A suitable place to speak: On turn-taking for a conversational computer

Mattias Heldner & Jens Edlund
KTH Centre for Speech Technology (CTT)
Seminar given at KTH 2004-09-21
Structure of conversation

- Conversation is characterized by turn-taking
- One participant, A, talks, stops; another participant, B, starts, talks, stops
- Gaps and overlaps minimized
- How is this achieved?
Turns and turn-taking

- “In the abstract, the phenomenon of turn-taking seems quite easy to define. The talk of one party bounded by the talk of others constitutes a turn, with **turn-taking** being the process through which the party doing the talk of the moment in changed.”
Conversation analysis (CA)
theory of turn-taking


- Turn-taking is (1) an **emergent property** of (2) **local decisions** based on (3) **prediction** by the participants

- Turn-taking rules
TCUs and TRPs

- Turns are composed of smaller turn-constructional units (TCUs)
- The end of a TCUs is a transition-relevance place (TRP)
- TRPs are predictable to the listeners
- A set of rules that govern the transition of speakers come into play at the TRP
Turn-taking rule 1a

For any turn, at the initial TRP of an initial TCU:

If the turn-so-far is so constructed as to involve the use of a ‘current speaker selects next’ technique, then the party so selected has the right and is obliged to take next turn to speak; no others have such rights or obligations, and transfer occurs at that place.
Rule 1b

If the turn-so-far is so constructed as not to involve the use of a ‘current speaker selects next’ technique, then self-selection for next speakership may, but need not, be instituted; first starter acquires rights to a turn, and transfer occurs at that place.
Rule 1c

If the turn-so-far is so constructed as not to involve the use of a ‘current speaker selects next’ technique, then current speaker may, but need not continue, unless another self-selects.
Rule 2

( If, at the initial transition-relevance place of an initial turn-constructional unit, neither 1a nor 1b has operated, and, following the provision of 1c, current speaker has continued, then the rule-set a–c re-applies at the next transition relevance place, and recursively at each next transition relevance place, until transfer is effected.
Predictions by the rules

- One speaker at a time
- Overlaps occur either as competing first starts, or, where TRPs have been misprojected
What exactly is a TCU?

- Syntactic unit
- TRPs occur at possible completion points of sentences, clauses, phrases, and one-word constructions
- Intonation also important
Problems with CA theory of turn-taking

- Syntactic (and semantic and pragmatic) categories can be very difficult to segment in spoken dialogue
- Spontaneous conversation is not always well-formed – fragmentary and/or ungrammatical utterances common
- Non-verbal signals not included
Psychologists working on conversation

- Turn-taking is regulated by explicit signals
- A current speaker signals when he/she intends to hand over the floor – turn-yielding
- No single cue is required to display a signal
Turn-taking rules

- The listener may take his speaking turn when the speaker gives a turn-yielding signal.
- An attempt-suppressing signal displayed by the speaker maintains the turn for him, regardless of the number of yielding cues concurrently being displayed.
- Back-channel communication does not constitute a turn or a claim for a turn.
Turn-yielding signals

- Intonation: Rising or falling terminal junctures (boundary tones)
- Paralanguage: Drawl on the final syllable (final lengthening)
- Body motion: Termination of any hand gesticulation
- Sociocentric sequences: e.g. “but uh”, “or something”, “you know”
- Paralanguage: Drop in pitch and loudness
- Syntax: Completion of a grammatical clause
Gaze

- A speaker will break mutual gaze while speaking, returning gaze to the addressee upon turn completion.
Problems with signaling view

- Simultaneous speaking occurs either because the listener attempts to take his speaking turn in the absence of a turn-yielding signal by the speaker or if the speaker displays a yielding signal, and the listener acts to take his turn, and the original speaker then continues to claim his speaking turn.
Synthesis of CA and psychotheories

- Signals indicating the completion of turn-constructional units do indeed occur.
- Signals are the features that conversants use to identify the turn-constructional units and their boundaries.
- Much subsequent work on turn-taking has tried to analyze what features are used to signal a TRP.
Final major accents

- Define the TRP as the space between the TRP-projecting accent of the current turn and the onset of the next turn
- TRP-projecting accent = final major accent (focal accent)
Boundary tones

Boundary tones...

