

# NPFL123 Dialogue Systems

## 1. Introduction

<https://ufal.cz/npfl123>

**Ondřej Dušek**, Mateusz Lango, Simone Balloccu, Jan Cuřín

21. 2. 2024



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



unless otherwise stated

# Organizational: NPFL123 – 2/2 C+Ex – 5 Credits

- Lecture (Wed 3:40pm) + labs (Wed 5:20pm, mostly 45 mins or less)
- Lecture: intro, theory
- Labs: practical hands-on exercises
- To pass the course:
  - **60%+ written exam** – 10 freeform questions, covered by the lectures
    - list of questions available on the web (may be updated slightly)
  - **50%+ points from lab exercises** – weekly homework assignment
    - implementing your system for a domain
    - other small dialogue-related exercises
- Slides, news etc. at [ufal.cz/npfl123](https://ufal.cz/npfl123)
- Slack channel for discussions (<https://ufal-dsg.slack.com/>)
  - you got an invite per email, let me know if not

# About Us

## **Ondřej Dušek:** lectures, course guarantor

- PhD at ÚFAL '17, 2 years at Heriot-Watt Uni Edinburgh, back '19
- mostly language generation, also chatbots (Alexa Prize)

## **Mateusz Lango:** labs

- post-doc, PhD at Uni Poznań, data imbalance & accurate NLG

## **Simone Balloccu:** data & eval lecture

- post-doc, PhD at Uni Aberdeen, health NLG & LLMs

## **Jan Cuřín:** dialogue authoring lecture

- PhD at ÚFAL, IBM Research, founded THE MAMA.AI in '21



# Course Syllabus

1. Introduction (today)
2. What happens in a dialogue?
3. Dialogue system data & how to evaluate
4. Language understanding (NLU)
5. NLU + Dialogue state tracking
6. Dialogue management (DM)
7. DM + Language generation
8. Dialogue authoring/tooling systems
9. Voice assistants (Alexa, Siri, Google etc.) + question answering
10. Speech recognition
11. Speech synthesis
12. Open-domain systems (chitchat, instruction-tuned LMs)

# Recommended Reading

## Primary:

- Jurafsky & Martin: Speech & Language processing. 3rd ed. draft 2024, Chap. 14-16 (<https://web.stanford.edu/~jurafsky/slp3/>) – brief, good intro
- McTear: Conversational AI. Morgan & Claypool 2021. (<https://doi.org/10.2200/S01060ED1V01Y202010HLT048>) – bit more advanced, very slightly outdated

## Other (see also website):

- Gao et al.: Neural Approaches to Conversational AI, 2019 (<http://arxiv.org/abs/1809.08267>)
- McTear et al.: The Conversational Interface: Talking to Smart Devices. Springer 2016.
- Janarthanam: Hands-On Chatbots and Conversational UI Development. Packt 2017.
- Skantze: Error Handling in Spoken Dialogue Systems. PhD Thesis 2007, Chap. 2 (<http://www.speech.kth.se/~gabriel/thesis/chapter2.pdf>)
- Jokinen & McTear: Spoken dialogue systems. Morgan & Claypool 2010.
- Psutka et al.: Mluvíme s počítačem česky. Academia 2006.
- Lemon & Pietquin: Data-Driven Methods for Adaptive Spoken Dialogue Systems. Springer 2012.
- Rieser & Lemon: Reinforcement learning for adaptive dialogue systems. Springer 2011.

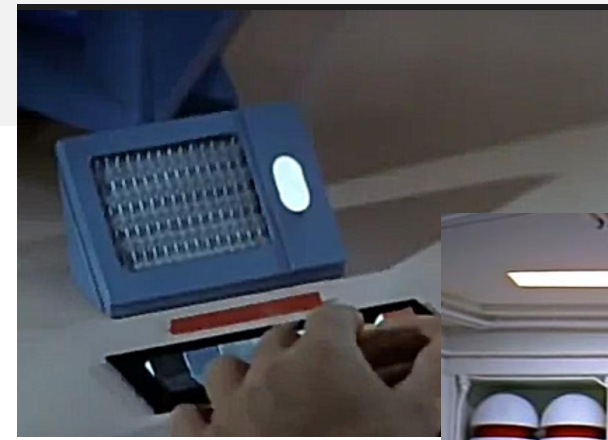
# What's a dialogue system?

