Gestural Linguistic Context Vectors Encode Gesture Meaning

Carl Vogel, Maria Koutsombogera, Anaïs Claire Murat, Zohreh Khosrobeigi and Xiaona Ma

Computational Linguistics Group, Trinity College Dublin, College Green, Dublin 2, Ireland

vogel@cs.tcd.ie, {koutsomm,murata,khosrobz,max4}@tcd.ie

Abstract

Linguistic context vectors are adapted for measuring the linguistic contexts that accompany gestures and comparable co-linguistic behaviours. Focusing on gestural semiotic types, it is demonstrated that gestural linguistic context vectors carry information associated with gesture. It is suggested that these may be used to approximate gesture meaning in a similar manner to the approximation of word meaning by context vectors.

Index Terms: gesture interpretation, semiotic types, computational paralinguistics, context vectors

1. Introduction

Distributional approaches to natural language semantics, such as informed by context vectors, have made impressive progress in capturing salient aspects of word meaning. We propose a means of associating linguistic context vectors with co-occurring gestures, and demonstrate that on relevant measures, gestures display internal homogeneity and comparative distinctiveness. We focus here on the semiotic types of gestures; however, our next steps are in applying the method to corpora for which hand-shapes are recorded, irrespective of annotations of semiotic types. Our analysis is anchored in the manual annotations of semiotic types gestures as used in fourteen dialogues of a multi-modal corpus and the words that are used in the company of those gestures. We conclude from the patterns reported that gestural linguistic context vectors encode linguistic content, approximating gesture meaning in the same way that traditional context vectors approximate word meaning.

2. Background

It is tempting to think of gestures as linked to individual words in the utterances that accompany them – "lexical affiliates". However, it has been noted that this is problematic (see e.g., Kirchhof, 2011), even if, for certain gestures, affiliates may be located (Zhang and Kender, 2012; Hughes-Berheim et al., 2020). Apart from the problem of identifying which single word is at stake, given the loose temporal coupling between words and gestures, more than one accompanying word may be salient to a gesture (Han et al., 2017). One might therefore conclude that there is merit in considering all the words that overlap with gestures, as well as words that precede and follow.

The work of Eisenstein and Davis (2007) involves a similar intuition to the one we explore (i.e., quantifying linguistic context of a gesture conveys systematic information about gesture meaning). They consider gesture strokes and entire gesture durations, use bag of word and bag of part of speech n-gram models for classification of accompanying gesture semiotic types as one of three categories – deictic, iconic, or other. With trigram models, they obtained 65.9% accuracy, greater than for unigram models (55.1%). This work provides evidence of systematicity in the language that accompanies these gesture types.

Word vectors are frequently used in computational linguistics. Given an ordering of lexical types, such as rank frequency, as the dimensions of a vector, individual **texts** are measured in relation to the appropriate relative frequency of each word (or n-gram), recorded at the relevant dimension of the vector. Word vectors provide a direct means of encoding bag of word (n-gram) models. Such a vector is illustrated in (1), where w^i represents the count of items of the ith most frequent type. For many purposes, the value of n is less than the total number of types in the corpus. For some purposes, such as in a χ^2 contingency table test of word counts in distinct categories of texts, it is useful to know the total count of the complement category to the enumerated types of interest – call this $\overline{W}$ – the vector illustrated in (2) includes this complement category count. Where it is important, for example, because of varying text lengths, the counts may be relativized to the total number of tokens witnessed (Σ), resulting in vectors of relative frequency, as in (3).