- High boundary tone associated with obligatory aspects of turn-taking
  - change of turn, e.g. answer to a question
  - turn holding, e.g. continued speech after a pause following an incomplete message

- Low boundary tone associated with optional aspects of turn-taking
  - completeness of a domain
Incomings

- **Turn-competitive incomings** i.e. interruptions – before final major accent,
- **Non-turn competitive incomings** i.e. backchannels/asides – after final major accent
Resolution of overlap

- One speaker drops out rapidly
- Recycling of the part obscured by overlap
- Competitive allocation – the speaker who ‘upgrades’ (increases intensity, slows tempo, etc.) most wins the floor
Summarizing

- Turn-endings predictable
- Gap and overlap minimized
- Syntactic, semantic, pragmatic completeness
- Gaze, head nods, hand gestures, facial expressions
- Prosody! Boundary tones, accents, speaking rate, silent pauses, voice quality etc.
A suitable place to speak...
Ultimately a conversational computer should be able to:

- perceive turn-keeping and turn-yielding signals
- initiate turns after turn-yielding signals
- to make non-competitive and turn-competitive incomings
- react to incomings from other participants
- avoid interrupting human participants – it must be unobtrusive!
Prosodic boundaries

- Turns where the speaker is allowed to finish end in a prosodic boundary
  
  Prosodic boundary ≠ turn-taking position

- Prosodic boundaries predictable to listeners from left-hand context only

- Prosodic rather than lexico-grammatical information the primary cue

- To some extent detectable using prosodic feature vectors and statistical classifiers
Goal

- Ultimate goal: Online prediction of acceptable places for turn-takings, as well as of impossible ones, for a conversational computer
- A step towards this goal: Exploring the relation between turn-taking and prosodic boundaries
- Two experiments: A listening test and a production experiment
Listening test

- Made-up turn-takings
- Fragment of a seminar followed by fragment of a question:
  “what about <um> could you give us some <hrm> rough idea what”
- Turn-takings in no boundary, weak boundary and strong prosodic boundary positions
- Task: to rate whether the questions occurred in appropriate places on a five-point scale
1. Turn-taking at a strong boundary
2. Turn-taking at a weak boundary
3. Turn-taking at a no boundary
Figure 1. The distribution of judgments on a five-point scale (where 1 represents an inappropriate and 5 an appropriate place for asking a question) for turn-takings at no boundaries, weak boundaries and strong boundaries.
Results by stimuli

- All strong boundary stimuli got higher means than the total of the experiment
- Nine out of ten no boundary stimuli got lower means than the total
- More variation in the weak boundaries
Production experiment

- Same speech material as in listening test
- Subjects pressed a button when they thought it was appropriate to take the turn
- Demo...
Results of production experiment

- Clear preference for strong boundaries (77%)
- Most of the strong boundaries (84%) used for turn-taking at least once
Timing differences

• Silence before question dependent on boundary type:
  Strong boundary 1 s, weak boundary 0.6 s, no boundary 0.02 s.
• Future work: Check whether the length of the silence should be governed by the prosodic boundary strength.
Conclusions

- Turn-taking closely related to prosodic boundaries
- Appropriate to take the turn after strong boundaries in this communicative situation
- If we can predict strong boundaries, we can predict possible places for turn-taking
Your turn to work...
Finding a suitable place to speak

- How to identify prosodic boundaries to find strong boundaries and to avoid weak and no boundaries?
- Preliminary results of acoustic analysis in the rhymes of the last words before the turn-takings
Boundary tones

- Level (less than 1 ST)
- Fall
- Rise
- Fall-rise
- Rise-fall
F0 range

- Cumulative mean ±2 standard deviations based on semitone transformed F0 data
- High, mid and low registers
- Stabilizes after about 20 seconds
Other measures

