A Theory-Driven Model of Handshape Similarity

July 22, 2015

Abstract

The Articulatory Model of Handshape (Keane 2014b) makes predictions about the phonetic and phonological similarity of handshapes. These predictions are in line with previous work, but are derived from a theory-driven approach instead of a data-driven one. We propose two methods for calculating phonetic similarity: a *countour difference* method that assesses the amount of change between handshapes within a fingerspelled word, and a *similarity* method that compares similarity between pairs of letters in the same position across two fingerspelled words. Both methods are validated with psycholinguistic evidence based on similarity ratings by deaf signers. The results indicate that the *similarity* method more reliably predicted native signer intuition judgements about handshape similarity. This new similarity metric fills a gap (the lack of a theory-driven similarity metric) in the literature that has been empty since effectively the beginning of sign language linguistics.

1 Introduction

Phonetic and phonological similarity has been a topic of exploration for linguists for quite some time (including the seminal Miller & Nicely (1955) study as well as many subsequent studies on spoken languages). Although it has been well explored for spoken languages, signed languages have seen much less research. This work is a further contribution in an effort to change that. We propose a novel method of quantifying similarity between handshapes that is theoretically driven. We then test this method against signers' subjective similarity ratings and find that our method is significantly correlated with signers' intuitions about form similarity. Although much of the previous work has not looked at fingerspelling, we use

fingerspelling to isolate handshape similarity from other aspects of American Sign Language (ASL). The reason for this will be discussed further in section 2.

At the beginning of systematic research into signed languages, there were a number of attempts to quantify handshape similarity within signs (Locke 1970; Weyer 1973; Lane *et al.* 1976; Stungis 1981; Richards & Hanson 1985). Most of these studies relied on signer judgements of similarity or confusion between two stimuli. The researchers then produced clusters of handshapes based on this data. In this way, these researchers are using psycholinguistic data to produce a linguistic model of similarity, rather than using psycholinguistic data to confirm the validity of a linguistic model. The lack of a theory-driven similarity metric making it impossible to use psycholinguistic data to test linguistic models was mentioned explicitly by Lane *et al.* (1976) as a necessity because there simply were not appropriate linguistic models to test: "The present study, then, undertakes to see what sort of featural analysis for ASL results when, using certain specific statistical techniques, we proceed from psychological data to a linguistic model, rather than the reverse".

All of the studies mentioned above came to the conclusion that there are (at least) two distinct categories of handshapes: open handshapes with the fingers of the hand extended, and closed handshapes with the fingers of the hand flexed. Individually, each study developed more finely grained distinctions. For example, Lane *et al.* (1976) found clusters of handshapes that they then used to separate handshapes into groups defined by distinct features. Moreover, Stungis (1981) proposed that this clustering could be turned into a continuous feature space. He found that handshapes could be decomposed along two dimensions: extension (open or closed) and uniform breadth (simplistically this is whether or not all of the fingers have the same configuration).

There has been much more work on phonological models of signed languages (Mandel 1981; Liddell & Johnson 1989; Sandler 1989; van der Hulst 1995; Brentari 1998; Eccarius 2002; Sandler & Lillo-Martin 2006). More recently, there has been work on the phonetics of sign languages (Tyrone *et al.* 2010; Johnson & Liddell 2011a; Johnson & Liddell 2011b; Liddell & Johnson 2011a; Liddell & Johnson 2011b; Whitworth 2011; Mauk & Tyrone 2012; Keane 2014b). Of these, Tyrone *et al.* (2010), Mauk & Tyrone (2012) (for location, and contact), and Keane (2014b) (for handshape) adopted the framework of Articulatory Phonology which explicitly links phonological representations of signs with articulatory gestures that produce those signs, which are phonetic in nature.

The models that have been proposed are exactly the kinds of models that Lane *et al.* (1976) observed were missing at the time of their studies on handshape sim-

ilarity. Most of these models divide the hand into subcomponents, each of which can take categorical values (via binary features, dependency models, etc.). For example, the Prosodic Model (Brentari 1998) represents handshapes using a branching feature system. It consists of specifications indicating which fingers are active (selected) and which fingers are inactive (nonselected), as well as what the flexion-extension configuration is of the base (metacarpophalangeal) and the non-base (proximal interphalangeal and distal interphalangeal) joints.

Keane (2014b), and his Articulatory Model of Handshape, furthers Brentari's model by developing an explicit connection between the phonological specification for a handshape, and target joint angles for each joint of the (phonetic) hand configuration. His model can produce continuous (as well as categorical) measures of hand configuration which have been shown in previous studies to better match data on handshape similarity and confusability (Stungis 1981). Additionally, these continuous measures provide a straightforward way to compare two handshapes. Other phonological models could, in principle, be used, although each would require the development of a translation from categorical phonological features to continuous joint angles or an independent method of comparing the categorical features directly to each other. For these reasons, we will use Keane's model as a start of our theory-driven measure of phonetic similarity. The nature of this similarity will be described in detail in the next section, and then tested with psycholinguistic evidence in section 3.

2 Metrics for similarity for ASL fingerspelling?

Handshapes in sign languages do not occur in a vacuum: they are just one component that makes up lexical signs, along with the other major parameters: location, movement, orientation, and non-manual markers (Stokoe *et al.* 1965; Battison 1978). In ASL, fingerspelling is a loanword system used to borrow (written) English words into the language. In the fingerspelling system, each orthographic letter is mapped onto a set of 22 unique handshapes plus, in a limited number of cases, a non-default palm orientation or with an added movement. These handshapes are executed in quick succession, in the sequence of the letters of the written word. Broadly speaking, fingerspelling has been found to conform to many aspects of the ASL phonological system (Padden 1998; Brentari 1998; Brentari & Padden 2001; Cormier *et al.* 2008). Because the main contrast between letters in fingerspelling is, for the most part, only a handshape contrast, fingerspelling is a perfect place to test theories of the representation of handshape independent of the possible con-

founds of movement or location that would be inherent in using lexical signs or nonce signs that conform to the phonological structure of lexical signs. Similarity across these three parameters has been studied by Hildebrandt & Corina (2002), who found that signs that had identical movements (or identical locations, for the (sign naive) hearing subjects only) were rated as more similar than signs that had identical handshapes by native signers as well as by hearing subjects. Late learners of ASL, however rated signs that had identical handshapes as much more similar than signs that had identical locations or movements. Hildebrandt & Corina (2002) did not compare (and did not purport to compare) the relative similarities between different handshapes, however. Instead they were comparing pseudo-signs with 4 other pseudo-signs that had one or two parameter(s) (movement, location, handshape) that were identical, but all other parameters were different and asking signers to pick one psuedo-sign as the most similar in order to determine which parameter impacted similarity the most. Our study, on the other hand, delves into the relative similarities of handshapes, rather than looking at just identical or not identical handshapes in a study of sign similarity.

