Improving phonological distance measures for signs: The CatFormCompare tool

Hope Morgan
hope.morgan@uni-hamburg.de
Amy Isard
amy.isard@uni-hamburg.de
Anh Dang

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SemaSign

Background

- Many studies have compared phonology of signs from datasets of signs coded for units of form[1,2,3,4]
- Due to the simultaneity of form, it is possible to get *proportional* results from these comparisons
- Yet currently no coding/notation system for representing units of form in sign languages (SLs) has been calibrated well enough to **automatically find minimal pairs**
- The gap is probably not only methodological, but also in linguistic documentation, analysis, & theory

Motivating question

How to advance research on sign language phonology by identifying contrastive units of form in large datasets?

Proposed solution

1. Create a pipeline: phonological coding + tool to find signs with one unit of difference
2. Test pipeline against existing set of "ground truth" minimal pairs
3. Evaluate output for errors
4. Address problems:
   - modify tool
   - modify coding
   - further linguistic analysis & theorising

FOCUS OF THIS STUDY: CatFormCompare tool in the pipeline

PIPELINE

Coding schema → CatFormCompare tool

SL CatForm v.1

The SL CatForm Coding Schema[5] contains 42 variables for all parameters of sign form: articulator, handshape, location, orientation, core movement, and manner of articulation. The variables apply across sign languages, but the values in each variable are currently 'tuned' to the inventory of Kenyan Sign Language (KSL). Examples of 6 KSL signs in string form are shown in Figure 1.

	VARIABLES																																						
GLOSS	ARTICULATOR						HANDSHAPE													ORIENTATION		LOCATION				CORE MOVEMENT								MANNER OF MOVEMENT							
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
EXPENSIVE-4	M	1	0	0	0	0	0	23	0	0	B	H	N	O	N	0	0	0	0	0	C	U	H	20	0	I	0	0	0	0	0	0	0	0	0	1	0	0	0	0	U	0
EXPERIENCE	M	2	S	U	0	C	0	2	0	2	E	S	N	R	0	E	S	N	R	0	C	C	2	14	22	I	P	T	V	0	D	0	0	0	0	M	E	0	0	0	U	0
EXPLAIN	M	2	D	U	0	U	0	34	27	1	A	U	U	O	C	A	S	U	U	0	U	A	A	23	0	N	P	T	M	0	A	H	O	0	0	M	E	0	0	0	U	1
EXPLICIT	M	2	S	B	S	U	0	34	27	88	A	U	U	O	C	A	U	U	O	C	C	C	N	1	0	S	0	0	0	0	0	H	O	0	0	M	E	0	0	0	U	1
EXPLOIT-1	M	2	D	U	0	U	0	27	5	1	A	B	S	O	N	A	S	U	U	0	T	C	2	8	5	S	P	T	M	0	A	H	H	0	0	E	E	0	0	0	U	1
EXPOSE-1	M	2	S	B	S	U	0	27	11	88	A	B	S	O	N	A	B	S	O	N	T	D	T	21	0	S	P	T	M	0	T	H	C	0	0	M	0	0	0	0	U	1

Figure 1. *Six examples of KSL signs (rows) from the SL CatForm coding schema with 42 variables (columns)*

Conditionalities in CatFormCompare tool

- A simple match/no match comparison between strings is not sufficient to find minimal pairs
- Why? Dependencies between phonological types, phonetic constraints, other generalisations about articulation of signs; for example:
  - *handshape contour vs. contrast*
  - *lateral symmetry features depend on location*
  - *free variation in # of hands and # repetitions*
- 22 custom conditionalities included in the CatFormCompare tool

CatFormCompare tool

The **CatFormCompare tool** inputs strings of signs and compares the values in each variable in the string using **conditionalities** to match only phonologically relevant information. Options for comparing pairs of signs are provided in the interface (Figure 2), including the target number of differences between pairs, and constraints on variables. E.g., handshapes can be compared as only features (ignore variables 7 and 9) or only as whole handshapes (ignore 10 to 19). Sample output is shown in Table 2.

