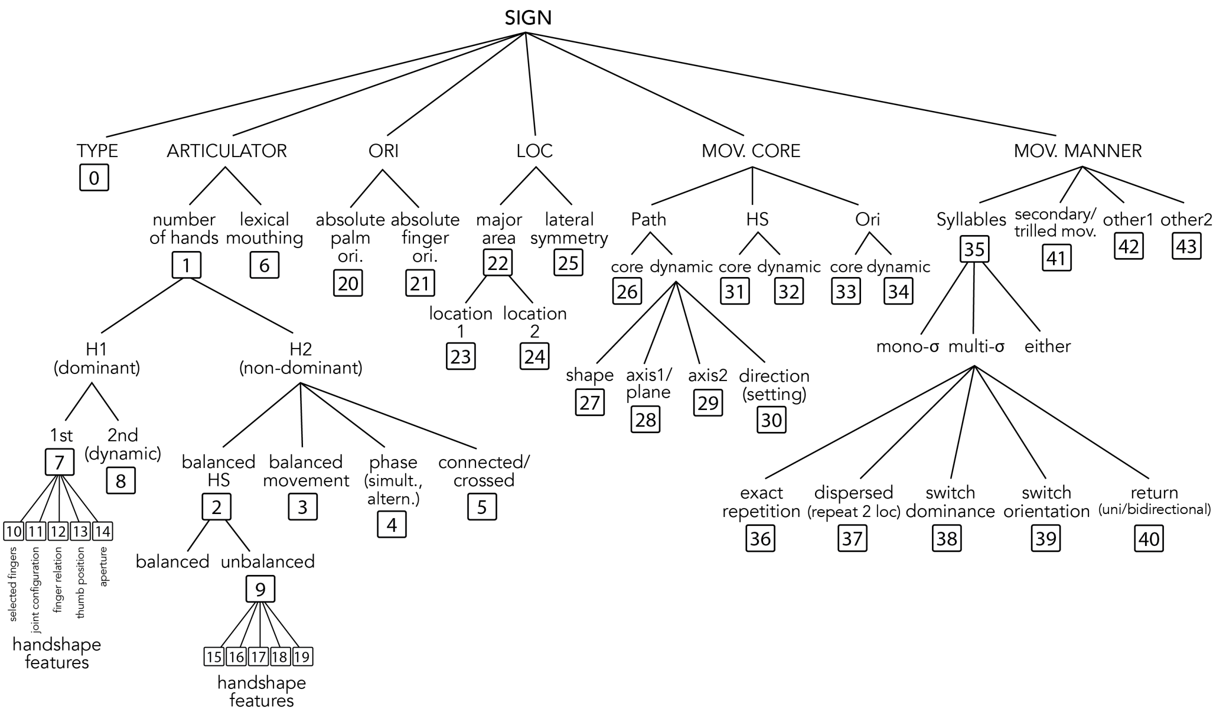


Figure 1: Tree diagram of SL CatForm variables

to calculate the number of differences between all pairs of words in a lexicon, typically done on a lexicon transcribed in the IPA. All changes or 'edits' of characters between pairs of words are tallied segment-by-segment, left to right; i.e., each deletion, insertion, and substitution of segments from one word to the next. The position of segments relative to each other in a sequence is intrinsic to the edit distance operation in spoken languages because it reflects the sequentiality of sounds in spoken words.

In contrast, the phonological content of signs is organised differently. First, there are more types of primary feature classes, and these classes are substantially different from each other.[3] Second, each feature class most often has only one specified value per sign; e.g., a value for *selected fingers*, for *path shape*, for *repetition*, etc. Thus, nearly all minimal contrasts in sign languages appear to be made by the substitution of one feature for another. Minimal pairs on the basis of sequential (re-)ordering are rare and theoretically debated (Corina, 1993; Channon, 2002; Morgan, 2022).[4]

This level of simultaneity and the nature of the different feature classes in signs has both helped and arguably hindered sign language research. It enables researchers to get measurable results using only a sub-set of a sign's phonological content, such as evaluating only handshape and location while leaving out aspects of movement, orientation, etc. This would never work for spoken languages; imagine trying to claim meaningful results about the phonology of a spoken language using only the first syllable of all words or all phonemes except high vowels and voiced consonants.

Many previous studies have used quantitative approaches to compare phonological content of signs, including cross-linguistic comparisons (McKee and Kennedy, 2000; Rozelle, 2003; Parks, 2011; Yu et al., 2018), research on sign neighbourhoods within one language (Caselli et al., 2021), and vector-based comparisons that also integrate semantic vector space (Börstell et al., 2020; Martinez del Rio et al., 2022).

