ANNOTATIVE PRACTICE (under perpetual revision) — Due to Susan Duncan (Revision of D. McNeill (2005) *Gesture & Thought*, Appendix.)

The following talking points are meant to serve as a basis for tutoring sessions with novice gesture/speech/discourse analysts. They are based on certain observation and analytic heuristics commonly employed in the McNeill Lab at the University of Chicago. They attempt to sketch current (evolving) 'best-practice', based on much analytic experience by several cohorts of annotator-analysts working in the McNeill Lab. As a careful reading of the points will show, the procedure is very much one of hypothesis formulation, testing, revision, further testing, and finally (provisional) acceptance of every analytic judgment concerning phases of gesture, the timing of gestures in relation to speech, and gesture meanings. Such is the process required for accurate and reliable gesture analysis. The procedures are presented in eight successive 'passes' through an interval of audio-video recorded data.

One goal of the descriptive-analytic method employed in the McNeill Lab is to observe speech-gesture synchrony to a degree of accuracy that permits assessment of how meaningful gestural movements co-occur with speech, syllable by syllable. Such analysis requires the ability to play the audio-video data at varying slow motion speeds; crucially, with access to the concurrent audio track at all playback speeds, including frame-by-frame. Consumer grade VCRs do not provide the requisite playback capabilities. With tape media we have used Sony EVO 9650 Hi-8 VCRs (which are no longer manufactured), because they have excellent jog/shuttle tape control and the required audio playback functionality. In more recent years we have migrated to working with digitized ("rich" compressed) audio-video data in interfaces designed for media editing. The two such interfaces we have so far identified that have the functionality we require are Apple Final Cut Pro/Express on Mac computers and Adobe Premiere Pro on PC Windows computers. With digitized media viewed in these interfaces, we have used the jog/shuttle controller mouse manufactured by Contour Design: 'ShuttlePro' or 'ShuttleXpress'.

A further goal is to annotate speech-co-occurring gesturing with sufficient clarity, depth, detail, and consistency that:

1) other analysts who make use of the annotated transcript, or add to it, later will be able to accurately infer previous analysts' decision-making process, in regard to parsing gesture phrases and phases, and inferring gesture meanings.

2) the annotated transcript will serve as a "visualization tool" for multi-modal analyses of language that focus on how speech and gesture mesh, both at moment-to-moment and extended discourse levels of analysis.

PASS 1.

Watch the complete product of the elicitation (for example, a cartoon or movie narration, a lecture, or conversation) once, all the way through. This pass permits the

analyst to develop an initial sense of speaker 'style.' This facilitates both speech transcription and interpretations of gesture productions on later passes.

PASS 2.

Create a 'verbatim' transcript of all the words (including partials and unintelligibles) spoken in the discourse, from beginning to end, making little attempt to annotate sentence grammatical structuring or production characteristics (pauses, intakes of breaths, and so on). The 'end' of a discourse, for example in the case of one elicited with a movie or cartoon stimulus, is when the speaker or listener speaks his/her final utterances pertaining to the content of the stimulus, or pertaining to whatever elaborated, stimulus-related discussion of it speaker and listener may have developed between them. That is, a complete transcription also includes any responses either interlocutor may make to an investigator's prompts for further information about the stimulus, but need not include added-on conversation about other topics that may have been captured on the tape by accident.

PASS 3.

Organize the speech into short utterances, reflecting the (sentence-approximating) grammatical structuring of the speech sequences (and/or larger intonational contouring of intervals of speech; a matter of analyst preference); that is, break up the stream of speech onto separate lines in units such as 'sentences,' 'clauses,' or intonation units. Use the typographic speech-annotation conventions given in section # of this appendix, or some other system that captures dimensions of speaking of interest to the analyst. In all instances typographic conventions are a matter of analyst preference and different schemes have different virtues.

Annotate:

A. the passage of time by periodically inserting, on the left-hand side of the typewritten transcript, the time stamp that appears on the video image;

B. Pauses ('unfilled' and 'filled'), breaths (intakes and exhalations), non-speech sounds such as laughter and audible mouth noises, and so on;

C. 'listener' (in the case of a quasi-monologic speaker narration) contributions ("mm-hm", %laugh, [nod], and so on.)

Save this non-gesture-annotated speech transcript separately for use in analyses for which only the speech is of interest, taking care to keep it updated, when repeated listenings during subsequent, gesture annotation, passes reveal errors of speech transcription.

THE FOLLOWING PASSES ARE EXECUTED RECURSIVELY ON MULTILINE CHUNKS RATHER THAN ON THE DISCOURSE AS A WHOLE

PASS 4.

