

NPFL123 Dialogue Systems

2. What happens in a dialogue?

<https://ufal.cz/npfl123>

Ondřej Dušek, Vojtěch Hudeček & Jan Cuřín

9. 3. 2021



Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated

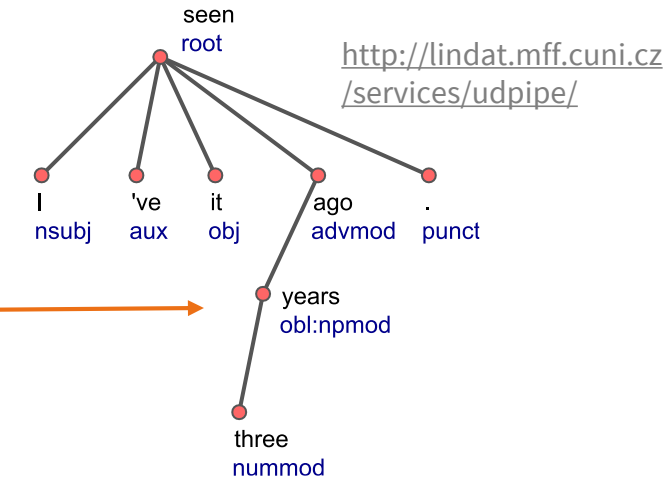
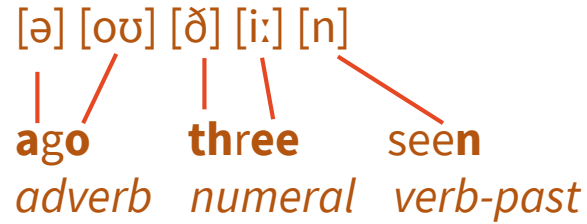
How do you “define” dialogue?

- Spoken/written conversational (interactive, collaborative) communication between two or more people
- **verbal** + (possibly) non-verbal
 - can be multimodal (language + gestures, pitch, expressions...)
- **collaborative**, social
 - participants aim at communicative goal(s)
 - involves inference about intended meanings
- **practical**, related to actions
- **interactive**, incremental, messy!

Dialogue systems – simpler than that

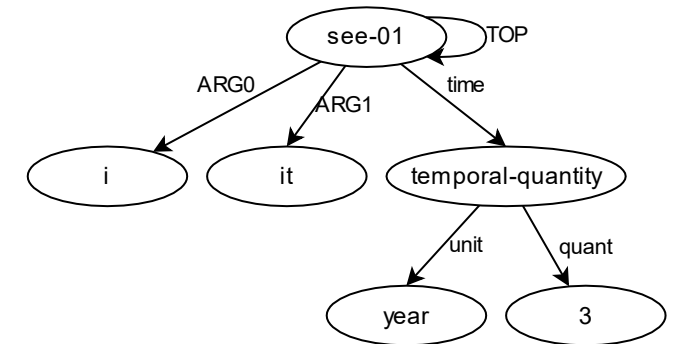
Describing a dialogue

- Levels of linguistic description
 - **phonetics / phonology** – sounds
 - **morphology** – word forms
 - **syntax** – sentence structure
 - **semantics** – sentence (propositional) meaning
 - **pragmatics** – meaning in context, communication goal



- This lecture is (a lot) about **pragmatics**

(I don't remember it well)



<http://cohort.inf.ed.ac.uk/amreager.html>

Turn-taking (interactivity)

- Speakers **take turns** in a dialogue
 - **turn** = continuous utterance from one speaker
- Normal dialogue – very fluent, fast
 - minimizing **overlaps & gaps**
 - little silence (usually <250ms), little overlap (~5%)
 - (fuzzy) rules, anticipation
 - cues/markers for turn boundaries:
 - linguistic (e.g. finished sentence), voice pitch
 - timing (gaps)
 - eye gaze, gestures (...)
- overlaps happen naturally
 - ambiguity in turn-taking rules (e.g. two start speaking at the same time)
 - **barge-in** = speaker starts during another one's turn

Turn-taking (example)

<https://youtu.be/BZF9eg35IXI?t=91>

20 seconds of a semi-formal dialogue (talk show):

S: um uh , you're about to start season [six ,]

J: [yes]

S: you probably already started but [it launches]

J: [yes thank you]

A: (*cheering*)

J: we're about to start thank you yeah .. we're starting , we- on Sunday yeah , we've been eh- we've been prepping some [things]

S: [confidence] is high . feel good ?

