Data annotation for parallel investigation of prosody and gesture at turn boundaries: categorization of form and function

Meg Zellers\textsuperscript{1}, Jan Gorisch\textsuperscript{2}, David House\textsuperscript{3}, Benno Peters\textsuperscript{1}

\textsuperscript{1}Institute for Scandinavian Studies, Frisian Studies, and General Linguistics, University of Kiel, Germany
\textsuperscript{2}Leibniz-Institute for the German Language, Mannheim, Germany
\textsuperscript{3}Department of Speech, Music & Hearing, KTH, Stockholm, Sweden

22 February 2019
Identifying Intersections between Prosody, Gesture, and Conversation

- Investigating 3-way meeting point of these aspects, with broad definition
- Interdisciplinary network-building to further promote research
  - Research visits in Kiel and Stockholm
  - Bilateral (German-Swedish) workshop held in November 2018 in Kiel
  - Future activities in connection with Fonetik 2019 in Stockholm
Fundamental spontaneous use of language: everyday conversation
Conversation flows with minimal gaps and overlaps\(^1\)
Across languages, similar systematicity with minor timing differences
Average silent gap between turns $\approx 200\text{ms}^2$
Time required to start articulation for speech $=$ minimum $500\text{ms}^3$

Listeners must have some way of predicting upcoming turn ends

\(^1\text{Sacks, Schegloff, and Jefferson (1974)}\)
\(^2\text{Stivers et al. (2009); Heldner and Edlund (2010)}\)
\(^3\text{Indefrey (2011)}\)
Regularly used to facilitate speaker change/floor hold:

- **Linguistic features** e.g. syntactic/semantic completion
- **Phonetic/prosodic features** e.g. intonation, phonation quality/spectral characteristics
- **Gestural/embodied features** e.g. eye gaze, hand movements
- **Multiple systems employed simultaneously**

---

4 Schaffer (1983); Auer (1996); Local, Kelly, and Wells (1986); Koiso, Horiuchi, Tutiya, Ichikawa, and Den (1998); Gravano and Hirschberg (2009, 2011); Kane, Yanushevskaya, de Looze, Vaughan, and Ní Chasaide (2014), *inter alia*.
Phonetics at turn boundaries

- F0 extremes can signal end of turn\textsuperscript{5}
- Slowing down of speech (beyond normal phrase-final lengthening) can indicate that speaker wishes to hold floor\textsuperscript{6}
- Coarticulation may be greater when speaker wishes to hold floor, influencing spectral characteristics\textsuperscript{7}

\textsuperscript{5}Local et al. (1986); Selting (1996); Gravano and Hirschberg (2009, 2011); Heldner and Włodarczak (2015)
\textsuperscript{6}Hjalmarsson and Laskowski (2011); Zellers (2017)
\textsuperscript{7}Niebuhr, Görs, and Graupe (2013); Kane et al. (2014)
**Gesture at turn boundaries**

- Turn-taking behavior changes based on eye gaze of on-screen avatars\(^8\)
- Gestures at possible turn boundaries can project upcoming behavior or response\(^9\)
- Hand gestures, head movements correlated with prosodic phrase boundaries and may contribute to parsing\(^{10}\)

---

\(^8\) Edlund and Beskow (2007, 2009)  
\(^9\) Streeck and Hartge (1992); Mondada (2007); Sikveland and Ogden (2012)  
\(^{10}\) Guellaï, Langus, and Nespor (2014); Frid, Svensson Lundmark, Ambrazaitis, Schötz, and House (2018)
• Relevant corpora already exist but not fully annotated and/or annotations are not sufficient for our research goals

• Work-in-progress: development of annotation schemes
  • Allow for parallel analysis of prosody and gesture at turn boundaries
  • Successful in conversational speech
  • Applicable to corpora with different content/structures
  • Enriching gestural annotations so function can be taken into account
Comparable data?

Spontal:\textsuperscript{11}

- Five 5-minute segments of spontaneous Swedish conversation
  - In process: $\approx 30$ more minutes
- Two-party conversations, no required topic; participants sat face-to-face
- Audio, video and motion capture data; some orthographic transcriptions available

\textsuperscript{11}Edlund et al. (2010)
Comparable data?

FOLK.\textsuperscript{12}

- Three 20-minute segments of German interaction
  - 40 minutes: mock job interviews (mostly two-party, participants sit face-to-face across table)
  - 20 minutes: interview of bird expert (two-party, participants stand and move around)
- Audio and video data; orthographic transcriptions available

\textsuperscript{12}Schmidt (2014)
Comparable data?

- Selection of FOLK data as similar as possible to Spontal
  - Two-party interactions
  - Hands visually available
  - No highly formal situations (e.g. televised debate, planned speech)
- Still not very similar
  - Free conversation vs. more structured interviews
  - FOLK excerpts recorded with one video camera and microphone
- Annotation system must be applicable to both (and ideally others)
Research foci and operationalizations

- Conversation: locations where speaker change may become relevant
- Prosody: phonetic variation in speech preceding these locations
- Gesture: (strokes of) hand gestures in vicinity of these locations
## Fields of annotation: minimum requirements