- Silent intervals
- Final lengthening
  - Average z-score normalized duration of the segments in the word-final rhyme
  - Z-score normalized duration of the word-final segment
<table>
<thead>
<tr>
<th>Mean score</th>
<th>Boundary type</th>
<th>Boundary tone</th>
<th>Range start</th>
<th>Range end</th>
<th>Silence duration (s)</th>
<th>Final lengthening (std devs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.70</td>
<td>Strong</td>
<td>Rise</td>
<td>Mid</td>
<td>Mid</td>
<td>4.489</td>
<td>0.2</td>
</tr>
<tr>
<td>3.90</td>
<td>Weak</td>
<td>Level</td>
<td>Mid</td>
<td>Mid</td>
<td>1.173</td>
<td>2.1</td>
</tr>
<tr>
<td>3.80</td>
<td>Strong</td>
<td>Rise</td>
<td>Low</td>
<td>Low</td>
<td>0.712</td>
<td>1.4</td>
</tr>
<tr>
<td>3.60</td>
<td>Strong</td>
<td>Level</td>
<td>Low</td>
<td>Low</td>
<td>0.878</td>
<td>-0.7</td>
</tr>
<tr>
<td>3.50</td>
<td>Strong</td>
<td>Level</td>
<td>Low</td>
<td>Low</td>
<td>3.412</td>
<td>-0.6</td>
</tr>
<tr>
<td>3.50</td>
<td>Strong</td>
<td>Fall</td>
<td>Mid</td>
<td>Low</td>
<td>1.362</td>
<td>0.4</td>
</tr>
<tr>
<td>3.50</td>
<td>Strong</td>
<td>Rise+fall</td>
<td>High</td>
<td>Mid</td>
<td>0.370</td>
<td>0.1</td>
</tr>
<tr>
<td>3.40</td>
<td>Weak</td>
<td>Fall+rise</td>
<td>Mid</td>
<td>Mid</td>
<td>0.000</td>
<td>2.1</td>
</tr>
<tr>
<td>3.15</td>
<td>Strong</td>
<td>Fall+rise</td>
<td>High</td>
<td>High</td>
<td>1.515</td>
<td>-0.6</td>
</tr>
<tr>
<td>3.15</td>
<td>Weak</td>
<td>Rise</td>
<td>Mid</td>
<td>Mid</td>
<td>0.162</td>
<td>2.3</td>
</tr>
<tr>
<td>3.10</td>
<td>Strong</td>
<td>Level</td>
<td>Low</td>
<td>Low</td>
<td>0.718</td>
<td>3.4</td>
</tr>
<tr>
<td>2.95</td>
<td>No</td>
<td>Fall</td>
<td>High</td>
<td>High</td>
<td>0.000</td>
<td>-1.8</td>
</tr>
<tr>
<td>2.85</td>
<td>Strong</td>
<td>Fall</td>
<td>Mid</td>
<td>Mid</td>
<td>1.344</td>
<td>0.6</td>
</tr>
<tr>
<td>2.70</td>
<td>Strong</td>
<td>Rise</td>
<td>High</td>
<td>High</td>
<td>0.798</td>
<td>1.3</td>
</tr>
<tr>
<td>2.55</td>
<td>Weak</td>
<td>Rise</td>
<td>High</td>
<td>High</td>
<td>0.460</td>
<td>0.4</td>
</tr>
<tr>
<td>2.30</td>
<td>No</td>
<td>Rise+fall</td>
<td>Mid</td>
<td>Mid</td>
<td>0.400</td>
<td>-0.6</td>
</tr>
<tr>
<td>2.05</td>
<td>Weak</td>
<td>Level</td>
<td>Mid</td>
<td>Mid</td>
<td>0.371</td>
<td>0.6</td>
</tr>
<tr>
<td>1.85</td>
<td>No</td>
<td>Level</td>
<td>Mid</td>
<td>Mid</td>
<td>0.800</td>
<td>-1.0</td>
</tr>
<tr>
<td>1.85</td>
<td>Weak</td>
<td>Rise</td>
<td>Mid</td>
<td>High</td>
<td>0.444</td>
<td>-0.1</td>
</tr>
<tr>
<td>1.80</td>
<td>Weak</td>
<td>Level</td>
<td>High</td>
<td>High</td>
<td>0.400</td>
<td>1.5</td>
</tr>
<tr>
<td>1.60</td>
<td>No</td>
<td>Fall</td>
<td>Mid</td>
<td>Low</td>
<td>0.800</td>
<td>0.0</td>
</tr>
<tr>
<td>1.75</td>
<td>No</td>
<td>Level</td>
<td>Low</td>
<td>Low</td>
<td>0.868</td>
<td>-1.0</td>
</tr>
<tr>
<td>1.70</td>
<td>No</td>
<td>Level</td>
<td>Low</td>
<td>Low</td>
<td>0.800</td>
<td>-1.1</td>
</tr>
<tr>
<td>1.60</td>
<td>Weak</td>
<td>Level</td>
<td>Mid</td>
<td>Mid</td>
<td>1.190</td>
<td>0.4</td>
</tr>
<tr>
<td>1.60</td>
<td>No</td>
<td>Level</td>
<td>Mid</td>
<td>Mid</td>
<td>0.800</td>
<td>-1.1</td>
</tr>
<tr>
<td>1.55</td>
<td>Weak</td>
<td>Fall</td>
<td>Mid</td>
<td>Mid</td>
<td>0.180</td>
<td>0.4</td>
</tr>
<tr>
<td>1.50</td>
<td>No</td>
<td>Rise+fall</td>
<td>Mid</td>
<td>Mid</td>
<td>0.800</td>
<td>0.0</td>
</tr>
<tr>
<td>1.45</td>
<td>No</td>
<td>Rise+fall</td>
<td>Mid</td>
<td>Mid</td>
<td>0.800</td>
<td>-1.0</td>
</tr>
<tr>
<td>1.40</td>
<td>Weak</td>
<td>Level</td>
<td>Mid</td>
<td>Mid</td>
<td>0.881</td>
<td>0.0</td>
</tr>
<tr>
<td>1.20</td>
<td>No</td>
<td>Fall</td>
<td>Mid</td>
<td>Mid</td>
<td>0.800</td>
<td>-1.6</td>
</tr>
</tbody>
</table>
Turn-keeping

- Duncan, 1972; 2 2 | (English)
- Selting, 1996; level pitch before pause (German)
- Caspers, 2003; level boundary tone (Dutch)
- Noguchi & Den, 1998; flat intonation at the end of pause bounded phrases (Japanese)
Thank you!