Natural Language Generation for Embodied Conversational Agents

Day 2

Kristina Striegnitz

ESSLLI 2008
Hamburg, Germany

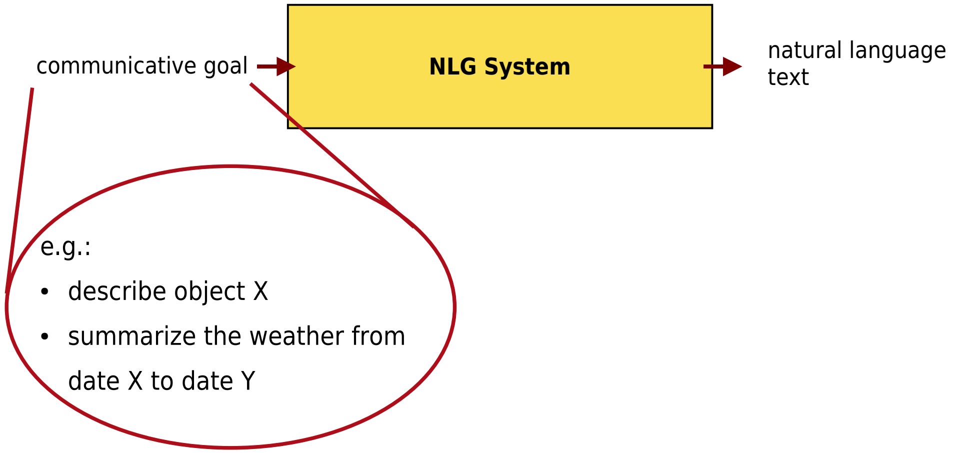
Today

- Overview of Natural Language Generation (NLG)
- Realizing Multimodal Utterances
- Where do the representations come from?
  - BEAT – a text-to-embodied-speech system
  - a grammar based approach

Today

- Overview of Natural Language Generation (NLG)
- Realizing Multimodal Utterances
- Where do the representations come from?
  - BEAT – a text-to-embodied-speech system
  - a grammar based approach

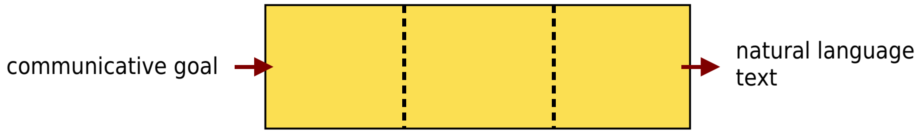
Natural Language Generation (NLG)



e.g.:

- describe object X
- summarize the weather from date X to date Y

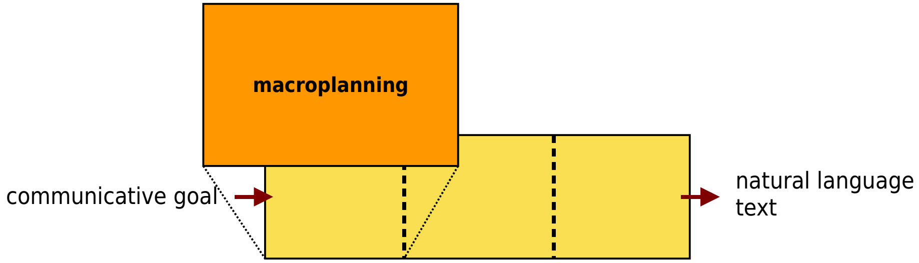
Natural Language Generation (NLG)



communicative goal

natural language text

Natural Language Generation (NLG)



macroplanning

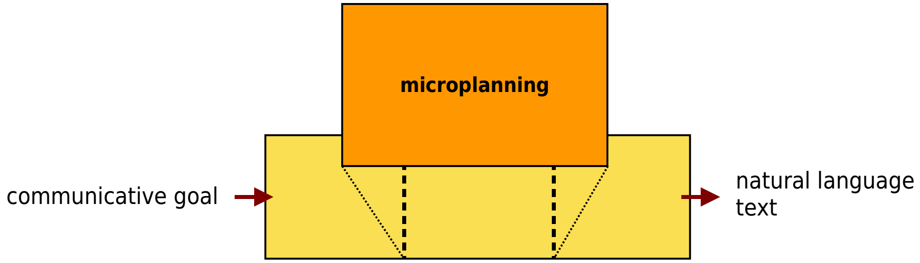
communicative goal

natural language text

also called: document planning, text planning

- selects the content that needs to be expressed (content determination)
- organizes it into a structure based on relations between pieces of content (document structuring)
- produces a text plan