Yet, in order to locate minimal pairs from a dataset of signs, all aspects of form must be accounted for because all values except one in the pair must be the same. If there is too little phonological information included for each sign, the search function will over-generate pairs; it will find those that actually have more than one contrastive difference and incorrectly report them as only one. If the wrong types of information is encoded, then the search function will under-generate; it will miss true minimal pairs.

How then to create a search function that compares only the relevant information for each feature type between two pairs of signs? As mentioned,

[3] For instance, spoken languages have two main classes, consonants and vowels, while sign languages have several, including handshape, location, mouth actions, etc. The phonetic properties of each class are so different that features in one class are almost never shared with another.

[4] Contrasts can also be found that depend on the presence or absence of a feature (Morgan, 2022, 252–254).

GLOSS	VARIABLES																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
EXPENSIVE-4	M	1	0	0	0	0	0	23	00	00	B	H	N	O	N	0	0	0	0	0	C
EXPERIENCE	M	2	S	U	0	C	0	02	00	02	E	S	N	R	0	E	S	N	R	0	C
EXPLAIN	M	2	D	U	0	U	0	34	27	01	A	U	U	O	C	A	S	U	U	0	U
EXPLICIT	M	2	S	B	S	U	0	34	27	88	A	U	U	O	C	A	U	U	O	C	C
EXPLOIT-1	M	2	D	U	0	U	0	27	05	01	A	B	S	O	N	A	S	U	U	0	T
EXPOSE-1	M	2	S	B	S	U	0	27	11	88	A	B	S	O	N	A	B	S	O	N	T

GLOSS	VARIABLES																				
	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
EXPENSIVE-4	U	H	20	00	I	0	0	0	0	0	0	0	0	0	1	0	0	0	0	U	0
EXPERIENCE	C	2	14	22	I	P	T	V	0	D	0	0	0	0	M	E	0	0	0	U	0
EXPLAIN	A	A	23	00	N	P	T	M	0	A	H	O	0	0	M	E	0	0	0	U	1
EXPLICIT	C	N	01	00	S	0	0	0	0	0	H	O	0	0	M	E	0	0	0	U	1
EXPLOIT-1	C	2	08	05	S	P	T	M	0	A	H	H	0	0	E	E	0	0	0	U	1
EXPOSE-1	D	T	21	00	S	P	T	M	0	T	H	C	0	0	M	0	0	0	0	U	1

Figure 2: Six example KSL signs from the SL CatForm coding schema (index number not shown)

our proposal is to use a pipeline with two components, a coding schema and tool with customised scripting; each of these are described in turn in Sections 3 and 4.

3. SL CatForm Coding Schema and SL CatForm String

The input for the CatFormCompare tool is a string of comma-separated values encoded in the SL CatForm schema described in Morgan (2026). SL CatForm contains 42 **variables** (0–41) containing discrete phonological information; these are shown in a tree diagram in Figure 1.[5] Thus, each variable is like a field, with a finite list of **values** for each field. This schema assigns one or two regular ASCII characters to each value in order to facilitate efficient computational processing. The characters (values) are ordered in a string starting with an index number, ID-gloss, and then 42 alphanumeric values corresponding to the phonology of each sign. Sample data for six signs are shown in Figure 2.

The theoretical foundation of this schema is the Dependency Model (van der Kooij, 2002) with modifications inspired by the Prosodic Model (Brentari, 1989); see Morgan (2022) for a detailed description using KSL data. Since a version of the schema has been used for phonological coding of two unrelated sign languages, KSL and Israeli Sign Language (Morgan et al., 2022), and because an analysis of KSL in Morgan (2022) makes explicit comparisons to other sign languages, we posit that the main categories (variables) will be largely representative of phonological form cross-linguistically. However, it is expected that specific values within these categories (e.g., specific handshapes, locations, path shapes) will vary between sign languages. Cross-linguistic applicability of the pipeline is addressed further in Section 7.

In order to compare only the relevant information between pairs, we developed the CatFormCompare tool. Before presenting the tool in Section 4, we explain why such a tool is necessary.

3.1. What is needed in an edit distance for sign languages?

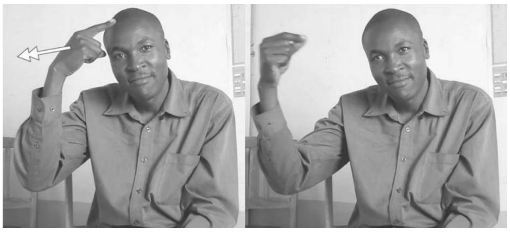
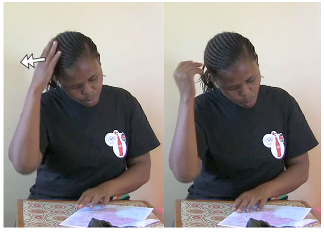
One way to calculate phonological distance from field-based coding is to apply a match/no-match operation to the value in each field for each pair of signs, resulting in a total number of similarities (matches) and difference (mismatches) for that pair. For example, the values in the signs EXPLAIN and EXPLICIT in Figure 2 match on variables 0, 1, 5, etc. but do not match on 2, 3, 4, etc. This has been the basic approach in some cross-linguistic lexicostatistical research, such as Yu et al. (2018). Further, a zero value in either pair can result in that field being excluded from the comparison, which reduces the denominator of overlapping variables between the pair; e.g., see Börstell et al. (2020).