On a copy of the speech transcript developed in PASSES 2-3, annotate points of primary peak prosodic emphasis (and secondary emphasis if the opportunity arises), assessed by ear (preferably, native-speaker ear). Use enlarged font (not capitals) for

these annotations. Limit the enlargement to the syllable(s) that your ear tells you is/are prosodically emphasized. (N.B.: lexical stress can complicate these judgments).

PASS 5.

Square-bracket the gesture phrases. Do this exhaustively across the discourse from beginning to end. That is, leave un-annotated no intervals of speech that co-occur with hand gestures or gestural movements of other body parts. Use the gesture-annotation conventions given in section # of this appendix, or another system that captures dimensions of speaking of interest to the analyst.

PASS 6.

Annotate the within-phrase phase-structure of gestures.

A. Locate gesture stroke phases through a process of comparing meanings the hands (or other body parts) appear to express, with meanings conveyed in co-occuring speech (considering individual words and phrases, but also more comprehensive discourse units). Take into account also the gesture movement dynamics, as the stroke phase of a gesture will typically (but not always!) be the interval of apparent greatest gestural effort; determination of 'effort' made with reference to parameters such as relative forcefulness of movement or apparent tenseness of handshapes, and so on. Annotate the extent of a stroke phase in relation to speech using bold face font.

1. Assess the location of the stroke first at full and 1/5th playback speeds
2. Fine-tune at 1/10th and frame-by-frame playback speeds (N.B: always while listening to the co-occurring speech).

   N.B. Step (6.A.2) is important not only if the analyses planned for the data are to take precise speech-gesture synchrony into account, but also generally, in that fine-grained observation often spurs reassignment of phases.

   A weird empirical fact: Gestures' apparent locations in relation to the speech stream can migrate very slightly right-to-left (in relation to the type-written sequence of spoken syllables), or backward in 'speech time' when viewed at progressively slower and slower playback speeds; e.g., a gesture stroke that, viewed at full speed, appears to synchronize with "down" in the phrase "rolls down", may appear at frame-by-frame speed to synchronize instead with "rolls". Furthermore, fine-grained analysis, dependent upon multiple viewings at slow motion speeds, tends to make additional, distinct gesture phases visible that are too small to be observable at faster speeds.

B. With annotation of gesture phrases ([...] demarkating the co-occurring speech) and stroke phases (**bold** font for the speech with which the stroke co-occurs), preparation and retraction phases of gesture are *de facto* also annotated. That is, the interval of speech between a left bracket and onset of bold font is what co-occurs with the gesture preparation phase and the interval between offset of bold font and the right bracket is the interval of gesture retraction. Bear in mind that there is nothing to prevent a speaker from launching a new stroke phase immediately upon termination of the preceding stroke phase. That is, a gesture phrase may lack preparation and retraction phases.

N.B. The question often arises: Where does one gesture phrase end and another begin? That is, in regard to a movement phase between gesture strokes, there is often a question of whether to interpret the movement as a retraction phase of the preceding gesture or the preparation phase of the following gesture. Recommendation: absent (gesture featural) evidence to the contrary, decide most times in favor of preparation phase of the following gesture. This recommendation reflects the working assumption that gesturing is a largely forward-directed activity, reflective of idea units yet to come in speech.

C. It seems to be at this point of dividing phrases into phases (preparation, stroke, holds, retraction), that the 'nested' nature of some gesture productions becomes apparent. Phrase-within-phrase nesting may be annotated using outer and inner brackets, like this: [[…][…]]. Two or more gesture phrases may be considered to function as part of a more encompassing gesture phrase when there is some maintained imagistic feature (realized, for instance, as a handshape, a body orientation, a marked spatial location, or similar) that they share and are unified by (on some level).

D. Annotate hold phases with underlines, distinguishing full holds (solid underline: no detectable movement) from 'feature' or 'virtual' holds (dotted umder1ine: some movement, but maintenance of handshape and/or general location in gesture space, for instance).

E. Below the line of gesture-annotated speech, enter information concerning assessment of and interpretation of the gesture, including:

1. Dimensions of semiosis the gesture manifests; e.g., iconicity, metaphoricity, deixis, rhythm-marking (beat), interactive, emblem, pragmatic, and so on. In one commonly accepted analytic framework this step is reduced to classifying a gesture as being of one 'type' or another, e.g., an iconic or a pointing gesture. That approach is based on a different concept of gestures than this guide assumes.