<table>
<thead>
<tr>
<th>Turn-Taking</th>
<th>Phonetics</th>
<th>Gesture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness in context</td>
<td>Phone segmentation</td>
<td>Gesture units</td>
</tr>
<tr>
<td>Transition type</td>
<td>Syllabification &amp; stress</td>
<td>Gesture phrases</td>
</tr>
<tr>
<td>Conversational activity</td>
<td>F0 contour</td>
<td>Gesture phases</td>
</tr>
<tr>
<td></td>
<td>Creak/breathiness</td>
<td>Referentiality</td>
</tr>
<tr>
<td></td>
<td>Inbreaths</td>
<td>Pragmatic function</td>
</tr>
</tbody>
</table>

Ideally theory-neutral and language-neutral
Turn-taking annotation

1. Before a silence, is utterance syntactically/semantically complete in context?
   - Response tokens/backchannels excluded

2. Classify turn transition type (see next slide)
   - Currently non-exhaustive; *Ambiguous* tag available

3. Future work: classify activity, e.g. storytelling, complaining, greeting, etc.
Four types of turn transitions

<table>
<thead>
<tr>
<th>Keep</th>
<th>M: idag är ju mitt barns födelsedag</th>
<th>today is of course my child’s birthday</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.18:21)</td>
<td>M: så jag är ju på supergott humör</td>
<td>so I am of course in a really good mood</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Backchannel</th>
<th>M: ...de som är lite svåra att uttala</th>
<th>...those which are a little hard to pronounce</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.44:36)</td>
<td>F1: ah yeah</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M: för honom i alla fall</td>
<td>for him in any case</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>M: ska vi se</th>
<th>shall we see</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1.62:30)</td>
<td>F1: ah vi kan titta</td>
<td>yeah we can look</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change</th>
<th>F1: jag borde ändå ha hört—hört det tidigare</th>
<th>I should anyway have heard—heard that earlier</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1.55:09)</td>
<td>M: och ändå är du inte övertygad</td>
<td>and anyway you aren’t convinced</td>
</tr>
</tbody>
</table>

Examples taken from DEAL Corpus (Hjalmarsson, Wik, & Brusk, 2007)
Phonetic/prosodic annotation
In Praat\textsuperscript{14}; no access to video

- Relatively straightforward: phone segmentation, syllabification, stress
- Somewhat straightforward: phonation quality, inbreaths
- Problematic: F0 contour in Swedish and German
  - Swedish: lexical word accent contrast, phrase-final pitch almost always falls to low\textsuperscript{15}
  - German: pragmatic pitch accent contrast, phrase-final rises frequent\textsuperscript{16}
- Our interim solution: label final F0 peak and valley (regardless of order), report size and directionality

\textsuperscript{14}Boersma and Weenink (2018)
\textsuperscript{15}Gårding (1989); House (2005)
\textsuperscript{16}Peters (2006); Grice, Baumann, and Benzmüller (2007)
Left: H L sequence in Swedish data; Right: L H sequence (interrupted by background noise) in German data
Phonetic measurements
To be taken during final 1000ms of speech before transition location, using Praat

- Pitch: F0 peak & valley height and alignment to segments/syllables
- Duration: average segment duration (seconds)
- Phonation quality: auditory judgment of creak and/or breathiness
Gesture annotation

In ELAN\(^{17}\); no audio for units/phases/rhythm

- Relatively straightforward: gesture units
- Somewhat straightforward: gesture phases\(^{18}\)
- Problematic: everything else!
  - Referentiality: can gesture stroke be considered as iconic, metaphoric, or deictic?\(^{19}\) (binary)
  - Pragmatic function: what is gesture doing if not referential, or in addition to it?
  - Rhythmic/prosodic function: is gesture stroke associated with stressed syllable?\(^{20}\)
- Current research prefers dimensional labelling approach\(^{21}\)

\(^{17}\) Max Planck Institute for Psycholinguistics (2018)
\(^{18}\) Kendon (1980, 2004)
\(^{19}\) McNeill and Levy (1982); McNeill (2005, 2006); Graziano and Gullberg (2018)
\(^{20}\) Prieto, Cravotta, Kushch, Rohrer, and Vilà-Giménez (2018)
\(^{21}\) McNeill (2006); Shattuck-Hufnagel and Ren (2018)
Gesture measurements

- Ongoing gesture phase (if available) at offset of speech
- Timing of offset of gesture in relation to offset of speech
- Function of gesture stroke at these locations
  - Do referential/non-referential gestures have different timing relationships with speech?
  - Is parallel phonetic variation same with different gesture functions?
  - Do rhythmic aspects differ with different conversational/pragmatic functions?
  - Does different gesture phrasing arise with turn-internal vs. turn-end phrase boundaries?
Some problematic cases

1. 10:53-10:56 from FOLK_E_00173
2. 4:42ff from Spontal 09-06
3. 11:33-11:35 from FOLK_E_00173
4. 4:11-4:19 from Spontal 09-22
Discussion

- Work still in progress
- Current system is designed for particular analysis, not exhaustivity
- Challenge: balance of accuracy with data manageability
  - Dimensional approach is promising but could turn out to be almost infinite in number of parameters
  - Prosodic/rhythmic aspects of gesture and speech probably require some kind of unified system (eventually)
Thank you!
Vielen Dank!
Tack så mycket!

Acknowledgements:
Portions of the work presented here were supported financially by the Deutsche Forschungsgemeinschaft (German Research Foundation), Vetenskapsrådet (Swedish Research Council) and the Riksbankens Jubileumsfond (Swedish Foundation for Humanities and Social Sciences). Many thanks to Simon Alexanderson, Jonas Beskow, and Jens Edlund for their assistance with processing Spontal.