When a good quality coding schema is used that captures all the contrastive phonological information in a sign, the results of a simple match/no-match approach will be informative as quantitative data, but will still fail to find true minimal pairs. This is due to dependencies between phonological types, phonetic constraints, and other generalisations about the articulation of signs.

3.2. Conditionalities

The dependencies between phonological values are operationalised in this project as **conditionalities** in the CatFormCompare tool. Here we provide an example of how a simple match/no-match approach is insufficient for finding true minimal pairs. It comes from the difference between handshape contours and contrasts (Battison, 1978; Mandel, 1981; Sandler, 1989). A well-known generalisation about handshape movements is that they are either produced by the flexion of one set of selected fingers, such as the index finger and thumb hinging closed in TO-MUSE in Table 1—i.e., a *handshape contour*, or it can be the result of a change from one hand configuration with one set of selected fingers to another configuration with another set of

[5] Forty-four variables are shown in Figure 1, but 42 and 43 are not currently specified for a feature type because they are reserved for new categories to be added, as addressed in Section 7

	TO-MUSE	SOMALIA
Variable 7	**59**	**27**
(starting handshape on h1)	(open-G)	(spray)
Variable 8	**30**	**11**
(ending handshape on h1)	(closed-G)	(flat-O)
Variable 32	**C**	**C**
(type of handshape change)	(closing)	(closing)

Table 1: Minimal Pair A: handshape contour

	IF	OKAY
Variable 7	**34**	**34**
(starting handshape on h1)	(O)	(O)
Variable 8	**10**	**19**
(ending handshape on h1)	(F)	(K)
Variable 32	**X**	**X**
(type of handshape change)	(HS contrast)	(HS contrast)

Table 2: Minimal Pair B: handshape contrast

selected fingers, as in OKAY in Table 2. Here, the hand configuration starts with all fingers selected in the O handshape and changes to the index and middle fingers plus thumb selected in the K handshape—i.e., a *handshape contrast*.

The two minimal pairs in Tables 1 and 2 pose a problem for the match/no-match approach. Because the signs in minimal pair B, IF and OKAY, contrast only by the second handshape in variable 8, a comparison of these two variables should find variable 8 as the minimal difference, with all other features being the same.[6] However, the minimal difference for the signs in minimal pair A, TO-MUSE and SOMALIA, are the starting handshapes in variable 7 because the ending handshapes are predictable based on the type of handshape movement in variable 32 (*closing*). In this case, variable 8 should be ignored so that only one difference is counted. Yet, a simple match/no-match calculation would find two differences in minimal pair A, not one.

Another use of conditional processing is to account for the fact that some characteristics of signs are phonetically important but are not themselves contrastive. That is, they play a role in phonology and should appear in a coding schema, but are not the basis for phonological contrast. An example is *major area*, which is important because it accounts for the distribution of two-handed signs, especially when laterality features are involved. For instance, two-handed signs in a contralateral location on the head are disallowed in many sign languages, but *head* is itself is not a unit of contrast.

In order to capture these kinds of contingent properties of signs, conditionalities are written into the CatFormCompare tool. In the case of minimal pairs A and B, the conditionalities for variable 8 (the ending handshape) are as follows: (i) if the handshapes in variable 7 are different then do not compare the values in variable 8 at all; (ii) if one value is null in variable 8 then do not compare them; (iii) if both have the same value do not compare them; (iv) if both are different then *do* count this as a difference. The full list of current conditionalities are listed in

[6]The sign OKAY can be produced with either one or two hands (value 'E' in variable 1), so the number of hands is not counted as a difference.

Table 3.[7] Any variables not listed in this table are evaluated using the simple match/no-match operation.

4. CatFormCompare tool

The core purpose of the CatFormCompare tool is to calculate the phonological distance between all pairs of signs in a dataset, applying conditionalities in order to make only the relevant comparisons. For each pair, it also removes any features from consideration that are not relevant, such as the ending handshapes in Table 1 (i.e., in variable 8). Therefore, it also functions as a way to correct one effect of using field-based coding in databases: the denominator of possible similarities is not all 42 variables, but only those that are relevant in each pair comparison.

First, the tool loads the data from a csv file containing the phonologically coded signs in SL CatForm format. It then performs comparisons between each pair of signs based on the conditionalities described in Section 3.2 and Table 3.