Natural Language Generation (NLG)



microplanning

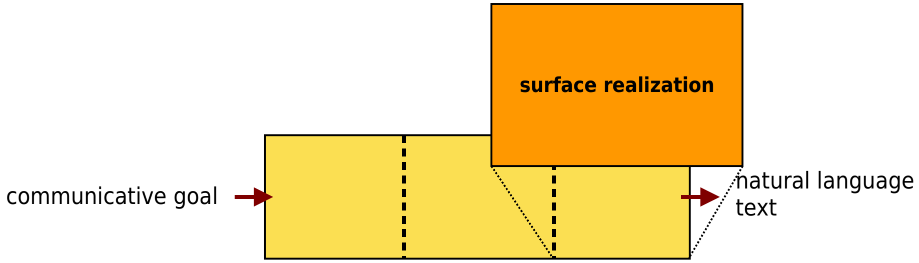
communicative goal

natural language text

also called: sentence planning, utterance planning

- decides how to distribute content over sentences (aggregation)
- decides how to refer to individuals (referring expression generation)
- produces a sequence of sentence plans

Natural Language Generation (NLG)



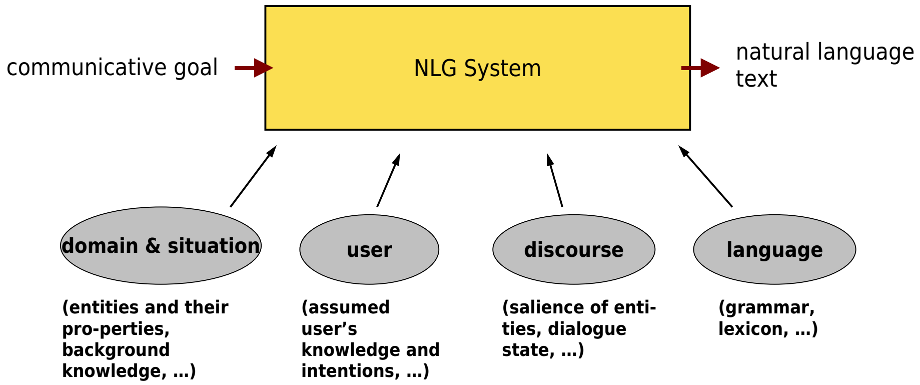
surface realization

communicative goal

natural language text

- uses grammatical constraints to specify sequence of words
- “formats” the output according to output mode
- produces the finished output

Natural Language Generation (NLG)



communicative goal

NLG System

natural language text

domain & situation

(entities and their pro-perties, background knowledge, …)

user

(assumed user's knowledge and intentions, …)

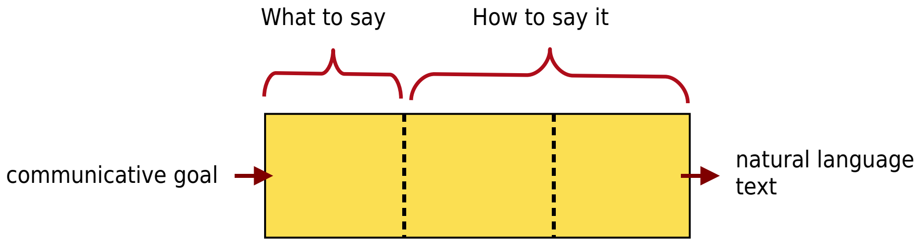
discourse

(salience of enti-ties, dialogue state, …)

language

(grammar, lexicon, …)

Natural Language Generation (NLG)



What to say

How to say it

communicative goal

natural language text

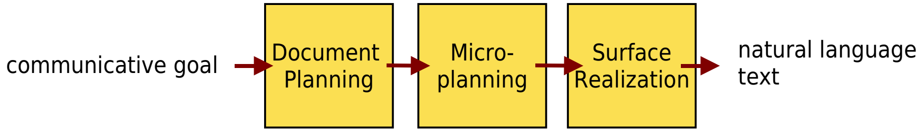
NL Generation vs. NL Understanding

David McDonald:

- Natural language generation is a process of making choices.
- Natural language understanding is a process of managing hypotheses.

Architectures of NLG systems

- Dale & Reiter's (standard) pipeline architecture:



communicative goal

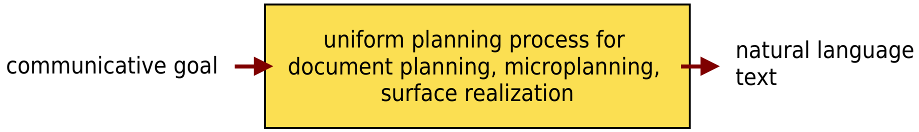
Document Planning

Micro-planning

Surface Realization

natural language text

- Integrated architecture (e.g., Appelt 1985)