$$\langle w^1, \ldots, w^n \rangle \tag{1}$$

$$\langle w^1, \ldots, w^n, \overline{W} \rangle \tag{2}$$

$$\Sigma^{-1} \times \langle w^1, \ldots, w^n, \overline{W} \rangle \tag{3}$$

Context vectors utilize the notion of word vectors, providing a means of measuring the contexts of use for each word, essentially by noting word vectors for each context position before and after the occurrences of a word. The context vector for a **word**, given a corpus, is normally constructed relative to a number m of context positions before the word and after the word, tracking the relative frequencies of the n most frequent items in the corpus as they occur in each of the context positions. Rank of frequency provides a natural order of word types held constant for each of the context positions. The structure of a context vector for a target word, noting the relative frequency of the n most frequent lexical types in each of the m contextual positions before and after the target, is depicted in (4). In the work we report here, we include among the vector positions the relative frequency of the complement lexical category (as in 5).

$$\Sigma^{-1} \times \langle w^1_{-m}, \ldots, w^n_{-m}, \ldots, w^1_{+m}, \ldots, w^n_{+m} \rangle \tag{4}$$

$$\langle \frac{w^1_{-m}}{\Sigma}, \ldots, \frac{w^n_{-m}}{\Sigma}, \frac{\overline{W_{-m}}}{\Sigma}, \ldots, \frac{w^1_{+m}}{\Sigma}, \ldots, \frac{w^n_{+m}}{\Sigma}, \frac{\overline{W_{+m}}}{\Sigma}, \rangle \tag{5}$$

By construction, context vectors have positive relative frequencies, and therefore the angle between context vectors *cosine* measures similarity on the interval [0,1] – the cosine of 0° is 1; the cosine of 90° is 0. Euclidean distance may also be used to calculate vector similarity. It has been established that with distributional data (positive examples) alone in a large corpus (c. 35m word tokens), clustering based on vectors of word co-occurrence counts is sufficient to identify similarity-based groupings that correlate strongly with syntactic categories

(Finch and Chater, 1992). In that work, $m = 2$ context positions before and after each target word were adopted, and the top $n = 150$ items were counted in each of the context positions. Clustering over the resulting vectors of length $2 \times 2 \times 150$, given the scale of the corpus, resulted in determiners clustering together, names of months clustering together, and so on. Thus, linguistic context vectors enable generalization of the contexts in which target words occur, and this in turn encodes aspects of syntactic categorization of words and aspects of word meaning.

3. Proposal

We adapt the idea of linguistic context vectors to **gesture**, using gestures as the target items, and explore the relative frequency of words in the contexts of gestures, given fixed context positions.[1] In relation to the use of context vectors for word meaning, the notion of left and right contexts requires generalization. With gestures, one is interested in the words that co-occur with the gesture, as well as the left and right contexts, but with a potential asymmetry not typically addressed for target word context vectors – given the sense that gestures tend to precede the linguistic content with which they are associated (Schegloff, 1985; Ferré, 2010), in this paper, we report on experiments that examine words that co-occur with gesture and words in right context positions. The fact that it is meaningful to consider words that co-occur with a target gesture and that the right context of a gesture is arguably more important than the left context are two generalizations required in approaching gesture with linguistic context vectors. In considering the n most frequent items in a corpus and their counts in each of the context positions, it is quite likely in a small corpus and small values of n that for some gestures, none of the most frequent items will be witnessed. Thus, it is helpful to add a vector position (as in (5)) that is used to count the total number of other tokens outside the n most frequent that appear in the context, or to binarize this count, setting it to 1 if it is non-zero.[2]

If the dimensionality of gestural linguistic context vectors were the same as for word linguistic context vectors, it would be possible to directly identify the target words most similar in distribution to gestures. However, the fact that gestures may be accompanied by simultaneous words while words may not (at least not by words of the same speaker) thwarts such alignment. On the other hand, it is possible to compute the similarity between linguistic context vectors that accompany gestures and comparable behaviours that occur in channels parallel to words – like gaze, head movement, etc. It also remains a possibility to identify an appropriate transformation that takes word linguistic context vectors and gestural linguistic vectors into a comparable dimensionality of feature space. In the meantime, evaluation is conducted with reference to systematicity within the categories of gestural linguistic context vectors computed.