In order to make the CatFormCompare tool easier to use, a lightweight interface was created using the Python tkinter library,[8] which allows the user to select the input file and perform the calculation of matching pairs, as shown in Figure 3. Some options are available in the interface: (i) the number of differences between pairs to search for (1 for minimal pairs); and (ii) whether the standard conditionalities should be used for the comparison. If the standard conditionalities are not used, the user can constrain the value of a particular variable or select to ignore certain pre-defined variables, as shown in the lower half of Figure 3.

The main output of the CatFormCompare tool is a comma-separated file containing all the matching sign pairs; sample shown in Table 4. A summary file report is also provided that lists the conditionalities that were chosen, the number of differences selected for comparison, and a count of how many pairs were found with each of the selected differences, as shown in Figure 4. The interface also optionally allows the input file to be checked for coding errors (based on a csv file of the coding schema variables and values). If any value is found that does not conform to the specifications—i.e., SL CatForm coding values—a file is generated with a list of warnings about values that do not match the master coding schema. Each warning specifies: the variable in question, allowed values specified in the coding schema, actual value in the input file, and ID of the gloss(es) with the non-matching value. Two example warnings can be seen in Figure 5.

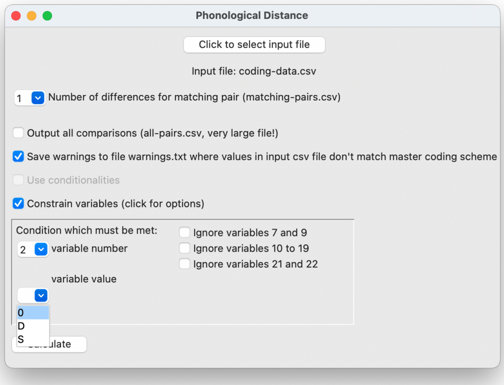


Figure 3: Screenshot of CatFormCompare tool

```
Comparisons with conditionalities
Ignoring variables []
Searching for pairs with 1 differences
Total pairs: 1075
Parameters which differ: [7] - number of pairs 512
Parameters which differ: [23] - number of pairs 189
Parameters which differ: [28] - number of pairs 74
Parameters which differ: [20] - number of pairs 72
Parameters which differ: [35] - number of pairs 62
Parameters which differ: [21] - number of pairs 36
Parameters which differ: [1] - number of pairs 31
Parameters which differ: [37] - number of pairs 16
Parameters which differ: [30] - number of pairs 13
Parameters which differ: [26] - number of pairs 10
Parameters which differ: [40] - number of pairs 9
Parameters which differ: [41] - number of pairs 8
```

Figure 4: Summary file report from the CatForm-Compare tool

```
WARNING: variable 4 "HANDS MOV. SIMULT./ALTERNAT.
                     (PHASE)"
 allowed values are ['0', 'A', 'S']
  unknown value "_" in 1 glosses ['403 CLASS-1']
  unknown value "O" in 1 glosses ['1672 PAKISTAN']

WARNING: variable 33 " + ORIENTATION MOVEMENT"
 allowed values are ['0', 'R']
  unknown value "C" in 1 glosses ['896 FROM-3']
```

Figure 5: CatFormCompare tool warnings about unknown values in variable codings

5. Preliminary Evaluation

The phonological analysis of KSL (Morgan, 2022) was largely based on a dataset of 1,880 non-compound signs coded for phonological characteristics in a FileMaker Pro database. There were 449 minimal pairs found among this set.[9] These

[7] Details of all of the possible values for each variable can be found in Morgan (2026)

[8] https://docs.python.org/3/library/tkinter.html#module-tkinter

[9] Morgan (2022) reports 461 minimal pairs, but 11 contain a sign that is not among the set of 1,880, so those pairs are not relevant for the evaluation here.