J: (*scoffs*)

S: think you're gonna [squeeze out the shows this time ? think you're gonna do it ?]

J: (*laughing*) [you're talking to me like I'm an a-] confidence high ? no !

S: [no]

J: [my confidence] is never high .

S: okay

J: self loathing high . concern astronomic .



Speech vs. text

- Natural speech is **very different from written text**
 - ungrammatical
 - restarts, hesitations, corrections
 - overlaps
 - pitch, stress
 - accents, dialect
- See more examples in speech corpora
 - <https://kontext.korpus.cz/> (Czech)
 - select the “oral” corpus and search for a random word

The screenshot shows a search interface for the word "řekni". The results are displayed in a list format, with each entry including the speaker's name and the full context of the utterance. The speakers listed are Linda_7158, Otakar_7651, and Dalibor_7582. The search results include:

- Linda_7158: maji* majitel Semlaru
- Linda_7158: no já sem to četla v novinách
- Linda_7158 + Otakar_7651: no
- Linda_7158 + Otakar_7651: hovno
- Dalibor_7582: si ho* v nedělu hodil mašlu ..
- Otakar_7651: ty vole
- Dalibor_7582: mně to říkal Martin že to četl v novinách já říkám no tak tady -
- Linda_7158: taji mám ty noviny .
- Otakar_7651: to von byl takovej divně ale
- Dalibor_7582: ale si mně řekni skrz peníze to určitě nebylo že by jako byl
- Dalibor_7582 + Otakar_7651: se mu jak to jelo ..
- Dalibor_7582 + Otakar_7651: tak to určitě ne
- Otakar_7651: dyť tam peněz bylo jako .
- Dalibor_7582: a to je ten jak byl na té železnici jak tam vjel
- Dalibor_7582 + Otakar_7651: sám
- Dalibor_7582 + Otakar_7651: no
- Otakar_7651: to

Turn taking in dialogue systems

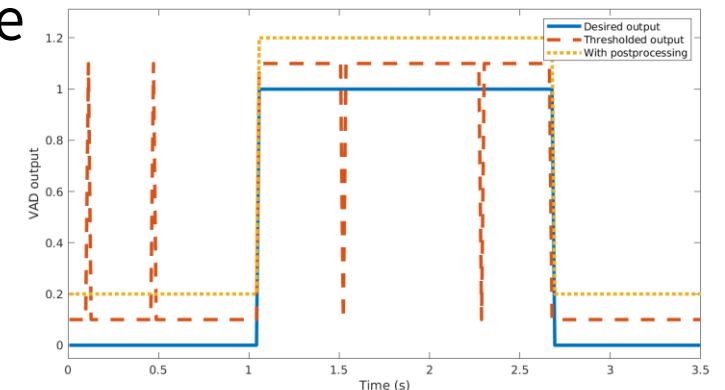
- consecutive turns are typically assumed
 - system waits for user to finish their turn (~250ms non-speech)
- **voice activity detection**
 - binary classification problem – “is it user’s speech that I’m hearing?”[Y/N]
 - segments the incoming audio (checking every X ms)
 - actually a hard problem
 - nothing ever works in noisy environments
- **wake words** – making VAD easier
 - listen for a specific phrase, only start listening after it
- some systems allow user’s barge-in
 - may be tied to the wake word

hey Siri
okay Google
Alexa

Voice activity detection

- **Overlapping windows of ~30ms + binary classifier**
- Features – actually similar to speech recognition itself
 - energy (loudness)
 - autocorrelation
 - checking for fundamental voice frequency
 - MFCCs (mel frequency spectrum)
 - deltas (trends over time)
- **Onset is easier to detect** than end of speech
 - they're louder, more pronounced
 - hard to detect speech towards the system vs. someone else
 - that's why wake words are used
 - how long can pauses/hesitations be?
- Postprocessing
 - smoothing out short-term errors