Definition:

- A (*spoken*) dialogue system is a **computer system designed to interact** with users **in** (*spoken*) **natural language**
- Wide definition – covers lots of different cases

# “AI”: sci-fi vs. reality

- Lots of talk about AI now
- Hype around LLMs (ChatGPT & co.)
- Sci-fi expectations – AI-complete
  - *Star Trek* – know-it-all
  - *2001 Space Odyssey* – mutiny
  - *Her* – personality
- We’re not there – probably for long
  - main bottleneck: understanding (not speech comprehension, meaning!)
  - ... more like the *Red Dwarf* talkie toaster



<https://youtu.be/qDrDUmuUBTo>



<https://youtu.be/1ZXugicgn6U?t=3>



[https://youtu.be/6QRvTv\\_tpw0?t=27](https://youtu.be/6QRvTv_tpw0?t=27)



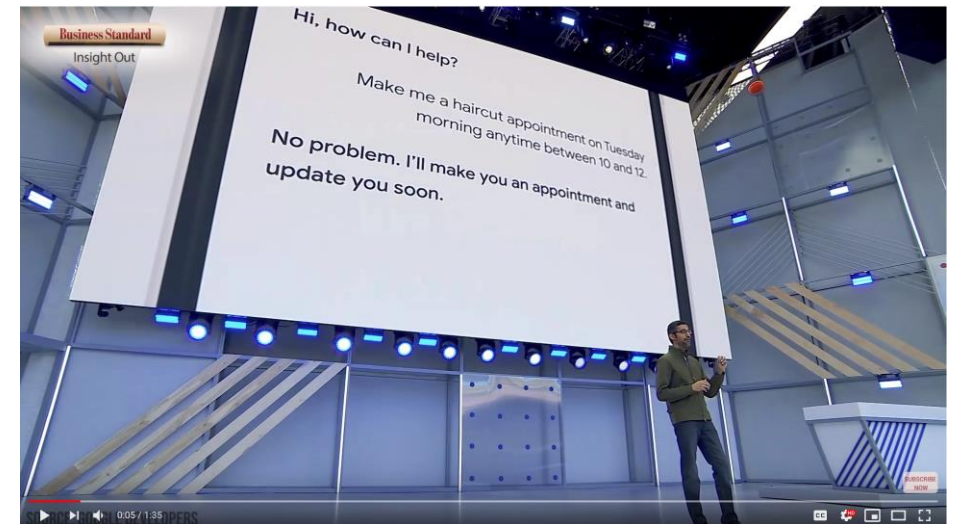
HOWDY DOODLY DOO! IM TALKIE TOASTER!  
YOUR CHIRPIE BREAKFAST COMPANION!

[https://youtu.be/LRq\\_SAuQDec?t=71](https://youtu.be/LRq_SAuQDec?t=71)



# Example: Google Assistant

- Handling call for a client (Google IO 2018 demo)
  - very natural speech
  - show's what's possible **in a limited domain**
  - redirects to a human if it can't handle a shop's request
- Deployed now in the US, but more limited
  - + some shops may just hang up