No.	Name	Description of conditionality operation
1	NUMBER OF HANDS	If a sign can be either 1-handed or 2-handed ('E' for either), then 'E' is counted as the same value as '1' or '2'
2	SAME SHAPE ON EACH HAND	Do not compare; this variable is only used to apply conditionality statements
3	BALANCED MOVEMENT	Do not compare; this variable is only used to apply conditionality statements
4	SIMULTANEOUS / ALTERNATING	Only compare values in this variable if the movement on the hands is balanced ('B') in Variable 3
5	HANDS CONNECTED, CROSSED	Only compare values in this variable if both signs are two-handed, either one- or two-handed, or the non-dominant hand is a base or grips the dominant hand; i.e., variable 1 = '2', 'E', or 'B'
8	H1 2ND SHAPE (DYNAMIC)	Only compare values in this variable if the handshapes in variable 7 are different (and not '0') and if both values in variable 8 are not '0', and if both values are different from each other
9	WHOLE HANDSHAPE (H2)	Only compare values in this variable if neither value is '00'; do not compare values in this variable if both signs in variable 2 have the value 'S'; i.e., if h2 is a copy of h1 in both signs in the pair
22	MAJOR AREA	Do not compare; this variable is only used to apply conditionality statements
24	2ND LOCATION	Only compare values in this variable if one sign in the pair has a value that is not '00'; if one sign in a pair doesn't have a second location, but the other sign does, then count this as a difference
25	LATERAL SYMMETRY	Only compare values in this variable if the major area is on the torso, neck, head, or in neutral space. Also, do not compare pairs for lateral symmetry when they have a location of cheek or ear (which are ipsilateral if one-handed by default) and those signs differ by number of hands
27	PATH SHAPE	Only compare values in this variable if both signs have path movement (variable 26 = 'P')
28	PATH AXIS 1, or PLANE	Only compare values in this variable if both signs have path movement (variable 26 = 'P')
29	PATH AXIS 2	Only compare values in this variable if both signs have a value that isn't null
30	PATH DIRECTIONALITY	Only compare values in this variable if the path axis is the same in both signs (both have same value in variable 28). Don't penalize a sign if it moves in both directions on an axis rather than only one direction: if either (or both) signs in the pair for variable 30 = 'B', then count these as the same
34	TYPE OF ORIENTATION CHANGE	Only compare values in this variable if both signs have orientation change ('R' in variable 33)
35	NUMBER OF SYLLABLES	If either sign for variable 35 = 'E' (is either mono- or multi-syllabic) , count this as the same as '1' or 'M' in a pair, but different from 'H' (hold); all other values are counted as different from each other
36	REPEATED, EXACT	Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic (variable 35 = 'M' or 'E')
37	DISPERSED (REPEAT 2 LOCATIONS)	Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic (variable 35 = 'M' or 'E'); if pairs differ in this variable, then ignore variable 36
38	SWITCH DOMINANCE	Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic (variable 35 = 'M' or 'E'); if pairs differ in this variable, then ignore variable 36
39	SWITCH ORIENTATION	Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic (variable 35 = 'M' or 'E'); if pairs differ in this variable, then ignore variable 36
40	RETURN (UNI/BI-DIRECTIONAL)	Only compare values in this variable if both signs are multi-syllabic or can be either mono- or multi-syllabic; and don't compare values if one or both is not relevant ('N') or a null value ('0')
41	SECONDARY / TRILLED MOVEMENT	Only compare values in this variable if neither one of the values = '0' (to prevent counting the information in variables 30 and 32 twice)

Table 3: Conditionalities in SL CatForm (first column is the variable number)

pairs were discovered in various ways: by chance, during the process of coding, and later by intentionally looking for minimal pairs with custom searches through the database once it was fully coded. This list serves as a 'ground truth' target for testing the entire pipeline, from phonological coding to transformation into SL CatForm to the mechanics of the CatFormCompare tool.

The dataset of 1,880 KSL signs was first transformed into SL CatForm characters by the first author, inside the KSL Lexical Database. It was then exported as a csv file and run through the CatFormCompare tool, which compares the strings in over 3.5 million pairs to run the calculation. Examples of correctly matched minimal pairs in the dataset are listed in Table 4 and shown in Figures 6 to 9; these include minimal pairs for absolute palm orientation (GOSSIP VS. MOTHER-2; Figure 6); for whole handshape on the dominant hand (MBITA VS. MOTHER-2 in Figure 7); for specific location (OIL-1 VS. SCIENCE in Figure 8); and for return (PAINT VS. UPPER in Figure 9).

The output in Table 4 shows how **phonological distance** can be computed from this data in different ways. Using a categorical measurement by variable, each pair differs by 1 (COUNTS: DIFF) meaning they have a phonological distance of 1. Yet, they also have different features in common (COUNTS: SAM); e.g. GOSSIP and MOTHER-2 have 17 variables that share the same value, while PAINT and UPPER have 19 variables with the same value.[10] By applying the conditionalities, we get a total number of relevant features for each pair, and this can function as another way to measure phonological distance, either by degree of difference or similarity. For example, GOSSIP and MOTHER-2 have a difference score of 0.056 (1/18) and a similarity score of 0.944 (17/18), while PAINT and UPPER have a differ-

[10] The VARIABLES column lists which variables contain the different or same values for each pair.