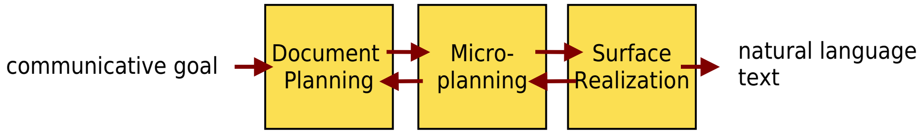


communicative goal

uniform planning process for document planning, microplanning, surface realization

natural language text

- feedback (e.g., Rubinoff 1992, Reithinger 1991, Hovy 1988)



communicative goal

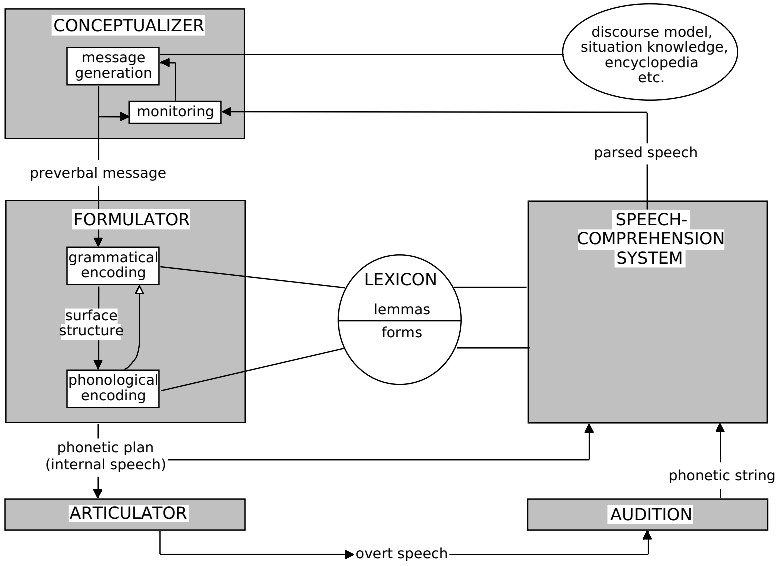
Document Planning

Micro-planning

Surface Realization

natural language text

A Psycholinguistically Motivated Architecture (Levelt 1989)



CONCEPTUALIZER

message generation

monitoring

discourse model, situation knowledge, encyclopedia etc.

preverbal message

parsed speech

FORMULATOR

grammatical encoding

surface structure

phonological encoding

LEXICON

lemmas

forms

SPEECH-COMPREHENSION SYSTEM

phonetic plan (internal speech)

phonetic string

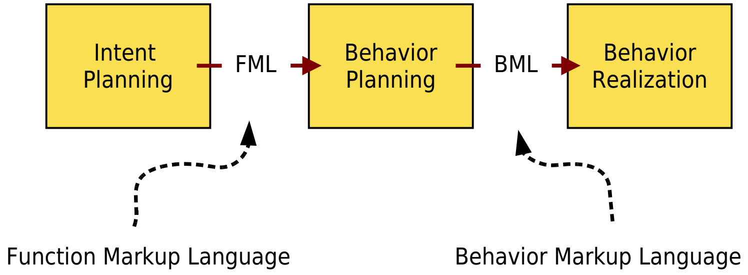
ARTICULATOR

AUDITION

overt speech

SAIBA Multimodal Behavior Generation Framework

(SAIBA = Situation, Agent, Intention, Behavior, Animation)



Intent Planning

FML

Behavior Planning

BML

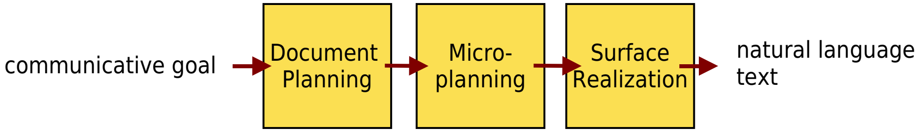
Behavior Realization

Function Markup Language

Behavior Markup Language

NLG for ECAs

- dialogue, not monologue
- output is not just words, also multimodal behavior
- When is the multimodal behavior generated?
  - text first then multimodal behavior, or
  - both together
- need to know what determines the use of different multimodal behaviors



communicative goal

Document Planning

Micro-planning

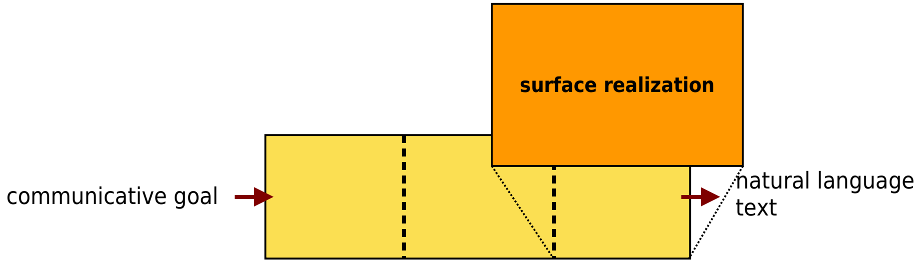
Surface Realization

natural language text

Today

- Overview of Natural Language Generation (NLG)
- Realizing Multimodal Utterances
- Where do the representations come from?
  - BEAT – a text-to-embodied-speech system
  - a grammar based approach

