FORM: An Extensible, Kinematically-based Gesture Annotation Scheme

Craig Martell[*†]

[*]Linguistic Data Consortium
University of Pennsylvania
3615 Market Street, Suite 200
Philadelphia, PA 19104
cmartell@unagi.cis.upenn.edu

[†]Department of Computer and Information Sciences
University of Pennsylvania
200 S. 33rd St
Philadelphia, PA 19104

Abstract

Annotated corpora have played a critical role in speech and natural language research; and, there is an increasing interest in corpora-based research in sign language and gesture as well. As examples, consider the tools Anvil and MediaTagger. These are excellent tools which allow for multi-track annotation of videos of speakers or signers. With tools such as these, researchers can create corpora containing, for example, grammatical information, discourse structure, facial expression, and gesture. The issue, then, is not the ability to create corpora containing gesture and speech information, but the type of information captured when describing gestures. We present a non-semantic, geometrically-based annotation scheme, FORM, which allows an annotator to capture the kinematic information in a gesture just from videos of speakers. In addition, FORM stores this gestural information in Annotation Graph format—allowing for easy integration of gesture information with other types of communication information, e.g., discourse structure, parts of speech, intonation information, etc.

1. Annotated Corpora and Multi-Modal Data

Annotated corpora have played a critical role in speech and natural language research; and, there is an increasing interest in corpus-based research in sign language and gesture as well. It was for this purpose that the FORM annotation scheme was developed[1]. This is a useful tool which allows for multi-tier gesture annotation of videos of speakers or signers. For example, Figure 1's four stills are from a video sequence of Brian MacWhinney teaching a research methods course at Carnegie Mellon University. These data were chosen because they are part of the TalkBank collection (http://www.talkbank.org).

Figure 1: Snapshots of Brian MacWhinney on January 24, 2000

The FORM annotation of the video, from timestamp 1:13.34 (1 minute 13.34 seconds) to timestamp 1:14.01 is shown in Figure 2.

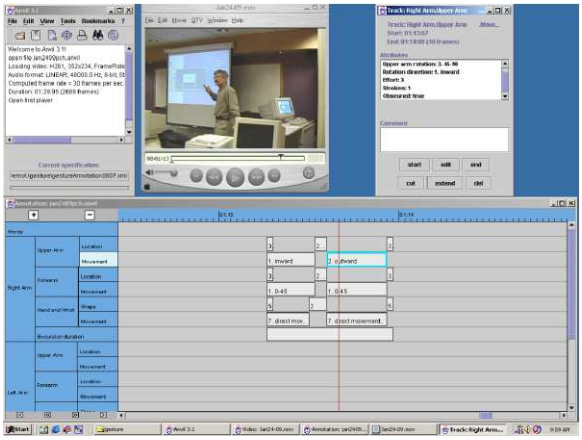

Figure 2: FORM annotation of Jan24.mov, using Anvil as the annotation tool

This is the view on the data that a particular tool, Anvil (Kipp, 2001), presents to the annotator. However, FORM uses annotation graphs (Bird and Liberman, 1999) as its logical representation of the data. So regardless of which annotation tool is used, FORM's internal view is the annotation graph given in Figure 3.

Both Annotation Graphs (AGs) and FORM are presented in greater detail, below. However, for now note, first, that an annotation graph is a directed acyclic graph (DAG) such that the nodes represent timestamps of some given signal and the arcs represent some linguistic event that spans the time between the timestamps. Second, note that FORM uses vectors of *attribute:value* pairs to capture the gestural

[1]The author wishes to sincerely thank Adam Kendon for his input on the FORM project. He has provided not only suggestions as to the direction of the project, but also his unpublished work on a kinematically-based gesture annotation scheme was the FORM project's starting point (Kendon, 2000).

