









Semantically Related Gestures Move Alike: Towards a Distributional Semantics of Gesture Kinematics

Wim Pouw^{1,2} , Jan de Wit³ , Sara Bögels⁴ , Marlou Rasenberg^{2,5} ,
Branka Milivojevic¹ , and Asli Ozyurek^{1,2,5} 

¹ Donders Centre for Cognition, Brain, and Behaviour, Radboud University Nijmegen,
Nijmegen, The Netherlands

w.pouw@donders.ru.nl

² Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

³ Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg,
The Netherlands

⁴ Department of Communication and Cognition, Tilburg University, Tilburg, The Netherlands

⁵ Center for Language Studies, Radboud University Nijmegen, Nijmegen, The Netherlands

Abstract. Most manual communicative gestures that humans produce cannot be looked up in a dictionary, as these manual gestures inherit their meaning in large part from the communicative context and are not conventionalized. However, it is understudied to what extent the communicative signal as such—bodily postures in movement, or kinematics—can inform about gesture semantics. Can we construct, in principle, a distribution-based semantics of gesture kinematics, similar to how word vectorization methods in NLP (Natural language Processing) are now widely used to study semantic properties in text and speech? For such a project to get off the ground, we need to know the extent to which semantically similar gestures are more likely to be kinematically similar. In study 1 we assess whether semantic word2vec distances between the conveyed concepts participants were *explicitly instructed* to convey in silent gestures, relate to the kinematic distances of these gestures as obtained from Dynamic Time Warping (DTW). In a second director-matcher dyadic study we assess kinematic similarity between *spontaneous* co-speech gestures produced between interacting participants. Participants were asked before and after they interacted how they would name the objects. The semantic distances between the resulting names were related to the gesture kinematic distances of gestures that were made in the context of conveying those objects in the interaction. We find that the gestures' semantic relatedness is reliably predictive of kinematic relatedness across these highly divergent studies, which suggests that the development of an NLP method of deriving semantic relatedness from kinematics is a promising avenue for future developments in automated multimodal recognition. Deeper implications for statistical learning processes in multimodal language are discussed.

Keywords: Manual gesture kinematics · NLP · Speech · Semantics · Time series comparison

1 Introduction

Humans exploit a multitude of embodied means of communication, where each mode of communication has its own semiotic affordances. Manual and whole-body communicative movements, such as co-speech gestures or signs in a sign language, have been suggested to leverage iconicity to convey meaning [1–3]. Iconicity is a special type of referential act, as the form of the message can inform more directly about the content of the message as compared to arbitrary symbols, by establishing a spatio-temporal resemblance between form and referent; for example, by moving in a way that resembles brushing one’s teeth (form), one can convey a meaning related to brushing one’s teeth (content). What is particularly astonishing is that during spoken language manual iconic references are spontaneously constructed, in a way that does not necessarily need to be repeated later when the referent is mentioned again [4], nor does it need to be replicated exactly when gestured about it in a similar context by someone else [5]. Thus even when two gestures have a similar meaning and occur in a similar speech context, they do not need to be replicated in form. This “repetition without repetition” [6]—a good characterization of human movement in general—is one of the reasons why the iconic meaning of gestures is generally held to be unformalizable in a dictionary-like way [7, 8], with the exception of more conventionalized emblem gestures (e.g., “thumbs up”; e.g., [8, 9]). To complicate matters further, gestures’ meaning is dependent on what is said in speech during gesturing, as well as the wider pragmatic context. All these considerations might temper expectations of whether information about the gesture’s content can be derived from the gesture’s form—bodily postures in motion, i.e., kinematics.

It is however an assumption that the kinematics of gestures are poorly informative of the meaning of a depicting or iconic gesture. Though it is undeniable there is a lot of variance in gestures’ form to meaning mapping, at some level there is invariance that allows depicting gestures to depict, some kind of abstract structural similarity at a minimum [10]. It is also possible that gestures are semantically associated by the mode of representation [11, 12] they share (which is not the same as, but related to certain kinematic properties such as handshape). For example, it has been shown that gestures for manipulable objects are likely to be of the type “acting” (e.g., moving your hand as if you are brushing your teeth to depict toothbrush) compared to gestures depicting non-manipulable objects (which are more likely to be “drawn”, e.g. tracing the shape of a house with the hands or index fingers) [3]. Gaining empirical insight in whether we can glean some semantic information from kinematics in a statistical fashion, is an important project as it would not only calibrate our deep theoretical convictions about how gesture kinematics convey meaning, but it would also pave the way for computer scientists to develop natural language processing (NLP) algorithms tailored for iconic gesture kinematics vis-à-vis semantics. Modern NLP procedures such as word embedding vectorization (word2vec) operate on the assumption of distributional semantics, holding simply that tokens that co-occur in similar contexts are likely semantically related. In the current study we will assess another assumption that could be powerfully leveraged by NLP procedures tailored to gesture semantics: Do gestures that semantically relate to one another move as one another?

If gestures do indeed show such statistical dependencies in form and meaning on the level of interrelationships, they offer a source of simplification of content that is similar in nature to statistical dependencies that characterize linguistic systems in general and are exploited by NLP [13]. Note though, that distributional semantics is something that simplifies the learning of a language for humans too, as for example an infant can leverage a language’s syntactic, semantic, and phonological co-dependencies via statistical learning [14]. Similarly, the current investigation of potential statistical co-dependencies between semantic and kinematic relatedness in gestures are key for coming to an understanding of how humans really learn and use language, which is a sense-making process steeped in a rich multimodal context of different forms of expression [15].

1.1 Current Investigation

In two motion-tracking studies we assess whether the semantic (dis)similarity between concepts that are putatively conveyed by gestures, are related to the (dis)similarity of the gesture’s kinematics. We computed word2vec distances between verbal labels of the concepts conveyed by gestures, and we computed kinematic distances using a well-known time-series comparison algorithm called Dynamic Time Warping (see e.g., [16–18]). By computing all possible distances between conveyed concepts, as well as gesture kinematics, we essentially map out a semantic and kinematic space that can be probed for covariances [13, 16, 19, 20].

For a large-scale charades-style study 1 with more than 400 participants, the concepts that were conveyed were defined from the outset, as participants were asked to convey in their own way a particular concept with a silent gesture (i.e., without speech) to a robot who was tasked to recognize the gesture [21]. Silent gestures are an idealized test case for us as they are designed to be maximally informative in that modality, and the structured nature of the interaction allows us to more definitively identify the semantic targets of the gestures.

However, silent gestures are not a common mode of expression in humans (note, signs in sign languages are not the same as silent gestures; for an introduction see [22]). Indeed, in most cases, gestures are generated spontaneously in the context of concurrent speech. There, speech often shares a communicative load with co-speech gestures, and verbally situates what is meant with a gesture [7]. Whatever semantic-kinematic scaling pattern we might find for highly communicatively exaggerated silent gestures, need thus not be replicated for co-speech gestures which perform their referential duties in a more speech-situated way.

