



Chapter 4

Annotating multimodal indicators of viewpoint shift in conversational narratives

Abstract: There is an increasing body of research on multimodal communication which uses corpora that are annotated for co-speech gestures and other bodily actions. In this chapter, we address the concern that the annotation schemes developed for such research tend to be presented as ready-to-use systems while the research practice in fact involves delicate processes of adaptation and learning from the data, which most publications do not describe. We discuss these processes as we encountered them when collecting a multimodal corpus (5.5 hours of video data) and annotating speech and bodily actions in 704 quotations. We focus on the systematic use of qualitative inter-annotator assessments, driven by discussion, which were used to optimize the annotation scheme applied to our corpus. We argue that approaching annotations as a process, rather than product, will aid in the creation of high-quality annotations and that increasing methodological transparency will help to establish sound methodological practices in the field of co-speech gesture studies.

4.1 Introduction

One of the principal concerns in research on multimodal communication is how to effectively represent and describe the actions of multiple articulators (such as the head, facial expression and gaze, the hands and arms, torso and legs) in coordination with speech so that these complex bodily actions may be fruitfully analyzed.^{13,14} Consequently, there are many different kinds of annotation schemes which attempt to capture, in as meaningful and objective a way as possible, the complex coordination of these behaviors. There are general-purpose, wide-scope schemes like the NEUROGES-ELAN system (Lausberg, in press), CoGesT (Trippel et al., 2004), and LASG (Bressem, Ladewig & Müller, 2013), which focus on the relationship of speech to manual co-speech gesture, or FACS (Ekman, Friesen & Hager, 2002) and Rossano's gaze notation system (Rossano, Brown & Levinson, 2009) which focus on the movements of particular non-manual articulators. Other annotation schemes focus more narrowly on particular phenomena to answer specific research questions concerning, for example, perspective taking via multi-articulator behavior in spoken and signed environments (Earis & Cormier, 2013), or the relationship between co-speech gesture and math ability (Alibali & Kita, 2010). In almost all cases, descriptions of these annotation schemes are partial in the sense that they describe in general terms the processes which were used but leave specific information about the variables and values of the annotation scheme implicit (see Chapter 3 of this dissertation). Although this may be sufficient for general-purpose schemes like NEUROGES-ELAN, for which additional publications and documentation exists, it is insufficient for the narrowly-defined and used annotation schemes which are prevalent in co-speech gesture research (e.g., Perniss & Ozyurek, 2015; Cormier, Smith & Sevcikova, in press, So, Kita & Goldin-Meadow, 2009).

In fact, examining the methodological literature in co-speech gesture research reveals a peculiar imbalance: One can find detailed information about a number of general-purpose annotation schemes such as NEUROGES-ELAN (Lausberg, 2013) or those used by the Grammar of Gesture project (e.g., Bressem, 2013; Ladewig & Bressem, 2013). Equally available is technical information about the software used to implement those schemes (e.g. ANVIL, ELAN), or the existence of certain public/private corpora and the specific way in which annotation schemes were applied to them (e.g., Chen et al., 2006; Blache et al., 2009; Johnston, 2010; Cormier et al., 2012; Brone & Oben, 2014). What is missing in this methodological picture, however, is information about the methodological processes used, for example, in developing an annotation scheme to suit a research question and adapting that scheme so that it is as intersubjectively-verifiable as possible. For example, NEUROGES-ELAN is often described as the outcome of “a process of repeated testing of

¹³ This chapter is adapted from: Stec, K., Huiskes, M., Cienki, A., & Redeker, G. (2015). Annotating bodily indicators of viewpoint shift in oral narratives. *Manuscript in preparation*.

¹⁴ This chapter introduces the procedure used to annotate our corpus. Information about the content of the annotation scheme is presented in the methods sections of the empirical investigations presented in Part II of this dissertation.

the categories and values” (Lausberg, 2013: 1024), but specific information about that process is not made available, whether for the development of the NEUROGES system itself or for the means by which it was tested and refined. As Chapter 3 of this dissertation reports, there is, in general, poor methodological reporting in co-speech gesture research: Enough information is presented to have a general idea about how a study was conducted, but specific information – the kind of information a junior researcher might use when conducting their first gesture project – is often lacking.

In co-speech gesture studies, information about how to collect corpus data (i.e. how participants are recruited; what text or stimuli they are presented; how actual data are collected) and how to annotate data once collected is scarce. When information is given, annotation work is treated as a final product rather than as a process which involves watching the data, creating and piloting an annotation scheme with multiple passes through one’s data, improving that scheme as one makes new observations or encounters unexpected difficulties, implementing it and obtaining (if desired) measures of inter-rater reliability. Moreover, however much we would like annotation schemes to be objective measures of “what’s in the data”, annotation schemes reflect theoretical choices and interpretations at every step of the way (cf. the discussions about transcription as a theoretical tool found in Ochs, 1979 or Bohle, 2013). After all, even the most quantitative of studies are based on qualitative interpretations insofar as they require the identification and use of labels at particular moments in the data, a process which is often rendered invisible by current methodological reporting practices.

It seems that externalizing and explicating the decisions made throughout annotation work can help shed light on this process. This is important for understanding how a study arrived at its findings. As annotations, and annotation schemes in particular, are often at the heart of co-speech gesture analyses, increasing the transparency of methodological processes would allow other researchers the chance to critique important methodological decisions. Although there are plenty of theoretical introductions to conducting co-speech gesture research (e.g., Gullberg, 2010 or Sweetser, 2006), only recently have there been specific methodological introductions which address the ethical and privacy concerns of collecting video data (see, e.g., Perniss, 2014 or Enfield, 2013), practical concerns such as recording facilities and equipment (Perniss, 2014; Enfield, 2013; Seyfeddinipur, 2011) and the tools and processes involved in performing analyses (Duncan, 2008 and Perniss, 2014; see the methods section of Müller, Cienki, Fricke Ladewig, McNeill & Tessendorf, 2013, Volume 1, for more information).

Methodological discussions such as these are important, especially insofar as they concern the training of annotators to use annotation schemes, the process of obtaining inter-rater reliability measures or the improvement and evolution of an annotation scheme over the life of a project. As far as we know, Duncan and Perniss are the only researchers to address the point of making multiple passes through one’s data and conducting finer grained analyses in subsequent passes through it; in a similar way, Ladewig & Bressem

(2013) highlight the need for the coding of gesture phases as a process which is built on a necessary interdependence of all phases of an annotation project. Stelma & Cameron (2007) provide an excellent discussion of the difficulties they experienced trying to apply a ‘standard’ transcription process to their data — and how, in the process, they found that creating one well-trained annotator who could work on the project’s entire dataset yielded more meaningful annotations than their attempts at obtaining inter-rater reliability measures with Cohen’s Kappa. To ensure intersubjective reliability, they made several systematic comparisons of files annotated by multiple annotators (their primary annotator and two external experts), and used observed differences to hone their primary annotator’s skills. Their attempts to train their primary annotator illuminated many blind spots in the literature, concerning especially the identification of boundaries for their items of interest (intonation units) and the way phonological factors interacted with them. Without their approach of ‘learning with the transcript’, rather than from measures of statistical reliability, this would not have happened.