<https://wiki.aalto.fi/pages/viewpage.action?pageId=151500905>



Speech acts (by John L. Austin & John Searle)

- each utterance is an **act**
 - intentional
 - changing the state of the world
 - changing the knowledge/mood of the listener (at least)
 - influencing the listener's behavior
- speech acts consist of:
 - a) utterance act** = the actual uttering of the words
 - b) propositional act** = semantics / “surface” meaning
 - c) illocutionary act** = “pragmatic” meaning
 - e.g. command, promise [...]
 - d) perlocutionary act** = effect
 - listener obeys command, listener's worldview changes [...]

X to Y: *You're boring!*

- a) [jʊr 'bɔ:ɪŋ]
- b) boring(Y)
- c) statement
- d) Y is cross

X to Y: *Can I have a sandwich?*

- a) [kæn aɪ hæv ə 'sændwɪtʃ]
- b) can_have(X, sandwich)
- c) request
- d) Y gives X a sandwich

Speech acts

- Types of speech acts:

- **assertive**: speaker commits to the truth of a proposition

- statements, declarations, beliefs, reports [...]

It's raining outside.

- **directive**: speaker wants the listener to do something

- commands, requests, invitations, encouragements

Stop it!

- **commissive**: speaker commits to do something themselves

- promises, swears, threats, agreements

I'll come by later.

- **expressive**: speaker expresses their psychological state

- thanks, congratulations, apologies, welcomes

Thank you!

- **declarative**: performing actions (“performative verbs”)

- sentencing, baptizing, dismissing

You're fired!

Speech acts

- Explicit vs. implicit

- explicit – using a verb directly corresponding to the act
- implicit – without the verb

explicit: *I **promise** to come by later.*
implicit: *I'll come by later.*

explicit: *I'm **inviting** you for a dinner.*
implicit: *Come with me for a dinner!*

- Direct vs. indirect

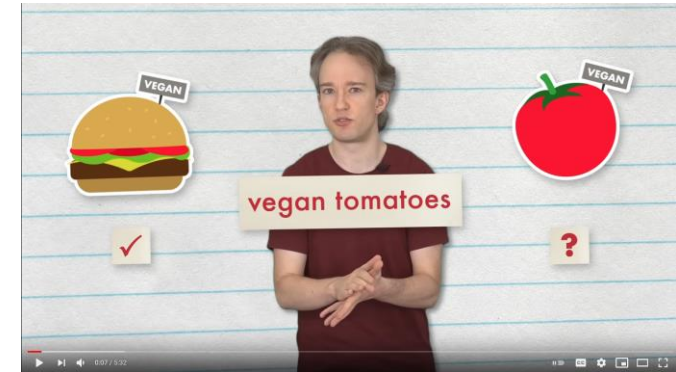
- **indirect** – the surface meaning does not correspond to the actual one
 - primary illocution = the actual meaning
 - secondary illocution = how it's expressed
- reasons: politeness, context, familiarity

direct: *Please close the window.*
indirect: *Could you close the window?*
even more indirect: *I'm cold.*

direct: *What is the time?*
indirect: *Have you got a watch?*

Conversational Maxims (by Paul Grice)

- based on Grice's **cooperative principle** (“dialogue is cooperative”)
 - speaker & listener cooperate w. r. t. communication goal
 - speaker wants to inform, listener wants to understand
- 4 Maxims (basic premises/principles/ideals)
 - M. of **quantity** – don't give too little/too much information
 - M. of **quality** – be truthful
 - M. of **relation** – be relevant
 - M. of **manner** – be clear
- By default, speakers are assumed to adhere to maxims
 - apparently breaking a maxim suggests a different/additional meaning



https://youtu.be/IJEaMtNN_dM

Conversational Implicatures

- **implicatures** = implied meanings
 - standard – based on the assumption that maxims are obeyed
 - maxim flouting (obvious violation) – additional meanings (sarcasm, irony)

John ate some of the cookies → [otherwise too little/low-quality information] not all of them

A: I've run out of gas.

