

# Speech recognition and synthesis

## 1 Dialog systems

- Introduction
- Turns
- Speech acts
- Minimal responses
- Conversations
- Automatic Dialog System basics
- Recognizer
- Speech Generator
- Dialog management
- Bibliography

Copyright ©2007-2008 R.J.J.H. van Son, GNU General Public License [FSF(1991)]



# Introduction

Speech recognition and synthesis are most useful if combined into a full Human-Machine dialog system

- Human conversations are extremely efficient and effective interactions
- Spoken dialogs are not like a command-line Question-Answer query session
- Conversations include “control” signals at *low* (pre-verbal) and *high* levels
- Humans speak in *turns*
- In simple automated systems, interactions must be restricted and well structured

Many pictures (and their copyrights) are from [Jurafsky and Martin(2000)]



# Introduction

## In conversations, timing is everything

- Human dialogs are composed of game-like *moves*
- *Turn* distribution is crucial for effective Human-Machine interactions
  - *who* speaks next
  - *when* should the next speaker start
- Central to human conversations is *projection*
- *Projection* is the ability to predict the
  - *timing* of turns
  - *type* of upcoming moves



# Turns

## What defines a turn?

- A *single* move in the conversation “game”
- Ends with the *end* of the last utterance
- Utterance *completes* a move
- Does *not* end in a level tone
- Does not end in a *filled* pause (eg, “uuhh”)
- Can be followed by a *silent pause*

The end of a turn is a *TRP*, a *Transition Relevance Place*.



# Turns: TRPs

Turns and Turn taking. At each TRP of each turn:

- If during this turn the current speaker has selected *A* as the next speaker then *A* must speak next
- If the current speaker does not select the next speaker, *any* other speaker may take the next turn
- If no one else takes the next turn, the *current* speaker may take the next turn



# Speech acts

Conversational *moves* are build from *speech acts*

## Basic speech acts

- **Assertives:** committing Sp. to something's being the case  
*suggesting, putting forward, swearing, boasting, concluding*
- **Directives:** attempts by Sp. to get addressee to do something  
*asking, ordering, requesting, inviting, advising, begging*
- **Commissives:** committing Sp. to some future course of action  
*promising, planning, vowing, betting, opposing*
- **Expressives:** expressing psychological state of Sp. about state of affairs  
*thanking, apologizing, welcoming, deploring*
- **Declarations:** changing the world by speech  
*E.g. "I resign", "You're fired"*



# Speech acts

## Basic control tasks, handle conversation flow

- **Attention** *someone is listening*
  - Visually, by looking
  - By using *minimal responses* whenever possible
- **Acknowledgment** *move is received*
- **Grounding** *move is integrated, or not*
  - Okay, etc.
  - By minimal responses
  - By (partially) repeating previous move
  - By a relevant next move
- **Assessing** *move is judged*
- **Relevant move** *just start a relevant turn*
- *New turn* can subsume *Assessing* can subsume *Grounding* can subsume *Acknowledgment* can subsume *Attention*

# Speech acts

## Timing of responses

- Respond immediately
- If a *complex* response cannot be given in time, switch to a *simpler, faster* response type
- If all else fails, start with an *Uhhhh* placeholder
- Signal problems with a *delayed* response
- Eg, an immediate repeat signals *acknowledgment*, a delayed repeat asks for *confirmation*
- If refusal or repair is dispreferred insert *significant silence*





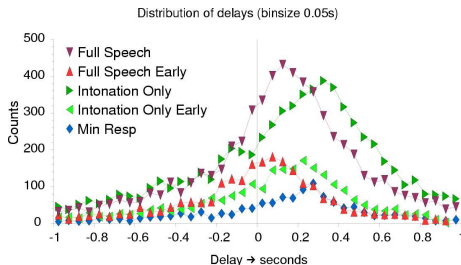
# Minimal responses

## Also: Backchannels, continuer, acknowledgment tokens

- *Uh, Uhm, HmmHmm, Yes, Sure*, etc.
- Perform the basic control tasks
- Do *not* take a turn
- Do *not* interrupt the speaker
- Are semantically, or even lexically, *empty*
- Keep the conversation going smoothly
- Without visual “feedback”, eg, on the phone, a lack of audible minimal responses interrupts the conversation



# Minimal responses: Timing



## Natural and elicited minimal responses

- Responses start directly after the TRP, even for the unintelligible signals ( $\approx 200ms$ ).
- Preparations (the *early responses*) start *before* the utterance ends

*Early responses* are laryngeal preparation signals. *Intonation Only* responses are unintelligible *uh* sounds [Wesseling and van Son(2005)][Wesseling and Van Son(2005)]