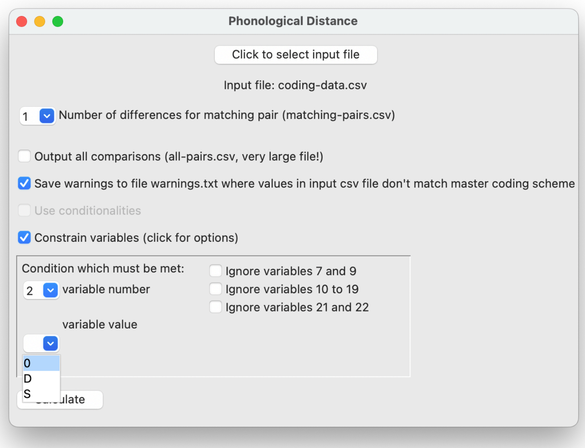


Figure 2. *CatFormCompare tool interface showing minimal pair search (number of differences = 1); in this case, the tool is set to only compare signs in which the value of variable 2 (handshape on both hands same/different) is '0' (one-handed or not a manual sign)*

EVALUATION

Question: *How well does the pipeline work to find the 449 minimal pairs?*

Methodology

- 1,880 non-compound signs in Kenyan Sign Language (KSL)[6] coded for phonemic form in a FileMaker Pro database
- Separate list of 449 minimal pairs from this set of 1,880 signs; these were collected one-by-one during a phonological analysis of KSL[6]
- Signs are transformed into SL CatForm coding values in the database, then output as strings in a CSV file
- CSV file run through CatFormCompare tool to find those pairs differing by only 1 variable
- From this output, identify which were in the list of 449 pairs

Running the CSV file of 1,880 strings of KSL signs through the CatFormCompare tool yielded 931 pairs of signs that differed by only one unit of form. However, in this output, only **96 pairs** matched the minimal pairs collected 'by hand'. Table 1 shows the distribution of matched minimal pairs by phonological parameter.

An analysis of the 353 missed pairs shows 33% differ by 2 instead of 1 difference, and 23% have 3 differences. Thus, with small modifications, it should be possible to capture more minimal pairs.

Table 1. *CatFormCompare tool matches compared to 'ground truth' pairs, by parameter*

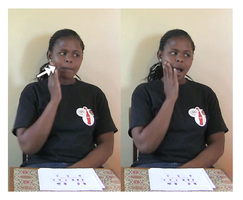
Parameter	Ground Truth Pairs	Matched Pairs	% Pairs Found	Missed Pairs	% Pairs Missed
Handedness	6	3	50 %	3	50 %
Handshape	149	50	34 %	99	66 %
Location	186	20	11 %	166	89 %
Movement	79	11	14 %	68	86 %
Orientation	16	9	56 %	7	44 %
Non-manuals	5	1	20 %	4	80 %
Other	5	2	40 %	3	60 %
Unsure	3	0	0 %	3	100 %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

	SIMILARITY SCORE	DIFFERENCE SCORE
CATEGORICAL	1 of 19	18 of 19
PROPORTIONAL	0.053	0.947

PHONOLOGICAL DISTANCE

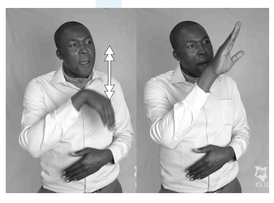
An important outcome of a pipeline that identifies true minimal pairs is that it can form the basis for an empirically-grounded way of measuring **phonological distance** between signs. The CatFormCompare output yields different ways of measuring distance, through either *similarity* or *difference scores*, and either *categorical* or *proportional* measures; see OIL-1 vs. SCIENCE-1.

GOSSIP with palm orientation contra



MOTHER-2 palm orientation towards body

MBITA (town in Kenya) with claw handshape



PAINT with bidirectional movement

UPPER with unidirectional movement

OIL-1 at mouth

SCIENCE at upper torso

Table 2. *Examples of four minimal pairs from the CatFormCompare tool output, showing the variables that are the same and different between each pair*

SIGN PAIRS			COUNTS		VARIABLES	
id	gloss1	gloss2	DIFF	SAM	DIFF	SAME
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

Discussion

- This appears to be the first time a dataset coded for phonological form has been evaluated for whether it correctly finds minimal pairs.
- Closer inspection of the mismatches is ongoing. It looks promising that the pipeline will eventually get close to locating a high number of minimal pairs, but there are challenges ahead, such as theoretical issue of what is a minimal unit especially for some aspects of form, like orientation, and making sure the conditionalities in the CatFormCompare tool work correctly.
- These results should also prompt questions about the nature of phonological coding in all sign language datasets: how 'phonemic' is it?

LREC 20 Palma 26

Poster presented at the *12th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC-COLING 2026),* Palma de Mallorca, Spain. 16 May 2026.

Find the paper here

References

[1] Jason Parks. 2011. Sign language word list comparisons: Toward a replicable coding and scoring methodology. Master's thesis, University of North Dakota. **[2]** 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). **[3]** 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. **[4]** 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). **[5]** Hope E. Morgan. 2026. *SL CatForm coding schema v.1: A field-based phonological coding for sign languages* (Version 1.0). **[6]** Hope E. Morgan. 2022. *A Phonological Grammar of Kenyan Sign Language*. De Gruyter Mouton, Berlin, Boston.

European Research Council
Established by the European Commission

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.