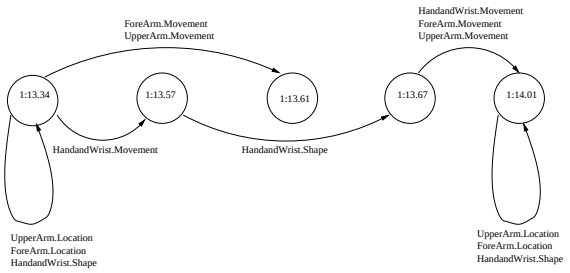

Figure 3: FORM/Annotation Graph representation of example gesture

information of each section of the arms and hands. In Figure 1, then, the arc labeled *HandandWrist.Movement* from 1:13.34 to 1:13.57 encodes the kinematics of Brian's moving his right hand or wrist during this time period, and the arc from 1:13.24 to 1:13.67 encodes a change in his right hand's shape.[2]

The particular advantage to using AGs to encode the kinematics of gesture, or any linguistic signal, is the ease with which the annotation can be extended to include other data. The only constraint is that all the data share the same timeline. As such, researchers can easily extend the FORM corpus to include, for example, grammatical information, discourse structure, facial expression, etc. Figure 4 is such an augmented AG. It is another representation of the video clip from Figure 1 (Jan24-09.mov) and is augmented with head/torso movement, speech transcription and syntactic information, and intonation/pitch information. Note that this is a conservative extension of the original AG from Figure 3, that is, the original AG remains unchanged and new information is simply added.

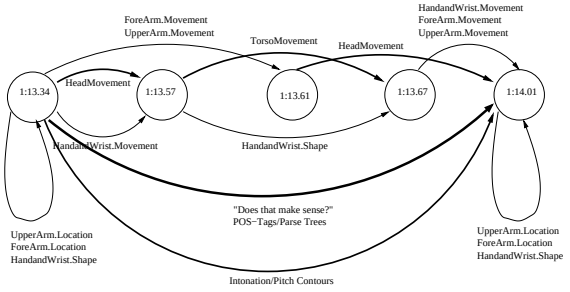

Figure 4: Augmented FORM annotation graph of Jan24-09.mov

The goal of the FORM project is to build such an extensible corpus of annotated videos in order to allow for general research on the relationship among the many different aspects of conversational interaction. Additionally, further tools and algoithms to add these annotations and evaluate inter-annotator agreement will be developed. The end result of this work will be a corpus of annotated conversational interaction, which can be:

- extended to include new types of information concerning the same conversations; as new tag-sets and coding schemes are developed—discourse-structure or facial-expression, for example—new annotations could easily be added;
- used to test scientific hypotheses concerning the relationship of the paralinguistic aspects of communication to meaning;
- used to develop statistical algorithms to automatically analyze and generate these paralinguistic aspects of communication (e.g., for Human-Computer Interface research).

[2]For the example given in Figure 1, Brian is only moving his right hand. Accordingly, the *Right.* which normally would have been prepended to the arc-labels has been left off.

2. Annotation Graphs

As described in (Bird and Liberman, 1999), annotation graphs are a formal framework for "representing linguistic annotations of time series data." AGs do this by extracting away from the physical-storage layer, as well as from application-specific formatting, to provide a "logical layer for annotation systems." An annotation graph is a collection arcs and nodes which share a common timeline, that of a video tape, for example. Each node represents a timestamp and each arc represents some linguistic event spanning the time between the nodes. The arcs are labeled with both attributes and values, so that the arc given by the 4-tuple (1,5,Wrist Movement,Side-to-side) represents that there was side-to-side wrist movement between timestamp 1 and timestamp 5. Again, the advantage of using annotation graphs as the logical representation is that it is easy to combine heterogeneous data—as long as they share a common time line. So, if we have a dataset consisting of gesture-arcs, as above, we can easily extend this dataset by adding more arcs representing discourse structure, for example, simply by adding other arcs which have discourse-structure attributes and values. Again, this allows different researchers to use the same linguistic data for many different purposes, while, at the same time, allowing others to explore the correlations between the different phenomena being studied.

3. FORM

The FORM annotation scheme was developed in order to capture the kinematic information of gesture from videos of speakers. This is done by annotating the video with geometric descriptions of the positions and movements of the upper and lower arms, and the hands and wrists. FORM uses AGs as its logical representation.