B: There's a gas station around the corner. → [otherwise irrelevant] the gas station is open

A: Will you come to lunch with us?

B: I have class. → [otherwise irrelevant] B is not coming to lunch

A: How's John doing in his new job?

B: Good. He didn't end up in prison so far. → [too much information] John is dishonest / the job is shady

Speech acts & maxims & implicatures in dialogue systems

- Learned from data / hand-coded
- **Understanding**
 - tested on real users → usually knows indirect speech acts
 - **implicatures limited** – there's no common sense
 - (other than what's hand-coded or found in training data)

*system: The first train from Edinburgh to London leaves at 5:30 from Waverley Station.
user: I don't want to get up so early. → [fails]*

- **Responses**
 - mostly strive for clarity – user doesn't really need to imply

Grounding

- dialogue is cooperative → need to ensure mutual understanding
- **common ground** = shared knowledge, mutual assumptions of dialogue participants
 - not just shared, but *knowingly* shared
 - $x \in \text{CG}(A, B)$:
 - A & B must know x
 - A must know that B knows x and vice-versa
 - expanded/updated/refined in an informative conversation
- validated/verified via **grounding signals**
 - speaker **presents** utterance
 - listener **accepts** utterance by providing evidence of understanding

Grounding signals / feedback


- used to notify speaker of (mis)understanding
- positive – understanding/acceptance signals:
 - **visual** – eye gaze, facial expressions, smile [...]
 - **backchannels** – particles signalling understanding *uh-uh, hmm, yeah*
 - **explicit feedback** – explicitly stating understanding *I know, Yes I understand*
 - **implicit feedback** – showing understanding implicitly in the next utterance

U: find me a Chinese restaurant

S: I found three Chinese restaurants close to you [...]

A: Do you know where John is?

B: John? Haven't seen him today.

- negative – misunderstanding:
 - **visual** – stunned/puzzled silence
 - **clarification requests**  *A: Do you know where John is?*
B: Do you mean John Smith or John Doe?
 - demonstrating ambiguity & asking for additional information
 - **repair requests** – showing non-understanding & asking for correction

Oh, so you're not flying to London? Where are you going then?

Grounding (example)

T: [...] And the ideology is also very against mixed-race couples. So that was also a target. Whenever we saw mixed-race couples, we attacked them.

E: Was there ever a moment back there where you felt a tiny bit bad about it?

T: No.

E: **No? So you were** absolutely convinced that you're doing the right thing...

T: Yeah, for quite some time **(nods), yeah.**

E: ... for the sake of the white race and et cetera?

E: No doubt at all?

T: Well I got **doubt** eventually, roughly a year before I left the movement [...]



<https://video.aktualne.cz/dvtv/cernoch-mi-miril-pistoli-na-hlavu-nevim-proc-me-nezabil-rika/r~d87679def2fd11e8a7f60cc47ab5f122/> (2:45 and onwards)

