2002

An Extensible, Kinematically-Based Gesture Annotation Scheme

Martell, Craig H.

ADVANCES IN NATURAL MULTIMODAL DIALOGUE SYSTEMS

Edited by
JAN VAN KUPPEVELT
The Netherlands

LAILA DYBKJÆR
NISLab, University of Southern Denmark

NIELS OLE BERNSEN
NISLab, University of Southern Denmark

Kluwer Academic Publishers
Boston/Dordrecht/London

Contents

Chapter 1

FORM

An Extensible, Kinematically-Based Gesture Annotation Scheme

Craig H. Martell
Department of Computer Science
and
The MOVES Institute[*]
Naval Postgraduate School, Monterey, CA, USA
cmartell@nps.navy.mil

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. 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]

Keywords: Gesture, annotation, corpora, corpus-based methods, multi-modal communication.

1. Introduction

FORM[2] 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.

[*]Much of this work was done at the University of Pennsylvania and at The RAND Corporation as well.
[1]This presentation is a modified version of [Martell, 2002].
[2]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 He has provided not only suggestions as to the direction of the project, but also his unpublished work on a kinematically-based gesture annotation scheme was the FORM project's starting point [Kendon, 2000].

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 additional 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

FORM[3] 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. 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

[3]The 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/.

(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

Figure 1.1 - Figure 1.4 are example stills with the appropriate values of the above two attributes given.

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.

Figure 1.1. UPPER ARM LIFT: approx. 90; RELATIVE ELBOW POSITION: outward.

Figure 1.2. UPPER ARM LIFT: approx. 45; RELATIVE ELBOW POSITION: front.

Figure 1.3. UPPER ARM LIFT: 0-45; RELATIVE ELBOW POSITION: behind.

BICEPS: INWARD/OUTWARD

none
inward (see Figure 1.5)
outward (see Figure 1.6)

Figure 1.4. UPPER ARM LIFT: no lift; RELATIVE ELBOW POSITION: outward.

BICEPS: UPWARD/DOWNWARD

none
upward (see Figure 1.7)
downward (see Figure 1.8)

BICEPS: FORWARD/BACKWARD

none
forward (see Figure 1.9)
backward (see Figure 1.10)

Figure 1.5. INWARD.

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

Figure 1.6. OUTWARD.

Figure 1.7. UPWARD.

Figure 1.8. DOWNWARD.

http://www.ldc.upenn.edu/Projects/FORM/. Listed there are all the Group, Subgroup, Track, Attribute and Value possibilities.

Figure 1.9. FORWARD.

Figure 1.10. BACKWARD.

3. Annotation Graphs

In order to allow for maximum extensibility, FORM uses annotation graphs (AGs) as its logical representation[4]. 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 time line, that of a video tape, for example. Each node represents a time stamp and each arc represents some linguistic event spanning the time between the nodes. In FORM, the arcs are labelled 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 time stamp 1 and time stamp 5.

[4]Cf. [Martell, 2002] for a more complete discussion of FORM's use of AGs.

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.

4. Annotation Example

To gain a better understanding of the process of FORM annotation, we present here a small visual example. The four stills of Figure 1.11 are from a video sequence of Brian MacWhinney teaching a research methods course at Carnegie Mellon University[5]. We show these four key frames, here, for illustrative purposes only. The character of the gesture is gleaned from viewing the continuous movement in the video. However, these key frames would be used to set the time stamp and locations of the beginning and end of the movement and in-between points that are important to capturing its shape. The arcs in the annotation graph described below (Figure 1.13) capture the information for the movement in between the key frames.

Figure 1.11. Snapshots of Brian MacWhinney on January 24, 2000.

The FORM annotation, then, of the video, from time stamp 1:13.34 (1 minute 13.34 seconds) to time stamp 1:14.01 is shown in Figure 1.12. This is the view on the data that a particular tool, Anvil [Kipp, 2001], presents to the annotator[6]. As described above, FORM uses annotation graphs as its logical

[5]These data were chosen because they are part of the TalkBank collection (http://www.talkbank.org). TalkBank was responsible for funding a large part of this project.
[6]Anvil is described in further detail in the appendix.

Figure 1.12. FORM annotation of Jan24.mov, using Anvil as the annotation tool.

representation of the data; so regardless of choice of annotation tool, FORM's internal view is the annotation graph given in Figure 1.13.