Their annotation procedure, and the difficulties they encountered, points to the difficulties involved in trying to annotate a supposedly well-defined linguistic feature (namely, intonation units). This raises the question of how much more difficult this process may be for annotation projects with co-speech gesture data, whether for established annotation schemes like CoGesT or for annotation schemes which are developed for use in specific projects (e.g., Earis & Cormier, 2013). Although developing generalizable coding schemes such as NEUROGES-ELAN may be useful, our experience with multimodal viewpoint studies rather points to the need for a standard for evaluating and describing general annotation practices. For example, while existing methodological literature presents annotation work as a linear process akin to the workflow presented in Figure 4.1, we found that we used a recursive process which is shown in the workflow presented in Figure 4.2, namely: Following data collection, the data were watched multiple times throughout the annotation process and the annotation scheme itself was revised as a result of annotation work or subsequent viewings of the data.

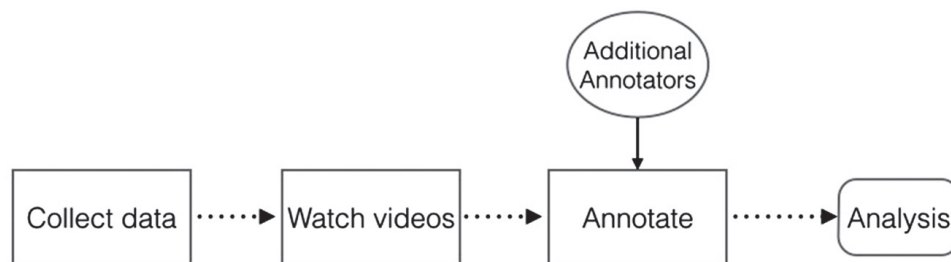


Figure 4.1: Annotating data as a linear process, where each stage of work leads directly to the next.

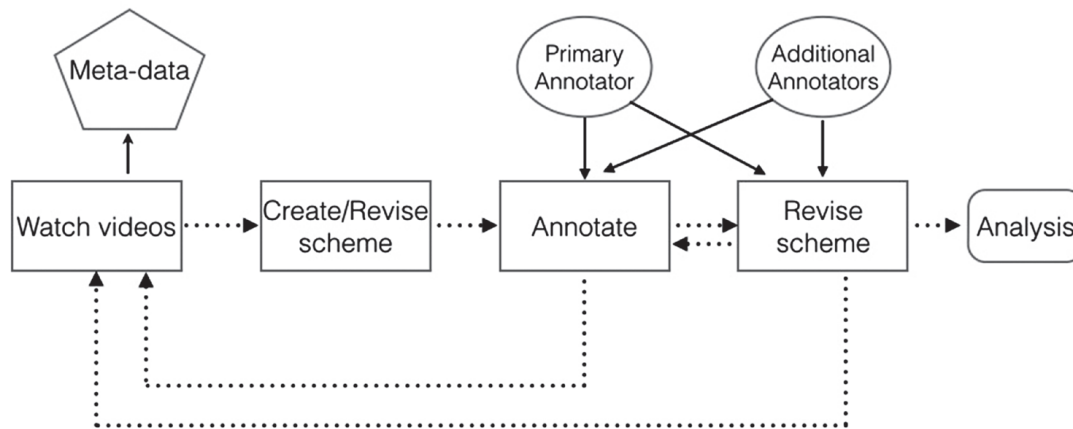


Figure 4.2: Annotating data as a recursive process.

Discussions about methodological practices take place in related fields (see, e.g., the 2013 special issue of *Dialogue & Discourse* on annotating discourse phenomena). However, we have not yet found such a nuanced discussion for co-speech gesture research. While the special issue on multimodal corpora published in *Language Resources & Evaluation* in 2007 does offer a theoretical overview of methodological concerns and provides several general-purpose annotation schemes, in our view, more attention needs to be given to the development and presentation of annotation schemes which are created for limited use. Although specific processes differ from project to project, they have certain methodological practices and approaches in common.

In the rest of this chapter, we discuss the procedures we used to obtain a skilled primary annotator. While the details are obviously specific to our project, we hope that the discussions of how to make multiple passes through one's data, with multiple revisions to an annotation scheme after discussion with multiple collaborators, will be useful to other researchers and projects as well. As others have noted (e.g., Perniss, 2014 and Chapter 3 of this dissertation), presenting scientific research as a process, rather than result, has the pedagogical advantage of aiding early stage researchers in designing and carrying out research. In addition, discussions on the strengths and weaknesses of specific methodological choices and challenges can help streamline methodological practices, e.g. by giving the community necessary information so that future research can avoid past mistakes. These methodological challenges may include technical difficulties of recording locations, the creation and use of an annotation scheme for multimodal behaviors, and establishing some kind of measure for determining the quality of those annotations — all of which, of course, depend on the aims of a particular study.

4.2 Project aims and corpus collection

As many researchers have pointed out, there is an inextricable link between research questions, theoretical framework, coding scheme and possible analyses (e.g., Ochs, 1979; Bohle, 2013). This study is part of a larger project on the use of the conversation frame¹⁵ in discourse in which special consideration is given to the role of direct speech quotation, particularly on the use of direct speech vs. fictive interaction. Direct speech is quoted speech which is reported in the first person, maintaining the quoted speaker's deictic center; fictive interaction identifies specific instances of direct speech as being functionally distinct, e.g., by voicing voice-less entities (such as animals, objects, situations, or attitudes) or quoting utterances that have not or could not have been witnessed (see Pascual, 2014 for more information about this difference).

Our contribution concerns the use of multimodal articulators – that is, visible bodily actions made by the eyes, face, hands and body, as well as paralinguistic information such as intonation – which act in coordination with direct speech utterances. As speakers and signers are increasingly shown to have similar expressive capabilities in the visual modality (e.g., Perniss & Vigliocco, 2014; Perniss, Özyürek & Morgan, 2015), especially concerning the coordination of manual and non-manual articulators (e.g., Chapter 2 of this dissertation) in narrative contexts, we aim to describe the extent to which manual and non-manual articulators are used by speakers of American English when quoting. Our research questions include: (1) How is multimodal viewpoint expressed? (2) Which multimodal articulators typically occur with quoted utterances? and (3) How similar are the multimodal actions used by speakers to the ones used during *role shift* sequences in sign languages?