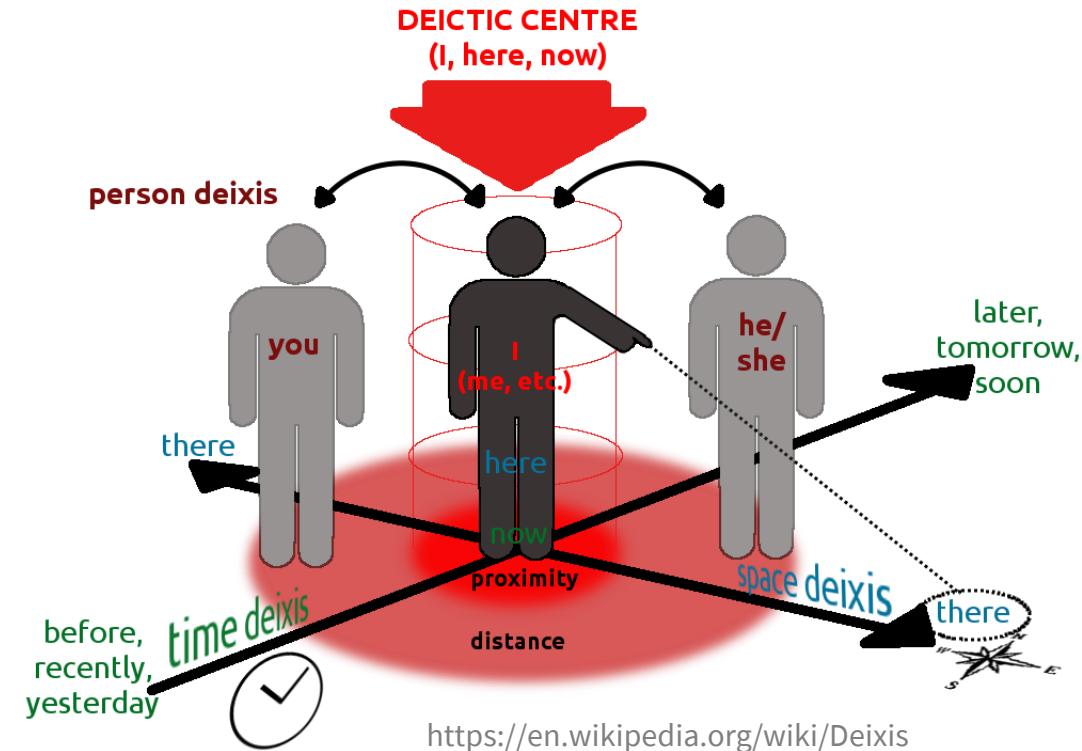
Grounding in dialogue systems

- Crucial for successful dialogue
 - e.g. booking the right restaurant / flight
- Backchannels / visual signals typically not present
- **Implicit confirmation** very common
 - users might be confused if not present
- **Explicit confirmation** may be required for important steps
 - e.g. confirming a reservation / bank transfer
- **Clarification & repair requests** very common
 - when input is ambiguous or conflicts with previously said
- Part of dialogue management
 - uses NLU confidence in deciding to use the signals

- **deixis** = “pointing” – relating between language & context/world
 - this is very important in dialogue
 - dialogue is typically set/situated in a specific context
- **deictic expressions** = words/grammar expressing deixis
 - their meaning depends on the context
 - who is talking, when, where
 - pronouns *I, you, him, this*
 - verbs: tense & person markers *goes* [3rd ps. sg.], *went* [past]
 - adverbs *here, now, yesterday*
 - other (lexical meaning) *come / go* [=here/away],
 - non-verbal (gestures, gaze...)

Deixis

- (typically) **egocentric**:
I – here – now is the center (**origo**)
- main types of deixis:
 - **personal** – *I/me/you/she...*
 - **temporal** (time) – *now, yesterday, later, on Monday...*
 - **local** (space) – *here, there...*
- other:
 - **social** (politeness)
 - formal/informal address (Cz. *ty/vy*, Ger. *du/Sie*), honorifics in Asian languages
 - **discourse/textual**
 - referring to words/portions of texts – *next chapter, how do you spell that?*



Anaphora/Coreference

- expression referring to something mentioned in context

- **anaphora** = referring back
- **cataphora** = referring forward

- avoiding repetition, faster expression

- can refer to basically anything

- objects/persons/events
- qualities
- actions/full sentences/portions of text

- used frequently in dialogue

- may be ambiguous

Susan dropped the plate. It shattered.
His friends describe John as smart and hard-working.

I don't like it as much as he does.

Her dress is green. So is mine.

– Shall I book a room for you? – Sure, I'd like that.

? Bill stands next to John. He is tall.
Bill tickled John. He squirmed.

Deixis & anaphora in dialogue systems

- systems typically assume a **single user**
 - this makes personal deixis much easier
- most systems are aware of time, location is more complicated
 - pronouns are often avoided – clearer, although less natural
- coreference resolution – separate problem
 - a whole area of research, specific resolution systems developed
 - some dialogue systems don't include it, some do, sophistication varies

Prediction

- Dialogue is a **social interaction**
 - people view dialogue partners as goal-directed, intentional agents
 - they analyze their partners' goals/agenda
- Brain does not listen passively
 - projects hypotheses/interpretations on-the-fly
- **prediction** is crucial for human cognition
 - people predict what their partner will (or possibly can) say/do
 - continuously, incrementally
 - unconsciously, very rapidly
 - guides the cognition
- this is (part of) why we understand in adverse conditions
 - noisy environment, distance