Keane's model (Keane 2014b) provides joint angle targets for each handshape used in ASL fingerspelling. This allows for a straightforward comparison of individual handshapes by taking the difference between the two sets of joint angle targets. This difference can then be thought of as the similarity between any given pair of handshapes. This difference is further refined by weighting each joint based on how proximal (or how close to the center of the body) it is. This weighting is supported by work that shows that movement of more proximal joints generates larger visual differences, which has been linked to visual sonority for signed languages (Brentari 1998). Additional support for this kind of sonority in sign languages can be found in (Hildebrandt & Corina 2002). Movement and location are parameters that in general use more proximal joints than handshapes, which results in larger visual differences. It is exactly those parameters where identical movements and locations were rated as more similar than identical handshapes (this pattern is found in general, and specifically with native signers).

2.1 Comparing two handshapes

To make this comparison explicit, consider the ASL fingerspelling handshapes -C- and -A-. Each handshape is made up of phonological features for each part of the hand that has been found to be phonologically contrastive, given in tables 1 and 2. These tables are a notational variant of the phonological features from (Brentari 1998), which were modified slightly in (Keane 2014b). The hand is subdivided into

three groups: the primary selected fingers, the secondary selected fingers, and the nonselected fingers. For each handshape, each finger must be assigned to one, and only one, of these groups. Within each of the selected finger groups, the joints (either the base joint: the metacarpophalangeal (MCP) joint, or the nonbase joints: the proximal interphalangeal (PIP) and the distal interphalangeal (DIP) joints) of the fingers all have the same configuration. The thumb additionally has a feature for opposition, to account for the additional degree of freedom that the carpometacarpal (CM) joint of the thumb has. Finally, the nonselected fingers are either all extended or all flexed.

group	feature	value
psf	members	index, middle, ring, pinky, thumb
	base (MCP) joint	ext
	nonbase (PIP and DIP) joints	mid
	abduction	adducted
ssf	members	none
	base (MCP)	NA
	nonbase (PIP and DIP)	NA
thumb	opposition	opposed
nsf	members	none
	joints	NA

Table 1: An image of the -C- handshape, generated by the Articulatory Model python module (Keane 2014a) and phonological specifications for the -C- handshape. The groups are the levels of selection: psf: primary selected fingers, ssf: secondary selected fingers, nsf: nonselected fingers, thumb: opposition features specific to the thumb. If a group has no members, the features for those members are not applicable (indicated here with NA).

group	feature	value
psf	members	index, middle, ring, pinky
	base (MCP) joint	flex
	nonbase (PIP and DIP) joints	flex
	abduction	adducted
ssf	members	thumb
	base (MCP)	mid
	nonbase (PIP and DIP)	ext
thumb	opposition	unopposed
nsf	members	none
	joints	NA

Table 2: An image of the -A- handshape, generated by the Articulatory Model python module (Keane 2014a) and phonological specifications for the -A- handshape. The groups are the levels of selection: psf: primary selected fingers, ssf: secondary selected fingers, nsf: nonselected fingers, thumb: opposition features specific to the thumb. If a group has no members, the features for those members are not applicable (indicated here with NAs).

The next step is to move from these phonological features to joint angles which represent the phonetic target for the handshape. To do this, the Articulatory Model of Handshape (Keane 2014b) uses translations for each phonological feature to (ideal) joint angle targets. For example, for the extension values of the base and nonbase joints, table 3 can be used. For more details about this translation mechanism see (Keane 2014b). It should be noted, that these angles are the angles formed by the bones on either side of the joint. For this reason, full extension (although not hyperextension) is 180°, full flexion (for most individuals) is 90°.

feature	joint angle target
ext	180°
mid	135°
flex	90°

Table 3: Translation from extension features to joint angles.

Using the computational implementation of the Articulatory Model of Handshape (Keane 2014a), we can calculate joint angle targets for our two example handshapes, -C- and -A-. These joint angle targets are given in tables 4 and 5.

From the tables of joint angle targets, comparing two handshapes is simple. Each joint angle from one handshape can be subtracted from the corresponding joint angle of the other handshape. For example, the DIP joint of the index finger for the -C- (135°) is subtracted from the DIP joint of the index finger for the -A- (90°) resulting in a difference of 45°; the PIP joint of the index finger for the -C- (135°) is subtracted from the PIP joint of the index finger for the -A- (90°) resulting in a difference of 45°; the MCP joint[1] of the index finger for the -C- (180°) is subtracted from the MCP joint of the index finger for the -A- (90°) resulting in a difference of 90°; and so on, for each joint angle. The full set of differences can be seen in table 6.

[1]Again, the MCP joint has two degrees of freedom (flexion and abduction) the angles here are just the flexion portion of the MCP joint. The abduction portion are labled as abd.

	flexion			abduction
	DIP	PIP	MCP	MCP
index	135°	135°	180°	0°
middle	135°	135°	180°	0°
ring	135°	135°	180°	0°
pinky	135°	135°	180°	0°
		IP	MCP	CM
thumb		135°	180°	(-22°,-27°,13°)

Table 4: Phonetic joint angle targets for each joint of the hand for the -C- handshape. Each of the interphalangeal joints (DIP and PIP) have a single degree of freedom: flexion-extension. The MCP joint has two degrees of freedom: flexion-extension and abduction. For the thumb, there is only one interphalangeal (IP) joint, the MCP joint only has one degree of freedom (flexion-extension), and the abduction column triplet of numbers is for the three degrees of freedom of the thumb's CM joint.

	flexion			abduction
	DIP	PIP	MCP	MCP
index	90°	90°	90°	0°
middle	90°	90°	90°	0°
ring	90°	90°	90°	0°
pinky	90°	90°	90°	0°
		IP	MCP	CM
thumb		180°	135°	(23°,0°,8°)

Table 5: Phonetic joint angle targets for each joint of the hand for the -A- handshape. The format of this table is the same as table 4.

	flexion			abduction
	DIP	PIP	MCP	MCP
index	45°	45°	90°	0°
middle	45°	45°	90°	0°
ring	45°	45°	90°	0°
pinky	45°	45°	90°	0°
		IP	MCP	CM
thumb		-45°	45°	(-45°,-27°,5°)

Table 6: Difference in the phonetic joint angle targets for each joint of the hand between the -C- handshape and -A- handshape. The format of this table is the same as table 4.

