Evaluation with real user in Real World

- Ethical issues
  - modality: voice data is easier than facial data
  - be transparent: this conversation will be recorded for future....
- How do you get answers?
  - Directly? Are you satisfied...
  - Observation? User behaviour, emotional cues etc.
- Two groups of users: expert-users (familiar with system), 'new'-users (not familiar with questions sequence) – solution: user-expectation

- solution: user response (positive/negative words – e.g. thanks)
- task-completion: (task-orientated dialogues) - successful
- confirmation: but it could be annoying
- Emotion-detection would help significantly
  - non-acoustical, barge-ins, bad words (f-word) only shows OBVIOUSLY angry people and it depends on personality – how about getting a common basis?
  - how the user say something?

  - but it's culture-based: some groups may not explicitly show negative emotions though they feel it

- Level of engagement:
  - measure success based on level of engagement (for non-task orientation – tutorial, casual dialogues)