Realization



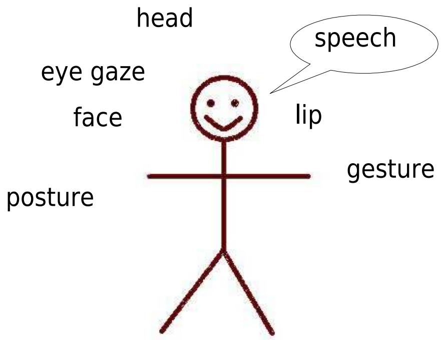
- produces the finished output
- uses grammatical constraints to specify sequence of words
- formatting if necessary

Exercise: Animate your friend

Volunteer: The class will give you instructions on how to behave: move, pose, speak ... Follow their instructions as closely as possible.

Class: You will see a video of a person speaking. “Animate” the volunteer to behave exactly like the person in the video. I.e., give him/her instructions on how to move, pose, speak, etc. so that in the end he/she will behave like the person in the video.

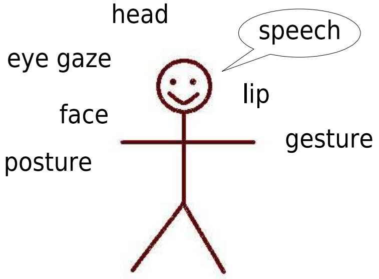
Components of a behavior specification



The Behavior Markup Language (BML)

- effort to create a standard XML interface between behavior planning and behavior realization for ECAs
- ECA researchers from Europe and the US
- work in progress

```
<bml>
  <gaze target="PERSON1"/>
  <speech>
     Welcome to my humble abode
  </speech>
</bml>
```



- goal is to be independent of a particular realizer
- provide a set of core descriptive elements and the possibility to add more detailed levels of description

http://wiki.mindmakers.org/projects:BML:main

Specifying gesture in BML (1)

type: POINT, BEAT, CONDUIT, GENERIC, LEXICALIZED

hand: LEFT, RIGHT, BOTH

amplitude: SMALL, MEDIUM, LARGE, EXTRA-LARGE

power: WEAK, NORMAL, FORCEFUL

Specifying gesture in BML (2 - lexicalized)

type: POINT, BEAT, CONDUIT, GENERIC, LEXICALIZED

lexeme: predefined animations

Specifying gesture in BML (3 - pointing)

type: POINT, BEAT, CONDUIT, GENERIC, LEXICALIZED

target: person or object in the environment

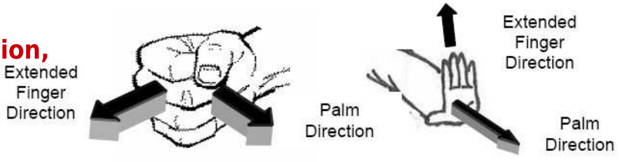
Specifying gesture in BML (4 - generic)

type: POINT, BEAT, CONDUIT, GENERIC, LEXICALIZED

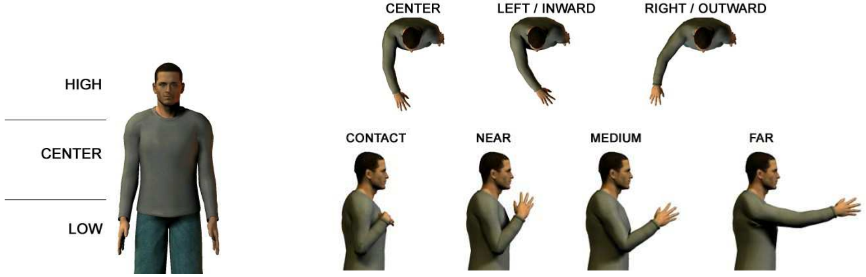
handshape: most common handshapes

orientation:

- **extended finger direction,**
- **palm direction**



location: vertical, horizontal, distance



Specifying gesture in BML (5 - movement)

type: POINT, BEAT, CONDUIT, GENERIC, LEXICALIZED

movement trajectory: straight, curved, circular, rectangular, triangular, wave-like, zigzag,…

movement direction: relative to speaker

repetition: number of times stroke is repeated

Specifying gesture in BML (5 – two handed)

type: POINT, BEAT, CONDUIT, GENERIC, LEXICALIZED

hand: LEFT, RIGHT, BOTH

two handed: coordination of the two arms; mirror, alternate, parallel, ...