Finally, to get a single number that represents how different the handshapes are from each other, we must sum them together. But first, we multiply each cell by a weighting factor[2] (where the more proximal the joint, the more weight is assigned). It is known that joints that are more proximal will result in larger parts of the body moving. Additionally, there is evidence from other research that shows that these larger visual differences, are a type of visual sonority for signed languages (Brentari 1998). For this reason, we weight the scores for similarity such that a 1° difference at the MCP joint will be quantified as more different than a 1° difference at the DIP joint. After the weights, we can sum the absolute values of each joint angle difference, to arrive at a single number that is a quantification of the difference between the -C- and -A- handshapes: 2031. Now that we have a quantification for the difference between two individual handshapes, we need to extend this method to account for handshapes sequences.

2.2 Handshape sequences (that is, fingerspelling)

The proposal explained above for individual handshape similarity must be extended to account for fingerspelled words that are composed of sequences of multiple handshapes.

All models of fingerspelling perception, except for the initial cipher model (Blasdell & Clymer 1978), posit that fingerspelling perception is not simply the identification of each handshape individually. Rather, the transitions play some role in perception (Wilcox 1992). The Movement Envelope Theory for fingerspelling (Akamatsu 1985) goes further and identifies that it is the overall shape of the hand opening and closing within a word that aides perception. Akamatsu (1982) shows that children acquiring fingerspelling first identify and mimic the overall movement of fingerspelling, and then master full execution. The Movement Envelopes of two words could be thought of as a proxy for similarity: if two words share similar or the same Movement Envelope, they will be more similar than two words that have very different Movement Envelopes. The Movement Envelope can be interpreted in two different ways:

The first interpretation is that the crucial aspect of fingerspelling that generates the perception of an overall Movement Envelope is the transition between different

[2]The weights were as follows: DIPs and IPs have a weight of 1, PIPs have a weight of 2, finally MCPs and CMs have a weight of 3. This is from (Keane 2014b) directly, who admits that these numbers are a first step towards the appropriate visual weigths. They capture the generalization that more proximal joints generally make more visually salient movements. The scale of the difference between them needs to be refined by further study of visual sonority.

handshapes (or letters). This view is supported by work on sonority and local lexicalization of fingerspelling (Brentari 1998). In local lexicalization, a fingerspelled word is reduced during a single discourse, from the full fingerspelled version to a reduced version that looks more like a loan sign. Which letters are preserved and which letters are omitted is not random: the transitions between letters that preserve the largest movements are kept, first those with a non-default orientation or movement, and then those that preserve an overall alternation of open and closed handshapes. This pattern has been linked to sonority, which is the relative strength (or salience) of a specific sound or syllable in spoken languages or of a specific movement or syllable in sign languages (Brentari 1998). Borrowing her example, when the word S-Y-N-T-A-X is being locally lexicalized, the output is S-Y-T-X, with an additional movement of the wrist downward between the -S- and -Y-, and an additional movement of the wrist sideways between -T- and -X-. The -N- and -A- are deleted because both N-T and T-A are transitions between closed handshapes, adding no salient twisting movements of the wrist or opening/closing movements of the fingers.

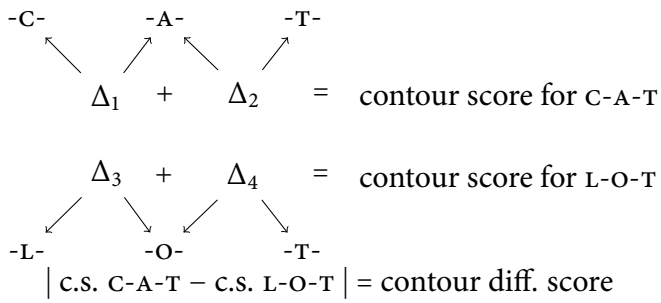
The second interpretation is that the crucial aspect of fingerspelling that generates the perception of an overall Movement Envelope is the overall shape of the whole word, including what position each of the different levels of openness or closedness of the hand occur in. Under this interpretation, it is not only the overall extension of the fingers that is important, but where within the word each class of handshape is. To make this concrete, and in the critical case where it differs from the previous interpretation: consider the two fingerspelled sequences A-B-A and B-A-B. In the first, the word starts with a closed handshape (-A-), then has an open handshape (-B-), and then ends with a closed handshape (-A- again). In the second, the word starts with an open handshape (-B-), then has a closed handshape (-A-), and then ends with an open handshape (-B- again). Using just the contours between the handshapes as a guide (as with the first interpretation), these words look similar: they each have the same sequence of transitions (just in a different order): A-B and B-A. In the second interpretation, despite the fact that there are the same transitions, the positions of each open or closed handshape is important in distinguishing these two sequences.

Based on these two disparate interpretations of the Movement Envelope, there are two possibilities for comparison. The first method, what we call the *contour difference* method, follows directly from the first interpretation of the Movement Envelope, and the second method, what we call the *similarity* method follows directly from the second interpretation of the Movement Envelope. It should be noted that although these different methods were inspired by these two different interpreta-

tions of the Movement Envelope theory, they stand fully independent of it. The *contour difference* method emphasizes the overall contour of the fingerspelled words: that is how one letter transitions to the next. Whereas the *similarity* method looks at the position of each letter within the word, and determines how similar the handshapes in that position are with handshapes in the same position of other words.

Figure 1: *Examples of* C-A-T *(above) and* L-O-T *(below).* Photos here are the canonical forms of each letter in both words. This pair of words is used in the diagrammatic descriptions of the two methods in figures 2 and 3

The first method results in what we call a *contour difference* score; this score is based on the general finding that there are (at least) two classes of handshapes (open and closed). In this method, each handshape in the word is compared to the one that follows it, that is, the differences between each sequential pair of letters is calculated and then summed together. Under this metric, a word that has a se-



For this pair, the contour difference score is:

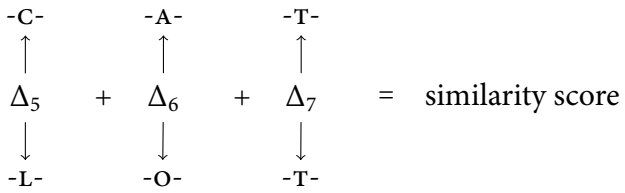
$$|(\ (\text{-C-};\text{-A-}) + (\text{-A-};\text{-T-})\ ) - (\ (\text{-L-};\text{-O-}) + (\text{-O-};\text{-T-})\ )| =$$
$$|(\quad \Delta_1 \quad + \quad \Delta_2 \quad) - (\quad \Delta_3 \quad + \quad \Delta_4 \quad)| =$$
$$|(\quad 2031 \quad + \quad 360 \quad) - (\quad 1521 \quad + \quad 1356 \quad)| = 486$$

Figure 2: *Contour difference* score calculation between the words C-A-T and L-O-T. Under this metric, a sequence of all open or all closed handshapes will have a low score, and a word with a sequence of open-closed handshapes will have a high score.

quence of all open or all closed handshapes will have a low score, and a word that has a sequence of open-close handshapes will have a high score. In order to arrive at a similarity score for a pair words using this method, a contour score for each word will be calculated, and then the difference between them will be calculated. See figure 2 for a diagram of an example pair of words.