The proposal is evaluated by comparing gestural linguistic context vectors. It is demonstrated that they encode aspects of gesture meaning if systematic differences and similarities emerge among all of the pairwise context vector representations for those comparisons that involve cognate elements. For example, if cluster analysis reveals that the context vectors for individual gesture types cluster together and away from other types, then this is evidence that the gestural linguistic context vectors encode aspects of gesture meaning. The extent to which such group together is indicative of meaning overlap. Similarly, evidence that the gestural linguistic context vectors are more similar for those who have spoken with each other than for those who have not supports the notion that the interlocutors natural mirroring of each other is manifest not just in physical movements but in the linguistic accompaniment of motions.

4. Method

4.1. Data

We use the data of the MULTISIMO corpus (Koutsombogera and Vogel, 2018). The corpus involves groups of three parties – two direct participants in a game and a moderator – communicating for several minutes each to achieve the cooperative game goal. Participants are each in only one such session, but moderators, more. The game in each session is modelled on *Family Feud*, in that with respect to three different questions posed by the moderator, the participants must first suggest and agree on what answers a separate group of 100 people might have suggested as answers, and in the same rank order (for example, "What is something one might *cut*?"). As each session involved the same questions, it is unsurprising that certain nouns appear among the top 25 most frequently used words (see Table 2).

The dialogues have several sorts of annotations, including the semiotic type of gestures (Beat, Iconic, Deictic, and Symbolic[3]) noted by McNeill (1992), following Peirce (1934), along with the onset and completion time of each gesture. An additional label, "N/A" (not-applicable) is used for visible hand movements lacking intrinsic relationship to the accompanying speech, namely self-adaptors (e.g. rubbing the face, fidgeting, stroking the other hand) or object-adaptors (i.e. participants playing with part of their clothing) (Ekman and Friesen, 1969). The full available data set contains 18 dialogues, but we analyze a 14 dialogue subset for which the temporal onset and offset of each word uttered by players is identified (Khosrobeigi et al., 2022). With these data annotations, it is possible to identify words that occur during gestures and between gestures, noting which speaker produced each. The 14 dialogues contain 12647 tokens of 1083 lexical types spoken by participants. Table 1 shows the distribution of participant gesture types.

Khosrobeigi et al. (2022) used the data of word and gesture time alignments to examine temporal relations between gestures according to semiotic type and syntactic categories of words and silences. The temporal relations considered were a subset of Allen relations (Allen, 1983) between durations. The most frequently attested were: gestures spanning the entirety of one or more words, gestures with onsets before words where the gesture completes during the word, and words with onsets before gestures where the word completes during the gesture. Contrary to what one might anticipate, the co-occurrence of deictics with nouns or iconics with verbs was not statistically significant. A purpose of the approach here is to examine a broader notion of linguistic context that might accompany gesture than is captured by temporal Allen relations with respect to the nearest word.

[1] An alternative approach would remain within the modality of gesture, and examine the gestures in the left and right context of a target gesture, rather than the words in the left and right context. Using an approach that addresses bodily shapes, as recorded using features highlighted by McNeill (2005, pp. 273–275) or less detailed MUMIN annotation scheme (Allwood et al., 2007), derived from McNeill's, rather than only semiotic types, where the inventory of possible types is larger than for semiotic types, this, too, would be very interesting.

[2] The reason this is helpful is in using cosine as a similarity metric but potentially being faced with zero vectors (which denote the origin) – the angle between a non-zero vector, and the origin is not defined.

[3] Emblem gestures as per the MUMIN coding scheme (Allwood et al., 2007).