4.2.1 Collection of co-speech gesture data

To answer our research questions, we decided to elicit semi-spontaneous data using a semi-structured, semi-spontaneous format in which participants told personal narratives to a friend. This is in part because previous work investigating the role of multimodal articulators used written narratives as prompts, focusing on the practices of experienced storytellers (Earis & Cormier, 2013 and Rayman, 1999). While this allowed for the manipulation of bodily actions used in co-speech gesture vs. sign, which both Rayman and Earis and Cormier were interested in, it lacks ecological validity in the sense that it does not tell us what ordinary people do when communicating. We choose to collect English data in order to compare findings from our corpus with findings from the work of Earis and Cormier and Rayman, and used a corpus-collection procedure which had previously been used to elicit semi-spontaneous interaction (Becker et al., 2011).

¹⁵ The *conversation frame* is a frame in the sense of Fillmore (1982), i.e. the structured representation of a concept. The conversation frame is evoked by speakers who produce utterances at a certain place and time, and is prototypically experienced in ordinary face-to-face communication.

The author of this dissertation (KS) collected video data on the West Coast of the US in January and February 2012. Following observations made by Henrich, Heine and Norenzayan (2010) about the overuse of WEIRD participants¹⁶ in behavioral research, we specifically aimed to collect data from participants who were outside of the age range of typical undergraduate students: namely, participants who were in their mid-20's or older. Although we did not collect education data, we know anecdotally that all participants had completed bachelor's degrees; many had also completed master's degrees. Participants were recruited using social media and word of mouth. All participants volunteered their time.

We used the data collection methods outlined by Becker et al. (2011) for their clarity and focus on the collection of semi-spontaneous interaction. Namely, (i) the 2-part consent form described below, (ii) pairs of participants whose chairs were positioned next to each other at about 45 degrees, with a wide-angle view on participants from the torso up, and (iii) a list of prompts which asked about events participants had either witnessed or experienced themselves. We made several necessary changes to Becker et al.'s methods: In order to increase accessibility to potential participants, two recording locations were used. In one location (Figure 4.3), participants were seated in chairs in the center of the room with the videocamera positioned on a table nearby. In the other location (Figure 4.4), participants were seated in chairs next to a table on which the videocamera was positioned; behind participants was another table. As both locations had many windows, actual locations of the chairs on which participants sat were altered depending on time of day. This maximized visibility of the recordings. Relative position of the chairs on which participants sat was kept at a constant 45 degrees in order to maximize visibility of co-speech gestures and give a sense of depth to the recording. However, we did not block or restrict chair movement, so sometimes, during the course of a recording, chair positions were changed by participants. Figure 4.3 shows one recording location, and Figure 4.4 shows the other.



Figure 4.3: Schematic of the first recording location, with a screenshot of data recorded there.

¹⁶ WEIRD participants are Westernized, Educated, Industrialized, Rich and Democratic participants. Although our participants are all of these things, they are at least older than the undergraduate students typically used in behavioral research (a secondary critique of Henrich et al. 2010).



Figure 4.4: Schematic of the second recording location, with a screenshot of data recorded there.

4

All participants were native speakers of American English. Participants were asked to bring a friend to one of two recording locations where they completed two consent forms: the first was completed before recording began and granted consent to participate in the study, and the second was completed immediately after recording ended and granted specific use of the materials just collected. We obtained consent in this way because of the personal, potentially sensitive nature of the narratives told. After completing the first consent form, participants were asked to tell each other personal narratives about which their friend had not yet heard.¹⁷ Some dyads improvised, and comfortably told and requested personal narratives from each other. Others used the topic sheet which had been prepared by the first author. All participants comfortably alternated the roles of telling and requesting.

Thirty-two participants (19 females) were recorded: 24 participants formed 12 dyads, while eight participated with the first author as their ‘friend’ thereby making 20 dyads. This was necessary as these eight participants were either unable to identify a friend willing to accompany them, or their friend had made a last minute cancellation. In these cases, the first author, who is a native speaker of American English, requested narratives and did not tell any of her own. Two dyads were excluded for technical issues (low light and low volume), one dyad was excluded for producing no quotatives and one dyad was excluded for producing no gestures. This resulted in 26 participants (16 females). As we had recorded approximately 15 minutes of narratives per participant, or about 30 minutes per dyad, this left us with a total usable corpus length of approximately 5.5 hours. We cut each recording into narrative-length clips, resulting in 85 narratives which range in length from 0:30 to 15:51, with an average length of approximately 5 minutes.

Immediately following data collection, we entered meta-data about each recording into a spreadsheet. An overview of this information is provided in Table 4.1. (Perniss, 2014, describes the meta-data noted by the Max Planck Institute for Psycholinguistics; their meta-data are more extensive than the meta-data we describe here.) These meta-

¹⁷ The text used to obtain informed consent is provided in Appendix A1 along with the text used to start recording (A2) and the optional topic sheets provided to participants (A3).

data concern participant information and clip information. Participant information tells us about the primary data file associated with each participant, as well as any privacy restrictions on the use of their data. Clip information provides specific information about each narrative, including technical information about the recording and the presence of particular features which are interesting for our research questions.

Table 4.1: Meta-data noted for each video file.

Variable	Value
Participant info	
Participant name	<i>Full Name</i>
Participant ID number	<i>Number</i>
Participant gender	Male, Female
Recording location	Location 1, Location 2
Restrictions on data use	Yes (<i>specify</i>), None
Clip info	
Filename	<i>dyad id numbers . clip name . mov</i>
Video OK	Yes, Profile, Dark
Audio OK	Yes, Low
Direct speech	Yes, No
Fictive interaction	Yes, No
Gesture	Yes, No
CVPT gestures	Yes, No
Role shift	Yes, No

4.2.2 Collection of psychological data

At the time of recording, we had no specific hypotheses about individual variation in behavior. However, as our familiarity with the corpus increased, we began to hypothesize about causes for the individual variation we saw in speakers' use of bodily actions. For example, some participants used relatively large gesture spaces and others used relatively small gesture spaces; some seemed to fully embody the quoted characters using their entire body – similar to descriptions of role shift in signed languages (e.g., Cormier et al., in press) – whereas others made minimal use of one or two bodily actions. Therefore, approximately one year after recording, participants were asked to complete a psychological survey. We decided to use the Interpersonal Reactivity Index (Davis, 1980; hereafter IRI) which has been widely used for assessing perspective-taking abilities, most notably in work connecting mirror neurons to empathy (Iacoboni, 2009). We specifically chose the IRI for its link to mirror neuron research, as well as for its use of sub-scales: perspective taking, fantasy, empathic concern, personal distress. We hypothesized that there would be a link between IRI scores and non-verbal behaviors, namely: participants with higher scores on the perspective taking and fantasy sub-scales would be more likely to produce multimodal quotations. (However, we have not found a significant link between IRI scores

and multimodal articulation during quotation.) Participants were contacted by email and asked to complete the 5-minute IRI Questionnaire online using Survey Gizmo (www.surveygizmo.com). At the time, Survey Gizmo was offering free accounts for scholarly research. Items were presented as a list on one page, and were scored using a 5-point Likert scale, with ‘does not describe me well’ and ‘describes me very well’ as the low and high values, respectively. All participant scores fall within the long-established ranges associated with the IRI Sub-Scales.