The second method results in what we call *similarity* score. In this method, each pair of letters in the same position within the two words are compared to each other and their difference is calculated. The differences for each position in the word are then summed together. With this metric, words that have the same or similar handshapes in the same positions will be scored as more similar than those that have dissimilar handshapes in the same positions. Under this metric words that are similar will have a low score, and words that are dissimilar will have a high score. See figure 3 for a diagram of an example pair of words.

Although the contour difference score can easily compare two words of different lengths, the similarity score as described above, is limited to words that have the same number of letters in them. In the experimental data described below, words



For this pair, the similarity score is:

$$(\text{-C-}; \text{-L-}) + (\text{-A-}; \text{-O-}) + (\text{-T-}; \text{-T-}) =$$
$$\Delta_5 + \Delta_6 + \Delta_7 =$$
$$1656 + 1356 + 0 = 3012$$

Figure 3: *Similarity* score calculation between the words C-A-T and L-O-T. Under this metric words that have similar handshapes in corresponding positions will have a low score, and words that are dissimilar will have a high score.

with either 3 or 4 letters were compared in pairs that were the same length, as well as in pairs that differed in length. In order to attain a similarity score, a composite metric was developed: The shorter word was held constant, but then compared to all possible strings of the longer word where one of the letters was deleted. The mean of this score resulted in the final similarity score for mismatched lengths. For example, to generate an overall similarity score for the pair of words L-O-T and L-E-A-N, a score was calculated for each of the following pairs: (L-O-T ; E-A-N), (L-O-T ; L-A-N), (L-O-T ; L-E-N), and (L-O-T ; L-E-A). The mean of these four individual scores was taken as the similarity score for δ(L-O-T ; L-E-A-N). Though there are other methods that could be used to compare words with mismatched lengths, this method is a first step in that direction, which deserves further research.

3 Psycholinguistic experiment

Previous studies relied on data from psycholinguistic experiments to develop clusters of handshapes that are similar and then proceeded from their psycholinguistic data to a linguistic model of handshape similarity rather than the reverse. The two proposed methods here, the *contour difference* method and the *similarity* method for calculating the similarity of two fingerspelled words (which could also be applied directly to any sequence of handshapes) are the opposite: they use a linguis-

tic model of the phonetics-phonology interface for handshape (the Articulatory Model of Handshape (Keane 2014b)) to generate a theory-driven metric of similarity. We conducted two separate fingerspelling similarity judgment studies to determine which of the two methods of similarity estimation – contour or similarity method – better predicts the signers' subjective ratings of similarity. For the first rating study, manually similar words contained compact handshapes (following Hanson *et al.* (1984)). For the second rating study, similar and dissimilar word pairs were selected based on a theory-driven handshape similarity metric (Keane 2014b). Subjects' scores were then fit and compared using multiple hierarchical linear regressions.

3.1 Methods

3.1.1 Participants

Twenty-three deaf ASL signers participated in two separate online rating studies. In the first study, there were 11 deaf ASL signers (mean age = 32.4, SD = 9.8, 7 female) and in the second study, 13 deaf ASL signers (mean age = 36.2, SD = 13.6, 11 female) participated. All participants acquired ASL before age 7 and reported using ASL as their primary and preferred language. All participants were congenitally deaf and had severe (71–90 dB) to profound (90–120 dB) hearing loss. The experiment was administered online and all participants received gift certificates upon completion.

3.1.2 Stimuli and procedure

In the first rating study, 214 pairs of manually similar and dissimilar FS words were selected based on psychological theories of handshape similarity; the manually similar words contained consonant handshapes that were argued to be confusable by native signers (e.g. -M-, -N-, -S-, and -T-; (Hanson *et al.* 1984; Richards & Hanson 1985)) and vowel handshapes that use the same compact hand configurations and are easily confusable (e.g. -A-, -O-, and -E-; (Lane *et al.* 1976)), as in M-E-A-T, S-O-N, T-E-N, and E-A-S-T. Examples of dissimilar words under these previous studies include K-I-N-G, F-A-R-M, B-U-G, and T-A-X. One participant only completed half of the experiment, so we removed all of their responses (although including them does not impact the results). There were four observations where participants reported difficulties viewing the videos, and thus did not report a similarity. No other data was removed for the analysis.

In the second rating study, 132 word pairs were selected based on a theory-driven handshape similarity metric (Keane 2014b), that is, the manually similar words all proceeded from open to closed (one movement open to closed), there was one change in selected fingers between letters, and no orientation changes (e.g. to or from -H-), as in C-A-T, V-A-N, R-E-N-T, and L-E-A-N. The dissimilar word pairs contained more than one movement and change in selected fingers and orientation change, examples include L-O-V-E, S-I-C-K, B-O-X, and H-A-T. Words in the similar and dissimilar groups were matched on length (3 or 4 letters), frequency, concreteness, and all had an ASL translation equivalent and no phonological or orthographic overlap. All participants responded to all the stimuli and no data was removed for the analysis.

In the first rating study, fingerspelled word pairs were produced by a deaf native ASL signer who was filmed at a frame rate of 29.97 frames per second. The edited video clips were uploaded to an online survey tool for rating and were divided into two blocks containing 107 pairs each to allow for a break. In the second study, word pairs were presented as print. Participants were asked to rate all word pairs for manual similarity based on how similar the words in the pair feel to each other when they fingerspelled them to themselves on a 1–5 scale based (1–do not feel similar at all; 5 – feel very similar). Ratings from the online surveys were exported as a comma-separated text file for further analysis.