# Conversations: Implicatures

Conversations contain rules of inference

## Conversational Maxims of Grice

- **Quantity:** Be *exactly* as informative as required
  - Not *less* informative
  - Not *more* informative
- **Quality:** Speak the *truth*
  - Do not say what you believe is *false*
  - Do not say that for which you lack *evidence*
- **Relevance:** Be relevant
- **Manner:** Be *perspicuous*
  - Avoid *obscurity*
  - Avoid *ambiguity*
  - Be *brief*
  - Be *orderly*

# Conversations: Practical dialogs

General conversations are much too complex. Limit *Automatic Dialog Systems* to practical dialogues

Dialogues that are focused on a concrete task, eg,

- Task-oriented
- Information seeking
- Advice and tutoring
- Command and control

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]



# Conversations: Adjacency pairs

Practical dialogues contain many controlled turn switches, called Adjacency pairs

- Question  $\Rightarrow$  Answer
- Proposal  $\Rightarrow$  Acceptance/Rejection
- Apology  $\Rightarrow$  Acceptance/Rejection
- Summons  $\Rightarrow$  Answer



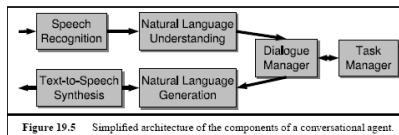
# Conversations: Example dialogue

C<sub>1</sub>: ...I need to travel in May.  
A<sub>1</sub>: And, what day in May did you want to travel?  
C<sub>2</sub>: OK uh I need to be there for a meeting that's from the 12th to the 15th.  
A<sub>2</sub>: And you're flying into what city?  
C<sub>3</sub>: Seattle.  
A<sub>3</sub>: And what time would you like to leave Pittsburgh?  
C<sub>4</sub>: Uh hmm I don't think there's many options for non-stop.  
A<sub>4</sub>: Right. There's three non-stops today.  
C<sub>5</sub>: What are they?  
A<sub>5</sub>: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time.  
The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the  
last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.  
C<sub>6</sub>: OK I'll take the 5ish flight on the night before on the 11th.  
A<sub>6</sub>: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air  
flight 115.  
C<sub>7</sub>: OK.

Figure 19.4 Part of a conversation between a travel agent (A) and client (C).

- No real minimal responses
- *Uh Hmm* as an *Acknowledgment*
- *OK, Right*, and repeating dates as *Grounding*
- A lot of *Question-Answering* pairs
- A lot of *Implicatures* (licensed inferences)

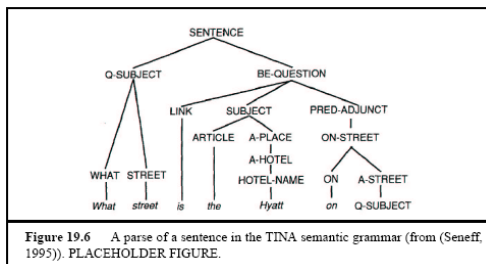
# Automatic Dialog System basics



## Three part system

- Speech recognition and understanding
  - ASR front end with adapted language model
  - NLP back end for task related semantic parsing
- Language generation and speech synthesis
  - TTS output, can be simple phrase concatenation
  - Frame based or simple grammar sentence generator
- Dialog management
  - Task related manager
  - Task Database back-end

# Recognizer



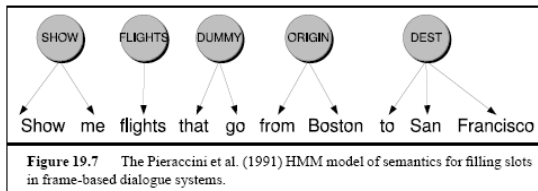
## Recognizer must deliver semantic message

- Semantic context-free grammar (SCFG) for TINA
- Mixes words and concepts
- Hand written rules

[Jurafsky and Martin(2000)]



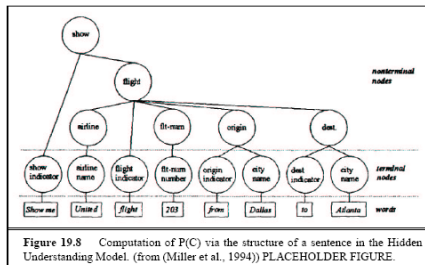
# Recognizer



## HMM concept grammar

- $\underset{C}{\operatorname{argmax}} P(C|W) = \underset{C}{\operatorname{argmax}} P(W|C) \cdot P(C)$
- $P(W|C) = \prod_{i=2,N} P(w_i | w_{i-N+1}, \dots, w_{i-1}, c_i)$
- $P(C) = \prod_{i=2,M} P(c_i | c_{i-M+1}, \dots, w_{i-1})$
- Trained on a concept-labeled corpus