4.3 Corpus annotation

Various tools are available for annotating video data (see Perniss, 2014 for an overview). We choose to use ELAN (see Wittenburg, Brugman, Russel, Klassmann & Sloetjes, 2006 and <http://tla.mpi.nl/tools/tla-tools/elan/>) because it was specifically designed for multimodal analysis of language-based interaction and is widely used in the gesture community. Before beginning our project, some researchers shared their ELAN templates with us and discussed specific concerns we had about our data and using ELAN. These discussions shaped the annotation steps we took, such as using hierarchical tiers and using direct speech utterances to anchor all annotations. While these choices facilitated and streamlined the annotation process, they also restricted the questions we were able to answer. For example, although we can say what happens during direct speech utterances – based on the categorical presence or absence of certain articulators – we cannot say whether those same behaviors occur outside of quotative contexts, or anything about the time course of the different behaviors.

Our goal for the annotated corpus was two-fold: We wanted to create an annotated corpus which would reflect the variety of multimodal behaviors used by participants. These annotations were intended to serve as the basis for qualitative and quantitative analyses. We also wanted to do this by training the first author in the creation and execution of an annotation scheme. Many people contributed to this project; we consider the range and variety of experience to be a strength. We based our choices on available literature whenever possible — though specific methodological information was sparse, both for exact definitions of the variables and values used during the annotation and concerning the development and testing phases of annotation schemes used in co-speech gesture research. In particular, we noticed differences between reported annotation schemes and the schemes as they were actually used and implemented in ELAN (or Excel, ANVIL, etc.). While this convinced us that we should report our final annotation scheme – as implemented in ELAN – in the methods sections of articles stemming from this corpus, we found that, as part of the editorial process, we had to substantially revise its presentation. The annotation scheme we report is presented in Table 4.2.

Table 4.2: The reported coding scheme.

Category	Tier	Controlled Vocabulary
Linguistic Information	Transcript	Text
	Quote type	Direct speech (the utterance voices a character) Fictive interaction (the utterance is a direct speech quote which voices a character's thoughts, voices an entity which cannot speak, or refers to a future, pretend or counterfactual scenario) Unclear
	Quote sequence type	Quoted dialogue (two or more characters are quoted over successive utterances) Quoted monologue (one character is quoted over several utterances) Quote island (one quoted utterance) Other
	Quote position	Initial (first quoted utterance in a multi-utterance sequence or quoted utterance that is a Quote Island) Continuing (non-initial quoted utterance)
	Quoting predicate	Bare (no quoting predicate) Be like Say Think Other
Multimodal Articulators	Quoted character	Speaker (the speaker quotes themselves) Addressee (the speaker quotes their addressee) Speaker + Addressee (the speaker quotes themselves + their addressee) A-F (the letters A-F are used to identify other quoted characters in the narrative, e.g., all quotes attributed to "the official" are identified as A)
	Role shift	Present (the speaker demonstrates the quoted character, e.g. by showing how they looked or felt) Absent (the speaker doesn't demonstrate the quoted character) Unclear
	Character Intonation	Present (speaker's voice altered to demonstrate the quoted character) Absent (speaker's voice unchanged) Unclear
	Hands	Character viewpoint gesture (speaker's hands demonstrate a manual action performed by another entity) Other gesture (including beats, iconic gestures which are not character viewpoint, deictic gestures, emblems, etc.) No gesture
	Character Facial Expression	Present (speaker's facial expression changes to demonstrate the quoted character) Absent (speaker's facial expression is unchanged) Unclear
	Gaze	Maintained with addressee (speaker's gaze is directed to addressee throughout the quote) Away from addressee (speaker's gaze is not directed to the addressee throughout the quote) Late change (speaker's gaze moves away from the addressee after the quote started) Quick shift (speaker's gaze jumps around throughout the quote) Unclear

	Posture	Horizontal (the speaker moves in a horizontal direction)
	Change	Vertical (the speaker moves in a vertical direction)
		Sagittal (the speaker moves in a sagittal direction)
		Unclear
		None (the speaker's body does not move)
Notes	Notes	Notes

In the rest of this chapter, we discuss the process by which we created this annotation scheme and rendered it “annotator-friendly”. Although we spent considerable time developing and implementing our initial annotation scheme (Sections 4.3.1 and 4.3.2), multiple rounds of annotating and assessing this scheme (Section 4.3.3) were needed before we arrived at the final annotation scheme (Section 4.3.4) which was implemented on our dataset.

4.3.1 Developing the initial annotation scheme

We began our project with a literature review (Chapter 2 of this dissertation) and a pilot study of TED Talk videos. The literature review identified features of interest, and the pilot study suggested to us that, in line with findings from Rayman (1999) and Earis & Cormier (2013), both manual and non-manual actions played a role in the production of multimodal direct speech utterances. We were especially interested in the role that the organization and use of gesture space might play, but later had to amend this aspect of our research questions to accommodate the annotators’ ability (or lack thereof) in coding this information.¹⁸ Given this dual interest in bodily action and linguistic action, we developed an annotation scheme which we hoped would capture these features of interest. Our scheme is primarily based on an overview of the manual and non-manual means by which viewpoint is expressed by speakers and signers (Table 2.1). As part of this process, we watched the entire corpus several times, thereby familiarizing ourselves with the data and allowing us to identify new features of interest which were included in our initial ELAN template (Table 4.3).

Annotation of the corpus was done by the thesis author (KS) in collaboration with her supervisors (AC, MH and GR). KS, designed the annotation scheme and annotated the entire corpus — this was her first major annotation task as a gesture researcher, hence our focus on the development of her skills. MH, an expert in conversation analysis, annotated a subset of the data and contributed to the design of the annotation scheme. This was his first gesture project. AC is an expert in gesture studies and cognitive linguistics, and GR is discourse researcher with extensive experience using corpora. They consulted with KS and MH in the design and implementation of the annotation scheme described here.

¹⁸ We specifically did not want to apply a geometric algorithm (such as, e.g., the script used by Paxton and Dale (2013), to extract frame-by-frame movements) to calculate actual differences; and the tools used to calculate motion trajectories with motion capture data, while interesting, were not possible with our data.