3.1.3 Analysis

It should be noted that our two experiments were collected independently. The second experiment was conducted to directly test that our *contour difference* method and our *similarity* method were quantifying something that is psychologically real, and to see which method matches the data better. The first experiment, however, was run as subset of a separate experiment, well before the formulation of either method by our research team. This being the case, those ratings can be thought of as a type of independent verification since they were collected in a kind of double-blind condition where neither the subjects nor the experimenters knew the hypotheses being tested (and thus neither could influence the outcome of the experiment). Additionally, as described in the methods above, the two experiments used slightly different methodologies: in the first signers were presented with videos of fingerspelled words and they were asked to rate the similarity between the pair. In the second study, signers were presented with printed words, and asked to think about fingerspelling them, and rate the similarity with how they felt when fingerspelled. Thus, in the second experiment, pure visual similarity of the stimuli it-

self was not influencing the signers' ratings. Models were fit both using subject and experiment as hierarchical grouping variables, as well as using experiment as a predictor variable. Although ratings in the second experiment were significantly lower (only by about 0.4 points), there was not variation in which predictors were significant when the experiment was used as a predictor or grouping variable in the model. Because using experiment as a predictor variable or as a grouping variable (where the intercept and slope of the predictor variables are allowed to vary) does not change the results or interpretation of the other predictor variables, we will use experiment as a grouping variable as that accurately represents the hierarchical structure of the data (individual signers are nested within experiments). Figure 4 shows predictions broken down by experiment (and length), and shows that the effects in these groups are extremely similar.

In order to test which of the two methods of scoring (*contour difference* scores or *similarity* scores) predicts signers' ratings, multiple hierarchical linear regressions were fit and then compared. All models were fit with the `lme4` package version 1.1-8 (Bates 2010) in R. All scores were divided by the length of the words in order to not unfairly penalize long words. In cases with mismatched word lengths, the scores were divided by 3, since the similarity score for mismatched pairs is the mean of the set of comparisons across the three letter word and all combinations of the four letter word minus one letter. For both scores, a higher score is less similar, and a lower score is more similar (i.e. perfect similarity is zero). If the scores are predictive of the signers' similarity ratings, we expect a negative correlation (this is because the signers rated on a scale where higher ratings were **more** similar, whereas both the *contour difference* and the *similarity* scores are higher if the words are **less** similar). All scores were scaled to z-scores for comparison of effect sizes. In each model, the subject and the experiment the subject's data was collected in were included as hierarchical grouping variables. This allows us to see if there were systematic differences between the two sets of data collection. The models were:

1. *Null model* with no predictor variables, which had varying intercepts (AKA mixed effects) for subject group, subject, first word of the pair, and second word of the pair.

2. *Contour difference score model* with predictor variables of the contour difference score for the word pair, the length of the words (3 letters, 4 letters, or mismatched), and the two way interaction of these. There were varying intercepts and slopes for subject group, subject, first word, and second word.

3. *Similarity score model* with predictor variables of the similarity score for the word pair, the length of the words (3 letters, 4 letters, or mismatched), and the two way interaction of these with the same varying intercepts and slopes as the previous

model.

4. *Full model* which included predictor variables of the similarity score, contour difference score, the length of the words (3 letters, 4 letters, or mismatched), and all possible two and three way interactions with the same varying intercepts and slopes as the previous model.

3.2 Results

Each model will be discussed in detail below, but the predictor variable for *similarity* score was significantly correlated with signers' ratings in the predicted direction in every model that it was included in. As the *similarity* score went up (with our model predicting that the word pairs were less similar) the signers' ratings went down (meaning the signers thought these words were less similar). The relationship between the *similarity* score and the signers' ratings can be seen in figure 4. Figure 4 has the signers' rating data plotted against *similarity* score data broken down by experiment and length as well as plotting the regression line (and confidence intervals) for the *similarity* score model. The *similarity* score model is displayed here because, as discussed in the model comparison section, that is the simplest model that is justified given the data. Additionally, plots for the full model, as well as models where experiment is a predictor variable instead of a grouping variable show the exact same patterns. For model comparison, which will be discussed in detail in the next session, figure 5 shows predictor coefficients for all models except the null model. In this plot the coefficients for each predictor variable in each model are plotted along with their confidence intervals. Thus, for each predictor, the dot is the coefficient estimate, the thick line is the 95% confidence interval, the thin line is the 99% confidence interval. We can have confidence that a coefficient is a true effect within the population, and is not attributable to noise in the sample, when the confidence intervals do not overlap zero. This plot additionally allows us to determine not only if we have confidence that an effect is statistically significantly different from zero, but also in which direction: positive or negative (as well as the magnitude of the effect size, this kind of analysis of significance follows Gelman *et al.* (2012); Gelman (2013); Gelman & Carlin (2014)).

The null model serves as a baseline of comparison to see if the complexity associated with adding predictors to the model is justified given the data. Because there are no predictors in this model, there are no significant effects to report.

In the contour difference score only model, the contour difference score alone does not significantly predict the signers' ratings. There is a significant effect of length, where four letter words are more similar than mismatched words or three