N.B. In the approach advocated by this guide, 'type' designations are meant solely as 'convenient handles,' designed to further <u>one target analysis</u> or another. We do not consider these semiotic dimensions (iconicity, metaphoricity, and so on) to function as mutually exclusive categories. Indeed, any minimally close observation of natural gesturing reveals that they do not *function this way. An essential descriptive-analytic concept that governs this aspect of our work is that these semiotic dimensions 'layer' in gestures. For example, any gesture (whether labeled 'iconic', 'deictic', or 'metaphoric') whose stroke coincides with a point of speech prosodic emphasis is analyzed as being also, underlyingly, a beat (Tuite 1993; see also Loehr 2004). Also, gestures are typically placed at particular locations in gesture space. They therefore manifest deixis, or, we say, all gesturing is 'deictically framed' (Liddell's analysis, e.g. 2003, of indicating verbs in ASL exploits a similar assumption). Consider, too, that the type categories that constitute coding schemes for much gesture research cannot really be understood independently of one another; for example, a metaphoric gesture is an iconic gesture, in that its form is a depiction of the base of some conceptual metaphor. In summary, though it may be true that some gestural productions may be accurately construed to be 'loaded' more on one semiotic dimension than another, virtually every gesture production is assumed to manifest*

multiple dimensions. A gesture analytic strategy too dependent on 'type'-ing gestures is likely to obscure many phenomena of interest in natural discourse data.

2. Indication of whether the gesture is, in overall form or in some feature, a repeat of, or related in form, location, motion, or some other feature, to another gesture in the preceding discourse.

3. A description of the physical form of the gesture, including handshape, location, and movement characteristics. Use the coding conventions described in section # of this appendix, or some other scheme that captures dimensions relevant to the target of the particular analysis.

4. The inferable meaning of the gesture.

5. If necessary, for difficult-to-analyze cases, notes about the process of inference that resulted in the descriptive hypothesis (see no.9, below), recorded in the transcript, about a gesture's phrase and phase structure, and/or meaning.

6. Notes to support various specific analytic purposes; for example, the gesture's use of space, specifics of speech-gesture synchrony, character of speech prosodic patterning in relation to the gesture's stroke phase, connections between this speech-plus-gesture production and the larger discourse frame, and so on.

F. The exercise of analyzing and annotating speech-co-occurring gesture (because it is dependent on repeated, slow-motion viewing) causes previously-overlooked aspects of speech production to become evident. Therefore, on PASS 6, one adds in or modifies all the many:

1. speech pauses missed on PASS 3;

2. intervals of dysfluent speech missed on PASS 3;

3. words and phrases that are now heard differently;

4. listener productions overlooked or misheard on PASS 3; and so on.

N.B. Be sure to make all such modifications as well to the separate, speech-only transcript, saved earlier.

PASS 7.

Reorganize the manner in which the transcript was earlier broken up into short utterances (PASS 2) in accord with what the gesture phraseology reveals about the organization of speech/gesture 'production pulses'. (A 'pulse' is a unit of speaker effort, encompassing prosodic highlighting, discourse highlighting, a gesture phrase; also, gaze, posture, and other dynamic factors – clearly, then, a judgment reflecting the analyst's final hypothesis concerning the organization of the example under analysis.)

N.B. There are analytic purposes for which the ideal, final, speech-gesture annotated transcript is organized – as far as is possible – as one production pulse per line, even when some of the pulses are quite short in terms of utterance length (a phrase or even single word). Often, these short utterances will not correspond to grammatical units (such as a clause or phrase).

PASS 8.

The exercise of gesture analysis and annotation is necessarily backward-adjusting. As the analyst moves forward through the narration from segment to segment, insights accumulate about how the particular speaker typically executes certain types of gestures, the speaker's handshapes, what is typical of the speaker's gestures during intervals of dysfluency (for instance, holding *versus* repeating gestures across such intervals); on and on. Multitudes of tiny insights accumulate. An interval of gesturing at discourse segment no.47 may require annotation that calls into question how an interval at segment no.33 was annotated (at any level: gesture 'type', gesture meaning; any aspect). The analyst studying segment no.47 is obliged to return to segment no.33 and re-do the annotations or add a note of some kind.

A. When an analysis is to be based on a sample size of $n > 1$, if the insights gained from annotating speaker no.25 call into question annotations for speakers nos.3, 4, and 15, the annotator is obliged to go back and adjust the annotations on those speakers' transcripts.

B. Item (8.A.) may be especially important when $n > 1$ analyses incorporate a subject-grouping variable; for example, language (e.g., English/Spanish/Chinese) or brain-language pathological condition (e.g., non-brain-damaged *versus* left hemisphere stroke). Insights gained from annotating transcripts of speakers in group 1 may be relevant to some dimension that crucially distinguishes group 1 speakers from groups 2 and 3. Annotations on transcripts for multiple speakers within all three groups may need to be adjusted to reflect the new insights.