In study 2, we opportunistically analyze dyadic interactions from a smaller lab study [23]. Dyads performed a director-matcher task, in which they took turns to describe and find images of novel 3D objects (‘Fribbles’ [24]). For each Fribble, we analyzed the gestural movements produced by both participants in the context of referring to that Fribble. Before and after the interaction, participants were individually asked to come up with a verbal label/name (henceforth “name”) for each Fribble (1–3 words) that would enable their partner to identify the correct Fribble. This allows us, similarly to study 1, to relate gesture kinematic differences to possible semantic word2vec differences of the Fribble names before as well as after the interaction task. Importantly, with regards to

study 1, we will analyze kinematic and semantic distances between individuals in a pair, such that we assess how gesture differences between Fribble i and j between participants in a pair relate to naming differences between participants for those Fribbles i and j . We thus analyze shared semantic and kinematic spaces, in search for covariances in their geometry.

2 General Approach

In both studies we computed the semantic (\mathbf{D}^s) and kinematic spaces (\mathbf{D}^g). Semantic spaces comprised semantic distances between concepts (study 1) or object names (study 2). Kinematic spaces comprised kinematic distances between the sets of gestures produced for two concepts (study 1) or two objects (study 2).

We used word2vec to compute semantic distances (1 - cosine similarity) between concepts that were (putatively) conveyed in gesture. To determine semantic dissimilarity between concepts we used SNAUT [25] to compute cosine similarity based on a Dutch model CoNLL17 [26]¹.

For the kinematic distance computation, we use Dynamic Time Warping (DTW). DTW is a well-known time series comparison algorithm, and it measures the invariance of time series under variations in time shifts. It does this by finding a warping line between time series, by constructing a matrix containing all distances between time series' values. The warping line is a trajectory over adjacent cells of the matrix which seeks the lowest distances between the time series values (see for details, [17, 18]). Conceptually, this amounts to aligning the time series through warping and then calculating the distances (or error) still remaining. The distance score is then normalized for the lengths of the time series, so that the possible amount of accumulated error is similar for time series of different lengths. The time series can be multivariate (e.g., horizontal and vertical position of a body part through time) such that the DTW is performed in a multidimensional space (for a visual explanation see, [16]). In essence the distance scores that are computed provide a summary value of the differences between two time series. In our case a time series defined the kinematic x, y, and z trajectory of a body part. We used an unconstrained version of DTW [17] implemented in R-package 'dtw', whereby beginning and trailing ends were not force aligned, thereby circumventing issues of discrepant errors that can be produced when the start and end points of the meaningful part of an event in a time series are not well defined [27]².

Given the exploratory nature of the current analysis, and given that we will be testing our hypothesis in two datasets, we will treat kinematic-semantic effects as statistically reliable at an Alpha of $<0.025(0.05/2)$.

Anonymized data and scripts supporting this report can be retrieved from our Open Science Framework page (<https://osf.io/yu7kq/>).

¹ The model used for word2vec can be downloaded here: <http://vectors.nlpl.eu/repository/>.

² For a visual example of how time series are compared by Dynamic Time Warping, see our supplemental figure <https://osf.io/dz9vx/>. This example from study 1, shows the vertical displacement of the left hand tip for three compared gestures that conveyed the concept "airplane".

2.1 Study 1

Study 1 utilizes the ‘NEMO-Lowlands iconic gesture dataset’ [21] for which 3D kinematic data (Microsoft Kinect V2, sampling at 30 Hz) was collected for 3715 gestures performed by 433 participants (children and adults) conveying 35 different concepts (organized within 5 themes containing 7 concepts each: e.g., animals, musical instruments). Participants were tasked with conveying a concept to a robot with a silent gesture, much like playing charades. The robot was tasked with recognizing the gesture via a kinematic comparison with a stored lexicon. If it could not recognize the gesture, the participant was asked to perform the gesture again and such trials were also included in the final gesture dataset. Importantly, participants were not instructed how to gesture, and creatively produced a silent gesture for the requested concepts³.

We computed the semantic distance for each pair of concepts using word2vec, ranging from a semantic dissimilarity or distance of 0 (minimum) to 1 (maximum). These semantic dissimilarity scores filled a symmetrical 35×35 semantic distance matrix \mathbf{D}^s (without diagonal values) containing comparisons between each concept c_i and concept c_j :

$$D_{i,j}^s = 1 - \text{cosine_similarity}(c_i, c_j), i \neq j$$

Gesture kinematic distance scores filled a similar 35×35 matrix, \mathbf{D}^g , with distances between all combinations of gestures belonging to concepts i and j , calculated using dynamic time warping:

$$D_{i,j}^g = \text{ave} \sum_{k_i, l_j}^{n_i, m_j} \text{ave} \sum_{o=1}^p \text{dtw}(t_{k_i o}, t_{l_j o}), i \neq j$$

Kinematic distances ($D_{i,j}$) were computed between all combinations of gestures k_i for concept i , and gestures l_j for concept j , except not for when $i = j$ (i.e., no diagonal values were computed). The computations were performed for all combinations of gesture set n_i and gesture set m_j , and then averaged. A dynamic time warping algorithm [‘dtw(query, referent)’] was used, where for each referent gesture k_i and each query gesture l_j a multivariate time series \mathbf{t} was submitted, containing the x, y, and z trajectories for key point o (e.g., $o =$ left wrist x, y, z). The computed distances were averaged over the total of $p = 5$ key points. We have previously observed that these body parts (as indexed by key points), left/right hand tip, left/right wrist, and head, are important for assessing the variance in silent gesture [22]. Note that each time series submitted to DTW was first z-scaled and centered, and time series were smoothed with a 3rd order Kolmogorov-Golai filter with a span of 2 frames (a type of Gaussian moving average filter).

Since we use an unconstrained version of DTW, computations can yield asymmetric results depending on which time series is set as the referent, so for each DTW distance calculation we computed the distance twice by interchanging the referent and query time series and then averaging, yielding a single distance score. Please see our OSF

³ Due to time constraints, participants only performed gestures for five randomly selected concepts. The repetition rate due to the robot’s failure to recognize the gesture was 79%.

page (<https://osf.io/39ck2/>) for the R code generating the kinematic matrix from the time series.

In sum, our analyses yielded a semantic distance matrix D^s and a similarly formatted kinematic distance matrix D^k containing information about semantic and kinematic (dis)similarity between each combination of 2 of the 35 concepts. This then allows us to assess whether semantic dissimilarity between concepts is related to the kinematic dissimilarity of the associated gestures. Figure 1 provides a geometric representation of the procedure’s logic.