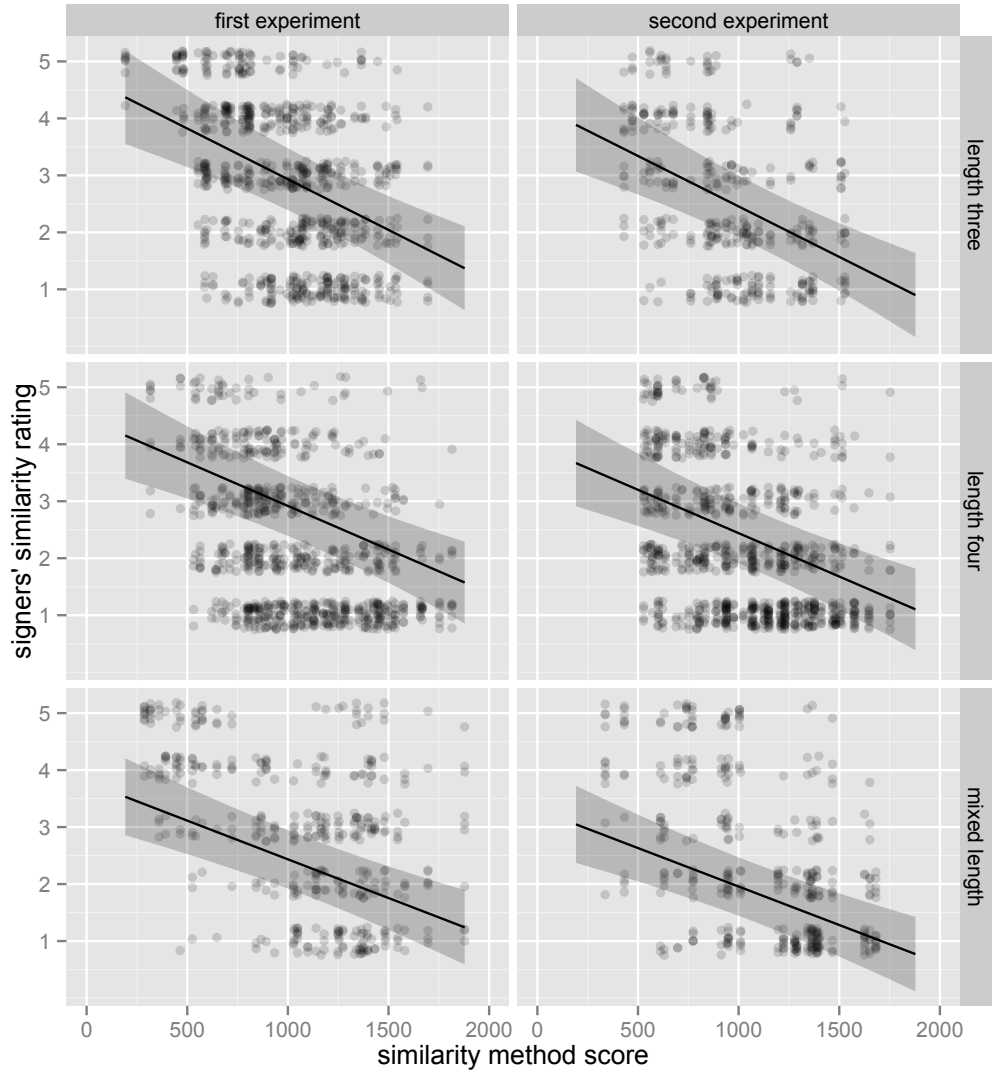


Figure 4: Signers' similarity rating data, along with model predictions from the *similarity score model* broken down by experiment (facet columns, labeled *first* or *second* (experiment)) and word length (facet rows, labeled *three*, *four*, or *mixed*). The dots are the similarity score of the pair of words (x axis) plotted against signers' ratings (y axis). They are jittered along the y (but not x) axis, the signers ratings were whole numbers 1–5. The lines are the fit of the linear regression model, with shaded areas are 95% confidence intervals.

letter words. The interaction of length and contour difference score is significant, when the words are both three letters, then the smaller the contour difference score, the more similar the signers rated the pair. No other predictors had significant effects.

In the similarity score only model, the similarity score significantly predicts the signers' ratings in the expected direction (the lower the similarity score, the higher the signers' ratings). Additionally, word pairs that had the same lengths (either both 3 letters or both 4 letters long) were rated significantly more similar than word pairs that were mismatched. No other predictors had significant effects.

Finally, in the full model, the similarity score significantly predicts the signers' ratings in the expected direction (the lower the similarity score, the higher the signers' ratings). Word pairs that were both four letters long were rated significantly more similar than word pairs that were mismatched. The effect of the similarity score was (marginally) magnified when both words were four letters. No other predictors had significant effects, which includes predictors for contour difference score, as well as all interactions with the contour difference score.

It stands out that in no model does the contour difference score alone significantly predict signers' ratings of the similarity of fingerspelled words. In contrast, the similarity score, does significantly predict signers' similarity ratings and in the predicted direction. Again, see figure 5 for a visualization of the predictor coefficients for all models except the null model.

For the two predictors that are the center of this paper, the two different methods for comparing two fingerspelled words, we expect both to have a negative correlation because the signers rated on a scale where higher ratings were **more** similar, whereas both the *contour difference* and the *similarity* scores are higher if the words are **less** similar. A negative coefficient in a hierarchical linear model shows exactly this negative correlation between predictor variables and outcome variables. As described in the text above, the only methods that are significantly different from zero (and in the correct direction: negative) are those for the *similarity* score in the *similarity* score only model and the full model. The *contour difference* score is not significantly different from zero (and thus we cannot asses the direction, or sign of the effect) in either the *contour difference* only model or the full model.

3.3 Model comparison

Although there is not a single, best method for model comparison, especially for hierarchical models like those used here, a number of methods have been proposed and have seen some acceptance.