Example specification

type: generic

hand: both

two handed: mirror

handshape: open hand

location: center, center, medium

orientation: palm inward, finger forward

Movie

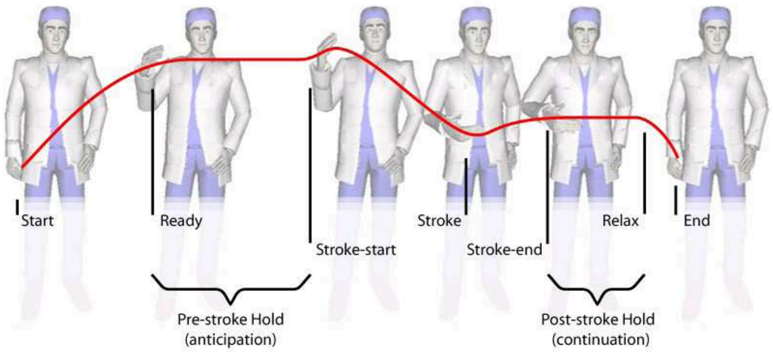
Synchronization

- Many non-verbal behaviors follow the “rhythm” of speech.
- They often depend crucially on their timing wrt. words and other non-verbal behaviors.

Synchronization in BML

- all behaviors are associated with 7 sync-points (in some cases several sync-points fall together, e.g., for gaze ready=stroke start)
- additional sync-points can be specified (e.g., in speech to synchronize with arbitrary words)

```
<speech id="s1"><text>This is a complete core level BML
<sync id="tm1"/> speech description.</text></speech>
<gesture id="g1" stroke="s1:tm1" type="BEAT">
```



Example specification

```
<speech id=”s”>
  and now take <sync id="t1"/> this bar and make it <sync id="t2"/> this
  big <sync id="t3"/>
</speech>
<gesture id=”g1” type=”POINT” target=”obj” stroke="s:t1"/>
<gesture id=”g2” type=”GENERIC” stroke-start=”t2” stroke-end=”t3”
    hand=”both”
    two handed=”mirror”
    handshape=open hand”
    location=”center, center, medium”
    orientation=”palm inward, finger forward”
/>
```

BML realization: requirements

- blending of behaviors, e.g., head shakes and gaze
- tight synchronization
  - length of non-verbal behaviors needs to adapt to timing constraints
  - starting and/or end phase may disappear or merge with starting/end phase of previous or next gesture

for more:

Kopp & Wachsmuth (2004). *Synthesizing multimodal utterances for conversational agents.*

Thiebaux et al. (2008). *SmartBody: Behavior Realization for Embodied Conversational Agents.*

Today

- Overview of Natural Language Generation (NLG)
- Realizing Multimodal Utterances
- Where do the representations come from?
  - BEAT – a text-to-embodied-speech system
  - a grammar based approach

BEAT: the Behavior Expression Animation Toolkit

[Cassell, Vilhjalmsson, Bickmore 2001]

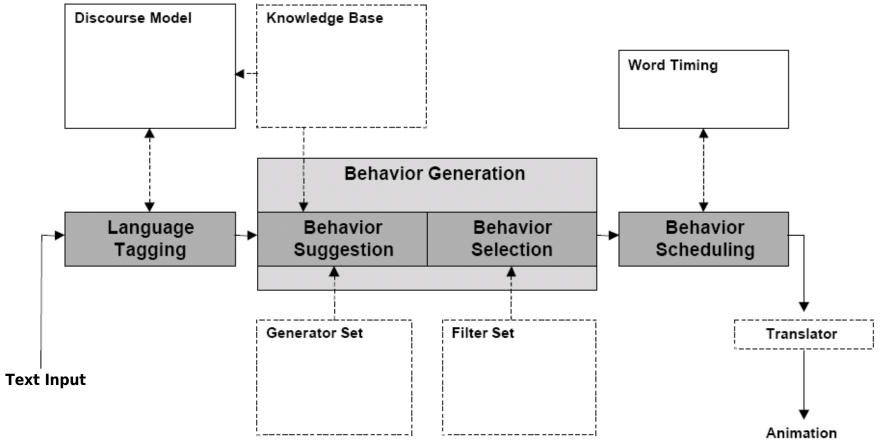
a text-to-embodied-speech system

input: text

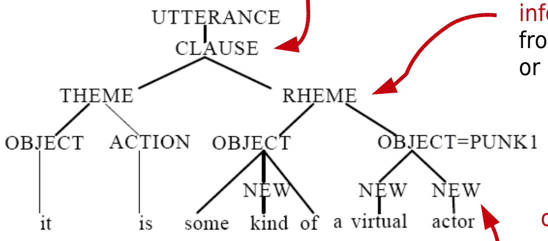
output: – a sequence of instructions that can be sent to different animation and speech synthesis systems