Table 1: *Distribution of gesture types in the data sample. The middle column indicates the count of participants that use the gesture type, and the final column indicates the total count of instances of the gesture type.*

Gesture Semiotic Type	Participants use	Count
Beat	25	374
Deictic	18	64
Iconic	25	251
N/A	27	300
Symbolic	8	15
Sum	103	1004

Table 2: *Ranks (r) of the 40 most frequent lexical and vocalization types, including silence. Items in square brackets are normalized transcriptions of non-lexical vocalizations.*

Type	r	Type	r	Type	r	Type	r
	1	[ah]	11	think	21	laugh	31
yeah	2	and	12	like	22	or	32
[laugh]	3	not	13	would	23	violin	33
i	4	a	14	meat	24	then	34
ok	5	you	15	paper	25	first	35
the	6	that	16	hair	26	in	36
be	7	we	17	do	27	one	37
[eh]	8	[un]	18	oh	28	to	38
so	9	is	19	cut	29	hospital	39
it	10	no	20	of	30	say	40

4.2. Processing

We count the instances of lexical types that occur in each context position of a gesture, and relativize those counts to the total number of tokens in the data. Silence is treated as a word type. Words whose durations contain a gesture start or completion are counted as occurring during the gesture. We have developed a method that allows for parameterized processing – varying the number of linguistic context positions outside the gesture $m \geq 0$, the number of types to consider in each context position $n \geq 1$, attention to gesture's left or right contexts (or both), retention of possible zero vectors or preserving a feature within each context position for the absolute count of non-ranked types that occur there, choice of measuring Euclidean distance between vectors or cosine similarity. We obtain count vectors for each participant and gesture type used, thus, 103 vectors,[4] and, therefore, (103×102)/2 = 5253 vector comparisons. The count vectors are relativized to total lexical token counts and then normalized using either vector-local minimum and maximum values or data-set global minimum and maximum values.

In the experiments reported here, we counted items that occurred during gestures and $m = 1$ context position after the gesture (right context only), counting occurrences of the $n = 50$ most frequent items in the corpus in each of those contexts, binarizing the count of items outside the most frequent n – thus each context vector has length 102. Using the notation introduced above, we report on experiments using gestural linguistic context vectors of the form in (6), where the 0th context position is the duration of the gesture. A word partially spanned by the gesture is treated as "in" it.

$$\langle \frac{w_0^1}{\Sigma}, \ldots, \frac{w_0^{50}}{\Sigma}, \frac{\overline{W_0}}{\Sigma}, \frac{w_{+1}^1}{\Sigma}, \ldots, \frac{w_{+1}^{50}}{\Sigma}, \frac{\overline{W_{+1}}}{\Sigma}, \rangle \qquad (6)$$

We report on results using global norms in the min-max vector normalization. The values chosen here are selected somewhat arbitrarily rather than with attention to optimal performance. This is because our goal here is to demonstrate the efficacy of the idea with arbitrary values, leaving it to later work to optimise. Non-arbitrary qualities are visible in ignoring left-contexts here (because of the mentioned intuition that gestures relate to words that follow more often than those that precede) and, attentive to the overall corpus size, not choosing a large value for n. Similar reasoning applies to the context positions: were we considering linguistic contexts of words, we would look at context lengths of at least $m = 2$, but here, with words uttered during gestures, and with a relatively small corpus, we consider only one additional context position after the gesture.

4.3. Tests

We consider four sorts of tests, having computed linguistic context vectors associated with each gesture type produced by each of the dialogue participants analyzed. Firstly, we ask whether there is an interaction between the gesture types and the categories that arise from hierarchical clustering of the context vectors. Secondly, for each gesture type, we ask whether gestures of that type have context vectors with greater similarity to others of that type than to context vectors of other gestures. Thirdly, noting the potential for interlocutors in dialogue to influence each other, we ask whether overall and for each gesture type, the aggregate similarity of gesture context vectors produced by individuals who spoke with each other is greater than among those who did not speak with each other. Fourthly, we ask whether the contexts of use of the gesture types are self-similar for a person – whether, on average, the gesture types are interchangeable, as given by the evidence that without attending to which gesture type, the gestural linguistic context vectors for the same person show smaller differences than comparisons associated with different people.