Our goal was to be as comprehensive as possible without attempting to record everything — hence our reliance on controlled vocabularies and linguistic segmentation. The first author wrote a codebook which was used and updated throughout the study. This codebook described the annotation scheme’s variables and values, and was used to generate an ELAN template. Our initial annotation scheme is presented in Table 4.3. The variables of interest were classified in five categories on 23 ELAN tiers: Linguistic content (transcript plus six variables), role shift indicators (two variables, one of which, “Present”, is a subjective assessment of the presence of a role shift), articulator activation¹⁹ (four variables), and spatial indicators (nine variables). Finally, the annotation scheme contained a space for notes.

Table 4.3: The original coding scheme.

Category	Tier	Variable	Values
Linguistic Info	Speech	Transcript	<i>Text</i>
		Quote	Yes, No, Unclear
		Quote type	Direct quote, Fictive interaction, Unclear
		Quote sub-type	1st person, 2nd person, 3rd person
		FI sub-type	Self, Other, Hypothetical, Unclear
		Quote head	Say, Think, Be like, Be all, Go, Other
		Character id	Speaker, Addressee, <i>Letter</i> (range: a-f)
Role shift	Gesture	Present	Yes, No, Unclear
		Start	Quote head, Quote, Other, None
Articulators		Hands	CVPT, OVPT, 2xVPT, Point, Discourse, No gesture, Unclear
		Handedness	Left, Right, Both, No gesture
		Face	Yes, No, Unclear
		Gaze	Yes, No, Unclear
Space		Eyes (Start,End)	9 point space from McNeill (1992)
		Head (Start,End)	9 point space from McNeill (1992)
		Shoulders (Start,End)	9 point space from McNeill (1992)
		Hands	Away/Towards addressee, Left/Right of addressee, Up/Down, Neutral
		Body	Away/Towards addressee, Left/Right of addressee, Up/Down, Neutral
		Change	Horizontal, Sagittal, Vertical, None, Unclear
Notes		Notes	<i>Notes</i>

¹⁹ “Activation” indicates that an articulator, e.g. facial expression, is “actively” being used to represent the quoted character. For example, consider a speaker who produces a CVPT gesture for rock climbing by successively gripping empty spaces with both hands with either a neutral facial expression or a facial expression which depicts determination or fear. In both cases McNeill (1992) would classify the speaker’s entire body as contributing to CVPT, but only in the case of the co-articulated facial expression can we say that the speaker’s face is “actively” used in the expression of CVPT.

4.3.2 Implementing the annotation scheme

ELAN offers several means for creating templates. One of the fundamental choices is whether to use hierarchical tiers, which are nested. By default, tiers in the same hierarchy use the same time segment. We chose to use hierarchical tiers since this would simplify the annotation process and let us easily distinguish linguistic annotations from annotations for bodily actions. Because of our specific interest in bodily actions accompanying direct speech utterances, we chose to use the linguistic transcript as an anchor for all annotations. This means that, in our corpus, whenever an utterance is a quote, a number of annotations are made; whenever an utterance is not a quote, no annotations are made. This is shown in Figure 4.5.²⁰ We knew we would have a large corpus to annotate (5.5 hours of data, with an unknown number of quotes) so we also developed a set of controlled vocabularies and linguistic types, two features which work together to provide user-defined drop down menus for every tier in an annotation scheme. This is also shown in Figure 4.5. The combination of these choices streamlined the workflow, and let us work quite quickly once we were familiar with the annotation scheme.²¹

The top-most level of segmentation in any corpus needs to be chosen with care. As noted earlier, one of the advantages of using hierarchies is that once a time segment is selected in ELAN (see the highlighted portion of Figure 4.5), all dependent tiers use the same time segment. This makes it possible to identify a large set of features for any segment of video, which makes spreadsheet work more efficient as values are time-locked.²² We had hoped this would also make identifying action features more efficient, as most gesture annotation schemes focus on annotation of gesture stroke phases (see, e.g., Kita, Van Gijn & Van Der Hulst, 1998), which would normally require independent annotation of speech and manual gesture data. While such annotations may give accurate information about manual actions (though see Kita et al., 1998 for a discussion of relevant difficulties), they are not appropriate for the other bodily actions we were interested in, e.g. head, shoulder or torso movements, or the use of gaze. In addition, it was unclear how, after exporting the data, we would be able to identify all instances of behaviors accompanying direct speech utterances. Given our research questions, it seemed more straightforward to use hierarchies which were time-locked to specific direct speech utterances.

²⁰ This choice contrasts with choices made by some gesture researchers (e.g., Ladewig & Bressem, 2013; Lausberg, 2013) who annotate gesture actions independently of linguistic actions. Because our research questions specifically concern the link between quotations and bodily actions, we first annotated video files for quotes and later used the time sequences identified as quotes to anchor annotations made on visible bodily actions.

²¹ Once created and linked to media files, ELAN templates do not auto-update. This means that if a template is created and a corpus is fully annotated using it, any new features (e.g., a controlled vocabulary or tier) must be manually added to each file. New tiers will always be added to the bottom of the template; this will affect any export of annotations. Because of this, several rounds of piloting are highly recommended before committing to an annotation scheme.

²² Another consequence is the format of the exported annotation files. In our view, an annotation scheme with hierarchies is more easily cleaned for analysis than one which does not make use of hierarchies.