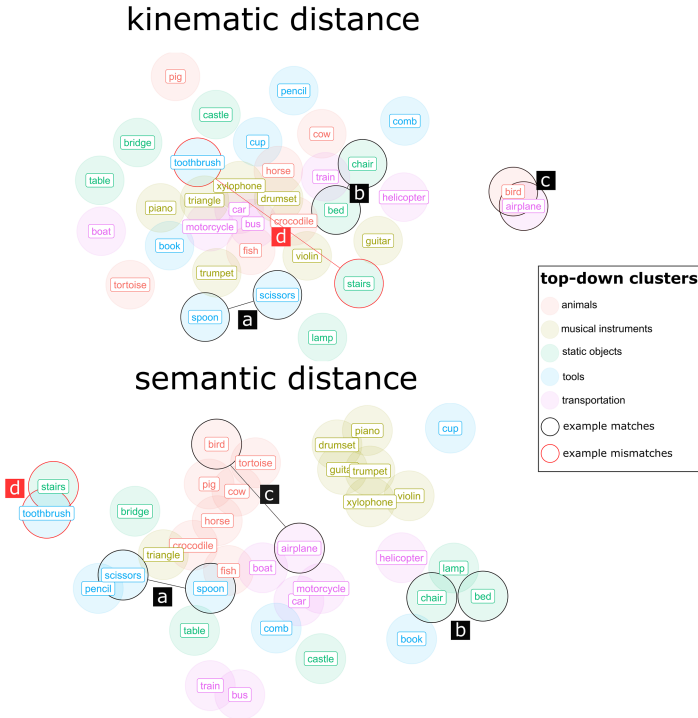


Fig. 1. Here the geometric/network representation is shown (using t-distributed stochastic neighbor embedding, for 2D projection through multidimensional scaling [28]) of the kinematic (above) and semantic distances between concepts conveyed by participants in the NEMO-Lowlands dataset. Examples of matches and a mismatch are highlighted, where matches (black boxes a-c) indicate that concepts that were kinematically alike were also semantically alike (e.g., spoon and scissors), and two red boxes (d) showing examples where concepts were kinematically dissimilar but semantically similar (e.g., stairs and toothbrush). Note that it could also be the other way around, such that there is high kinematic similarity but low semantic similarity (though we did not find this in the current dataset). (Color figure online)

2.2 Results Study 1

We performed mixed linear regression (see Fig. 2; analysis script: <https://osf.io/kvmfc/>) to assess whether semantic distances would scale with kinematic distances, with random intercepts for the concept that is used as reference (models with random slopes for the effect of semantic distance did not converge). Relative to a base model predicting the overall mean of kinematic distance, a model including semantic distance was reliably better in explaining variance, Chi-squared change (1) = 16.23, $p < .001$; Model coefficient semantic distance $b = 0.033$, $t(560) = 191.43$, $p < .001$, Cohen's $d = 0.34$.

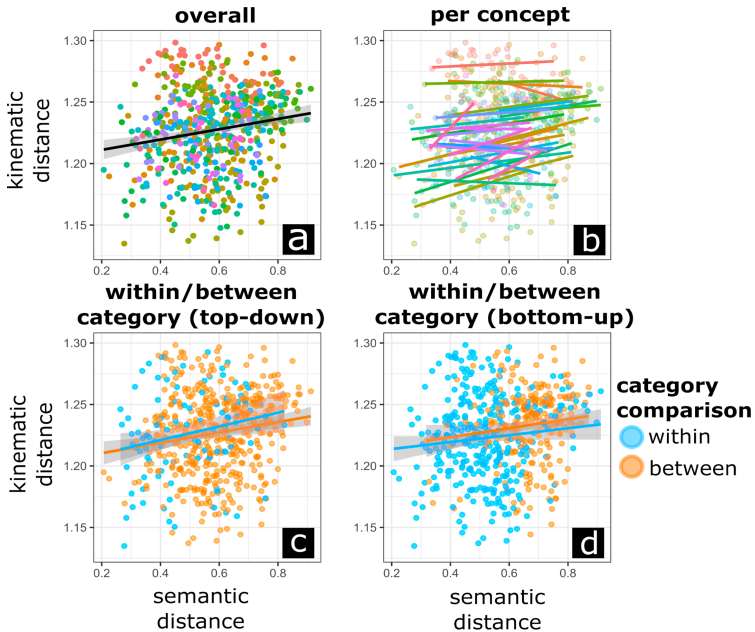


Fig. 2. Relations between semantic and kinematic distances are shown, overall slope and the simple correlation coefficient is given with colored points indicating the referent object (e.g., plane, bird) (panel a). Panel (b) shows separate slopes for each concept. Panel (c) shows different colors and slopes for the within top-down category (e.g., transportation-transportation) or between category comparisons (e.g., static object-transportation), and panel (d) shows different colors and slopes for within bottom-up category (e.g. cluster1-cluster1) and between category (e.g., cluster1-cluster2) comparisons. We can see that there is a positive relation between semantic and kinematic distance, which is globally sustained, such that within and between categories that positive relation persists. This indicates that gesture comparisons within a similar domain (either defined through some thematization by the researcher, or based on the structure of the data) are as likely to be related to semantic distance as when those comparisons are made across domains. Note further that it seems that semantic distances in panel (c) are lower for within category comparisons, suggesting that top-down categories are reflected in the semantic word2vec results (this aligns with Fig. 1 showing that categories tend to cluster in semantic space).

It is possible that this general weak but highly reliable relation between semantic vs. kinematic distance mainly relies on comparisons between concepts that are highly dissimilar, so that, say, the kinematic distance between two concepts that are within the same category (e.g., bus and train are in the category transportation) does not scale with semantic distance. To assess this, we compared the relation between kinematic vs. semantic distance for comparisons that are within a defined category versus between different categories. Firstly, we can use the top-down categories (e.g., transportation, musical instruments) that were used to group the stimulus set for the original study [21]. Secondly we used a bottom-up categorization approach, by performing *k*-means clustering analysis on the semantic distance matrices, where the optimal cluster amount was pre-determined by assessing the cluster amount with the highest average silhouette (i.e., silhouette method; yielding 2 clusters).