5. Results

5.1. Hierarchical Clustering

We use Ward agglomerative clustering of context vectors, which is designed to minimize within-cluster variance (Murtagh and Legendre, 2014).[5] Because we have five gesture type annotations, we cut the cluster dendrogram also to five clusters, and ask the question of whether the cross-tabulation given by gesture types and cluster assignment demonstrates a non-random interaction (see Table 3 for the observed counts). However, because the total count of gestures in each category cannot lead to satisfaction of the assumptions of a χ^2 test (which would require at least 125 instances rather than the attested 103), we also conducted a smaller test of the interaction between the three most attested types (Beats, Iconics, N/A – see Table 1) and a three cluster cut. With the caveat that the χ^2 assumptions are not satisfied, we note that for the five by five contingency table test, $\chi^2 = 52.365, df = 16, p < 0.00001$, and examine the standardized Pearson residuals (see Table 4).

Given that standardized residuals follow a normal distribution the cut-off value for $\alpha = 0.05$ is ± 1.96. Thus, there is not

[4] The total participant use of gesture types (Table 1) is 103 not 28×5.

[5] We use the R's hclust with the ward.D2 method.

Table 3: *Cross-tabulation of gesture types and clusters of gestural linguistic context vectors.*

	Cluster index				
Gesture Semiotic Type	**1**	**2**	**3**	**4**	**5**
Beat	13	1	7	4	0
Deictic	5	4	2	5	2
Iconic	8	2	9	6	0
N/A	26	0	1	0	0
Symbolic	3	1	0	4	0

Table 4: *Standardized residuals from the χ^2 test applied to the values in Table 3.* **Bold** *values are significant (given $N(0,1)_{0.05}$). Negative values indicate fewer observations than would be expected with no interaction between the categories; positive values indicate more observations than would be expected with no interaction.*

	Cluster index				
Gesture Type	**1**	**2**	**3**	**4**	**5**
Beat	-0.161	-0.809	1.420	-0.362	-0.809
Deictic	**-2.400**	**2.520**	-0.883	1.120	**3.100**
Iconic	**-2.460**	0.050	**2.600**	0.823	-0.809
N/A	**5.200**	-1.760	**-2.300**	**-2.880**	-0.851
Symbolic	-0.939	0.521	-1.400	**2.400**	-0.414

a significant interaction ($p < 0.05$) in terms of a greater than expected number of observations between beats and any of the clusters, but there is for deictics and the second and fifth clusters, iconics and the third cluster, general hand movement and the first cluster, and symbolics and the fourth cluster.

Table 5: *Cross-tabulation of three most frequent gesture types and clusters of gestural linguistic context vectors; and accompanying standardized residuals.* **Bold** *values are significant (given $N(0,1)_{0.05}$). Negative values indicate fewer observations than would be expected with no interaction between the categories; positive values indicate more observations than would be expected with no interaction.*

	Observations Cluster index			Standardized residuals Cluster index		
Gesture Type	**1**	**2**	**3**	**1**	**2**	**3**
Beat	13	5	7	-1.130	0.506	0.869
Iconic	8	8	9	**-3.620**	**2.460**	**2.040**
N/A	26	0	1	**4.660**	**-2.910**	**-2.860**

Table 5 shows the counts of each of the three most frequent gesture types and clustering into three categories minimizing within-category variance. A χ^2 test of the interaction between the gesture types that give rise to the associated linguistic context vectors and the clusters that minimize variance in Euclidean distance is significant ($\chi^2 = 24.094, df = 4, p < 0.0001$). The standardized residuals reveal significance ($p < 0.05$) in the greater than expected witnessing of iconics in the second and third clusters and general hand movements in the first cluster.

5.2. Gesture type similarity

We also examine the linguistic context vectors associated with each participant's gestural types by considering the pairwise comparison of each of the 103 vectors with each of the other vectors, and then explore whether the Euclidean distance between vector comparisons is smaller within comparisons of the same type (i.e., the linguistic context vector for iconics produced by participant i compared with linguistic context vector for iconics produced by participant j, $i \neq j$) than within comparisons of distinct types, as would be expected if the contexts within which distinct speakers use the same semiotic type are more similar to each other than contexts of distinct types are.