Before FORM, recording kinematic information required special laboratory equipment in order to motion-capture the data. Usually this requires special sensors to be attached to the speaker in 10-15 key points on the body. These sensors have cables which physically impede motion, as well as create a very unnatural feel to the conversation itself. Subjects are self-conscious of this equipment, and of being in a laboratory. In this way, it is difficult to get natural communication. However, by combining a geometrically-based tag set with annotation techniques which have been very successful in other fields (e.g., NLP), FORM has succeeded in being able to record the kinematics of gesture without the expense and unnaturalness associated with conventional methods. Additionally,

with motion-capture methods, one is only able to capture conversations staged for a specific lab session. The FORM method allows a researcher to capture kinematic information from any video-recorded conversation. This allows us to take advantage of vast amounts of already-recorded conversational data. This aspect is invaluable if we want to study the gestures of other cultures or linguistic groups in natural, unimpeded settings. And, even when doing controlled experiments, the setting can be almost natural. The only addition to the natural setting would be video cameras, which are now quite small and can be easily hidden.

3.1. Structure of FORM[3]

FORM is designed as a series of tracks representing different aspects of the gestural space. Generally, each independently moved part of the body has two tracks, one track for Location/Shape/Orientation, and one for Movement. When a part of the body is held without movement, a Location object describes its position and spans the amount of time the position is held. When a part of the body is in motion, Location objects with no time period are placed at the beginning and end of the movement to show where the gesture began and ended. Location objects spanning no period of time are also used to indicate the Location information at critical points in certain complex gestures.

An object in a movement track spans the time period in which the body part in question is in motion. It is often the case that one part of the body will remain static while others move. For example, a single hand shape may be held throughout a gesture in which the upper arm moves. FORM's multi-track system allows such disparate parts of single gestures to be recorded separately and efficiently and to be viewed easily once recorded. Once all tracks are filled with the appropriate information, it is easy to see the structure of a gesture broken down into its anatomical components.'

At the highest level of FORM are groups. Groups can contain subgroups. Within each group or subgroup are tracks. Each track contains a list of attributes concerning a particular part of the arm or body. At the lowest level (under each attribute), all possible values are listed. The structure, then, is as follow:

Group

Subgroup

Track

ATTRIBUTE

Value

The following description will follow this structure. Described are the tracks for the Location of the Right or Left UpperArm.

Right/Left Arm

Upper Arm (from the shoulder to the elbow).

Location

UPPER ARM LIFT (from side of the body)

no lift
0-45
approx. 45
45-90
approx. 90
90-135
approx. 135
135-180
approx. 180

RELATIVE ELBOW POSITION: The upper arm lift attribute defines a circle on which the elbow can lie. The relative elbow position attribute indicates where on that circle the elbow lies. Combined, these two attributes provide full information about the location of the elbow and reveal total location information (in relation to the shoulder) of the upper arm.

extremely inward
inward
front
front-outward
outward (in frontal plane)
behind
far behind

The next three attributes individually indicate the direction in which the biceps muscle is pointed in one spatial dimension. Taken together, these three attributes reveal the orientation of the upper arm.

BICEPS: INWARD/OUTWARD

none
inward
outward

BICEPS: UPWARD/DOWNWARD

none
upward
downward

BICEPS: FORWARD/BACKWARD

none
forward
backward

OBSCURED: This is an binary attribute which allows the annotator to indicate if the attributes and values chosen were "guesses" necessitated by visual occlusion. This attribute is present in each of FORM's tracks.

Again, we have only presented the **Right/LeftArm**.***UpperArm***.*Location*.ATTRIBUTE.Value options here. The full "Code Book" can be found at http://www.ldc.upenn.edu/Projects/FORM/. Listed there are all the **Group**.***Subgroup***.*Track*.ATTRIBUTE.Value possibilities.

[3]The author wishes to acknowlege Jesse Friedman and Paul Howard in this section. Most of what is written here is from their "Code Book" section of http://www.ldc.upenn.edu/Projects/FORM/.