			COUNTS		VARIABLES	
id	gloss1	gloss2	DIFF	SAM	DIFF	SAM
1071019	GOSSIP	MOTHER-2	1	17	20	1 6 7 21 23 26 27 28 30 31 33 35 36 37 38 39 40
1384822	MBITA	MOTHER-2	1	17	7	1 6 20 21 23 26 27 28 30 31 33 35 36 37 38 39 40
1509121	OIL-1	SCIENCE-1	1	18	23	1 6 7 20 21 25 26 27 28 30 31 33 35 36 37 38 39 40
1542671	PAINT	UPPER	1	19	40	1 6 7 20 21 23 25 26 27 28 30 31 33 35 36 37 38 39 41

Table 4: Examples of four minimal pairs from the CatFormCompare tool output

Parameter	Ground Truth Pairs	Matched Pairs	% pairs found	Missed pairs	% pairs missed
Handedness	6	3	50%	3	50%
Handshape	149	50	34%	99	66%
Location	186	20	11%	166	89%
Movement	79	11	14%	68	86%
Orientation	16	9	56%	7	44%
Non-manuals	5	1	20%	4	80%
Other	5	2	40%	3	60%
Unsure	3	0	0%	3	100%
All	449	96	21%	353	79%

Table 5: CatFormCompare tool matches compared to 'ground truth' pairs, grouped by phonological parameter

ence score of 0.050 (1/20) and a similarity score of 0.950 (19/20). It remains to be seen which of these calculations are most useful, and in what types of research applications.

The results of the minimal pair search using our pipeline finds that only 96 of 449 minimal pairs, or 21%, were listed as differing by 1 variable in the output. Table 5 shows the matches and mismatches by phonological parameter. The worst performing parameter was location. On first impression, this seems like a rather poor result. However, each of the successful matches is possible only if all the variables in both pairs are exactly the same except one value in the string. This is also the first study of its kind so there is no baseline for what might be expected. Subsequent (and on-going) investigation shows that each step—from phonological coding to transformation into SL CatForm to the mechanics of the CatFormCompare tool—all reveal actionable places for improvement. However, due to the breadth of issues and the level of detail, the full results are beyond the scope of this paper. Therefore, we report here only on data from one domain: the 353 missed minimal pairs that were expected but not found.

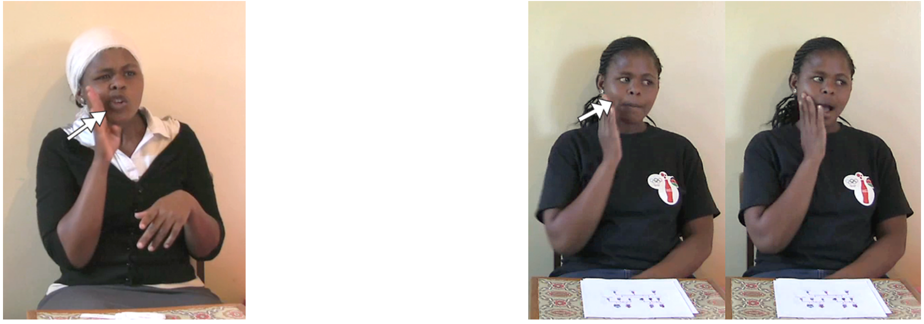
6. Results

A detailed analysis of a quarter of these missed minimal pairs, 87 pairs, was performed by the first author, showing that in all cases except one, the expected value of contrast in the pair was selected as a difference.[11] This means the pipeline did find the expected minimal value; yet, for these signs, it finds too many differences, leading to the question: how far off is it? That is, how many differences beyond one were in the output? The results show that around a third of the missed pairs, 33%, differed by only two instead of one difference, 23% had three differences, 21% had 4 differences, 16% had five differences, and 4% had six differences. Thus, around half of the pairs might be matched correctly as having the single expected difference if only one or two variables or values were corrected. Yet, a few pairs show a very high number of mismatched variables/values, so more work is needed.

When the specific phonological features responsible for the incorrect differences were tracked, two culprits were overwhelmingly most responsible. *Absolute palm orientation* (variable 20) and *absolute finger orientation* (variable 21) were incorrectly counted as differences in half of 87 pairs (43 and 44 pairs, respectively). This was followed by *path axis 1* (variable 28) in 26 pairs, *handshape* (variable 7) in 19 pairs, *number of syllables* in 13 pairs, and *handedness* (variable 1) and *lateral symmetry* (variable 25) each had 10 pairs. Fifteen other variables were problematic in seven or fewer pairs.

In the case of the two orientation variables, the problem appears to lie within the nature of the phonological parameter itself and how it is represented rather than the CatFormCompare tool. For analytical reasons, discussed in Morgan (2022, 408–413), absolute orientation was used in SL CatForm instead of relative orientation. However, the current output of the pipeline shows this is not working for some reason. While this requires more research to solve, it already demonstrates the useful feedback function of having the pipeline.

[11] The one that was missed pointed to a modification needed in the CatFormCompare tool, since implemented.