Table 6: *Mean Euclidean distance between gestural linguistic context vectors for gestures of distinct types and the same types.* **Bold** *figures are significant (Wilcox, $p < 0.05$) and align with our hypotheses;* Italicized *figures are significant but un-aligned.*

Gesture Type	**Distinct Types**	**Same Type**
Beat	0.3959073	*0.4329663*
Deictic	0.3315388	**0.1584853**
Iconic	0.3607134	0.3386864
N/A	0.4083339	**0.3478011**
Symbolic	0.3127734	**0.1097768**

Table 6 shows that for each of the semiotic types except beats, the mean distance between comparisons of linguistic context vectors for gestures of the same type (by different participants) is smaller than the mean for mis-matched gesture types. Wilcox tests of the difference in rank sums between the two groups for deictics, arbitrary hand movements (N/A), and symbolics reveals that the differences are significant (deictics: $W = 196751, p < 0.000001$; arbitrary hand movements (N/A): $W = 303373, p < 0.000001$; symbolics: $W = 19116, p < 0.000001$). The smaller dissimilarity for matched iconics is not significant. The greater dissimilarity for matched beats than gestures of other types compared with beats is significant ($W = 253875, p < 0.001$).

5.3. Interlocutor similarity

One might expect that alignment effects in conversation lead to the use of similar gestures and similar contexts of use for gestures. Thus, one might anticipate smaller mean Euclidean distances for each gesture type for individuals who participated in the same session than for participants in distinct sessions. Table 7 shows that the observed effect conforms to this expectation for beats, deictics, and arbitrary hand movements, but without statistical significance in any of those three cases.[6] In none of the fourteen conversations did both participants use symbolics.

5.4. Self similarity

A related question is whether the contexts in which participants use any gesture at all are more similar to each other than the use contexts of each gesture type as used by anyone. That is,

[6] A Wilcox rank sum test was used. Noting the disparity in the number of measurements in each category – in one case, gestures of the same type produced in the same conversation, where the upper-bound on the number of observations per gesture type is 14 – and the other case, the same gesture but within distinct conversations, where the count upper-bound is 364 (28×26)/2, repeated testing with random sampling (without replacement) of 14 items from the larger set was applied.

Table 7: *Mean Euclidean distance between gestural linguistic context vectors for each type as produced by individuals who participated in distinct sessions or the same session.*

Gesture Type	Distinct Sessions	Same Session
Beat	0.4332659	0.4250950
Deictic	0.1586156	0.1546293
Iconic	0.3385742	0.3413798
N/A	0.3482588	0.3359012
Symbolic	0.1097768	NA

regardless of who produced an iconic, the contexts of iconics might be more similar to each other. Alternatively, it might be that regardless of the gesture type, the contexts in which each individual gestures are more similar to each other.

Restricting attention for each gesture type to the contrast between the distances between gestural linguistic context vectors for the same gesture as produced by distinct participants and the distances between all other gestural context vectors produced by a participant compared with the vector for that gesture type, greater distance is visible among distinct types for one participant than for the type among all participants for all but beats – see Table 8. The differences are significant for beats ($W = 12561, p = 0.03167$), deictics ($W = 1456, p < 0.000001$) arbitrary hand movements ($W = 9937, p = 0.002553$) and symbolics ($W = 104, p < 0.000001$).