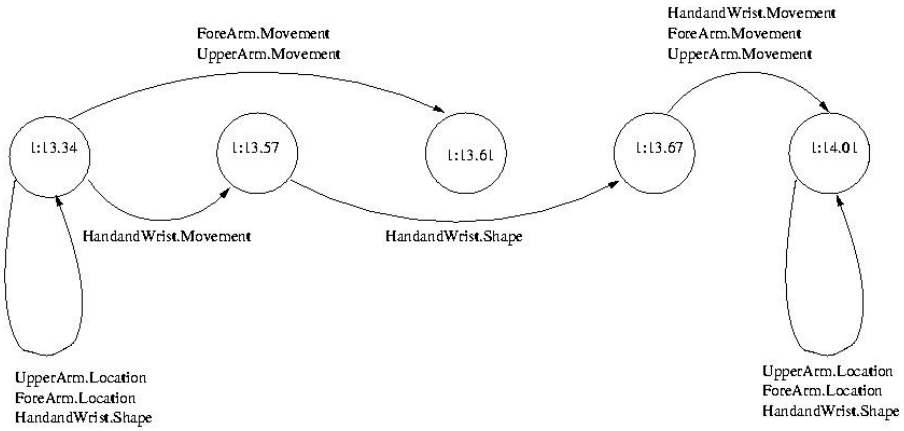

Figure 1.13. FORM/Annotation Graph representation of example gesture.

Again, FORM uses vectors of *attribute:value* pairs to capture the gestural information of each section of the arms and hands. In Figure 1.13, then, the arc labelled *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.[7]

5. Preliminary Inter-Annotator Agreement Results

Preliminary results from FORM show that with sufficient training, agreement among the annotators can be very high. Table 1.1 shows preliminary interannotator agreement results from a FORM pilot study.[8] The results are for two trained annotators for approximately 1.5 minutes of Jan24-09.mov, the video from Figure 1.11. 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, 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 1.2 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:

- the less precise the time-stamps, the better the results;

[7] For the example given in Figure 1.11, Brian is only moving his right hand. Accordingly, the *Right.* which normally would have been prefixed to the arc-labels has been left off.

[8] 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.

Table 1.1. Inter-Annotator Agreement on Jan24-09.mov.

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 1.2. Intra-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

- *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. Also, it is not clear as of the time of writing how these results might generalize. In particular, the relationship between inter- and intra-annotator agreement needs to be further explored. In addition, comparison studies with other methods of judging agreement are necessary. For example, how does FORM's method compare with Cronbach's alpha evaluated at discrete time-slices[9]? And, what would be the result adding a kappa-score analysis to the bag-of-arcs technique?

6. Conclusion: Applications to HLT and HCI?

We are augmenting FORM to include richer 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 and other Human-Language Technologies (HLT). As an example of similar research, consider the work of Francis Quek et al. [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, we believe 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 modalities.

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

[9] I am indebted to an anonymous reviewer for this suggestion.

Appendix: Other Tools, Schemes and Methods of Gesture Analysis

FORM has been designed to be *simultaneously* useful for both

1 capturing the kinematics of gesture; *and*

2 developing a corpus of annotated videos useful for computational analysis and synthesis.

Prior research along each of these dimensions that has contributed to, or has motivated, FORM. In this section we briefly review this prior work.

A.1 Non-computational Gesture Analysis

Two important figures in linguistic and psychological (read: non-computational) analysis of gesture are David McNeil and Adam Kendon. Each has develop annotation schemes and systems to analyze and annotate the gestures of speakers in video. However, their respective levels of analysis are quite different.

A.1.1 David McNeill: Hand and Mind. David McNeil [1992] uses as scheme which divides the gesture space into four basic types:

- Beat gestures;
- Iconic gestures;
- Metaphoric gestures; and
- Deictic gestures.

These categories are not meant to be mutually exclusive, although McNeill [1992] has been blamed for making it appear so. According to the McNeill Lab web site:

> A misconception has arisen about the nature of the gesture categories described in *Hand and Mind*, to wit, that they are mutually exclusive bins into which gestures should be dumped. In fact, pretty much any gesture is going to involve more than one category. Take a classic upward path gesture of the sort that many subjects produce when they describe the event of the cat climbing up the pipe in our cartoon stimulus. This gesture involves an iconic path-for-path mapping, but is also deictic. . . . Even "simple" beats are often made in a particular location which the speaker has given further structure (e.g. by setting up an entity there and repeatedly referring to it in that spatial location). Metaphoric gestures are de facto iconic gestures. . . . The notion of a type, therefore, should be considered as a continuum–with a given gesture having more or less iconicity, metaphoricity, etc.[10]

[10] http://mcneilllab.uchicago.edu/topics/type.html, as of 12/15/2003.