3.2. Annotation Complexity

An experienced annotator can create approximately 3 seconds of annotation per hour. He/she can annotate at most for 6 hours per day, generating 18 seconds/day. Accordingly, it will take an experienced annotator 5 work days to annotate a 90-second video of conversational interaction.

Generating only 90 seconds of annotation per work week makes such an annotation project seem a daunting task. However, the amount of information contained in conversational gesturing is substantial—on the order of 3500 distinct *attribute:value* arcs per minute. This underscores the potential value of such a corpus, viz. there is seemingly much more information in 90 seconds of communicative interaction than we are currently capturing by only transcribing speech.

4. Preliminary Inter-Annotator Agreement Results

Preliminary results from FORM show that with sufficient training, agreement among the annotators can be very high. Table 2 shows preliminary interannotator agreement results from a FORM pilot study.[4] The results are for two trained annotators for approximately 1.5 minutes of Jan24-09.mov, the video from Figure 1. For this clip, the two annotators agreed that there were at least these 4 gesture excursions. One annotator found 2 additional excursions. Precision refers to the decimal precision of the time stamps given for the beginning and end of gestural components. The *SAME* value means that all time-stamps were given the same value. This was done in order to judge agreement with having to judge the exact beginning and end of an excursion factored out. *Exact* vs. *No-Value* percentage refers to whether both the attributes and values matched exactly or whether just the attributes matched exactly. This distinction is included because a gesture excursion is defined as all movement between two rest positions of the arms and hands. For an excursion, the annotators have to judge both which parts of the arms and hands are salient to the movement (e.g., upper-arm lift and rotation, as well as forearm change in orientation and hand/wrist position) as well as what values to assign (e.g., the upper-arm lifted 15-degrees and rotated 45-degrees). So, the *No-Value%* column captures the degree to which the annotators agree just on the structure of the movement, while *Exact%* measures agreement on both structure and values.

The degree to which inter-annotator agreement varies among these gestures might suggest difficulty in reaching consensus. However, the results on *intra*-annotator agreement studies demonstrate that a single annotator shows similar variance when doing the same video-clip at different times. Table 3 gives the intra-annotator results for one annotator annotating the first 2 gesture excursions of Jan24-09.mov.

For both sets of data, the pattern is the same:

[4]Essentially, all the arcs for each annotator are thrown into a bag. Then all the bags are combined and the intersection is extracted. This intersection constitutes the overlap in annotation, i.e., where the annotators agreed. The percentage of the intersection to the whole is then calculated to get the scores presented.

Gesture Excursion	Precision	Exact%	No-Value%
1	2	3.41	4.35
	1	10.07	12.8
	0	29.44	41.38
	SAME	56.92	86.15
2	2	37.5	52.5
	1	60	77.5
	0	75.56	94.81
	SAME	73.24	95.77
3	2	0	0
	1	19.25	27.81
	0	62.5	86.11
	SAME	67.61	95.77
4	2	10.2	12.06
	1	25.68	31.72
	0	57.77	77.67
	SAME	68.29	95.12

Table 1: Inter-Annotator Agreement on Jan24-09.mov

Gesture Excursion	Precision	Exact%	No-Value%
1	0	5.98	7.56
	1	20.52	25.21
	0	58.03	74.64
	SAME	85.52	96.55
2	2	0	0
	1	25.81	28.39
	0	89.06	95.31
	SAME	90.91	93.94

Table 2: Intra-Annotator Agreement on Jan24-09.mov

- the less precise the time-stamps, the better the results;
- *No-Value%* is significantly higher than *Exact%*.

It is also important to note that Gesture Excursion 1 is far more complex than Gesture Excursion 2. And, in both simple and complex gestures, inter-annotator agreement is approaching intra-annotator agreement. Notice, also, that for Excursion 2, inner-annotator agreement is actually better than intra-annotator agreement for the first two rows. This is a result of the difficulty for even the same person over time to precisely pin down the beginning and end of a gesture excursion. Although the preliminary results are very encouraging, all of the above suggests that further research concerning training and how to judge similarity of gestures is necessary. Visual information may need very different similarity criteria.