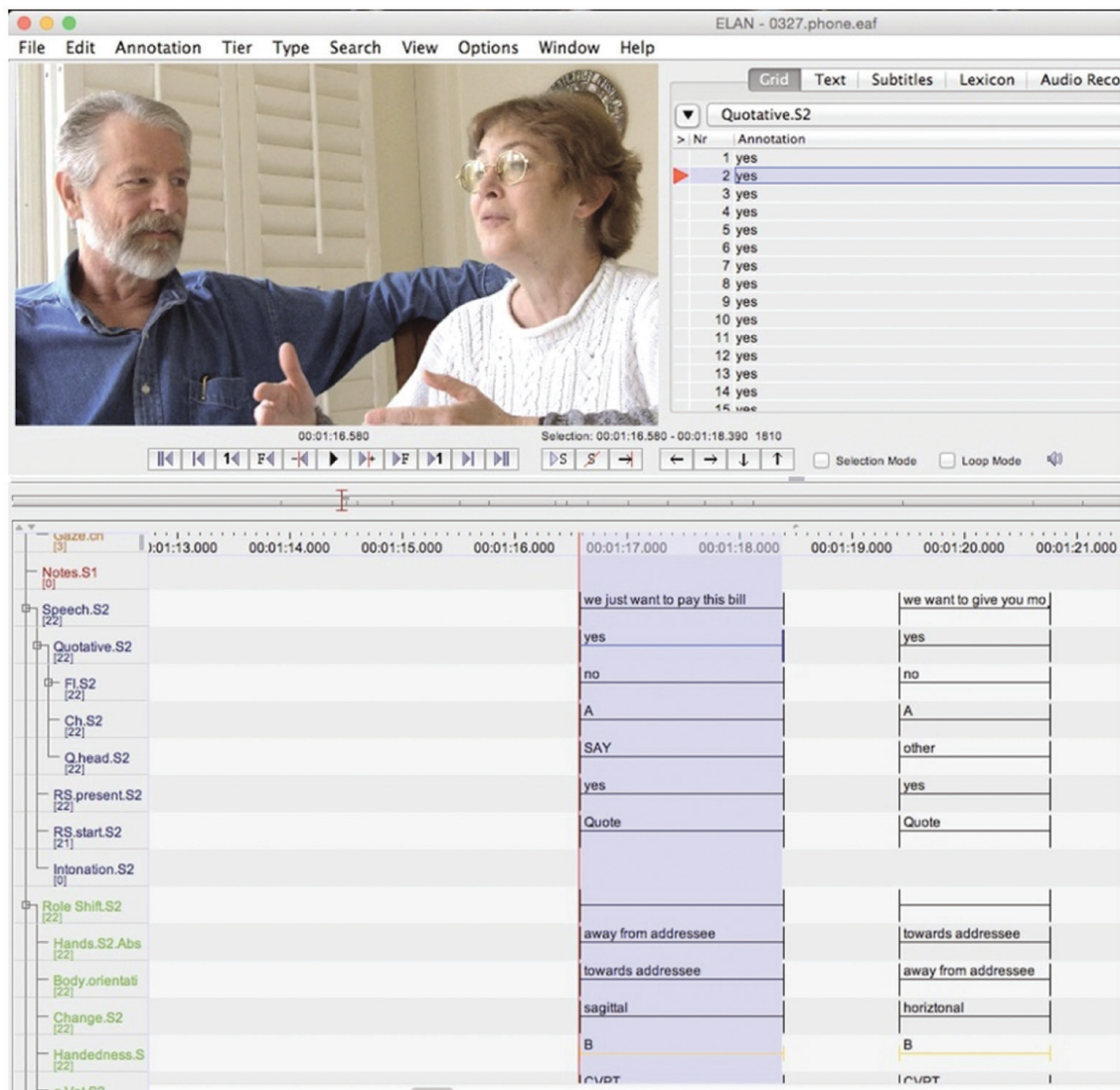


Figure 4.5: Screenshot of ELAN with a selected quote (highlighted) and an active tier, identified by the drop-down menu.

Following advice from Duncan (2008) and others, the corpus was annotated with several passes through the data: First, each clip was watched. Second, a linguistic transcript was made – initially, we made a complete transcript for five narratives, but as this proved too time-consuming, we decided to only transcribe instances of direct speech (see Buchstaller, 2013 for the identification of direct speech utterances in discourse). Annotations were only made on direct speech utterances. Third, linguistic annotations were made. Finally, annotations for bodily actions were made in chunks; often a clip was watched multiple times for articulator items, and then watched again for spatial items. In most cases, bodily actions started at the onset of the quote (the left-edge boundary, in terms of Sidnell, 2006). However, in some cases bodily actions preceded the quote – these were only included when the ‘stroke’ of the action occurred in the quote. Otherwise, they were not annotated.

4.3.3 Assessing and improving the annotation scheme

To arrive at the final annotation scheme (and annotations) for our corpus, we used a procedure which involved multiple assessments with multiple experts, indicated below by their initials. These assessments served two purposes: they (1) contributed to the skill development of the first author, and (2) ensured validity insofar as multiple experts attempted to apply our annotation scheme to our data, thereby strengthening it with their observations and comments.

Our annotation scheme was assessed three times. All comparisons were made between files annotated by our primary annotator (KS) and files annotated by someone else (MH, AS, CD or KK). AS, a sign language scholar, is experienced in sign language annotation, having worked on the development of a major national sign language corpus using ELAN. CD and KK, both gesture researchers, were familiar with gesture annotation, but this was the first time either had used ELAN. AS and CD worked remotely on their annotation task with only brief discussions with KS by email. KK annotated one narrative and discussed it in detail with KS (in person); KK and KS then annotated a second narrative together (also in person). These discussions lead to the final changes in the annotation scheme used in our project.

Our procedure was as follows, and is schematized in Figure 4.6:

- Corpus work: KS annotated the entire corpus.
- Assessing the annotation scheme:
- KS and MH completed an inter-rater reliability check using 10% of the entire corpus. This inter-rater check, for which measures of Cohen's Kappa for each variable were obtained, pointed to strengths and weaknesses of the coding scheme, and led to the adoption of different definitions and new/excluded variables for the second stage of assessments. In general, linguistic variables obtained high measures of Cohen's Kappa (all above 0.8) while non-linguistic variables obtained low measures of Cohen's Kappa (0.2 to 0.6). As will become clear, we found discussions about annotation issues to be more insightful than particular values of Cohen's Kappa.
- KS, AS and CD annotated the same 10% of the data as in the first stage, and discussed difficulties which arose. KS then generated an error report, which led to the adoption of different definitions and new/excluded variables for the third stage of assessments.
- KS and KK annotated one narrative, and discussed in detail decisions each had made. KS and KK then jointly annotated a second narrative, discussing each annotation choice. This led to the adoption of the final coding scheme, which is given in Table 3. Items which are different (compared to the initial coding scheme, given in Table 2) are underlined.
- Corpus work: KS then re-annotated the entire corpus with the final set of changes.

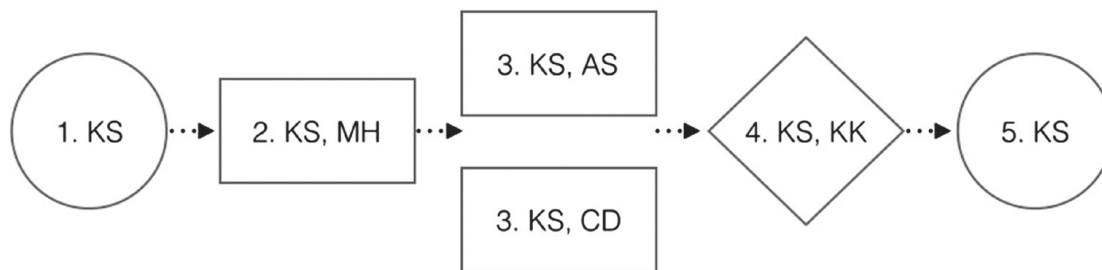


Figure 4.6: Workflow of annotations on the corpus. Each shape signifies a subset of corpus which was annotated (i.e., 1 and 5 identify the same subset of corpus, as do 2 and 3).