This work has been very influential, and has been the basis for at least one major computational project (see the BEAT toolkit, below). However, this level of analysis only serves to categorize the gesture. It provides no useful computational information for either automatic gesture analysis or for the automatic generation of gestures in computational agents.

A.1.2 Adam Kendon: The Kinetics of Gesture. Adam Kendon's approach, best articulated in "An Agenda for Gesture Studies" [Kendon, 1996], is to annotate and analyze at a more fine-grained, level. His goal is to develop a "kinetics" of gesture, analogous to the "phonetics" of speech. As such, he develops in [Kendon, 2000] a scheme which captures how joints are bent, how the different aspects of the arm move, and even how these different dimensions of gesture align with speech. This system describes positions and changes in position of the speakers arms, hands, head and torso. Unfortunately, from our perspective, the annotation scheme was designed to be written on paper or to be used with a word processor. As such, there is not a sufficient way to do fine-grained time alignment of gesture to speech. FORM's original motivation was to computerize this scheme so that fine-grained time alignment was possible. Kendon's work is the fundamental starting point for FORM.

A.2 Computer-based Annotation Tools and Systems

A.2.1 CHILDES/CLAN. The CHILDES/CLAN system [MacWhinney, 1996] is a suite of tools for studying conversational interactions in general. The suite allows for, among other things, the coding and analyzing of transcripts and for linking those transcripts to digitized audio and video. CLAN supports both CHAT and CA (Conversational Analysis) notation, with the alignment of text to the digitized media at the phrase level.

The CHILDES/CLAN system has the major advantage of being one of the first of its kind. The CHILDES database of transcripts of parent-child interactions has dramatically pushed forward both the theory and science of linguistics and language-acquisition. Additionally, it appears possible—in the future—to integrate FORM data with that developed by CHILDES/CLAN into a unified data set. This is due to the open-ended nature and extensibility of both systems. However, from the perspective of actually annotating videos with fine-grained, time-aligned gesture data, CLAN presents a problem. It is possible to describe the gesture that occurred during an utterance, but, given that time alignment is only at the phrasal level, we are unable to finely associate the parts of the gesture with other aspects of conversational interaction.

A.2.2 SignStream. SignStream [Neidle et al., 2001] allows users to annotate video and audio language data in multiple parallel fields that display the temporal alignment and relations among events. It has been used most

extensively for analysis of signed languages. It allows for annotation of manual and non-manual (head, face, body) information; type of message (e.g. Wh-question); parts of speech; and spoken-language translations of sentences.

Although SignStream would work with the FORM annotation scheme, and there has been some attempt at integrating the two projects, its interface is too comprehensive. Anvil, described below, allows an annotator to quickly see the relationship among all the aspects of left arm, right arm, head and torso movement.

A.2.3 Anvil. Anvil [Kipp, 2001] is a Java-based tool which permits multi-layered annotation of video with gesture, posture, and discourse information. The tags used can be freely specified, and can easily be hierarchically arranged. See Figure 1.12 as an example.

Anvil is the tool of choice for the work done in the FORM Lab. It works well for creating multi-tiered, hierarchical, time-aligned annotations. In the beginning of FORM, we toyed with the idea of building FORMTool, our own annotation tool. However, we soon realized that we were just duplicating the benefits of Anvil. Additionally, the extensible nature of Anvil will allow for the development of an Annotation Graph plug-in, so our data can be directly exported to AG format. Currently, we save the data in Anvil XML format and convert to AG format. This future plug-in will avoid this step.

A.3 Systems for Computational Analysis and Generation

A.3.1 VISLab: Francis Quek. The VISLab project[11] is a large-scale, low-level-of-analysis research project developed and led by Francis Quek at Wright State University. It has achieved significant results in understanding the relationship of speech to gesture. See [Quek et al., 2001], for example. The long-term intent of the project is to create a large-scale dataset of videos annotated with information about gesture, speech and gaze.

