Interaction Behavior Annotation Protocol

This document contains the description of an annotation protocol for expressions used in human interaction. This protocol is thought to obtain the most accurate annotations for such expressions with the least amount of time possible. The annotations obtained have several purposes going from pure phenomena analyses to building machine learning-based systems.

A template for annotating with the ELAN software can also be obtained along with a processing toolkit in projects like: CBA toolkit.

In case you use this document please cite:

Kevin El Haddad, Sandeep Nallan Chakravarthula, and James Kennedy. 2019. Smile and Laugh Dynamics in Naturalistic Dyadic Interactions: Intensity Levels, Sequences and Roles. In 2019 International Conference on Multimodal Interaction (ICMI '19). Association for Computing Machinery, New York, NY, USA, 259–263. DOI:https://doi.org/10.1145/3340555.3353764

ROLES:

Modality:

Criteria: role of the guest with respect to the dialog: is the guest a speaker

Values: speaker, listener

Comments: a guest can give verbal and nonverbal feedback to the interlocutor like “really?” or “I see”. In this case the guest is not classified as speaker.

Detailed Description:
The focus here is the segments in which the guests behave as “Speaker” in a conversation. We consider a “Speaker” to be the interlocutor who is uttering a full sentence (short or long), while the “Listener” is the interlocutor to whom the Speaker is talking. The Listener is thus the one giving feedback to the “Speaker” or waiting for a response or, more generally, a message from his interlocutor. A “Listener” segment will be annotated when a guest is waiting for a response or message and the Speaker is not providing it (this is because the “Listeners” will be annotated automatically with respect to the interlocutor’s “Speaker” segments, except in this case since the interlocutor will not have a “Speaker” label). The segments will start when the guest starts speaking/sending messages and finish at the end of the sentence or any nonverbal expression (audio, facial expression or body movement) directly following the sentences if it started during the sentences like laughter or head movements.

Tips:

1. Speaker roles can overlap when one person trails off while the other person starts their own sentence, i.e. both interlocutors can be speakers for an overlapping duration of time.
2. A listener can continue a word or phrase the speaker started. In this case the listener becomes the speaker (see the first point).
3. When the speaker waits for an answer or reply, they become listeners.

S&L:

The S&L tier is the parent tier of the “Smiles” and “Laughs” tiers described below. This format is used to avoid overlapping between these two tiers as will be explained later on.

a-Smiles:

Modality: Primarily visual but also audio can be used

Criteria: smiles can be perceived through the lips but also from other part of the face such as the eyelids and the cheeks. It does not have to involve the lips (although it does most of the time).

Values: intensity levels: subtle, low, medium, high

Comments:

- Cannot overlap with laughs (it is either a laughter or a smile or none).
- Even very subtle smiles should be annotated. Even when it feels like the guest is smiling the whole time.

Detailed Description:
Smiles are annotated when they are perceived. The annotation is not based on a specific expression or facial movement. A smile does not have to be expressed exclusively with the lips. Other facial features can be a reference for smiling too, such as eyelids, cheeks, eyebrows etc. Smiling is primarily a facial expression but the audio can also be used to recognize it.

Smiles detected will be given intensity values. The intensities are chosen in a subjective way based on the intensity perceived from the expression itself (how intense and/or noticeable are the traits used by the guest to express the smile) and not on the emotion behind it. Even very subtle smiles should be annotated. A smile can be split (or not) into several consecutive smiles of different intensity.
The intensities categories are:

- Subtle: very low level smiles barely expressed. For example when very low level smiles are expressed by a subject while talking (smile is covered by articulations of some vowels like “o” for example) or when we feel the presence of a smile but not 100% sure what facial features/expression create them.
- Low: low level smiles. The difference with subtle is that these smiles are expressed enough to be sure they are present but at low intensity
- Medium: A flagrant smile
- High: A highly intense expressed smile.

The choice of the smile’s intensity, even though based on the expression, is somewhat subjective and the annotator is free to choose it as perceived.

So the smiles will first be detected and then segmented based on their intensity (see Fig.1). The choice of the intensity is somewhat subjective. **The segments will start and end from the beginning of a level to the beginning of the next level respectively. Except when a smile transitions from or to a non-smile expression (neutral or other), in this case, segment ends (or starts) at the end (or beginning) of the transition to (or from) the non-smile expression (like in the first and last part of Fig.1 where the transitions from and to the neutral states respectively are annotated as part of the smile segments)**

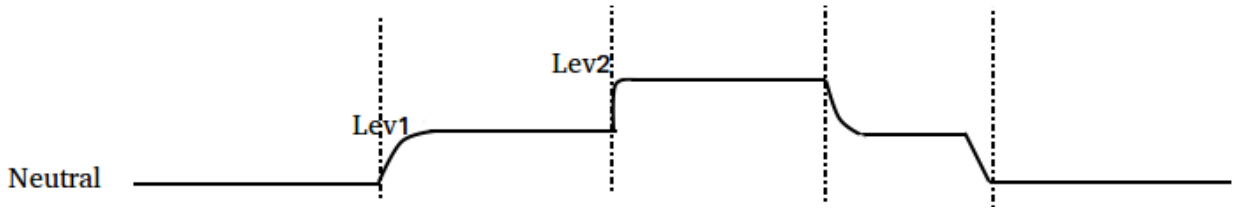

Fig1: This figure shows a continuous smile segmented into several levels. The segment starts at the beginning of the smile and stops at the beginning of the following level. Except if the following level is neutral, in that case it stops at the end of the transition to neutral.