Entropy (Claude Shannon)

- Information theory: dialogue is information transfer
 - **communication channel** – speaker to listener (in the given situation)
 - **entropy** – expected value of information conveyed (in bits)

$$H(\text{text}) = - \sum_{\text{word} \in \text{text}} \frac{\text{freq}(\text{word})}{\text{len}(\text{text})} \log_2 \left(\frac{\text{freq}(\text{word})}{\text{len}(\text{text})} \right)$$

over vocabulary →

XXXX : entropy = 0
WXYZ : entropy = 2

- Plays well with the social interaction perspective
 - people tend to **use all available channel capacity**
 - limiting factors: noise, listener's hearing ability, mental capacity
 - people tend to **spread information evenly**
 - words carrying more information are emphasized

Conditional entropy

- how hard it is to guess the next word in the sentence?
 - given preceding context (n-gram)
 - related to Shannon entropy, but may differ
 - typically much lower than Shannon entropy
 - better estimate of prediction difficulty
 - although humans work with “unlimited” preceding context and reevaluate using following context

<s> The cat sat on the mat .

P(cat | <s> The)

P(sat | the cat)

P(on | the cat sat)

P(the | the cat sat on)

$$H_{\text{cond}}(\text{text}) = - \sum_{(c,w) \in \text{text}} \frac{\text{freq}(c, w)}{\text{len}(\text{text})} \log_2 \left(\frac{\text{freq}(c, w)}{\text{freq}(c)} \right)$$

of times word w occurs after context c

context (preceding n-gram)

word

means # of n-grams here (not just words)

total # of times context c occurs, with or without word w

Prediction in dialogue systems

- Used a lot in speech recognition
 - **language models** – based on information theory
 - statistical, trained on a text corpus (bunch of texts)
 - predicting likely next word given context
 - weighted against acoustic information
- Not as good as humans
 - may not reflect current situation (noise etc.)
 - (often) does not adapt to the speaker
- Less use in other DS components

Alignment/entrainment

- People subconsciously **adapt/align/entrain** to their dialogue partner over the course of the dialogue

- wording (lexical items)
- grammar (sentential constructions)
- speech rate, prosody, loudness
- accent/dialect

pram → *stroller* [BrE speaker
lorry → *truck* talking to AmE speaker]

S: [...] *Confidence is high, feel good?*
[...]

J: **Confidence high?** No!

S: No.

J: My **confidence is** never **high**.

S: Okay.

J: **Self loathing high**, concern astronomical.

- This helps a successful dialogue
 - also helps social bonding, feels natural

Alignment in dialogue systems

- Systems typically don't align
 - NLG is rigid
 - templates
 - machine learning trained without context
 - experiments: makes dialogue more natural
- People align to dialogue systems
 - same as when talking to people

(Dušek & Jurčiček, 2016)

<http://www.aclweb.org/anthology/W16-3622>

context *is there a later option*
response DA `implicit_confirm(alternative=next)`
base NLG Next connection.
+ alignment You want a later option.

context *I need to find a bus connection*
response DA `inform_no_match(vehicle=bus)`
base NLG No bus found, sorry.
+ alignment I'm sorry, I cannot find a bus connection.

*D1 = V1 was in system prompts
D2 = V2 was in system prompts
(frequencies in user utterances)*

Words	D1 Freq. (% rel. Freq)	D2 freq (% rel. Freq)
V1: next	13204 (99.9%)	492 (82.9%)
V2: following	3 (0.1%)	101 (17.1%)
V1: previous	3066 (100%)	78 (44.8%)
V2: preceding	0 (0%)	96 (55.2%)
V1: now	6241 (99.8%)	237 (80.1%)
V2: immediately	10 (0.2%)	59 (19.9%)
V1:leaving	4843 (98.4%)	165 (70.8%)
V2: departing	81 (1.6%)	68 (29.2%)
V1: route/schedule	2189 (99.9%)	174 (94.5%)
V2: itinerary	2 (0.1%)	10 (5.5%)
V1: okay/correct	1371 (49.3%)	48 (27.7%)
V2: right	1409 (50.7%)	125 (72.3%)
V1: help	2189 (99.9%)	17 (65.3%)
V2: assistance	1 (0.1%)	9 (34.7%)
V1: query	6256 (99.9%)	70 (20.4%)
V2: request	3 (0.1%)	272 (79.6%)