Further mixed regression modeling onto kinematic distance was performed by adding within/between category comparisons to the previous model containing semantic distances, as well as adding an interaction between semantic distance and within/between category. For the top-down category, neither a model adding within/between category as a predictor, Chi-squared change (1) = 0.0005, $p = .982$, nor a model with category x semantic distance interaction, Chi-squared change (1) = 0.113, $p = .737$, improved predictions. For the bottom-up category, adding within/between category as a main effect improved the model relative to a model with only semantic distance, Chi-squared change (1) = 8.50, $p = .004$. Adding an interaction did not further improve the model, Chi-squared change (1) = 0.17, $p = .674$. The statistically reliable model coefficients, indicated the main effect of semantic distance, $b = 0.020$, $t(559) = 166.29$, $p < .001$, Cohen's $d = 0.18$, as well as a main effect of category, $b_{\text{within vs. between}} = -0.006$, $t(559) = -2.92$, $p < .001$, Cohen's $d = -0.25$. The main effect of bottom-up category, indicates that when comparisons are made between concepts that are within a semantic cluster, those gestures are also more likely to have a lower kinematic distance. The lack of an interaction effect of category with semantic distance, indicates that the kinematic-semantic scaling effects holds locally (within categories) and globally (between categories), suggesting that there is no clear overarching category that drives the current effects. If this would be the case we would have found that the semantic-kinematic scaling relation would be absent for within category comparisons.

To conclude, we obtain evidence that silent gestures have a weak but reliable tendency to be more kinematically dissimilar if the concepts they are supposed to convey are also more semantically dissimilar.

3 Study 2

In study 2, there were 13 pairs, consisting of 26 participants (11 women and 15 men, $M_{\text{age}} = 22$ years, $\text{Range}_{\text{age}} = 18\text{--}32$ years). This is subset of the original data (20 pairs), as we only included data for which we also have some human gesture annotations for, which we could relate to our automatic processing. The participants were randomly grouped into 13 pairs (5 female dyads, 3 male dyads, and 5 mixed dyads) who performed a director-matcher task. The interlocutors took turns to describe and find images of novel 3D objects ('Fribbles' [24]). In each trial, a single Fribble was highlighted for the director,

and participants worked together so that the matcher could identify this object among a set of 16 Fribbles on their screen (note that the order in which Fribbles were presented was not the same for the director and matcher). Matchers indicated their selection by saying the corresponding position label out loud, and used a button box to move to the next trial. There were six consecutive rounds, consisting of 16 trials each (one for each Fribble). Participants switched director-matcher roles every trial. Participants were instructed to communicate in any way they wanted (i.e., there was no explicit instruction to gesture). Figure 3 provides an overview of the 16 Fribbles used and the setup of the experiment.

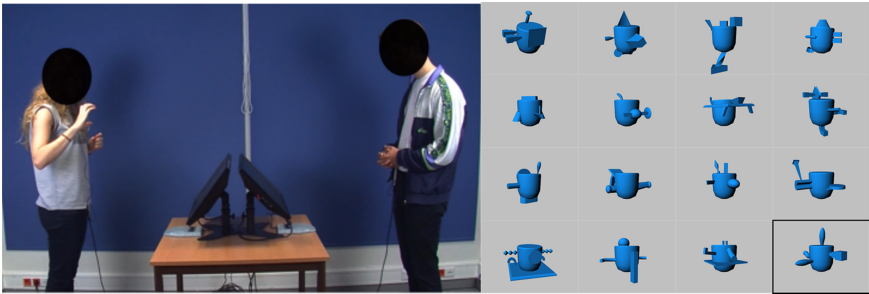


Fig. 3. This participant was explaining how this Fribble (the one with a black rectangle around it on the right) has “on the right side sort of a square tower”, producing a gesture that would be a member of the set of gestures she would produce for that Fribble.

During each trial we have information about which Fribble was the object to be communicated and thus all gestural kinematics that occurred in that trial are likely to be about that Fribble (henceforth target Fribble). Before and after the interaction, participants were individually asked to come up with a verbal label/name for each Fribble (1–3 words) that would enable their partner to identify the correct Fribble (henceforth ‘naming task’). In order to enable word2vec processing for these names, spelling errors were corrected and compounds not available in the word2vec corpus were split up (see <https://osf.io/x8bpq/> for further details on this cleaning procedure).

Similar to study 1, Kinect collected motion tracking data at 25 Hz, and traces were similarly smoothed with a Kolmogorov-Golai filter (span = 2, degree = 3).

Since we are now working with spontaneous, interactive data (where people move their body freely, though they are not constantly gesturing), we need an automatic way to detect potential gestures during the interactions. We used a custom-made automatic movement detection algorithm to identify potential iconic gestures, which was developed for this dataset (also see Fig. 4). This is a very simple rule-based approach, similar in nature to other gesture detectors [25], where we used the following rules:

1. A movement event is detected when the body part exceeds 15 cm per second speed (15 cm/s is a common movement start threshold, e.g., [26]).
2. If the movement event is next to another detected movement event within 250 ms, then they are merged as a single movement event. Note that for each gesture movement

two or multiple velocity peaks will often be observed: as the movement initiates, performs a stroke, potentially holds still, and detracts. The appropriate time interval for merging will treat these segments as a single event.

3. If a movement lasts less than 200 ms, it is ignored. This way very short movements were filtered out (but if there are many such short movements they will be merged as per rule 2 and treated as a relevant movement).
4. Gesture space is confined to movement above the person-specific -1 SD from the mean of vertical displacement. Participants in our study never to raise their hands to show their gestures to their interlocutor. This also prevents that button presses needed to move between trials were considered as gestures.

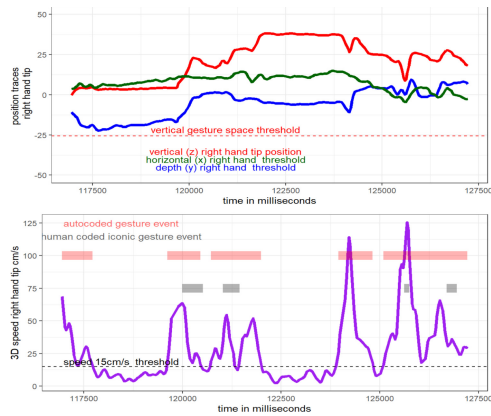


Fig. 4. Example automated gesture coding from time series. The upper panel shows for the right hand tip the three position traces (x = horizontal, y = depth, z = vertical), with the vertical axis representing the cm space (position traces are centered), and on the horizontal axis time in milliseconds. The vertical threshold line shows that whenever the z -trace is above this threshold, our autocoder will consider a movement as potentially relevant. In the lower panel, we have the 3D speed time series which are derived from the upper panel position traces. The vertical axis indicates speed in centimeters per second (cm/s). The autocoding event detections are shown in light red and the iconic gesture strokes as coded by a human annotator are shown in grey. The autocoder detects 5 gestures here, while the human coder detected 4 gesture strokes (note that they do not all overlap).

Note further, that for this current analysis we will only consider a subset of detected movements that were at least 500 ms in duration, as we ideally want to capture movements that are likely more complex and representational gestures, in contrast to gestures that are of a beat-like or very simple quality, which are known to take often less than 500 ms [29, 30]. Further we only consider right-handed movements, so as to ensure that differences in kinematics are not due to differences in hand used for gesturing, as well as for simplicity.