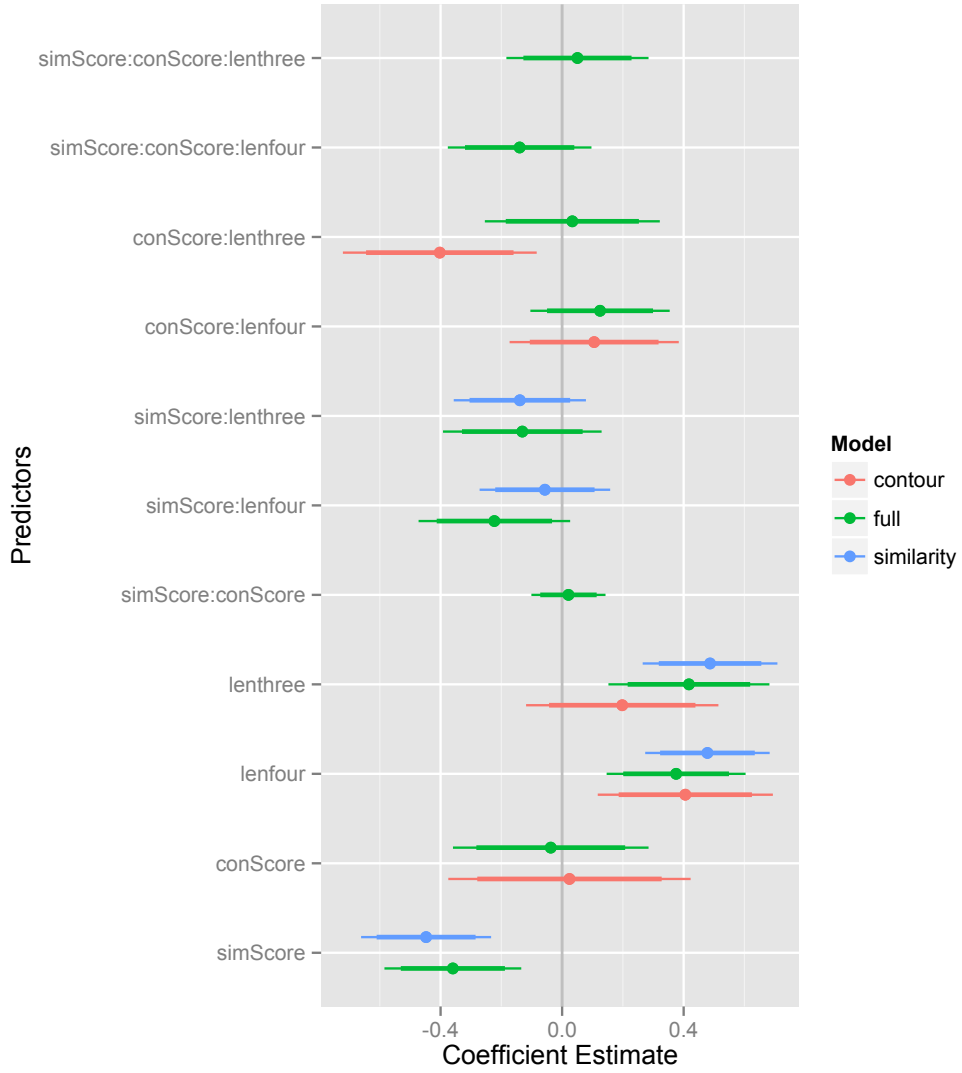


Figure 5: Coefficient plot for contour difference score, similarity score, and full models. Thick lines are 95% CI, thin lines: 99% CI, and dots: estimates of the predictor coefficients. If a particular predictor's confidence intervals do not overlap with zero, we can have confidence that the effect of that predictor is statistically significant. This plot allows us to evaluate not just simple significance, but also the sign or direction of the effect (positive or negative), as well as see the relative magnitudes of each effect. conScore: contour diff. score; len: length of word with levels four, three, and mismatched (reference level); simScore: similarity score.

The first kind of comparison is the use of information theoretic measures to determine if the extra complexity of adding predictors is justified by the data. In other words, does adding a given parameter give us enough predictive power to justify the added complexity it introduces to the model. There are two mainstream information theoretic measures: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Both methods can be fit to different non-nested models applied to the same underlying data set (how we are using them here) (Burnham & Anderson 2004; Anderson & Burnham 2006). For both AIC and BIC, lower numbers indicate a better fit of the model to the data. In the most conservative recommendations, a difference of 10 or more indicates that the models differ significantly and the model with the lower score should be preferred (all differences between the AIC and BIC for our models were larger than this threshold). Using the AIC, the simplest model that is justified given the data is the full model (AIC: 11381.47) that includes both the similarity score and the contour difference score (however, it should be reiterated that the contour difference score does not have a significant effect in this model). Additionally, the similarity score only model (AIC: 11438.1) is significantly more well supported than the contour score only model (AIC: 11868.27). Using the BIC, the simplest model that is justified given the data is the similarity score only model (BIC: 11556.97). Additionally, the full model (BIC: 11612.95) is significantly more well supported than the contour score only model (BIC: 11987.15). See table 7 for AICs and BICs.

The second kind of comparison is to use a new method for calculating R^2, or the variance of the data explained by the model. Traditionally, calculating R^2 for hierarchical models has not been straightforward. However recent work (Nakagawa & Schielzeth 2013; Johnson 2014) has developed a method that gives a marginal R^2, which corresponds to the R^2 of the predictors alone, and a conditional R^2, which corresponds to the R^2 of the predictors along with the varying intercepts and slopes. With both traditional R^2 and with this new calculation, R^2 ranges from zero (no variance of the data is explained by the model) to one (all of the variance of the data is explained by the model). We will only discuss the marginal R^2 here, because we are concerned with the variance explained by the predictors, and not the the varying intercepts or slopes. Under this metric, the model that explains the most variance of the data is the similarity score only model ($R^2 = 0.16$) and in a close second, is the full model ($R^2 = 0.13$). Both the contour difference score only model and the null model explain very little variance of the data ($R^2 = 0.02$ and $R^2 = 0$, respectively).

Although the model comparisons do not all agree on one specific model, it is clear that two stand out: the similarity score only model (the simplest model jus-

model	AIC	BIC	R^2
null	12194.00	12231.54	0.00
contour diff. score	11868.27	11987.15	0.02
similarity score	11438.10	11556.97	0.16
full	11381.47	11612.95	0.13

Table 7: Model comparison using AIC, BIC, and marginal R^2

tified given the data using BIC, and marginal R^2, and the second simplest model justified given the data using AIC) and the full model that includes both the similarity score and the contour difference score (the simplest model justified given the data using AIC, and the second simplest model justified given the data using BIC and marginal R^2). Additionally, even when the contour difference score is included in the full model, it does not significantly predict signers' similarity ratings.

4 Conclusion

We have demonstrated that *similarity* score is the theory-driven description of handshape similarity that best matches signers' intuitions when asked to rate the similarity of fingerspelled words. The similarity metric proposed here is exactly the kind of theory-driven metric that was recognized as missing from the similarity research in the 1970s and 80s, which has also been independently confirmed with signers' intuitions of similarity.

It is clear that the *similarity* approach is a superior fit to the data when compared with the *contour difference* approach. In order to define similarity in sequences of handshapes (e.g., fingerspelling), it is more important to look at the positional configuration of handshapes than it is to concentrate solely on the transitions between handshapes. Similarity between stimuli (words, sentences, etc.) is known to affect many psycholinguistics process (e.g., the phonological similarity effect in short term memory (Wilson & Emmorey 1997), form priming effects, syntactic priming effects). Our *similarity* metric is an easy to use metric that can be used to evaluate and control for the similarity of stimuli in other experiments.

The similarity method seems to be best supported by our psycholinguistic data. This approach is one of the two possible interpretations of previous work on the perception of fingerspelling (i.e. the Movement Envelope), and it matches with signers' ratings of similarity.

Although we presented convincing evidence that the *contour difference* method is not supported by the data from signers' similarity ratings, there might be other areas where handshape contours are important. As discussed above, signers' have shown sensitivity to handshape contours in local lexicalization (Brentari 1998). Additionally, signers show a tendency to chose the variant of -E- (open or closed) in order to make a more contrastive (larger) contour difference with surrounding handshapes (Keane *et al.* 2013; Keane & Brentari forthcoming).

In our studies, we used signers' ratings of similarity, but previous methods, such as handshape confusion under visual noise (Lane *et al.* 1976; Stungis 1981), could be used to test our *similarity* method as well. Because our *similarity* method is a metric of phonetic similarity, a metric concerning the form of the hands making the handshapes, we predict that we would see similar results for other tasks to what we found with our experiment here. Confusion in visual noise might attenuate or magnify the differences between two handshapes as quantified by our metric. However, there is no evidence that, when given a sequence of handshapes (e.g. fingerspelling) in noise, the positionally-sensitive *similarity* method would do worse than the *contour difference* method. This, however, needs to be tested empirically, and that is left to future work.

Finally, our *similarity* method answers a call made nearly 40 years ago. The Articulatory Model of Handshape, combined with our *similarity* approach, is a phonetically and phonologically theory-driven similarity metric for comparing handshapes. This metric not only produces results that match intuitions from previous studies (Locke 1970; Weyer 1973; Lane *et al.* 1976; Stungis 1981; Richards & Hanson 1985), but also produces results that match signers' similarity ratings of fingerspelled words.