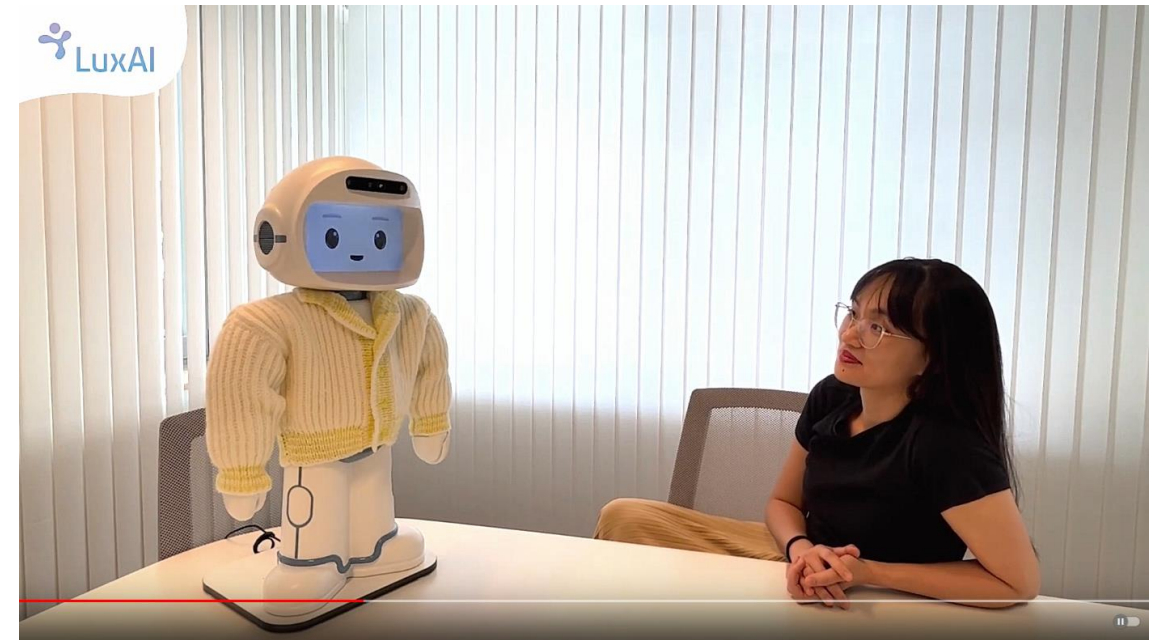


<https://youtu.be/d40jgFZ5hXk>



# Example: QTRobot / ChatGPT

- LLM running via a robot interface
  - remote API / locally
  - + additional handling for voice, gaze, gestures
- Assumes a given personality (e.g. chef, astronaut, fisherman)
- Open-domain chat about anything



<https://youtu.be/-mOYg8P5OaY>

# Why take interest in Dialogue Systems?

- It's *the* **ultimate natural interface** for computers
- Exciting & **active research topic**
  - some stuff works, but there's a long way to go
  - potential in many domains
  - integrates many different technologies
  - lots of difficult AI problems – **dialogue is hard!**
- **Commercially viable**
  - interest & investment from major IT companies

# Basic Dialogue System Types

## Task-oriented

- focused on completing a certain task/tasks
  - booking restaurants/flights, finding bus schedules, smart home...
- most actual DS in production
- “backend access” vs. “agent/assistant”

## Non-task-oriented

- chitchat – social conversation, entertainment
  - getting to know the user, specific persona
- gaming the Turing test

# Communication Domains

- “domain” = conversation topic / area of interest
- traditional: **single/closed-domain**
  - one well-defined area, small set of specific tasks
  - e.g. banking system on a specific phone number
- **multi-domain**
  - basically joining several single-domain systems
- **open-domain**
  - “responds to anything”
  - used to be mostly chitchat, now somewhat working via LLMs

# Application Areas

- **phone** (traditional)

- users call a phone number, a dialogue system picks up
- even DTMF systems belong here (e.g. banks, phone operators)
- information – buses (Let's Go), restaurants/tourist info

<http://www.speech.cs.cmu.edu/letsgo/example.html>

<https://youtu.be/lHfLr1MF7DI>

- **apps**

- assistant apps for your phone/computer
- language learning, navigation (Spacebook) <https://youtu.be/qQZnwrOyeTE?t=65>
- companions (Xiaolce)

- **smart speakers**

- home automation, assistants (Alexa/Google Home)

- **appliances**

- voice operated TVs
- other devices connect to smart speakers



<https://www.digitaltrends.com/mobile/5-things-you-need-to-know-about-microsofts-chinese-girlfriend-chatbot-xiaoice/>

# Application Areas

- **cars**

- hands-free car-specific functions
- Android Auto, Apple CarPlay, vendor-specific solutions



- **web**

[https://www.irozhlas.cz/zpravy-domov/ministerstvo-zdravotnictvi-web-sestra-anezka-chatbot-provoz-konec\\_2101181306\\_tzr](https://www.irozhlas.cz/zpravy-domov/ministerstvo-zdravotnictvi-web-sestra-anezka-chatbot-provoz-konec_2101181306_tzr)

- search assistants (IKEA Anna, ČS George, Anežka)
- ChatGPT & co. <https://chat.openai.com/>
- Facebook Messenger chatbots
- chit-chat chatbots (Pandorabots)

<https://george.csas.cz>



Dobrý večer, jsem chatbot George, virtuální bankéř a rád zodpovím Vaše dotazy. Kdykoliv Vás můžu spojit i s mým lidským kolegou, stačí napsat a hned Vás na něj přepojím.

S čím Vám můžu poradit?

Zaslání karty

Změna limitů ke kartě

Délka převodu

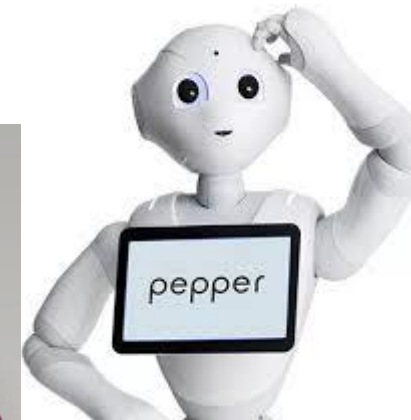
Změna osobních údajů

- **embodied (robots)**

- information assistants

- **virtual characters**

- computer games
- therapy, elderly care



<https://robots.nu/en/robot/Pepper>

(DeVault et al., 2014) <https://dl.acm.org/doi/10.5555/2615731.2617415>

# Modes of Communication

- **text**

- most basic/oldest
- easiest to implement, robust
- not completely natural

- **voice**

- more difficult, but can be more natural
- easy to deploy over the phone

- **multimodal**

- voice/text + graphics
- additional modalities: video – gestures, mimics; touch
- most complex