Table 8: *Mean Euclidean distance between gestural linguistic context vectors of the same gesture type produced by distinct participants vs. that gesture type compared with all other gestures produced by the same participants.* **Bold** *figures are significant (Wilcox, $p < 0.05$) and align with our hypotheses;* Italicized *figures are significant but hypothesis contradicting.*

Gesture Type	Distinct Participants	Same Participant
Beat	0.4329663	*0.3839854*
Deictic	0.1584853	**0.3296585**
Iconic	0.3386864	0.3560252
N/A	0.3478011	**0.4045158**
Symbolic	0.1097768	**0.3069025**

6. Discussion

The results of clustering presented in §5.1 show that the gestural linguistic context vectors for each participant's gesture types cluster into categories that have non-random interaction with the gesture type categories. If there were nothing systematic about the linguistic contexts of use of the gesture types, then it would be expected that the interaction with cluster categories would not show significant interactions. These tests support the inference that the gestural linguistic context vectors constructed encode distributional facts about the use-conditions of gestures. However, the two forms of categorization (gesture semiotic types and clusters that emerge from reasoning about raw similarity among the vectors) are revealed to be non-isomorphic.

The differences in Euclidean distances noted in §5.2 show, as expected, that gestural linguistic context vectors succeed in encoding aspects of the linguistic meaning of gestural types, in that pairwise comparisons of the context vectors involve smaller distances for vectors for the same gestural type (regardless of who produced it) than for distinct gestural types. The significance of the effect suggests contextual homogeneity for deictics, symbolics, and arbitrary hand movements. The effect was not significant for iconics. Significance in the opposite direction for beats suggests greater contextual heterogeneity for beats.

We hypothesized that given a set of context vectors arising from a data set in which some people have communicated with each other while most have not, that the people who communicated with each other will align with each other to the extent of using similar linguistic contexts for their gestures. In accord with this hypothesis, §5.3 shows that the Euclidean distances among gestural linguistic context vectors of those who have communicated with each other are less than for the vectors of the complement set for beats, deictics and arbitrary hand movements, but without statistical significance.

If gestural linguistic context vectors capture aspects of meaning of gesture types, then they should not be interchangeable (unless they all mean the same thing). Under the hypothesis that the distinct semiotic types have distinct meanings, one would expect that the Euclidean distances among all gestural context vectors for one participant should not be smaller than the comparisons for distinct participants. This is mostly what §5.4 reports: among distinct participants, gestural linguistic context vectors deictics, symbolics, and arbitrary hand movements have significantly less distance between them than the distance among all gestures produced by each participant considered only among themselves – deictics are more like deictics than gestures of the ith participant are like others of the ith participant. Exceptions are iconics (where the effect was not significant) and beats, where the effect is in the opposite direction.

These tests reveal distinctive contextual homogeneity for the gesture categories other than beats. Clustering §5.1 did not identify significant interaction between context comparisons of beats and the categories that arise from clusters, but did in each other case. From §5.2 we see the overall contexts of beats as less like other beats than other gestures and from §5.4 we see that contexts of beats are closer to the contexts in which each participant uses other gestures than to the contexts in which other participants use beats. Different behaviour for beats makes some sense if they are akin to natural language prosodic emphasis: an inclination to use such is not obviously linked to particular content. The measures of homogeneity of linguistic contexts associated with arbitrary hand movements might be understood in the opposite direction. Although they are not readily classified among the other semiotic types, there is visible relative homogeneity of contexts associated with arbitrary hand movements, as if, in certain contexts, participants are inclined to gesture meaningfully, but do so in a manner that is not readily interpreted by interlocutors. Thus, the N/A category may, more than the other categories of gesture, be speaker-centric, part of thought formation more than supporting hearer understanding (cf. McNeill, 1997 or Kita, 2000). These tests show strong tendencies in the contexts of use of the gesture types as measured using gestural linguistic context vectors.