– specifying words, intonation, non-verbal behaviors and synchronization

BEAT: architecture



BEAT: language tagging



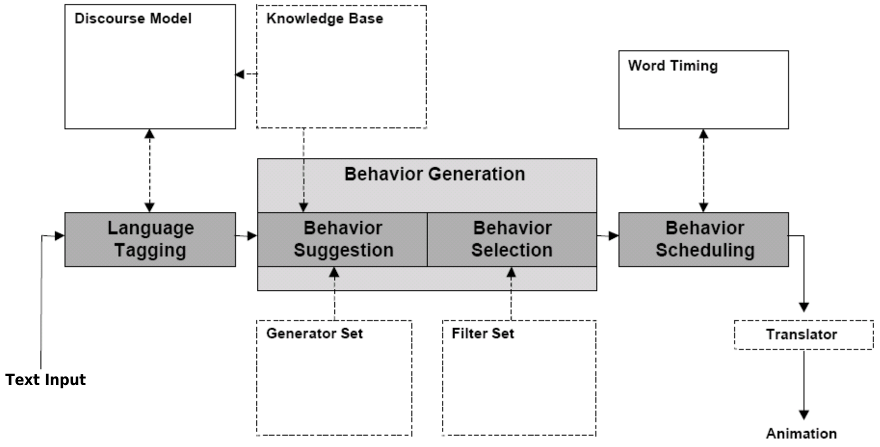
clause: single proposition; based on punctuation and syntactic parse

information structure: approximated from syntactic information and whether or not words have occurred before

objects/actions: verbs and NP chunks

newness: Has the word been used before?

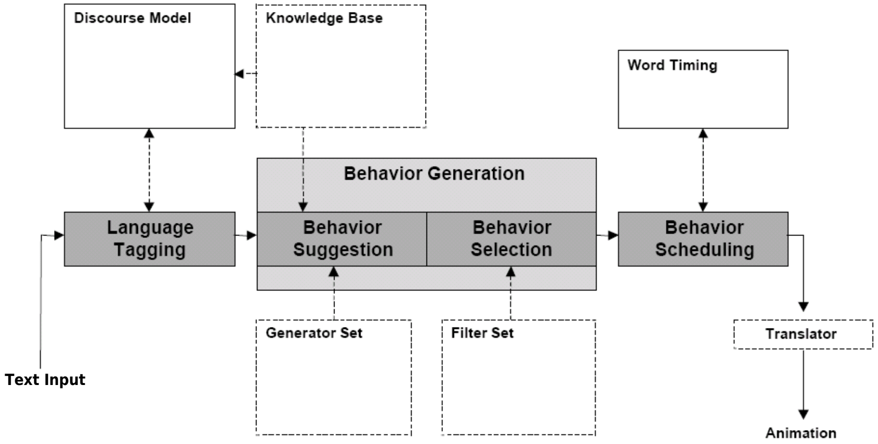
BEAT: architecture



BEAT: knowledge bases

- object knowledge
  - definitions of classes of objects and instances
  - possibly gesture specification for attributes/properties of object classes and instances
- action (verb) knowledge
  - gesture specifications for verbs

BEAT: architecture



BEAT: behavior generation

- phase 1: suggestion
  - rules that introduce non-verbal behavior → overgeneration

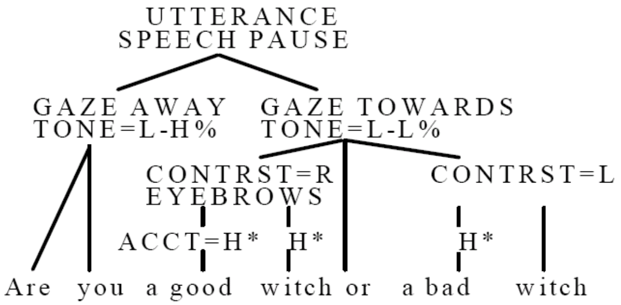
  e.g.: – associate a beat gesture with rhematic objects

  – associate an eyebrow raise with rhematic objects

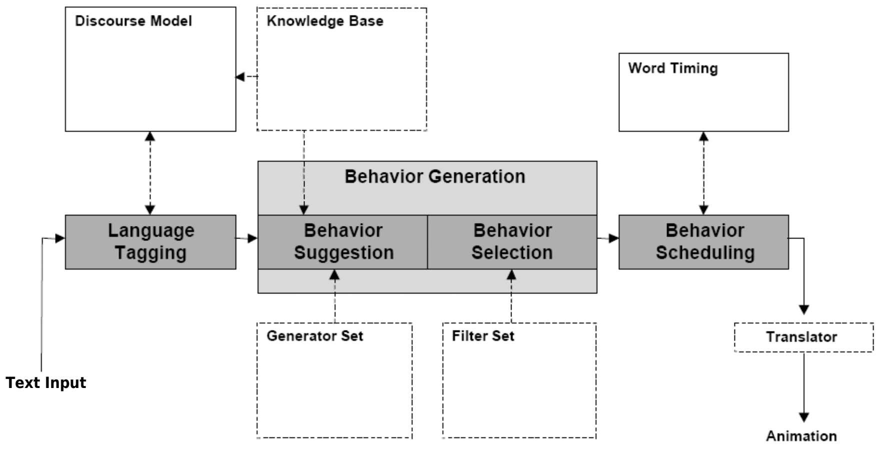
  – associate an iconic gesture with rhematic objects that have “unusual” features (as specified in the object knowledge base)
- phase 2: selection
  - rules for filtering out behaviors

  e.g.: for conflicting behaviors, keep the one with the higher priority

BEAT: behavior generation output



BEAT: architecture



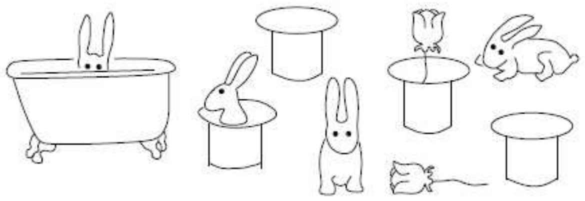
SPUD

[Stone et al. 2003]