(Johnston et al., 2002)

<https://www.aclweb.org/anthology/P02-1048/>



(Skantze & Al Moubayed, 2012)

<https://doi.org/10.1145/2388676.2388698>



# Dialogue Initiative

- **system-initiative**

- “form-filling” (*“Hello. Please tell me your date of birth.”*)
- system asks questions, user must reply in order to progress
- traditional, most robust, but least natural

- **user-initiative**

- user asks, machine responds (*“Alexa, set the timer for two minutes”*)

- **mixed-initiative**

- system and user both can ask & react to queries
- most natural, but most complex

*S: Hello. How may I help you?*

*U: I’m looking for a restaurant.*

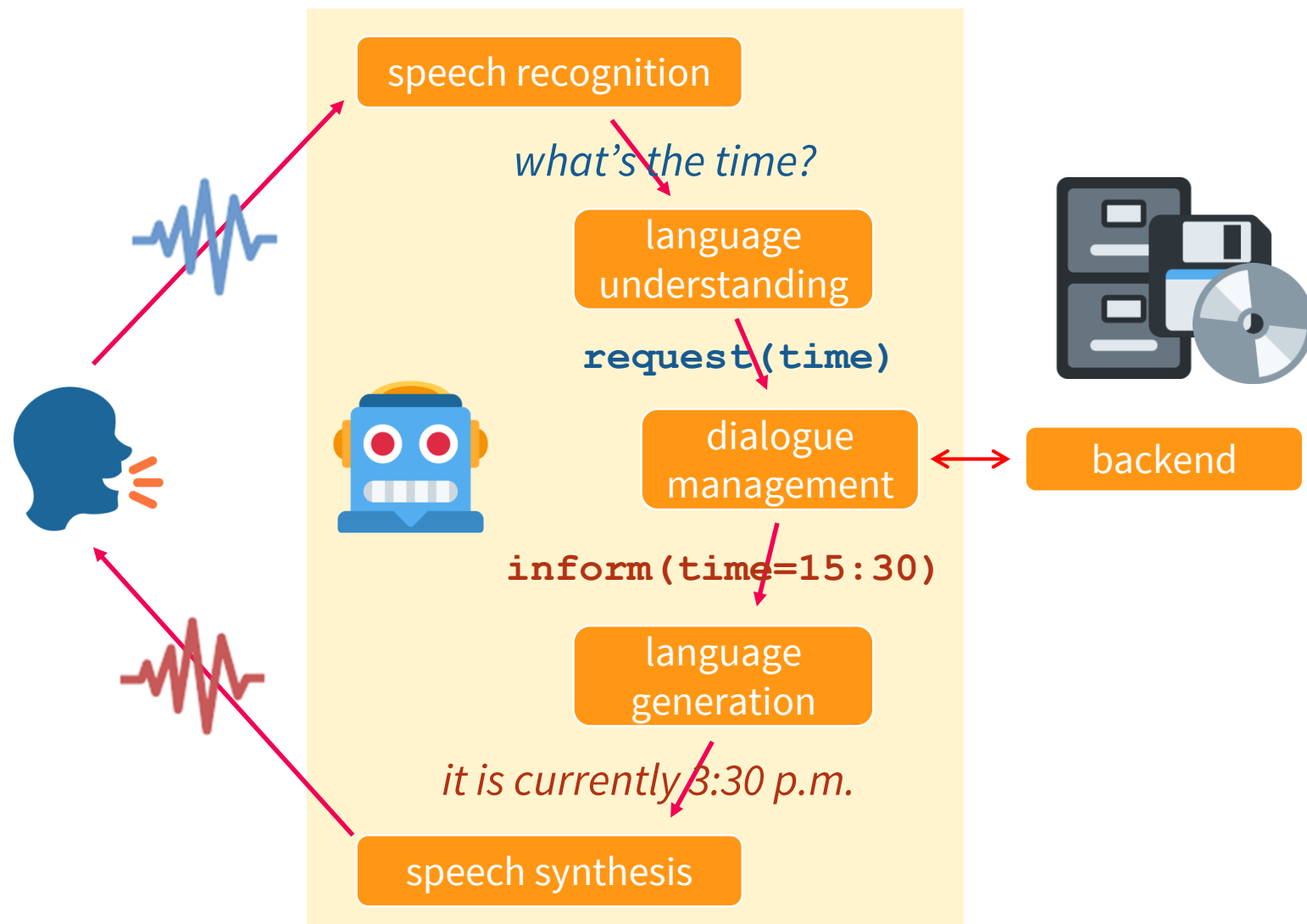
*S: What price do you have in mind?*

*U: Something in the city center please.*

*S: OK, city center. What price are you looking for?*

# (Task-oriented) Dialogue Systems Architecture

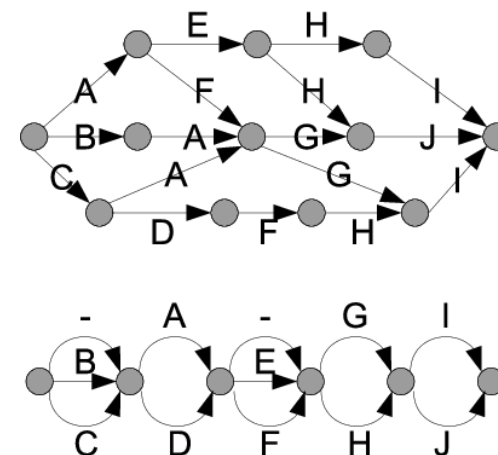
- main loop:
  - voice → text
  - text → meaning
  - meaning → reaction
  - reaction → text
  - text → voice
- access to backend
  - required to perform tasks
- multimodal systems:  
additional components



# Automatic Speech Recognition (ASR)

- Converting **speech signal** (acoustic waves) **into text**
- Typically produces several possible hypotheses with confidence scores
  - **n-best list**
  - lattice
  - confusion network