7. Conclusions

The work we report is limited by the small size of the corpus considered and the gesture annotations it makes available solely of semiotic type. The nature of the annotations available in this corpus are such that one could apply the approach to other channels that are simultaneous with linguistic content and ask questions like which laughter type or which gaze type is closest in meaning to a deictic gesture, for example, where the dimension-

ality of the linguistic context vectors can be directly matched. It remains an analytic challenge to identify how to match gestural linguistic context vectors with the linguistic context vectors that would be assigned to words in order to answer the question of which words are closest in meaning to each gesture. This is distinct from seeking lexical affiliates to gestures. The words closest in meaning to a gesture might not occur temporally near any gesture: synonymy is computed through similarity of contexts of use, not co-occurrence of synonymous terms. A natural next step in this research is to apply the approach to gesture annotations of a lower-level of abstraction than semiotic type, with focus on hand shapes, orientations, and trajectories, such as individuated in the MUMIN scheme (Allwood et al., 2007) or the more detailed framework of McNeill (2005). Even within the current inventory of gesture types, it would be useful to systematically explore the range of parameters open (i.e., number of context positions, number of items considered within each context position, etc.). Nonetheless, with this focus on semiotic types and distinctiveness of associated gestural linguistic context vectors in a number of comparisons involving them with somewhat arbitrary parameter settings, we have demonstrated that there is promise in this approach to estimating aspects of linguistic meaning associated with gestures.

8. Acknowledgements

This work was supported by the GEHM research network (Independent Research Fund Denmark grant 9055-00004B), the Science Foundation Ireland Centre for Research Training in Artificial Intelligence, Grant No. 18/CRT/6223, and the China Scholarship Council. We are grateful to anonymous GESPIN reviewers, who provided constructive responses to an earlier draft of this article.

9. References

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3):273–287.

Eisenstein, J. and Davis, R. (2007). Visual and linguistic information in gesture classification. In *ACM SIGGRAPH 2007 Courses*, SIGGRAPH '07, page 15–es, New York, NY, USA. Association for Computing Machinery.

Ekman, P. and Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98.

Ferré, G. (2010). Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French. In *Language Resources and Evaluation, Workshop on Multimodal Corpora*, volume W6, pages 86–91, Malta.

Finch, S. and Chater, N. (1992). Bootstrapping syntactic categories using statistical methods. In *14th Annual Conference of the Cognitive Science Society*, pages 229–235.

Han, T., Hough, J., and Schlangen, D. (2017). Natural language informs the interpretation of iconic gestures: A computational approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 134–139.

Hughes-Berheim, S. S., Morett, L. M., and Bulger, R. (2020). Semantic relationships between representational gestures and their lexical affiliates are evaluated similarly for speech and text. *Frontiers in Psychology*, 11.

Khosrobeigi, Z., Koutsombogera, M., and Vogel, C. (2022). Gesture and part-of-speech alignment in dialogues. In Gregoromichelaki, E., Hough, J., and Kelleher, J. D., editors, *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue*, pages 172–182.

Kirchhof, C. (2011). So what's your affiliation with gesture? In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*.

Kita, S. (2000). How representational gestures help speaking. In McNeill, D., editor, *Language and Gesture*, pages 162–85. Cambridge University Press.

Koutsombogera, M. and Vogel, C. (2018). Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus. In Calzolari, N. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2945–2951, Paris, France. European Language Resources Association (ELRA).

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago.

McNeill, D. (1997). Growth points cross linguistically. In Nuyts, J. and Pederson, E., editors, *Language and Conceptualization*, Language, Culture & Cognition, pages 190–212. Cambridge University Press.

McNeill, D. (2005). *Gesture & Thought*. University of Chicago Press.

Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295.

Peirce, C. S. (1934). *The Collected Papers of Charles Sanders Peirce, Vol. V: Pragmatism and Pragmaticism*. Harvard University Press, Cambridge.

Schegloff, E. A. (1985). On some gestures' relation to talk. In Atkinson, J. M., editor, *Structures of Social Action*, Studies in Emotion and Social Interaction, pages 266–296. Cambridge University Press.

Zhang, J. R. and Kender, J. R. (2012). Arm gesture variations during presentations are correlated with conjunctions indicating contrast. In *Proceedings of the 2012 ACM workshop on User experience in e-learning and augmented technologies in education*, pages 13–18.