- integrates aspects of microplanning with realization

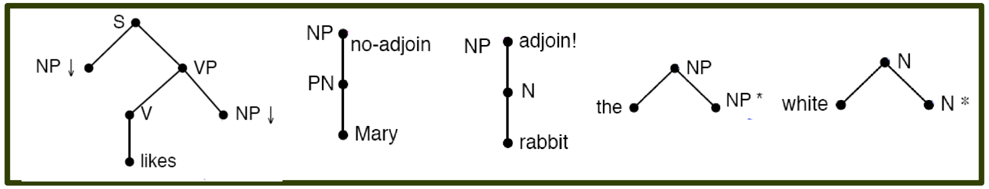
→ concise utterances

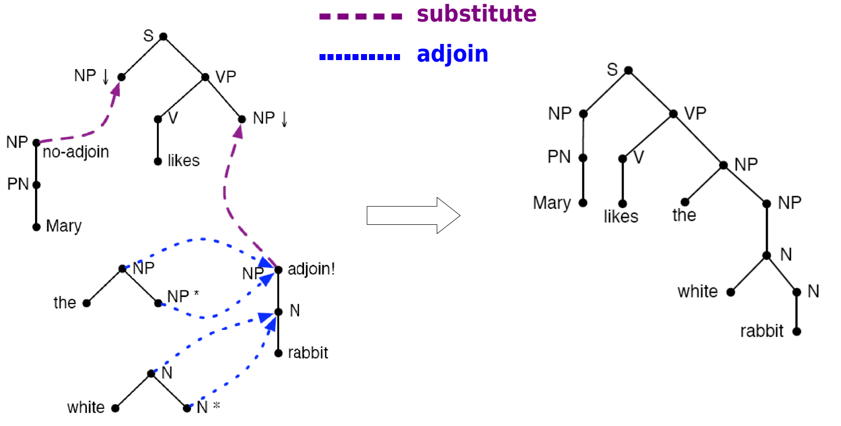
“remove the rabbit from the hat”



- general idea:
  – (parse) tree fragments associated with semantics and pragmatic constraints
  – build a tree from these fragments which is syntactically and pragmatically appropriate and fulfills all communicative goals

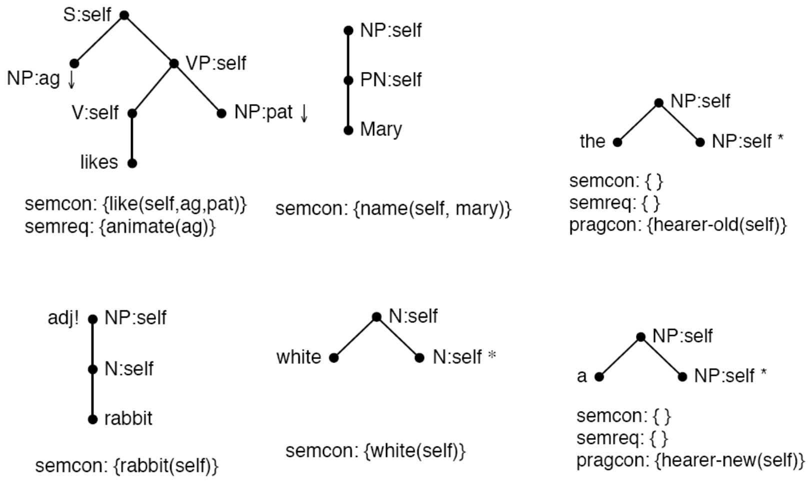
Excursion: LTAG – Lexicalized Tree Adjoining Grammar





SPUD - grammar

- LTAG with semantics and pragmatics



semcon: {like(self,ag,pat)}
semreq: {animate(ag)}

semcon: {name(self, mary)}

semcon: { }
semreq: { }
pragcon: {hearer-old(self)}

semcon: {rabbit(self)}

semcon: {white(self)}

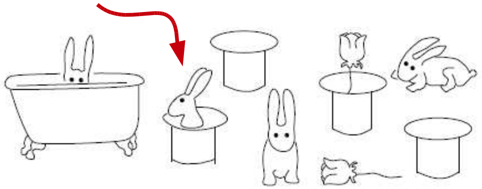
semcon: { }
semreq: { }
pragcon: {hearer-new(self)}