- Very good in ideal conditions
- **Problems:**
  - noise, accents, distance, channel (phone)...

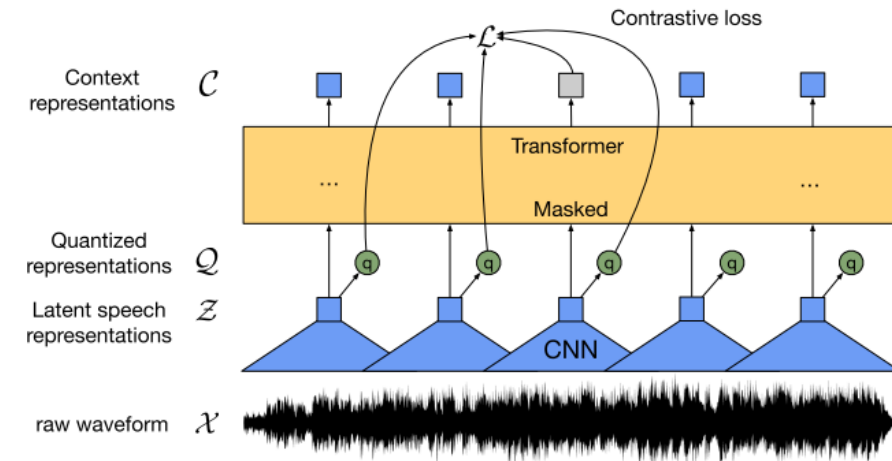
*0.8 I'm looking for a restaurant*  
*0.4 uhm looking for a restaurant*  
*0.2 looking for a rest tour rant*



Kazemian et al., ICMR 2008  
DOI 10.1145/1460096.1460112

# Speech Recognition

- Also: **voice activity detection**
  - detect when the user started & finished speaking
  - wake words (“OK, Google”)
- ASR implementation: mostly **neural networks**
  - take acoustic features (frequency spectrum)
  - compare with previous
  - emit letters
- Limited domain: use of language models
  - some words/phrases more likely than others
  - previous context can be used



(Baevski et al., 2020)  
<http://arxiv.org/abs/2006.11477>

# Natural/Spoken Language understanding (NLU/SLU)

- **Extracting the meaning** from the (now textual) user utterance
- Converting into a structured semantic representation
  - **dialogue acts:**
    - act type/intent (*inform, request, confirm*)
    - slot/attribute (*price, time...*)
    - value (*11:34, cheap, city center...*)
  - other, more complex – e.g. syntax trees, predicate logic
- Specific steps:
  - **named entity recognition** (NER)
    - identifying task-relevant names (London, Saturday)
  - **coreference resolution**
    - (“*it*” → “*the Athletic Arms bar*”)