CAUTIONS

A. The majority of speech-gesture co-productions display relatively transparent semantic co-expressivity.

B. However, some proportion of all gestures are vague or ambiguous; either:

   1. At the level of the totality of data we have to muster in support of competing hypotheses concerning a phrase parse or the meaning of a phase, within the universe defined by the audio-video data we collect on an individual discourse, or,

   2. At the level of speaker 'speech-thinking'. That is, at some moments in a narration a speaker's speech-thinking representations may simply be a bit indeterminate or confused. A speaker may have inadequately distinguished, competing notions in mind simultaneously. Such facts of speaker mental state will manifest in gesture, yielding behavior whose meaning is difficult to infer.

C. Gestures pattern in multiple levels simultaneously. They are multifunctional. Therefore, various hypotheses about them may all be supportable. The 'genius' of gesture is that it does not have to segregate meanings into temporally distinct units. Multiple meanings and functions can load up in a single production pulse; no problem. Some hypotheses about gesture parses and meanings that the analyst may consider, though, are truly in competition with one another, at a given level of analysis. For such, the hope is that evidence available from the discourse as a whole will aggregate in support of one hypothesis over any other.

D. It is important to retain access to all reasonable hypotheses (those not disconfirmed by all available evidence) about each gesture production. Currently, this is how we attempt to meet this requirement:

1. Type all hypotheses concerning a production into the transcript, with, if necessary, the reasoning underlying each.
2. Procedure for dealing with difficult-to-interpret gestures:
    a. Enter a note about the difficulty, take a stab at formulating an hypothesis about the phrasing, function, and/or meaning of the gesture, and move on with annotating the transcript, promising to return.
    b. Return, either:
        i. After annotating some more of the discourse yields insight about the problem case, or
        ii. After just annotating some more.
    c. Upon returning:
        i. Ponder the problem case, considering the speaker's discourse locally and narrowly as well as broadly. Include consideration of what is known, as well, from having analyzed other speakers speaking about similar content.
        ii. Fine-tune parsing and annotations, if necessary. Also, if the decision about how to parse a gesture and interpret it crucially hangs on small details of changes in motion or handshape, insert time stamp(s) from the video media above the line of speech-gesture, indicating where these articulations occur, and/or elaborate the descriptive text associated with the production to make clear the reasoning underlying the decision.
        iii. Refine, change, or eliminate individual hypotheses of (D.1.) and add a note to the transcript stating the evidence for the changes, or for retaining any hypothesis. The latter is necessary because subsequent analysts may have the same difficulties with the same production. They will find assistance in such notes; also, because without such notes, an analyst wanting to incorporate the production in an analysis may do so without noticing that there is something problematic about it.
    d. Know that it is not possible to interpret the meanings and functions of every individual gesture (See B.1&2, above).

E. Under-specified in the above is the foundational issue of how a stroke's meaning is inferred. Many of the steps in the procedure outlined above interact with this issue. The approach to gesture phraseology and phaseology, advocated in this guide, is **in its essence** meaning-driven. Locating the beginnings and ends of gesture phrases, or locating the gesture stroke among a movement's several phases of execution is a matter of how the phases (movement or hold phases) coordinate, in terms of meaning, with units of the co-occurring speech, and/or with larger-scale discourse meanings currently in play. One goal of the descriptive-analytic exercise is to try to observe where, in a sequence of movement-speech or hold-speech pairings, the two modalities

seem linked in meaning, at one or more levels of discourse analysis. Inferring the meaning of a gesture stroke is an act heavily influenced by considerations outside the particular speech-gesture production pulse the analyst is working on. To be adequate, the process must draw on the larger discourse frame(s) that the pulse is embedded in, what meanings are emerging sequentially in the speaker's utterances, what viewpoint the speaker is embodying, what this speaker typically does with his hands in gesture, and so on; also, in the case of stimulus-elicited narrative discourse, what stimulus-derived image the speaker likely has in mind at the moment of speaking. An assessment of gesture meaning based solely on physical features of the gestures in a single production (e.g., movement dynamics, handshape features) **will be inadequate**.

The experienced analyst expects and does not avoid dealing with whatever phenomena may emerge to complicate an interpretation or analysis. The essence of the approach advocated in this guide is analytic and annotative flexibility. Our analyses are conceived of as basic linguistic descriptive work rather than as 'coding' exercises.

RISKS INHERENT IN DEVIATING FROM THE ABOVE PROCEDURE

1. Time wastage.
2. Bogging down.
3. Commission of avoidable errors.
4. Production of transcripts that are sketchy and/or internally-inconsistent at the level of their annotated representations.

FINAL CAUTION: A SPEECH TRANSCRIPTION CAN APPROACH A STATE OF COMPLETION. GESTURE ANNOTATIONS TO IT (LIKELY) NEVER DO.