(a) GOSSIP with palm orientation contra

(b) MOTHER-2 with palm orientation towards body

Figure 6: Minimal pairs that differ in palm orientation (variable 20)

(a) MBITA (town in Kenya) with claw handshape

(b) MOTHER-2 with flat handshape

Figure 7: Minimal pairs that differ in whole handshape on dominant hand (variable 7)

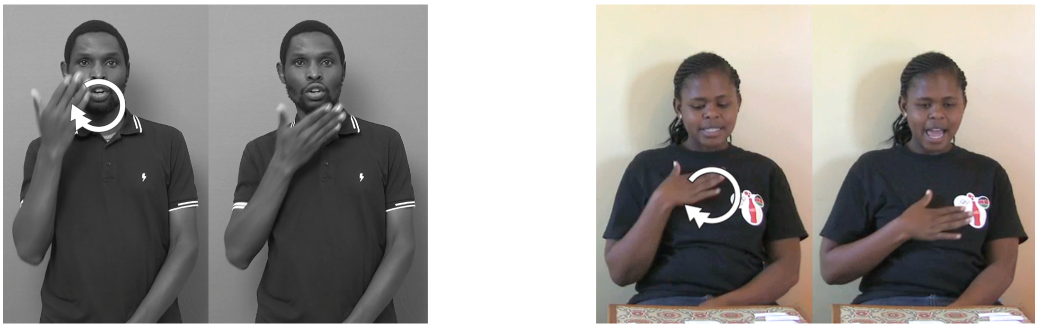
7. Discussion and Next Steps

The CatFormCompare tool and pipeline continue to be assessed through evaluations with the main test set of minimal pairs, as well as a second set of diagnostic sign pairs that has gradually accumulated while testing the pipeline. Tracking down mismatches in the output at a fine-grained level is painstaking work, but when the tool is correctly tuned and functioning, as shown by the successful matches, the benefits and applications should be substantial. With a measure of phonological distance grounded in linguistic data, further research advancements are possible, such as exploring phonological neighbourhoods, studying perception of phonological similarity with more precision, and better controlling stimuli in experimental settings.

Cross-linguistic analyses are also possible, and this brings up an obvious question about the SL CatForm coding schema and the CatFormCompare tool: how can it be used with other sign languages? First, it is important to state that SL CatForm is one of many ways of representing units of form, and its coding categories and values largely overlap with many other notation systems. We are not advocating here for the use of a new coding system. Our aim is more fundamental. If this pipeline works, it will show what needs to be encoded in *any* notation system in order to find minimal pairs—and/or it will highlight places where sign language linguists must do more descriptive and theoretical work.

That said, the coding schema and tool are modifiable based on one's own language data. Values can be added to the SL CatForm variables and the CatFormCompare tool can be be updated to reflect these new values, as well as new variables, if necessary. Two variables, 42 and 43, are kept open for this purpose in case new feature types should be added. In fact, recent evidence suggests that *path size* is contrastive in KSL signs, and this would require its own variable.

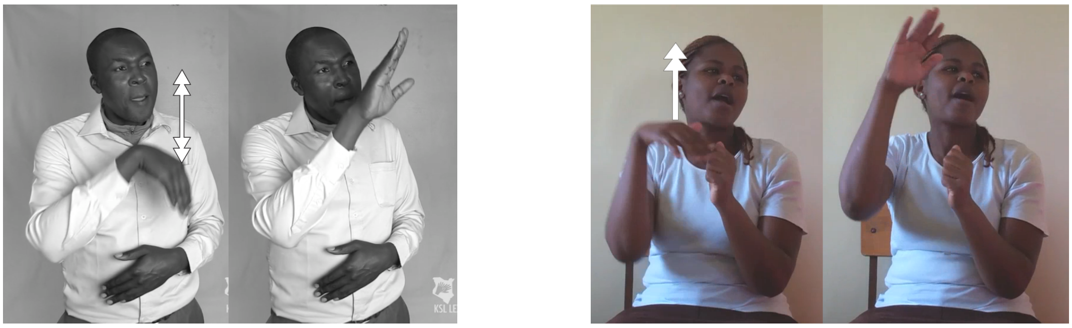
There are three basic approaches to using this pipeline with different phonological coding. First, all of the information required to replicate the process is provided in references cited in this paper, with the exception of detailed coding guidelines for the SL CatForm schema. In lieu of such documentation at present, one is directed to Morgan (2022) to understand the nature of the variables and values in the schema. A second approach is to transform existing coded data into SL CatForm. For instance, Loy and Morgan (2026) demonstrate how one well-known coding system, Hamburg Notation System (Hanke, 2004), can be transformed into SL CatForm. Lastly, one could download the CatFormCompare tool (see footnote 2) and modify it to fit one's own dataset.