*inform(food=Chinese, price=cheap)*  
*request(address)*

# Language Understanding

- Implementation varies
  - (partial) **handcrafting** viable for limited domains
    - keyword spotting
    - regular expressions
    - handcrafted grammars
  - **machine learning** – various methods
    - intent classifiers + slot/value extraction
- Can also provide n-best outputs
- Problems:
  - recovering from bad ASR
  - ambiguities
  - variation

*S: Leaving Baltimore. What is the arrival city?*

*U: fine Portland [ASR error]*

*S: Arriving in Portland. On what date?*

*U: No not Portland Frankfurt Germany*

*[On a Tuesday]*

*U: I'd like to book a flight from London to New York for next Friday*

*U: Chinese city center*

*U: uhm I've been wondering if you could find me a restaurant that has Chinese food close to the city center please*

# Dialogue Manager (DM)

- Given NLU input & dialogue so far, responsible for **deciding on next action**
  - keeps track of what has been said in the dialogue
  - keeps track of user profile
  - interacts with backend (database, internet services)
- Dialogue so far = **dialogue history**, modelled by **dialogue state**
  - managed by **dialogue state tracker**
- System actions decided by **dialogue policy**



# Dialogue state / State tracking

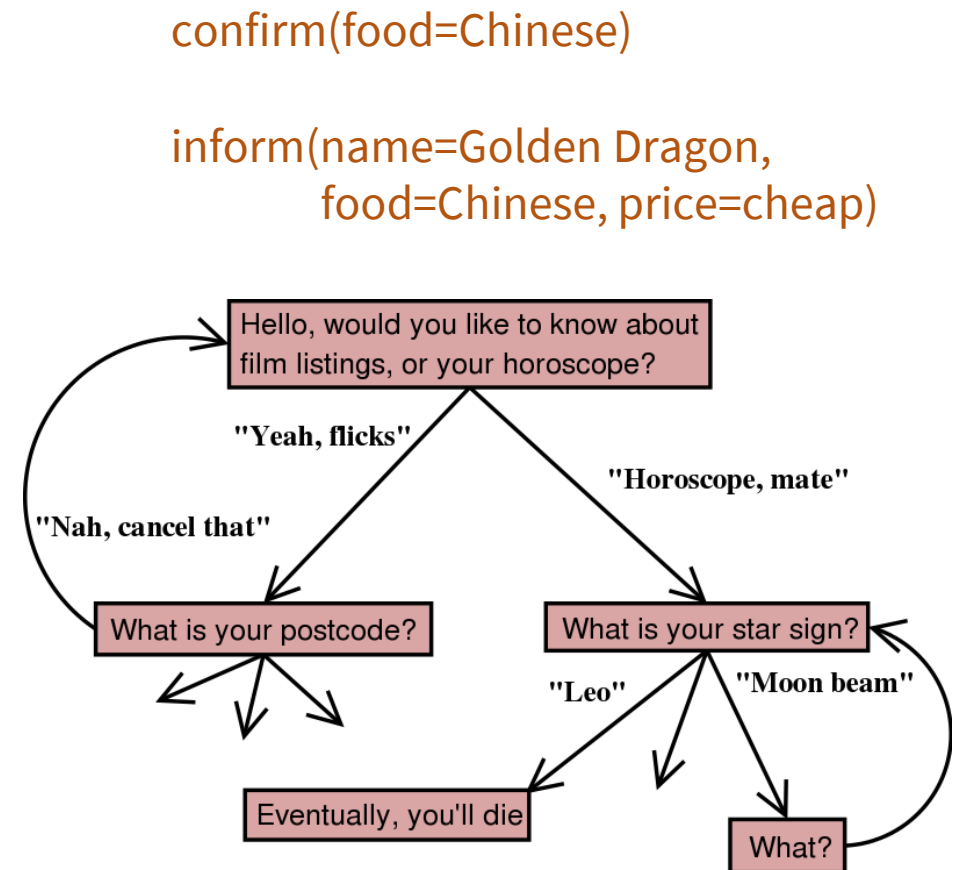
- Stores (a summary of) dialogue history
  - User requests + information they provided so far
  - Information requested & provided by the system
  - User preferences
- Implementation
  - **handcrafted** – e.g. replace value per slot with last-mentioned
    - good enough in some circumstances
  - **probabilistic** – keep an estimate of per-slot preferences based on SLU output
    - more robust, more complex

price: cheap  
food: Chinese  
area: riverside

price: 0.8 cheap  
0.1 moderate  
0.1 <null>  
food: 0.7 Chinese  
0.3 Vietnamese  
area: 0.5 riverside  
0.3 <null>  
0.2 city center