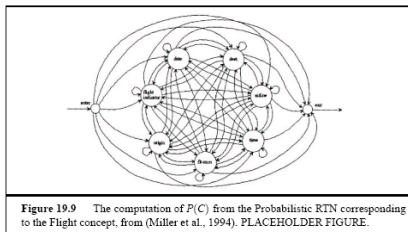
# Recognizer



## Data fragmentation problem

- Identical names can be different concepts
- Eg, cities as *origin* and *destination*
- Use a modified SCFG for  $P(C)$
- Add SCFG rules for concepts, i.e. non-terminals

# Recognizer



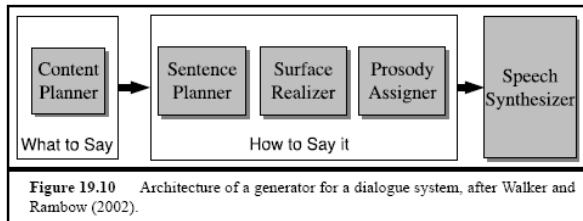
## $P(C)$ : Probabilistic finite state concept network

- Enter and Exit states
- Each arrow has a probability
- Circles indicate origin, destination, flight indicator, airline, etc.

[Jurafsky and Martin(2000)]



# Speech Generator



## Concept to speech

- The database manager generates an abstract message
- Modelled into a sentence structure
- Surface form, i.e. the words, are generated
- Prosody generated from words and content,
- Fed into a TTS system

# Dialog management

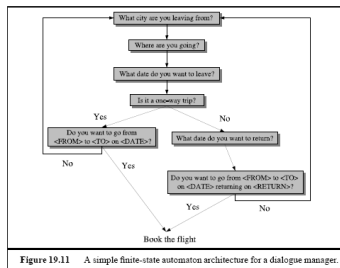


Figure 19.11 A simple finite-state automaton architecture for a dialogue manager.

## Finite state automata

- Simple dialog states
- Good for form filling dialogues (frames)
- Can handle frame switching (stochastically)

# Dialog management

Grammar	Prompt Type	
	Open	Directive
Restrictive	<i>Doesn't make sense</i>	System Initiative
Non-Restrictive	User Initiative	Mixed Initiative

**Figure 19.12** Operational definition of initiative, following Singh et al. (2002).

## Who takes the initiative

- Machine prompts all user actions  $\Rightarrow$  Finite state script
- User asks questions  $\Rightarrow$  Single frame
- Machine allows some user initiatives  $\Rightarrow$  Frame switching
- Negotiation  $\Rightarrow$  Plan based models

[Jurafsky and Martin(2000)][Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]



# Further Reading I



James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent.  
Toward conversational human-computer interaction.  
*AI Magazine*, Winter, 2001.  
URL <http://www.cs.rochester.edu/research/cisd/pubs/2001/allen-et-al-aimag2001.pdf>.



P. Boersma.  
Praat, a system for doing phonetics by computer.  
*Glott International*, 5:341–345, 2001.  
URL <http://www.Praat.org/>.



P. Boersma and D. Weenink.  
Praat 4.2: doing phonetics by computer.  
Computer program: <http://www.Praat.org/>, 2004.  
URL <http://www.Praat.org/>.



FSF.  
GNU General Public License.  
Web, June 1991.  
URL <http://www.gnu.org/licenses/gpl.html>.



Daniel Jurafsky and James H. Martin.  
*Speech and Language Processing*.  
Prentice-Hall, 2000.  
ISBN 0-13-095069-6.  
URL <http://www.cs.colorado.edu/~martin/slp.html>.  
Updates at <http://www.cs.colorado.edu/>



# Further Reading II



Helmer Strik, Albert Russel, Henk van den Heuvel, Catia Cucchiariini, and Lou Boves.

A spoken dialog system for the dutch public transport information service.

*Int. Journal of Speech Technology*, 2:121–131, 1997.

URL <http://lands.let.ru.nl/literature/strik.1996.4.ps>.

Link is to an older version.



W. Wesseling and R. J. J. H. van Son.

Timing of experimentally elicited minimal responses as quantitative evidence for the use of intonation in projecting TRPs.

In *Proceedings of Interspeech2005*, Lisbon, 2005.



Wieneke Wesseling and R.J.J.H. Van Son.

Early Preparation of Experimentally Elicited Minimal Responses.

In *Proceedings of SIGdial 2005*, September 2005.

URL <http://www.fon.hum.uva.nl/rob/Publications/ArtikelSIGdial2005.pdf>.





# Appendix A



# Copyright License

Copyright ©2007-2008 R.J.J.H. van Son, GNU General Public License [FSF(1991)]

*This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.*

*You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA.*