5. Future Directions

In order to build a useful, multi-modal corpus for human communication research, tools capable of not only annotation, but also of searching and verifying, must be built. The FORM project has the beginnings of a tool, FORMTool, which allows for easy input of gestural information while viewing video. The problem is that there is no easy way for the annotator to assess the "correctness" of his/her annotation. Inter-annotator agreement studies help assure that the annotators are all creating similar data, but this does not

guarantee that these data accurately represent the phenomena. With a strong annotation-manager, and a very specific coding manual, a team might achieve high inter-annotator agreement scores, but still not have sufficiently captured the phenomena in question.

To deal with these issues, we plan to concomitantly do research concerning the following.

- **Visualization and Animation Tools** which will "play back" an annotation. This will allow the annotator to better judge how well he/she has captured the linguistic phenomenon in question.
- **New Metrics for Inner-Annotator Agreement**. As mentioned in Section 4, above, our current numbers are based on the bag-of-arcs technique. However, as the scores there indicate, often annotators agree to a large degree on structure, but differ only on exact beginning or ending timestamp, or on the value of an attribute. Unfortunately, small differences in timestamp and value are judged incorrect to the same degree as large differences. Visual feedback, as just described, will allow us to discover whether small differences in coding actually have little difference visually. If this proves to be the case, then we will need to experiment with more geometrically-based measures of similarity, e.g., distance in n-dimensional space.
- **Augmented Search Algorthims for Annotation Graphs**. The annotation-graph community has already begun research into the most efficient ways to search AG data (Bird and Buneman, 2000). But, as we add richer information, we need to extend the search capabilities to allow researchers fast access to this complex data. An example would be the need to search for all gestures similar to the one given in our example. Further, the researcher might want to then search those results for gestures which accompany certain syntactic or intonational structures.

6. Conclusion: Applications to HLT and HCI?

The augmentation of FORM to include richer paralinguistic information (Head/Torso Movement, Transcription/Syntactic Information, and Intonation/Pitch Information) will create a corpus that allows for research that heretofore we have been unable to do. It will facilitate experiments that we predict will be useful for speech recognition, as well as other Human-Language Technologies (HLT). As an example of similar research, consider the work of Francis Quek et al. (Quek and others, 2001). They have been able to demonstrate that gestural information is useful in helping with automatic detection of discourse transition. However, their results are limited by the amount of kinematic information they can gather with their video-capture system. Further, an augmented-FORM corpus will contain much more specific data and will allow for more fine-grained analyses than is currently feasible.

Additionally, knowing the relationships among the different facets of human conversation will allow for more informed research in Human-Computer Interaction (HCI). If one of the goals of HCI is to have better immersive-training, then it will be imperative that we understand the subtle connections among the paralinguistic aspects of interaction. A virtual human, for example, would be much better if it were able to understand, and act in accordance with, all of our communicative quirks

Having an extensible corpus such as we describe in this paper is a first-step that will allow many researchers, across many disciplines, to explore these and other useful ideas.

7. References

S. Bird and P. Buneman. 2000. Towards a query language for annotation graphs. In *International Conference on Language Resources and Evaluation.*, Paris. European Language Resources Association. http://citeseer.nj.nec.com/298297.html.

Steven Bird and Mark Liberman. 1999. A formal framework for linguistic annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, Pennsylvania. http://citeseer.nj.nec.com/article/bird99formal.html.

Adam Kendon. 2000. Suggestions for a descriptive notation for manual gestures. Unpublished.

Michael Kipp. 2001. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of Eurospeech 2001*, pages 1367–1370.

Francis Quek et al. 2001. Gestural origo and loci-transitions in natural discourse segmentation. Technical Report VISLab-01-12, Department of Computer Science and Engineering, Wright State University. http://vislab.cs.wright.edu/Publications/QueBMH01.html.

Talkbank project. http://www.talkbank.org.