# Dialogue Policy

- Decision on next system action, given dialogue state
- Involves backend queries
- Result represented as system dialogue act
  - **confirm**(food=Chinese)
  - **inform**(name=Golden Dragon, food=Chinese, price=cheap)
- Handcrafted:
  - **if-then-else** clauses
  - **flowcharts** (e.g. VoiceXML)
- Machine learning
  - often trained with **reinforcement learning**
  - POMDP (Partially Observable Markov Decision Process)
  - recurrent neural networks



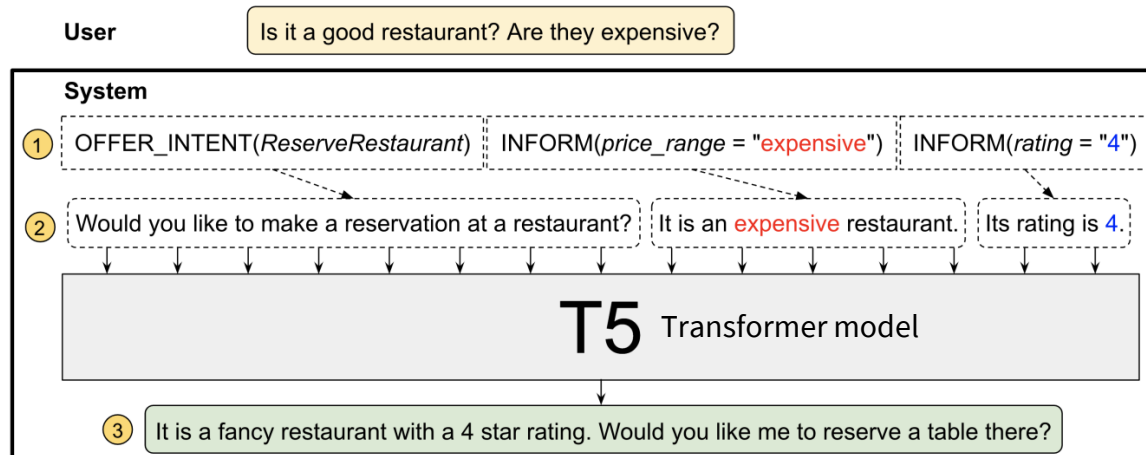
# Natural Language Generation (NLG) (Response Generation)

- Representing system dialogue act in natural language (text)
  - reverse NLU
- How to express things might depend on context
  - Goals: fluency, naturalness, avoid repetition (...)
- Traditional approach: **templates**
  - Fill in (=lexicalize) values into predefined templates (sentence skeletons)
  - Works well for limited domains

inform(name=Golden Dragon, food=Chinese, price=cheap)  
+  
<name> is a <price>-ly priced restaurant serving <food> food  
=  
Golden Dragon is a cheaply priced restaurant serving Chinese food.

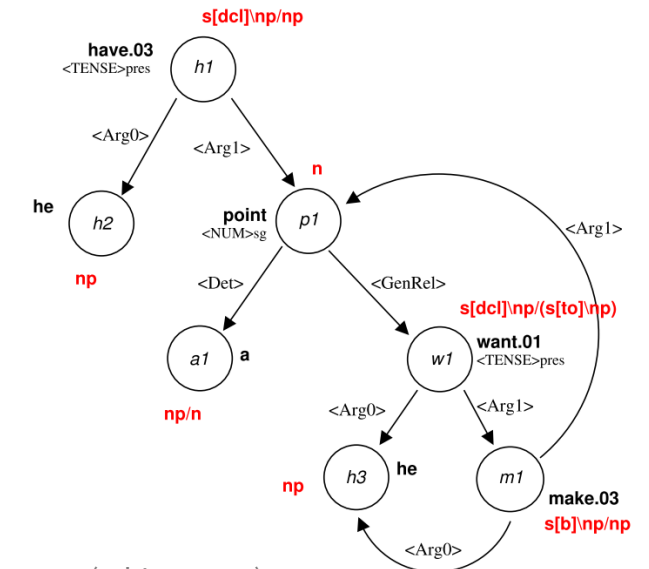
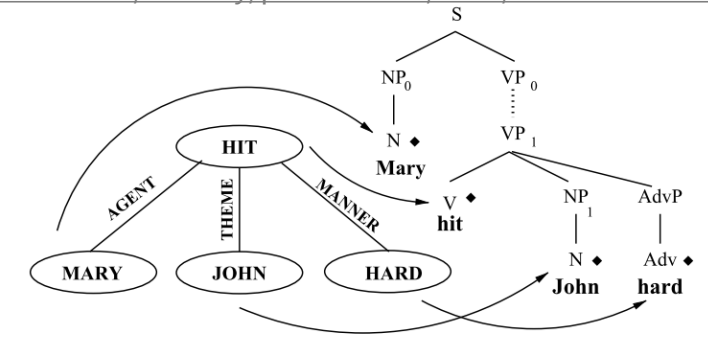
# Natural Language Generation