The difference between levels (the way a change of level is detected) is based on the amplitude of the change perceived by the annotator from one smile to the next. This difference is mainly based on the speed of the change and the intensities of the 2 different expressions. An important change would mean a bigger difference between the levels (from subtle to medium for example). A smaller change means a smaller difference (from medium to high for instance). A non-significant change means that we divide the annotation in two consecutive segments but that the segments are labeled with the same intensity (2 consecutive medium smiles for instance).

A facial expression like lip pressing for example could be considered as a point of transition from one level to another of the expression (in this case smiling).

If the annotator is not sure whether the expression in a segment is a smile or not due to the very low intensity, then this segment should be annotated as a smile with the level subtle.

Be aware that mouth movements like articulatory movements while speaking for example affects the shape of the lips and the area around the mouth for a short duration. So a smile of a certain level can be annotated across an entire sentence for example, even if an “o” occured at some point making the smile less intense than the rest of the sentence for 30 ms.

Tips:

1. When annotating, first watch a small time period of the video (~3-10 seconds) and decide how many smile levels are present in the period. Then, annotate the levels and identify the specific transitions.

2. If a facial movement such as a lip press occurs and the smile intensity after the movement remains the same as it was before the movement, then the smile segment does not need to be further segmented, i.e. it can remain as a continuous segment. There must be a clear transition, even if it is drawn out, in order to create a new segment. The lip press in this case can be thought of like a comma in a sentence, rather than a period which represents a segment division.

3. If there is a change in smile intensity after the lip press occurs, then the movement is a transition and thus belongs to a new subsequent segment.
   Example in video sess2_gst1_3_68: Lip press between 1:57 and 1:58

4. There are small variations within the four levels of subtle, low, medium, and high. Further, the variations of smile movements are not necessarily linear. A smile may not escalate continuously from one intensity to the next, but rather the changes can be nonlinear with multiple local maxima and minima values (or dips and rises) all within one smile level (ex. point A and B in Fig 2).

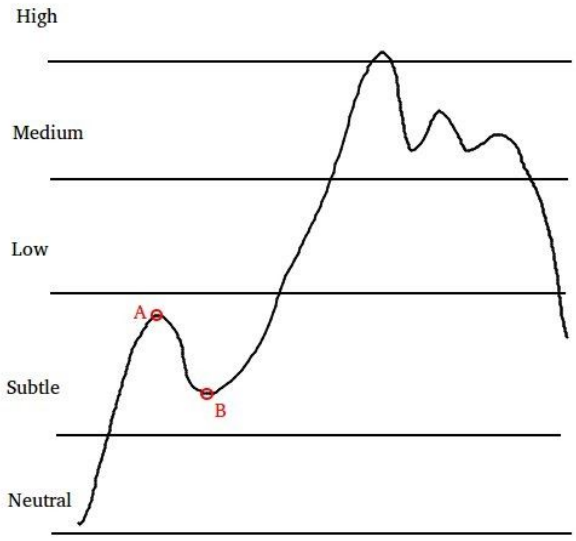

Fig 2: level division for smiles

b-Laughs:

Modality: audio and visual (facial expressions, head movement, body movements)

Criteria: laughter should differ from smile by the fact it contains either a body movement, a head movement or audible laughter related sounds.

Values: intensity: low, medium, high

Comments:

- Smiles or part of a smile at the start of laughs can be perceived as part of the laughs and so annotated as part of a laugh.
- Laughs and smile do not overlap

Detailed description:
For laughs, the segments start when an audio, facial expression or body movement event related to laughter is observed and stops when a breath intake is perceived whether audibly or visually (from the stomach, face, head etc.). The breath intake is considered part of the laugh. If no breath intake is perceived the end of the segment is considered to be when the movement stops.
In some cases breath intake sounds occur after a relatively long delay. In this case, during this delay, if the guest is perceived as laughing than the breath intake marks the end of the laugh and is part of it. Otherwise the end of the laughter is the end of the sound or movement.

So two consecutive laughs can be separated by eventual breath intake events (not only audio) happening in between them.

The intensities of the laughs are annotated subjectively based on the annotator's perception.

Laughter co-occurring with speech (speech-laughs) are also annotated as laughter and follow the same rules are above. The segments start when the laughter starts to happen (through a smile or not) and ends either with a breath intake sound or when the movement or sound ends. Speech-laughs can be as short as one vowel/consonant (30 ms) or as long as full sentences.

Tips:

1. If unsure about whether a noise or movement is a laugh, the segment label can be kept blank. To do this, select the segment, right-click, and select "remove annotation value." You can also press "Alt-Delete."

Interaction Question Prompts

These questions were used to trigger conversations and exchanges with the goal to collect naturalistic nonverbal expressions.

What is the most embarrassing situation you have been in?
What was your most disappointing experience?
What is your proudest memory?
What is the funniest story you have experienced?
What is the most disgusting experience you've had (food, odor, sight)?
Who or what do you admire the most?