ABSTRACT
Announced 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. 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

1. INTRODUCTION
FORM² is an annotation scheme designed both to describe the kinematic information in a gesture, as well as to be extensible in order to add speech and other conversational information.

Our goal is to build 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 algorithms 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 speech and 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. STRUCTURE OF FORM
FORM³ 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 begins and ends. 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.

¹This presentation is a shortened version of [4]
²The 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 [2].
³The author wishes to acknowledge 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/.
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. Described below 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 Location tracks for the Right or Left Arm UpperArm group. The full “Code Book” can be found at http://www.ldc.upenn.edu/Projects/FORM/. Listed there are all the Group, Subgroup, Track, Attribute and Value possibilities.

As an example, consider Figure 1. These four stills are from a video sequence of Brian MacWhinney teaching a research methods course at Carnegie Mellon University.

![Fig. 1. Snapshots of Brian MacWhinney on January 24, 2000](image)

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.

![Fig. 2. FORM annotation of Jan24.mov, using Anvil as the annotation tool](image)

This is the view on the data that a particular tool, Anvil [3], presents to the annotator. However, FORM uses annotation graphs [1] 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.

Annotation Graphs (AGs) and are presented in greater detail in [1]. However, here note, first, that an annotation

---

4These data were chosen because they are part of the TalkBank collection [6].
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 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.\(^5\)

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.

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

\(^5\)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.
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 2 gives the intra-annotator results for one annotator annotating the first 2 gesture excursions of Jan24-09.mov.

<table>
<thead>
<tr>
<th>Gesture Excursion</th>
<th>Precision</th>
<th>Exact%</th>
<th>No-Value%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>3.98</td>
<td>7.56</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>20.52</td>
<td>25.21</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>58.03</td>
<td>74.64</td>
</tr>
<tr>
<td>SAME</td>
<td>85.52</td>
<td>96.55</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>25.81</td>
<td>28.39</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>89.06</td>
<td>95.31</td>
</tr>
<tr>
<td>SAME</td>
<td>90.91</td>
<td>93.94</td>
<td></td>
</tr>
</tbody>
</table>

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

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

- 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.

4. CONCLUSION: APPLICATIONS TO HLT AND HCI?

We plan to augment FORM to include richer linguistic and paralinguistic information (Head/Torso Movement, Transcription/Syntactic Information, and Intonation/Pitch Information). This 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. [5]. 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.

5. REFERENCES