Note that the current automatic movement detection is a very crude and an imperfect way to identify communicative gestures, and rule-based approaches are known to have a relatively large number of false positives [31]. To verify the performance of our

algorithm, we compared its output to human-coded iconic gestures for this data; we test against human-coded iconic gestures rather than all gestures, as iconic gestures are the gestures that are of interest for the current analysis (rather than e.g., beat-like gestures). Iconic gestures were coded for a subset of the current data (i.e., for 8 out of the 16 Fribbles in the first two rounds of the interaction). Only the stroke phase was annotated, for the left and right hand separately. We found that the number of iconic gestures detected per participant by the human coder was positively related to the number of auto-coded gestures, $r = .60, p < .001$. In terms of overlap in time of human-coded and auto-coded gesture events there was 65.2% accuracy (true positive = 70%, false positive = 86%, true negative = 93%, false negative = 1%).

The total number of auto-detected gestures (henceforth gestures) that were produced was 1429, $M (SD, \min, \max) = 208.84 (75.35, 65, 306)$ gestures per participant (i.e., an average of 13 gestures per Fribble). The average time of a gesture was $M = 1368$ ms ($SD = 1558$ ms).

We used the same approach to construct semantic and kinematic matrices as in study 1, with some slight modifications. Semantic distances were computed for the names from the pre and post naming task separately, each matrix D_{pre}^s and D_{post}^s containing information about semantic distances between names of Fribble i to j (but not for identical Fribbles, i.e., $i \neq j$). There were 16 different Fribbles, yielding 16×16 distance matrices for each pair. These distance matrices were thus computed between participants in a dyad. See Fig. 5 for an explanation.

For the kinematics (see <https://osf.io/a6veq/> for script) we only submit right-hand related key points, with additional more fine-grained information about hand posture. Therefore, we selected x, y, z traces for key points of the hand tip and thumb, and the Euclidean distance over time between hand-tip and thumb. Again we z-normalized and centered the movement traces before submitting to DTW. The distance matrices for kinematics were also computed between participants in a dyad (as the semantic distance matrices). Further note, that when there were no gestures detected for a particular Fribble i , then no kinematic distance scores could be computed for any comparison that involved Fribble i , and the kinematic distance matrix would contain a missing value for that comparison.

Thus we will analyze the relation between the semantic distances and the kinematic distances between participants, both for naming in the pre as well as the post-test.

3.1 Results Study 2

We performed mixed regression analysis (analysis script: <https://osf.io/a657t/>), whereby we predict kinematic distances based on semantic distance of pre- and post-naming (in two different analyses). The names and kinematics were repeatedly generated per pair and between Fribbles, and therefore we added Pair nested in Fribble comparison (e.g., Fribble comparison 1:2) as random intercept. See Fig. 6 for the graphical results.

Between-participant kinematic distances were not better predicted by pre-interaction naming semantic distances, as compared to a base model predicting the overall mean, Chi-squared change (1) = 0.06, $p = .812$. However, post-interaction naming semantic distances as a predictor improved predictions as compared to a base model, Chi-squared change (1) = 6.32, $p = .012$. The resulting model showed that post-naming semantic

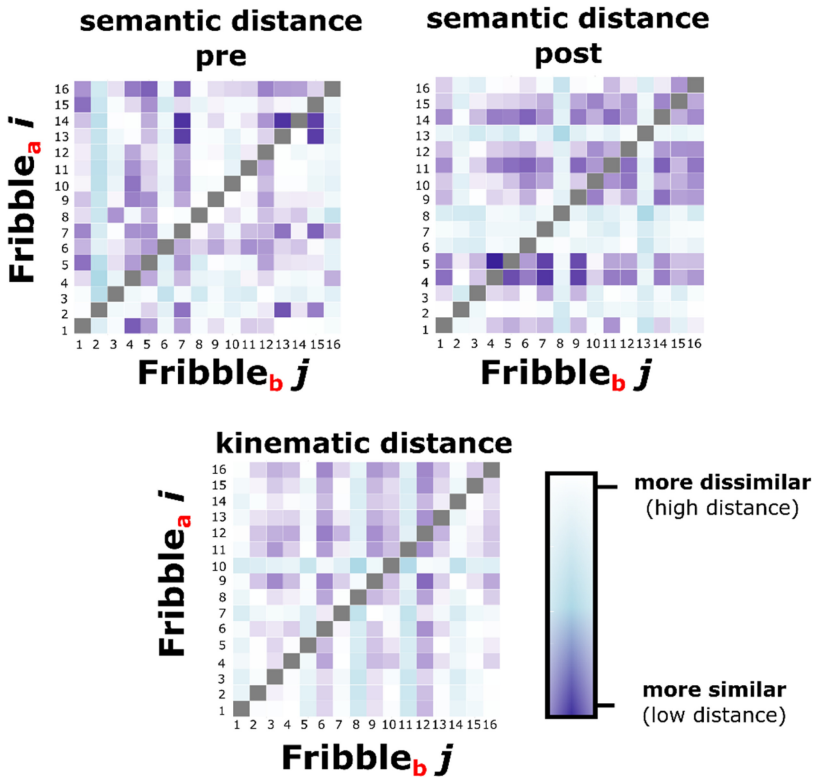


Fig. 5. Example of distance matrix data is shown as colored maps with lower distance scores in darker blue, with 16 rows and columns for each matrix, as there were 16 different Fribbles in total. Each comparison assesses for Fribble *i* for participant *a* (Fribble_a *i*), versus Fribble *j* for participant *b* (Fribble_b *j*) within a dyad the distances between the naming/kinematics between participants for each comparison between two Fribbles. This means that the upper and lower triangles of the matrix are asymmetrical and provide meaningful information regarding the distances in naming/kinematics between interlocutors within the dyad. For the analysis, similar to study 1, we only assess the relation between the off-diagonal cells of the pre and post naming distances with that of the off-diagonal of kinematic distances. Diagonals are in principle computable, and this would be measuring alignment between participants, but we are interested in the relation between gestures that convey different concepts and their semantic-kinematic relatedness.

distances reliably predicted kinematic distances between participants, $b = 0.045$, $t(2583) = 2.15$, $p = .012$, Cohen's $d = .10$. This means that Fribbles that had semantically more similar names produced after interaction by the interlocutors also were more likely to elicit gestures with similar gesture kinematics between interlocutors.