The evolution of the annotation scheme was driven by two types of changes: (a) different definitions — either changes to the definition itself or to the values associated with the variable, and (b) different variables — both the addition and subtraction of features of interest. Some of these changes were driven by technical challenges, e.g. the frames of the glasses some of our participants wore prevented accurate gaze information from being recorded by annotators.²³ Most were driven by challenges the annotators faced in applying the annotation scheme to the data. As a result of (re)assessment of our annotation scheme, we changed the values for three variables (Hands, Face and Gaze), added four variables (Quote context, Character intonation, Head and Change2), and removed three spatial variables (Eyes, Head and Shoulders, each with annotations made for the start and end of each direct speech utterance). One variable was doubled (Change, Change2) as discussions with annotators showed that complex trajectories (e.g., a diagonal movement with horizontal and vertical components) were difficult to agree on when only one option was possible. In the following sections, we discuss the types of changes which were made and then present the final annotation scheme which was applied to the dataset.

4.3.3.1 Refining definitions

We refined definitions in three ways: by making changes to the definition in the codebook, the controlled vocabularies, and both the definition and controlled vocabularies.

In our original codebook, we linked a number of our variables to interaction between speakers and addressees in our corpus, e.g. by noting that the speaker “[gazed] at the addressee” or “[moved] left of the addressee”. This seemed unproblematic until we realized that a number of the participants in our corpus treated the primary data collector (situated immediately behind the video-camera) as an addressee. Different annotators treated this in different ways: some kept the designated on-camera addressee as the anchor for these annotations; others were flexible in assigning ‘addressee-hood’ to either the on-camera or off-camera addressee, as appropriate. This ‘flexibility’ is a strength of speakers and the communicative situation, but problematic when trying to assign consistent coding values such as “[moved] left of addressee”. We operationalized this by editing the definitions for

²³ We accounted for this by adding an item to the controlled vocabulary for the gaze variable.

these variables, and adding a general guideline saying that, for the sake of consistency in coding practices and reference, the ‘addressee’ is always the on-camera addressee of the narrative in question.

In our initial viewing of the corpus, we observed a variation in gaze behavior which seemed to vary by speaker: some seemed to purposefully look towards their addressees during quoted utterances while others seemed to purposefully look away. We thought this would be an easy behavioral difference to capture, but it proved difficult — and not only for the technical challenge mentioned above. At each stage of the annotation process, we adjusted the values of this variable, using increasingly narrow definitions of possible gaze behavior. It was only at the final stage that we discovered that KS and KK were sensitive to similar differences in gaze behavior, but were systematically using different labels. Because of this, we made a final change which both KS and KK were comfortable with — and, crucially, were able to use comfortably, and with high agreement, in the final stage of annotation checks.

We were surprised to discover that annotators had difficulty agreeing on whether character viewpoint was expressed in manual gesture or facial expressions. We found that even using standard definitions (e.g., McNeill, 1992) annotators had strong opinions about whether character viewpoint had been expressed, were consistent about their own use of the controlled vocabulary items, were fairly confident that they were correct, and yet were inconsistent when compared to each other. To account for this, we made two changes:

- Manual gestures: We adopted a conservative definition of character viewpoint (CVPT) gestures, based on McNeill (1992), which focuses on manual actions performed by a character. Points (deictic gestures) and discourse gestures such as the Palm Up Open Hand can be attributed to characters, and thus constitute CVPT gestures, but our annotators were not able to agree on whether these gestures expressed character or narrator viewpoint (see Parrill, 2012); we therefore decided to exclude them, and changed the specification of the manual gesture values to CVPT, No gesture, Other.²⁴
- Facial expression: We adopted a conservative definition of character viewpoint expression, based on McNeill (1992), and changed the values to Yes, No so that we identify only the clearest, strongest displays of character facial expressions, e.g. fear, surprise, joy, sorrow, anger. This was done in order to minimize conflation of character and narrator viewpoint, or character viewpoint and the speaker’s meta-commentary on the quote (see Parrill, 2012 or Sweetser, 2012).

²⁴ We found this inconsistency (i.e. problems annotators faced in deciding whether or not a manual gesture should be classified as CVPT) the most puzzling, and intend to use data from the comparisons with AS and CD to inform a project about distinguishing CVPT gestures which enact characters from those which enact narrators.

4.3.3.2 Adjusting variables

We made two types of variable changes: removing variables and adding variables. These changes were motivated either by difficulties encountered in applying the variables to the data, or by new observations which emerged during the process of annotating the corpus.

Our initial coding scheme included several tiers which captured spatial information according to the nine-point spatial grid successfully used by McNeill (1992) and others. This grid uses terms like “upper left, upper center, upper right” to capture meaningful bodily actions; we had hoped to use it to capture movements of the head, shoulders and hands at the start and end of each quoted utterance. Although both KS and MH were confident in their abilities to use this system, they made fundamentally different choices in applying it: MH used the grid as is, and noted less spatial variation overall than KS — a consequence of some speakers having smaller gesturing spaces. KS, recognizing that limitation, had adapted the nine-point grid to fit each speaker’s personal gesture space. Consequently, she captured more spatial variation. Moreover, both annotators felt that the 45-degree filming angle – which added depth to the recording but prevented a direct shot of each participant – affected their coding choices. In discussion, KS and MH felt there was no satisfactory way to merge their approaches, so these variables were removed from our annotation scheme in favor of movement trajectories (e.g. “[moved] left/right of the addressee”), which were added and successfully used in subsequent checks. While this limited our ability to answer questions concerning the spatial locations used for character enactment, this did allow us to answer questions concerning the use of movement, or the direction of movement, co-occurring with quoted characters.

Following discussion with KK, who had observed with KS that intonation seemed to play a role in adopting character viewpoint, we added a variable to our coding scheme which captures character intonation in a very simple way: Yes, No. KK and KS piloted this together while jointly annotating one file, and KS later added this variable to the entire corpus. We made a subjective rather than a formal distinction; e.g. to qualify as Yes (“character intonation present”), the quote had to sound different than the speaker’s normal voice.

In summary, implementing our annotation scheme meant using a recursive process based on group ratification where each stage of assessment affected (and improved on) the existing annotation scheme. This necessitated a fluid approach, whereby features of interest, and the variables and values associated with them, were refined, adjusted or removed based on the observations, comments and annotation experiences of the multiple experts which worked on our data.

4.3.4 Final version of the annotation scheme

The final annotation scheme implemented in our corpus is presented in Table 4.4. Changes with respect to the initial annotation scheme presented in Table 4.3 are underlined.

Table 4.4: The final annotation scheme. Changes with respect to the initial annotation scheme presented in Table 4.3 are underlined.