(Parent & Eskenazi, 2010)

https://www.isca-speech.org/archive/interspeech_2010/i10_3018.html

Politeness

- Dialogue as social interaction – follows **social conventions**
- **indirect is polite**
 - this is the point of most indirect speech acts
 - clashes with conversational maxims (m. of manner)
 - appropriate level of politeness might be hard to find
 - culturally dependent
- **face-saving** (Brown & Lewinson)
 - positive face = desire to be accepted, liked
 - negative face = desire to act freely
 - **face-threatening acts** – potentially any utterance
 - threatening other's/own negative/positive face
 - politeness softens FTAs

Open the window.
Can you open the window?
*Would you be so kind as
to open the window?*
Would you mind closing the window?

threat to	positive face	negative face
self	<i>apology, self-humiliation</i>	<i>accepting order / advice, thanks</i>
other	<i>criticism, blaming</i>	<i>order, advice, suggestion, warning</i>

Politeness in dialogue systems

- Typically **handcrafted** by system design
 - does not adapt to situation very much
 - typically not much indirect speech, but trying to stay polite
- Learning from data can be tricky
 - **check your data** for offensive speech!
 - not just swearwords – problems can be hard to find

I already have a woman to sleep with.

(Experimental chatbot we trained at Heriot-Watt using Reddit data)

Microsoft Tay Twitter chatbot
(learning from users)
[https://en.wikipedia.org/wiki/Tay_\(bot\)](https://en.wikipedia.org/wiki/Tay_(bot))



Summary

- Dialogue is messy
 - **turn** overlaps, **barge-ins**, weird grammar [...]
- Dialogue utterances are acts
 - **illocution** = pragmatic meaning
- Dialogue needs understanding
 - **grounding** = mutual understanding management
 - **backchannels, confirmations, clarification, repairs**
- Dialogue takes place in context
 - lot of pointing – **deixis**
- Dialogue is cooperative, social process
 - **conversational maxims** ~ “play nice”
 - all while following **social conventions** (politeness)
 - people **predict & adapt** to each other
- Next week: where & how to get data, how to evaluate dialogue systems

Thanks

Contact us:

[https://ufaldsg.slack.com/
{odusek,hudecek}@ufal.mff.cuni.cz](https://ufaldsg.slack.com/{odusek,hudecek}@ufal.mff.cuni.cz)
Skype/Meet/Zoom (by agreement)

Get the slides here:

<https://ufal.cz/npfl123>

References/Inspiration/Further:

Apart from materials referred directly, these slides are based on:

- Pierre Lison's slides (Oslo University): <https://www.uio.no/studier/emner/matnat/ifi/INF5820/h14/timeplan/index.html>
- Ralf Klabunde's lectures and slides (Ruhr-Universität Bochum): <https://www.linguistics.ruhr-uni-bochum.de/~klabunde/lehre.htm>
- Filip Jurčiček's slides (Charles University): <https://ufal.mff.cuni.cz/~jurcicek/NPFL123-SDS-2014LS/>
- Arash Eshghi & Oliver Lemon's slides (Heriot-Watt University): <https://sites.google.com/site/olemon/conversational-agents>
- Gina-Anne Levow's slides (University of Washington): <https://courses.washington.edu/ling575/>
- Eika Razi's slides: <https://www.slideshare.net/eikarazi/anaphora-and-deixis>
- Wikipedia: [Anaphora \(linguistics\)](#) [Conversation Cooperative principle](#) [Coreference](#) [Deixis](#) [Grounding in communication](#) [Implicature](#) [Speech act](#) [Sprechakttheorie](#)

Next week:
Lab questions 9am
Lab assignment 9:50
Lecture 10:40