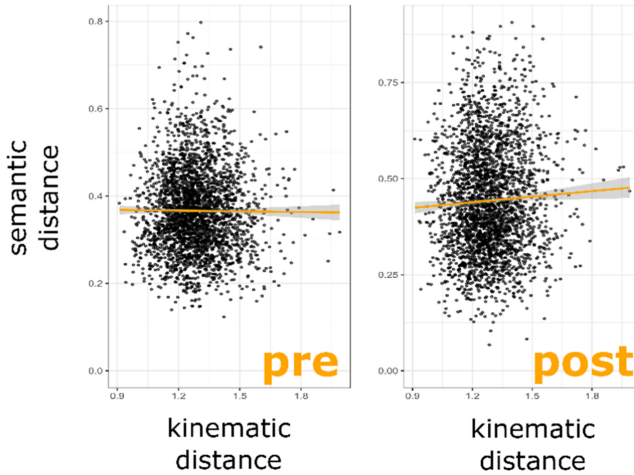


Fig. 6. Scatter plot for the relation between semantic distance between names of Fribble i versus j (pre- and post-interaction) and the kinematic distance between the set of gestures produced for Fribble i versus the set of gestures produced for Fribble j . This means that when a participant “a” showed a higher dissimilarity with “b” on the post naming for Fribble i_a versus j_b , then they also tended to have a more dissimilar set of gestures for Fribble i_a versus j_b . It can be seen that the pre-interaction names do not show any positive kinematic-semantic scaling relation, while the post-interaction names are related to the kinematic distances computed from gestures produced during the interaction.

4 Discussion

In this study we assessed whether gestures that are more similar in kinematics, are likely to convey more similar meanings. We provide evidence that there is indeed a weak statistical dependency between gestures’ form (i.e., kinematics) and their (putative) meanings. We show this form-meaning relation in two studies, which were highly divergent in design. In a charades-style study 1, participants interacting with a robot were explicitly instructed to convey one of 35 concepts using silent gestures (i.e., without any speech). In a director-matcher style study 2, participants were interacting in dyads, producing spontaneous co-speech gestures when trying to refer to novel objects. Participants were asked to verbally name these novel objects before and after interacting with their partner. In both studies we obtain that the difference in the gestures’ putative referential content (either the concepts to be conveyed, or the post-interaction naming of the objects) scales with the dissimilarity between the form of the gestures that certainly targeted (study 1) or were likely to target (study 2) that referential content. Thus in both silent gestures and gestures produced with speech, the kinematic space seems to co-vary with the putative semantic space.

There are some crucial caveats to the current report that need to be mentioned. Firstly, we should not confuse the semantics of a gesture with our measurement of the semantics, using word2vec distance calculations of the instructed (study 1) or post-interaction elicited (study 2) conceptualizations of the referential targets. Thus we should remind ourselves that when we say that two gestures’ meanings are similar, we should actually

say that the concepts that those gestures putatively signal show up in similar contexts in a corpus of Dutch webtexts (i.e., the word2vec model we used; [26]). Furthermore, there are also other measurements for semantic distance computations that possibly yield different results, e.g., [32], and it is an interesting avenue for future research to see how gesture kinematics relates to these different semantic distance quantifications [33]. This goes the other way too, such that there are different ways to compute the kinematic distances [e.g., 19, 34] for different gesture-relevant motion variables [e.g., 35] and more research is needed to benchmark different approaches for understanding semantic properties of communicative gesture kinematics.

Additionally, the way the putatively conveyed concept is determined in study 1 and 2 is dramatically different. In study 1 it is more clear and defined from the outset, but in study 2 participants are asked to produce a name for novel objects, such that their partner would be able to identify the object. This naming was performed before and after interacting about those objects with their partner. The kinematic space was only related to the names after the interaction, and these names were not pre-given but likely created through communicative interaction. Thus while we can say that in study 1 gestures that convey more similar concepts are also more likely to be more kinematically similar, for study 2 we must reiterate that kinematically similar gestures for two objects x and y produced by two interlocutors (in interaction), forges a context for those persons to name these two objects similarly. Thus it seems that gestures between participants can constrain—or are correlated to another process that constrains (e.g., speech)—the between-subject semantic space that is constructed through the interaction. We do not find this to be the case the other way around, as the semantic space verbally constructed before interaction (i.e., based on pre-interaction naming) did not relate to the kinematic space constructed gesturally.

It is clear that more research is needed to understand these effects vis-à-vis the semantic and kinematic relation of gestures in these highly different contexts of study 1 and 2. We plan more follow-up analyses taking into account semantic content of gestures' co-occurrent speech, as well as arguably more objective visual differences between the referential targets themselves (e.g., are Fribble objects that look alike also gestured about more similarly?). However, for the current report we simply need to appreciate the now promising possibility that gesture kinematic (dis-)similarity spaces are informative about their semantic relatedness. Implications are easily derivable from this finding alone.

For example, consider a humanoid whose job it is to recognize a gesture's meaning based on kinematics as to respond appropriately (as was the setup for study 1, [21]). The current results suggest that information about an undetermined gesture's meaning can be derived by comparing it to a stored lexicon of gesture kinematics of which the semantic content is determined. Though certainly no definitive meaning can be derived, the current statistical relation offers promise for acquiring some initial semantic gist of a semantically undefined gesture based on kinematic similarities computed against a library of representative set of gesture kinematics. The crucial importance of the current findings is that such a gesture kinematic lexicon does not need to contain a semantically similar or identical gesture to provide some minimal semantic gist about the semantically undefined gesture. It merely needs a computation of form similarity against its database of representative gesture kinematics. This also means that a humanoid without any such

lexicon, with enough training, can at least derive some information about which gestures are more likely to be semantically related. A humanoid can build its kinematic space from the bottom up, by detecting gestures in interaction, construct a kinematic similarity space over time, and infer from the distance matrices which gestures are likely to be semantically related (given the assumption that kinematic space and semantic space tend to align). Moreover, the humanoid's own gesture generation process may be tailored such that there is some weak dependency between the kinematics of gestures that are related in content, thus optimizing its gesture behavior to cohere in a similar way as human gesture does [36–38]. The current findings thus provide an exciting proof-of-concept that continuous communicative bodily movements that co-vary in kinematic structure, also co-vary in meaning. This can be exploited by the field of machine learning which is known to productively leverage weak statistical dependencies to gauge semantic properties of communicative tokens (e.g., word2vec).

Note further that the principle of distributional semantics is said to provide an important bootstrapping mechanism for acquiring language in human infants (and language learners in general), as statistical dependencies yield some information about the possible meaning of an unknown word given its contextual or form vicinity to other words for which the meaning is more determined [26, 39]. Word learning is facilitated in this way, as language learners do not need explicit training on the meaning of each and every word, but can exploit statistical dependencies that structure the language [40, 41]. Here we show a statistical dependency that is similar in spirit, but for continuous communicative movements: the similarity between the putative semantic content of one gesture and that of another, can be predicted to some extent based on their movement similarity alone. It thereby offers a promising indication that gestures' contents too are to some extent more easily learnable based on their covariance in form. It opens up the possibility that gestures, similar to other forms of human communication, are not simply one-shot communicative patterns, but to some statistical extent constellated forms of expressions with language-like systematic properties amenable to geometric/network analysis performed on the level of interrelationships between communicative tokens [13, 29, 42].