Category	Tier	Variable	Values
Linguistic Info	Speech	Transcript	<i>Text</i>
		Quote	Yes, No, Unclear
		Quote type	Direct speech, Fictive interaction, Unclear
		FI Subtype	Self, Other, Irrealis, ?
		Quote Head	Say, Think, Be like, Be all, Go, Other
		<u>Quote Context</u>	<u>Quoted dialogue, Quoted monologue, Quote island, Other</u>
		Character	Speaker, Addressee, Speaker + Addressee, Letters A-F
Role shift	Gesture	Role Shift Present	Yes, No, Unclear
		Role Shift Start	Head, Quote, Other, None
Articulators		<u>Character Intonation</u>	<u>Yes, No, Unclear</u>
		Hands	<u>CVPT, Other, No gesture</u>
		Handedness	Left, Right, Both, No gesture
		Face	<u>Yes, No, Unclear</u>
		Gaze	<u>Yes, Late change, Quick shift, No, ?</u>
		<u>Head</u>	<u>Towards/Away from addressee, Left/Right of addressee, Up/Down, Neutral</u>
Space		Hands	Towards/Away from addressee, Left/Right of addressee, Up/Down, Neutral
		Body	Towards/Away from addressee, Left/Right of addressee, Up/Down, Neutral
		Change	Horizontal, Vertical, Sagittal, None, ?
		<u>Change 2</u>	<u>Horizontal, Vertical, Sagittal, None, ?</u>
Notes		Notes	<i>Notes</i>

4.4 Discussion

This chapter has presented the development of a coding scheme for annotating quotations and multimodal features in semi-spontaneous oral narratives. It is intended both as a description of our project and as a contribution to the methodological discussion on the development and reporting of co-speech gesture corpus annotation. As researchers, we will be better able to understand and critique each other's work if we can be as methodologically transparent as possible. As far as we know, there are no widely accepted standards for the annotation of bodily actions during face-to-face communication; easily accessible

annotation information, such as the wide-scope, general schemes like NEUROGES-ELAN, tend to focus on easy and reliable identification of the coding categories, while neglecting the need for honing the scheme for each new corpus or research question, training annotators, and consolidating the annotations by discussion and post-hoc reconciliation of differences. Annotation schemes which are developed for particular projects (such as the Earis and Cormier study cited above) eschew all of these issues by presenting what is (currently) needed for publication, despite the complexity of the task they face. Moreover, when presented thoroughly, such methodological transparency can provide the amount of information needed for the purposes of replication. Some of the important methodological features we touched on are the collection and organization of video data, and the development and implementation of a usable annotation scheme. We will discuss each in turn.

First, the process of collecting data, from recruiting participants which meet a study's selection criteria, to actually recording data and the physical, technical and ethical sensitivities which are needed to do so: Chapter 3 of this dissertation points out, general research narratives abound in the literature, but specific methodological information is still lacking. For example, previous publications in this area primarily concerned the introduction of public/private corpora, descriptions of specific annotation schemes and examples of projects for which they were designed, and technical documents which introduce some of the software tools used in gesture research. Gradually, more publications addressing specific methodological concerns are beginning to appear. There are several papers now which describe practical aspects of gesture research, such as best practices for approaching participants and collecting data (both the ethical concerns involved therein, and practical concerns such as how to record data) — Seyfeddinipur (2011) and Enfield (2013) are particularly good examples for gesture; Perniss (2014) provides similar information for collecting sign language data, which is also relevant for research in gesture. We hope that the discussion of the adaptations made to our annotation scheme as well as the meta-data used in this project can help other researchers in designing and implementing their own projects.

Second, we have emphasized the importance of the process of developing and improving the annotation scheme, and the development of annotator skills. Like Stelma and Cameron (2007), our position is that the process of annotation is a central part of rigorous research — and the steps taken during that process should be discussed not only to make the annotation schemes themselves more transparent, but to inform and thereby enable future gesture research. Moreover, this discussion will externalize the process of applying labels to data and quantifying the use of those labels — an important process, but one which is often rendered invisible by current methodological reporting practices. We gave detailed descriptions of our process so that it may serve as a guide for future studies.

To that end, we provided flowcharts which describe the annotation phases of this project. These may serve as an overview or guideline, insofar as they make explicit the fact that

not only are multiple passes through the data needed to complete an annotated file, but that multiple rounds of annotation are needed in order to refine an annotation scheme into something which is both internally consistent and externally valid. These multiple rounds are also needed in order to train annotators. As Stelma and Cameron (2007) point out, even well-defined terms require expert training. This may be even more true for the variables used in gesture research, whether for general-purpose annotation schemes like NEUROGES-ELAN or schemes which are narrowly defined for particular projects, such as what we described here.

As part of the discussion of our annotation scheme, we provided information for the means by which it evolved – including the fact that, in articles which draw from this annotated dataset, we had to adapt our reporting of the annotation scheme, both reducing the number of variables we reported and changing the way in which it was presented. Part of our discussion included tables which showed how the variables, values and definitions in our codebook changed over time. These changes were made possible by three features of our assessment procedure: the in-depth discussions annotators had together, the decision documents in which annotators noted difficult annotation decisions and their outcomes, and the annotation notes which KS kept to document progress and decisions made throughout the project. Together, these documents helped to externalize annotators' decisions — often pointing to the fact that the supposedly “objective” codebook had masked certain intuitive or internalized choices and processes used by the primary annotator. Discussions between annotators, the error reports KS generated and the decision documents each annotator kept, helped us to learn with the annotations which had been made by each annotator. Like Stelma and Cameron (2007: 386) observe for their study, this eventually led to our primary annotator working within her skill level rather than at the limits of it.

In particular, we found discussions between annotators to be highly informative. Initially, we had thought that deciding whether an articulator was ‘active’ would be easy. In fact, we found that although annotators felt decisions were both easy and straightforward, there was a surprising amount of disagreement between what pairs of annotators considered to be ‘active’ with respect to character viewpoint. In most cases, the result of these discussions were more strict, more detailed definitions. In the case of gaze, it meant adding values to the controlled vocabulary. These discussions also pointed to the difficulty of applying supposedly objective spatial terms to our data. Work by McNeill (1992) and others led us to believe that this would be relatively straightforward to implement. Instead, we found that it was difficult to obtain reliable annotations of spatial features. We made several attempts at improving annotations for this feature and eventually discovered that the difficulties were due to different applications of spatial terms; annotators were seeing the same movement, but were applying different labels to it. Once realized, it was easy to correct — but noticing it without discussion would have been difficult. Together, these changes resulted in variables we were comfortable using, and which, in our view, capture the range of bodily actions in our data. In every case, we feel that having annotators

discuss and externalize their decisions led to better annotations – in part because, as we have argued throughout this chapter, it made the process transparent. In our view, it would help the field if coding templates and codebooks were made more widely available, especially for the smaller, specific-purpose, limited-use schemes such as the one described here. We hope that documentation concerning the process of collecting and annotating gesture data will increase. Such methodological transparency will improve the community’s methodological practices.