This project is in the same spirit as the FORM project, and there has been significant collaboration. There are plans for the VISLab to store the gesture aspects of their data in the FORM format. There are, however, major differences between the two projects. Firstly, FORM aims at developing a mid-level representation that humans can use to annotate gestures *and* that machines can use to analyze and generate gestures. The VISLab system's level of representation is much lower. They use multiple cameras to extract 3D information about position, velocity, acceleration, etc. concerning a gesture. They are doing the *physics* of gesture, where FORM is looking at something closer to the *phonetics* of gesture. Secondly, the VISLab system requires a complex set up

[11] http://vislab.cs.wright.edu/

of multiple, precisely-positioned cameras and proper placement of the subjects in order to gather their data. FORM allows any researcher with a notebook PC and a video camera to generate useful data. Thus, FORM can be used in the "field," where the VISLab system requires a laboratory setting.

A.3.2 BEAT Toolkit: Justine Cassell et al. The other important, large-scale project is the Behaviour Expression Animation Toolkit (BEAT) [Cassell et al., 2001]. It was developed at the MIT Media Lab in the Gesture and Narrative Language Research Group. This work is advanced and is, by far, the most influential to date. It allows for the easy generation of synchronized speech and gesture in computer-animated characters. The animator simply types in the sentence that he/she wishes the character to say, and the BEAT Toolkit generates marked-up text which can serve as the input to any of a number animation systems. The system is extensible to many different communicative behaviours and domains. The output generated for a given input string is domain specific, and the training data for that domain must be provided.

The main purpose of BEAT is to appropriately schedule gestures (and other non-verbal behaviours) so they are synchronized with the speech.

> The BEAT toolkit is the first of a new generation (the *beat* generation) of animation tool that extracts actual linguistic and contextual information from text in order to suggest correlated gestures, eye gaze, and other nonverbal behaviours, and to synchronize those behaviours to one another. For those animators who wish to maintain the most control over output, BEAT can be seen as a kind of "snap-to-grid" for communicative actions: if animators input text, and a set of eye, face, head and hand behaviours for phrases, the system will correctly align the behaviours to one another, and send the timings to an animation system. For animators who wish to concentrate on higher level concerns such as personality, or lower level concerns such as motion characteristics, BEAT takes care of the middle level of animation: choosing how nonverbal behaviours can best convey the message of typed text, and scheduling them.[12]

FORM's relationship to BEAT is more one of potential partners than as competitors. BEAT is most concerned with the automatic generations of the timings, and the higher and lower levels, as aforementioned, are left to the animator. In particular, the lower level of specifying motion characteristics is where FORM is most concerned. We see FORM as potentially a more robust way to specify the gestures for which BEAT schedules the timings. The typology of gestures that BEAT uses is based on the work of McNeill[13]. As such, it sees gestures through the eyes of his ontology. It is, then, left up to the animator to specify exactly how a beat or a deictic, for example, is to be animated.

[12] [Cassell et al., 2001, page 8].
[13] Cf. discussion of McNeill, above.

We believe the data generated by the FORM annotation system will allow for a more robust output from BEAT, which would further alleviate the work of animators.

References

Bird, S. and Liberman, M. (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.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). BEAT: The Behavior Expression Animation Toolkit. In Fiume, E., editor, *Proceedings of SIGGRAPH*, pages 477–486. ACM Press / ACM SIGGRAPH. http://citeseer.ist.psu.edu/cassell01beat.html.

Kendon, A. (1996). An Agenda for Gesture Studies. *Semiotic Review of Books*, 7(3):8–12.

Kendon, A. (2000). Suggestions for a Descriptive Notation for Manual Gestures. Unpublished.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of Eurospeech 2001*, pages 1367–1370, Aalborg, Denmark.

MacWhinney, B. (1996). The CHILDES System. *American Journal of Speech-Language Pathology*, 5:5–14.

Martell, C. (2002). FORM: An Extensible, Kinematically-based Gesture Annotation Scheme. In *Proceedings of International Language Resources and Evaluation Conference (LREC)*, pages 183–187. European Language Resources Association. http://www.ldc.upenn.edu/Projects/FORM.

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

Neidle, C., Sclaroff, S., and Athitsos, V. (2001). SignStream: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data. In *Behavior Research Methods, Instruments, and Computers*, volume 33:3, pages 311–320. Psychonomic Society Publications. http://www.bu.edu/asllrp/SignStream/.

Quek, F., Bryll, R., McNeill, D., and Harper, M. (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.vt.edu/Publications/2001/QueBMH01.html.