Additionally, the potential of multimodal learning should be underlined here, as co-speech gesture kinematic interrelationships are informative about semantic space and therefore also likely co-informative about co-occurrent speech which you may not know the meaning of. Thus when learning a new language, gestures can come to reduce the potential meaning space of the entire communicative expression (i.e., including speech), reducing the complexity of word learning too. This mechanism can be related to another putative function of iconicity in gestures as a powerful starting point in acquiring language [43], as kinematics are informative about a referent given the kinematics structures by association through form-meaning resemblance (e.g., a digging movement may signal the referent of the word DIGGING in its close resemblance to the actual action of digging). However, this particular way of constraining semantics via iconicity necessitates some basic mapping on the part of the observer, so as to complete the iconic reference between form and meaning. The current kinematic-semantic scaling provides in potential a more indirect or bottom up statistical route to reduce the semantic space to likely meanings, namely by recognizing similarities of a gesture's form with other forms previously encountered, one can reduce the meaning space if the kinematic space

and semantic space tend to be co-varying. Thus the current geometric relations between gesture kinematic and semantic space are a possible statistical route for constraining potential meanings from detecting covariances between form alone, at least in artificial agents, but potentially this is exploited by human infants and/or second-language learners too.

Though promising, the statistical dependency is currently underspecified in terms of how such dependencies emerge in the human ecology of gesture. It remains unclear which particular kinematic features tend to co-vary with semantic content. So we are not sure at what level of similarity or analogy gesture kinematics relate as they do semantically [44]. It is further not clear whether the semantic content co-varies with kinematics because the gestures are part of some kind of overarching movement type (e.g., static handshape, continuous movement, etc.) or mode of representation (acting, representing, drawing or personification; [11]) which may co-vary with semantic categories. Indeed, in previous research it has been shown that e.g., gestures representing manipulable objects are most likely to have ‘acting’ as mode of representation, while gestures depicting animals are more likely to recruit ‘personification’, as observed by human annotators [3]. We tried to assess in study 1 whether it is indeed the case that the reported effects might be due to local covariance of different gesture classes, leading to global kinematic-semantic differences between classes. Namely, if gestures are kinematically grouped by an overarching category, then within that class there should be no relation between gesture kinematic and semantic similarity. The results however, indicate that semantic-kinematic distance persisted both for comparisons within and between gesture classes, irrespective of whether we construct such classes based on human-defined themes, or empirically based kinematic cluster assignment. We hope the current contribution invites further network-topological study [13, 45] of the current geometrical scaling of gesture semantic and kinematic spaces so as to find the right level of granularity at which these spaces co-vary.

To conclude, the current results suggest a persistent scaling relation between gesture form and meaning distributions. We look forward to researching this more deeply from a cognitive science perspective, but we hope that the HCI as well as machine learning community could one day leverage covariances that we have identified between kinematic and semantic spaces, in the employment and development of an automatic detection of a gesture’s meaning via principles of distributional semantics.

Acknowledgements. For study 2, we would like to thank Mark Dingemans for his contributions in the CABB project to assess optimality of different word2vec models. For study 1, we would like to thank James Trujillo for his contributions to setting up the Kinect data collection. Study 2 came about in the context of a multidisciplinary research project within the Language in Interaction consortium, called Communicative Alignment in Brain and Behaviour (CABB). We wish to make explicit that the work has been shaped by contributions of CABB team members, especially (alphabetical order): Mark Blokpoel, Mark Dingemans, Lotte Eijk, Iris van Rooij. The authors remain solely responsible for the contents of the paper. This work was supported by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium and is further supported by the Donders Fellowship awarded to Wim Pouw and Asli Ozyurek.

References

1. Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., Kirby, S.: Evolving artificial sign languages in the lab: from improvised gesture to systematic sign. *Cognition* **192**, (2019). <https://doi.org/10.1016/j.cognition.2019.05.001>
2. Ortega, G., Özyürek, A.: Types of iconicity and combinatorial strategies distinguish semantic categories in silent gesture across cultures. *Lan. Cogn.* **12**, 84–113 (2020). <https://doi.org/10.1017/langcog.2019.28>
3. Ortega, G., Özyürek, A.: Systematic mappings between semantic categories and types of iconic representations in the manual modality: a normed database of silent gesture. *Behav. Res.* **52**, 51–67 (2020). <https://doi.org/10.3758/s13428-019-01204-6>
4. Gerwing, J., Bavelas, J.: Linguistic influences on gesture's form. *Gesture* **4**, 157–195 (2004). <https://doi.org/10.1075/gest.4.2.04ger>
5. Rasenberg, M., Özyürek, A., Dingemanse, M.: Alignment in multimodal interaction: an integrative framework. *Cogn. Sci.* **44**, (2020). <https://doi.org/10.1111/cogs.12911>
6. Bernstein, N.: *The Co-ordination and Regulations of Movements*. Pergamon Press, Oxford (1967)
7. McNeill, D.: *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago (1992)
8. Kendon, A.: *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge (2004)
9. Kolorova, Z.: *Lexikon der bulgarischen Alltagsgesten* (2011)
10. Gentner, D., Brem, S.K.: Is snow really like a shovel? Distinguishing similarity from thematic relatedness. In: Hahn, M., Stoness, S.C. (eds.) *Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society*, pp. 179–184. Lawrence Erlbaum Associates, Mahwa (1999)
11. Müller, C.: Gestural modes of representation as techniques of depiction. In: Müller, C. (ed.) *Body–Language–Communication: An International Handbook on Multimodality in Human Interaction*, pp. 1687–1701. De Gruyter Mouton, Berlin (2013)
12. Streeck, J.: Depicting by gesture. *Gesture* **8**, 285–301 (2008). <https://doi.org/10.1075/gest.8.3.02str>
13. Karuza, E.A., Thompson-Schill, S.L., Bassett, D.S.: Local patterns to global architectures: influences of network topology on human learning. *Trends Cogn. Sci.* **20**, 629–640 (2016). <https://doi.org/10.1016/j.tics.2016.06.003>
14. Gleitman, L.R.: Verbs of a feather flock together II: the child's discovery of words and their meanings. In: Nevin, B.E. (ed.) *The Legacy of Zellig Harris: Language and Information Into the 21st Century*, pp. 209–229 (2002)
15. Fowler, C.A.: Embodied, embedded language use. *Ecol. Psychol.* **22**, 286 (2010). <https://doi.org/10.1080/10407413.2010.517115>
16. Pouw, W., Dixon, J.A.: Gesture networks: Introducing dynamic time warping and network analysis for the kinematic study of gesture ensembles. *Discourse Processes* **57**, 301–319 (2019). <https://doi.org/10.1080/0163853X.2019.1678967>
17. Giorgino, T.: Computing and visualizing dynamic time warping alignments in R: the dtw package. *J. Stat. Softw.* **31** (2009). <https://doi.org/10.18637/jss.v031.i07>
18. Muller, M.: *Information Retrieval for Music and Motion*. Springer, Heidelberg (2007). <https://doi.org/10.1007/978-3-540-74048-3>
19. Beecks, C., et al.: Efficient query processing in 3D motion capture gesture databases. *Int. J. Semant. Comput.* **10**, 5–25 (2016). <https://doi.org/10.1142/S1793351X16400018>
20. Pouw, W., Dingemanse, M., Motamedi, Y., Özyürek, A.: A systematic investigation of gesture kinematics in evolving manual languages in the lab. *OSF Preprints* (2020). <https://doi.org/10.31219/osf.io/heu24>