- Grammar-based approaches
  - grammar/semantic structures instead of templates
  - NLG **realizes** them (=converts to linear text) by applying syntactic transformation rules
- Statistical approaches
  - most prominent: **Transformer neural networks**
  - generating word-by-word
  - input: encoded semantics + previous words



(Kozłowski, 2002)

<https://www.eecis.udel.edu/~mccoy/publications/2002/Kozlowski-ACL-Stu.ps>



(White, 2011)

<https://www.aclweb.org/anthology/W11-2827/>

(Kale & Rastogi, 2020)

<https://aclanthology.org/2020.emnlp-main.527/>

# Text-to-speech (TTS) / Speech Synthesis

- Generate a speech signal corresponding to NLG output
  - text → sequence of **phonemes**
    - minimal distinguishing units of sound (e.g. [p], [t], [ŋ] “ng”, [ə] “eh/uh”, [i:] “ee”)
  - + pitch/intonation, speed, pauses, volume/accents

- Standard pipeline:

- text normalization
  - abbreviations
  - punctuation
  - numbers, dates, times

*take bus number 3 at 5:04am*





take bus number three at five o four a m

tɛɪk bʌs nʌmbə θriː æt faɪv əʊ fɔːr eɪ ɛm

- pronunciation analysis (**grapheme → phoneme conversion**)
- intonation/stress generation
- waveform synthesis

# Speech Synthesis

- TTS Methods:

- Formant-based: phoneme-specific frequencies  <http://www.festvox.org/history/klatt.html> (example 33)
  - oldest, not very natural, but works on limited hardware
- Concatenative  <https://en.wikipedia.org/wiki/MBROLA>
  - record a single person, cut into phoneme transitions (diphones), glue them together
- Hidden Markov Models  <http://flite-hts-engine.sp.nitech.ac.jp/>
  - phonemes in context modelled as hidden Markov models
  - Model parameters estimated from data (machine learning)
- Neural networks  <https://google.github.io/tacotron/>
  - HMMs swapped for a recurrent neural network / end-to-end neural
  - can go directly from text, no need for phoneme conversion

# Organizing the Components

- Basic: pipeline
  - ASR → NLU → DM → NLG → TTS
  - components oblivious of each other
- Interconnected
  - read/write changes to dialogue state
  - more reactive (e.g. incremental processing), but more complex
- Joining the modules (experimental)
  - ASR + NLU
  - NLU + state tracking
  - NLU & DM (& NLG sometimes)



# Dialogue Systems Research

- Multi/open domains
  - reusability, domain transfer
- Joint models (“end-to-end”) – all-in-one neural network
- Multimodality
  - adding video (input/output)
- Context dependency
  - understand/reply in context (grounding, speaker adaptation)
- Incrementality
  - don't wait for the whole sentence to start processing

# Summary

- We're far from AI sci-fi dreams, but it still works a bit
  - dialogue is hard
- DSs have many forms & usage areas
  - **task-oriented vs. non-task-oriented**
  - **closed, multi vs. open domain**
  - system vs. user initiative
- Main components: **ASR → NLU → DM → NLG → TTS**
  - implementation varies
- It's an active and interesting research topic!
- Next week: linguistic background

### Contact us:

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

### Get the slides here:

<http://ufal.cz/npfl123>

### References/Inspiration/Further:

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

- Pierre Lison (Oslo University): <https://www.uio.no/studier/emner/matnat/ifi/INF5820/h14/timeplan/index.html>
- Oliver Lemon & Verena Rieser (Heriot-Watt University): <https://sites.google.com/site/olemon/conversational-agents>
- Filip Jurčíček (Charles University): <https://ufal.mff.cuni.cz/~jurcicek/NPFL099-SDS-2014LS/>
- Milica Gašić (University of Cambridge): <http://mi.eng.cam.ac.uk/~mg436/teaching.html>
- David DeVault & David Traum (Uni. of Southern California): <http://projects.ict.usc.edu/nld/cs599s13/schedule.php>
- Luděk Bártek (Masaryk University Brno): <https://is.muni.cz/el/1433/jaro2018/PA156/um/>
- Gina-Anne Levow (University of Washington): <https://courses.washington.edu/ling575/>