(a) OIL-1
at mouth

(b) SCIENCE
at upper torso

Figure 8: Minimal pairs that differ by location (variable 23)



(a) PAINT with bidirectional movement

(b) UPPER with unidirectional movement

Figure 9: Minimal pairs that differ by return (uni-/bi-directional movement) (variable 40)

Looking forward, the next steps are toward making improvements in the pipeline. First, we will continue to diagnose the minimal pair mismatches to discover exactly where the process breaks down, such as in the high number of minimal pairs for location being missed. This includes checking both the expected minimal pairs that were not found, described above, as well as signs that differed by one difference in the output, but were not already in the minimal pair dataset. Second, to catch problems with the tool which have not yet been identified and make sure it does not break with new updates, the diagnostic set of test signs will be expanded from the initial set. Third, the coding of the orientation parameter will be reconsidered and new variables tested in the pipeline.

8. Acknowledgements

Funded by the European Union (ERC, *SemaSign*, No.101117395). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Early work on this tool was done at Radboud University, supported by the following grant project: *Deaf Communication without a Shared Language*; VICI project funded by NWO.

9. Bibliographical References

Robbin Battison. 1978. *Lexical Borrowing in American Sign Language*. Linstok Press, Silver Spring, MD.

Carl Börstell, Onno Crasborn, and Lori Whynot. 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*, pages 21–26, Marseille, France. European Language Resources Association (ELRA).

Diane Brentari. 1989. *A Prosodic Model of Sign Language Phonology*. MIT Press, Cambridge, MA.

Naomi K. Caselli, Karen Emmorey, and Ariel M. Cohen-Goldberg. 2021. The signed mental lexicon: Effects of phonological neighborhood density, iconicity, and childhood language experience. *Journal of Memory and Language*, 121:104282.

Rachel Elizabeth Channon. 2002. *Phonological Representations and Temporal Sequencing in ASL and Other Sign Languages*. Ph.D. thesis, University of Maryland, College Park.

David P. Corina. 1993. To branch or not to branch: Underspecification in ASL handshape contours. In Geoffrey R. Coulter, editor, *Phonetics and Phonology: Current Issues in ASL Phonology*, pages 63–95. Academic Press.

Bart de Boer, Wendy Sandler, and Simon Kirby. 2012. New perspectives on duality of patterning: Introduction to the special issue. *Language and Cognition*, 4:251–259.

Thomas Hanke. 2004. HamNoSys – representing sign language data in language resources and language processing contexts. In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 1–6, Lisbon, Portugal. European Language Resources Association (ELRA).

Lisa Loy and Hope E. Morgan. 2026. Hns2cf: A mapping tool from hamnosys to sl catform. In *Proceedings of the LREC-COLING 2026 12th Workshop on the Representation and Processing of Sign Languages: Language in Motion*, Palma, Mallorca (Spain). ELRA Language Resources Association (ELRA).

Mark Alan Mandel. 1981. *Phonotactics and Morphophonology in American Sign Language*. Ph.D. thesis, University of California, Berkeley.

Aurora Martinez del Rio, Casey Ferrara, Sanghee J. Kim, Emre Hakgüder, and Diane Brentari. 2022. Identifying the correlations between the semantics and the phonology of American Sign Language and British Sign Language: A vector space approach. *Frontiers in Psychology*, 13:806471.

David McKee and Graeme Kennedy. 2000. Lexical comparison of signs from american, australian, british and new zealand sign languages. In Karen Emmorey and Harlan L. Lane, editors, *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima*. Psychology Press.

Hope E. Morgan. 2022. *A Phonological Grammar of Kenyan Sign Language*. De Gruyter Mouton, Berlin, Boston.

Hope E. Morgan. 2026. SL CatForm coding schema v.1: A field-based phonological coding for sign languages. Technical report, IDGS, University of Hamburg, Hamburg, Germany.

Hope E. Morgan, Wendy Sandler, Rose Stamp, and Rama Novogrodsky. 2022. ISL-LEX v.1: An online lexical resource of Israeli Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 148–153, Marseille, France. European Language Resources Association (ELRA).

Jason Parks. 2011. Sign language word list comparisons: Toward a replicable coding and scoring methodology. Master's thesis, University of North Dakota.

Lorna Grace Rozelle. 2003. *The Structure of Sign Language Lexicons: Inventory and Distribution of Handshape and Location*. Ph.D. thesis, University of Washington.

Wendy Sandler. 1989. *Phonological Representation of the Sign: Linearity and Nonlinearity in American Sign Language*. De Gruyter Mouton, Berlin, New York.

Yue Sun and David Poeppel. 2023. Syllables and their beginnings have a special role in the mental lexicon. *Proceedings of the National Academy of Sciences*, 120(36):e2215710120.

Els van der Kooij. 2002. *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity*. LOT Publications, Leiden, Netherlands.

Andrew Wedel, Abby Kaplan, and Scott Jackson. 2013. High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2):179–186.

Shi Yu, Carlo Geraci, and Natasha Abner. 2018. Sign languages and the online world online dictionaries & lexicostatistics. In *11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).