References

Akamatsu, Carol Tane, 1982. *The acquisition of fingerspelling in pre-school children*. University of Rochester dissertation.

——. 1985. Fingerspelling formulae: A word is more or less the sum of its letters. *SLR* 83.126–132.

Anderson, David, & Kenneth Burnham, 2006. aic myths and misunderstandings.

Bates, Douglas, 2010. *Linear mixed model implementation in lme4*. Department of Statistics, University of Wisconsin-Madison.

Battison, Robbin. 1978. *Lexical Borrowing in American Sign Language*. Silver Spring, Maryland: Linstock Press.

Blasdell, Richard, & William Clymer. 1978. An empirical study of cipher, phonological and syntactic models of fingerspelling production. *American annals of the deaf* 123.857–72.

Brentari, Diane. 1998. *A prosodic model of sign language phonology*. The MIT Press.

——, & Carol Padden. 2001. *Foreign vocabulary in sign languages: A cross-linguistic investigation of word formation*, chapter Native and foreign vocabulary in American Sign Language: A lexicon with multiple origins, 87–119. Mahwah, NJ: Lawrence Erlbaum.

Burnham, Kenneth P, & David R Anderson. 2004. Multimodel inference understanding aic and bic in model selection. *Sociological methods & research* 33.261–304.

Cormier, Kearsy, Adam Schembri, & Martha E. Tyrone. 2008. One hand or two? Nativisation of fingerspelling in asl and banzsl. *Sign Language & Linguistics* 11.3–44.

Eccarius, Petra, 2002. Finding common ground: A comparison of handshape across multiple sign languages. Master's thesis, Purdue University.

Gelman, Andrew. 2013. Commentary: P values and statistical practice. *Epidemiology* 24.69–72.

——, & John Carlin. 2014. Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9.641–651.

——, Jennifer Hill, & Masanao Yajima. 2012. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 5.189–211.

Hanson, Vicki L, Isabelle Y Liberman, & Donald Shankweiler. 1984. Linguistic coding by deaf children in relation to beginning reading success. *Journal of experimental child psychology* 37.378–393.

Hildebrandt, Ursula, & David Corina. 2002. Phonological similarity in american sign language. *Language and Cognitive Processes* 17.593–612.

Johnson, Paul CD. 2014. Extension of nakagawa & schielzeth's r2glmm to random slopes models. *Methods in Ecology and Evolution* 5.944–946.

Johnson, Robert E, & Scott K Liddell. 2011a. Toward a phonetic representation of hand configuration: The thumb. *Sign Language Studies* 12.316–333.

——, & ——. 2011b. Toward a phonetic representation of signs: Sequentiality and contrast. *Sign Language Studies* 11.241–274.

Keane, Jonathan, 2014a. amohs python module. software. Version 0.1.0 zenodo.org/record/11456.

——, 2014b. *Towards an articulatory model of handshape: What fingerspelling tells us about the phonetics and phonology of handshape in American Sign Language*. University of Chicago dissertation. Doctoral dissertation, defended 22 August 2014 Advisors: Diane Brentari, Jason Riggle, and Karen Livescu.

——, & Diane Brentari. forthcoming. *The Oxford Handbook of Deaf Studies in Language: Research, Policy, and Practice*, chapter Fingerspelling: Beyond Handshape Sequences. Oxford University Press.

——, Diane Brentari, & Jason Riggle, 2013. Dispelling prescriptive rules in asl fingerspelling: the case of -e-. poster. Theoretical Issues in Sign Language Research 11; London, uk.

Lane, Harlan, Penny Boyes-Braem, & Ursula Bellugi. 1976. Preliminaries to a distinctive feature analysis of handshapes in american sign language. *Cognitive Psychology* 8.263–289.

Liddell, Scott, & Robert E Johnson. 1989. American Sign Language: A phonological base. *Sign Language Studies* 64.195–277.

Liddell, Scott K, & Robert E Johnson. 2011a. A segmental framework for representing signs phonetically. *Sign Language Studies* 11.408–463.

——, & ——. 2011b. Toward a phonetic representation of hand configuration: The fingers. *Sign Language Studies* 12.5–45.

Locke, John L. 1970. Short-term memory encoding strategies of the deaf. *Psychonomic Science* 18.233–234.

Mandel, Mark, 1981. *Phonotactics and morphology in American Sign Language*. University of California, Berkeley dissertation.

Mauk, Claude E, & Martha E Tyrone. 2012. Location in asl: Insights from phonetic variation. *Sign Language & Linguistics* 15.128–146.

Miller, George A, & Patricia E Nicely. 1955. An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America* 27.338–352.

Nakagawa, Shinichi, & Holger Schielzeth. 2013. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4.133–142.

Padden, Carol. 1998. The asl lexicon. *Sign Language and Linguistics* 1.39–60.

Richards, John T, & Vicki L Hanson. 1985. Visual and production similarity of the handshapes of the american manual alphabet. *Perception & psychophysics* 38.311–319.

Sandler, Wendy. 1989. *Phonological Representation of the Sign: Linearity and Nonlinearity in American Sign Language*. Foris Pubs USA.

——, & Diane Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge Univ Press.

Stokoe, William C., Dorothy C. Casterline, & Carl G. Croneberg. 1965. *A Dictionary of American Sign Language on Linguistic Principles*. Washington DC: Gallaudet College Press.

Stungis, James. 1981. Identification and discrimination of handshape in American Sign Language. *Perception & Psychophysics* 29.261–276.

Tyrone, Martha E, Hosung Nam, Elliot L Saltzman, Gaurav Mathur, & Louis Goldstein. 2010. Prosody and movement in American Sign Language: A task-dynamics approach. In *Speech Prosody 2010*, 1–4.

van der Hulst, H. 1995. The composition of handshapes. *Working Papers in Linguistics* 1–17.

Weyer, Stephen A. 1973. Fingerspelling by computer. Technical Report 212, Institute for Mathematical Studies in the Social Sciences, Sanford University, Stanford, CA.

Whitworth, Cecily, 2011. *Features, Clusters, and Configurations: Units of Contrast in American Sign Language Handshapes*. Gallaudet University dissertation.

Wilcox, Sherman. 1992. *The Phonetics of Fingerspelling*. Amsterdam: John Benjamins Publishing Company.

Wilson, Margaret, & Karen Emmorey. 1997. Working memory for sign language: A window into the architecture of the working memory system. *Journal of Deaf Studies and Deaf Education* 2.121–130.