21. de Wit, J., Krahmer, E., Vogt, P.: Introducing the NEMO-Lowlands iconic gesture dataset, collected through a gameful human–robot interaction. *Behav. Res.* (2020). <https://doi.org/10.3758/s13428-020-01487-0>
22. Müller, C.: Gesture and sign: cataclysmic break or dynamic relations? *Front. Psychol.* **9** (2018). <https://doi.org/10.3389/fpsyg.2018.01651>
23. Rasenberg, M., Dingemanse, M., Özyürek, A.: Lexical and gestural alignment in interaction and the emergence of novel shared symbols. In: Ravignani, A., et al. (eds.) *Evolang13*, pp. 356–358 (2020)
24. Barry, T.J., Griffith, J.W., De Rossi, S., Hermans, D.: Meet the Fribbles: novel stimuli for use within behavioural research. *Front. Psychol.* **5** (2014). <https://doi.org/10.3389/fpsyg.2014.00103>
25. Mandera, P., Keuleers, E., Brysbaert, M.: Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. *J. Mem. Lan.* **92**, 57–78 (2017). <https://doi.org/10.1016/j.jml.2016.04.001>
26. Zeman, D., et al.: CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, pp. 1–19. Association for Computational Linguistics (2017). <https://doi.org/10.18653/v1/K17-3001>
27. Silva, D.F., Batista, G.A.E.P.A., Keogh, E.: On the effect of endpoints on dynamic time warping. Presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco (2016)
28. Donaldson, J.: tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE) (2016)
29. Pouw, W., Dixon, J.A.: Entrainment and modulation of gesture–speech synchrony under delayed auditory feedback. *Cogn. Sci.* **43**, (2019). <https://doi.org/10.1111/cogs.12721>
30. Pouw, W., Dixon, J.A.: Quantifying gesture–speech synchrony. In: *Proceedings of the 6th meeting of Gesture and Speech in Interaction*, pp. 68–74. Universitaetsbibliothek Paderborn, Paderborn (2019). <https://doi.org/10.17619/UNIPB/1-812>
31. Ripperda, J., Drijvers, L., Holler, J.: Speeding up the detection of non-iconic and iconic gestures (SPUDNIG): a toolkit for the automatic detection of hand movements and gestures in video data. *Behav. Res.* **52**, 1783–1794 (2020). <https://doi.org/10.3758/s13428-020-01350-2>
32. Kenett, Y.N., Levi, E., Anaki, D., Faust, M.: The semantic distance task: quantifying semantic distance with semantic network path length. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 1470–1489 (2017). <https://doi.org/10.1037/xlm0000391>
33. Kumar, A.A., Balota, D.A., Steyvers, M.: Distant connectivity and multiple-step priming in large-scale semantic networks. *J. Exp. Psychol. Learn. Mem. Cogn.* **46**, 2261–2276 (2020). <https://doi.org/10.1037/xlm0000793>
34. Beecks, C., et al.: Spatiotemporal similarity search in 3D motion capture gesture streams. In: Claramunt, C., Schneider, M., Wong, R.C.-W., Xiong, L., Loh, W.-K., Shahabi, C., Li, K.-J. (eds.) *SSTD 2015. LNCS*, vol. 9239, pp. 355–372. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22363-6_19
35. Trujillo, J.P., Vaitonyte, J., Simanova, I., Özyürek, A.: Toward the markerless and automatic analysis of kinematic features: a toolkit for gesture and movement research. *Behav Res.* **51**, 769–777 (2019). <https://doi.org/10.3758/s13428-018-1086-8>
36. Hua, M., Shi, F., Nan, Y., Wang, K., Chen, H., Lian, S.: Towards more realistic human-robot conversation: a Seq2Seq-based body gesture interaction system. [arXiv:1905.01641](https://arxiv.org/abs/1905.01641) [cs] (2019)

37. Alexanderson, S., Székely, É., Henter, G.E., Kucherenko, T., Beskow, J.: Generating coherent spontaneous speech and gesture from text. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, pp. 1–3 (2020). <https://doi.org/10.1145/3383652.3423874>
38. Wu, B., Liu, C., Ishi, C.T., Ishiguro, H.: Modeling the conditional distribution of co-speech upper body gesture jointly using conditional-GAN and unrolled-GAN. *Electronics* **10**, 228 (2021). <https://doi.org/10.3390/electronics10030228>
39. Romberg, A.R., Saffran, J.R.: Statistical learning and language acquisition. *WIREs Cogn. Sci.* **1**, 906–914 (2010). <https://doi.org/10.1002/wcs.78>
40. Saffran, J.R., Aslin, R.N., Newport, E.L.: Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996). <https://doi.org/10.1126/science.274.5294.1926>
41. Steyvers, M., Tenenbaum, J.B.: The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn. Sci.* **29**, 41–78 (2005). https://doi.org/10.1207/s15516709cog2901_3
42. Goldstein, R., Vitevitch, M.S.: The influence of clustering coefficient on word-learning: how groups of similar sounding words facilitate acquisition. *Front. Psychol.* **5** (2014). <https://doi.org/10.3389/fpsyg.2014.01307>
43. Nielsen, A.K., Dingemans, M.: Iconicity in word learning and beyond: a critical review. *Lang. Speech*, 0023830920914339 (2020). <https://doi.org/10.1177/0023830920914339>
44. Forbus, K.D., Ferguson, R.W., Lovett, A., Gentner, D.: Extending SME to handle large-scale cognitive modeling. *Cogn. Sci.* **41**, 1152–1201 (2017). <https://doi.org/10.1111/cogs.12377>
45. Siew, C.S.Q., Wulff, D.U., Beckage, N.M., Kenett, Y.N.: Cognitive network science: a review of research on cognition through the lens of network representations, processes, and dynamics. <https://www.hindawi.com/journals/complexity/2019/2108423/>. <https://doi.org/10.1155/2019/2108423>. Accessed 29 Jan 2021