SPUD – generation strategy

- generation happens with respect to knowledge bases encoding:
  - shared knowledge
  - speaker's knowledge
  - pragmatic/discourse information
- a tree fragment can be use if
  - all pragmatic constraints are satisfied by the pragmatic knowledge base
  - the semantics is completely entailed by shared and/or speaker's knowledge
- we are done when
  - all syntactic constraints have been satisfied (no open substitution nodes)
  - all entities from the shared knowledge are uniquely identified

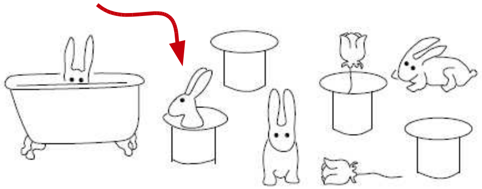
SPUD - example

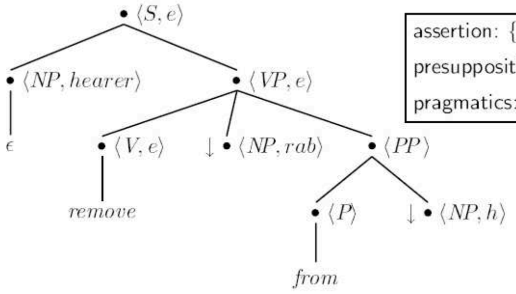
- speaker's intent: remove(e, hearer, rab, h)
- shared knowledge:



SPUD - example

- speaker's intent: remove(e, hearer, rab, h)
- shared knowledge:





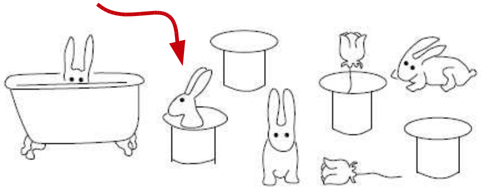
assertion: $\{remove(e, hearer, rab, h), do\_next(e)\}$

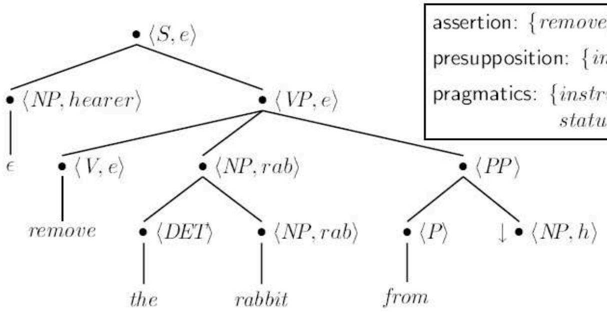
presupposition: $\{in(s, rab, h)\}$

pragmatics: $\{instruct(system, hearer)\}$

SPUD - example

- speaker's intent: remove(e, hearer, rab, h)
- shared knowledge:





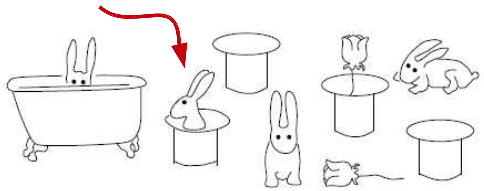
assertion: $\{remove(e, hearer, rab, h), do\_next(e)\}$

presupposition: $\{in(s, rab, h), rabbit(rab)\}$

pragmatics: $\{instruct(system, hearer),$
$status(rab, discourse\_old)\}$

SPUD - example

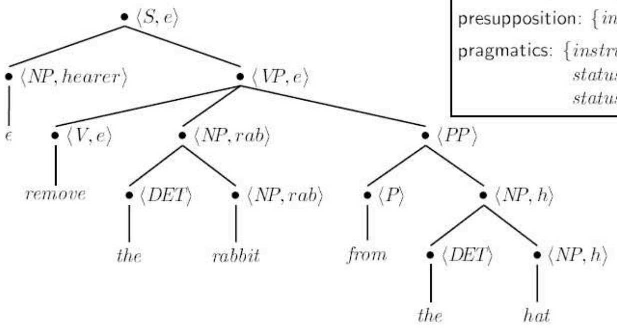
- speaker's intent: remove(e, hearer, rab, h)
- shared knowledge:



assertion: $\{remove(e, hearer, rab, h), do\_next(e)\}$

presupposition: $\{in(s, rab, h), rabbit(rab), hat(h)\}$

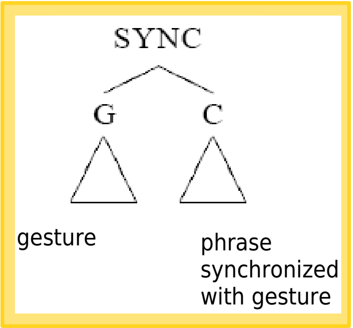
pragmatics: $\{instruct(system, hearer),$
$status(rab, discourse\_old),$
$status(h, discourse\_old)\}$



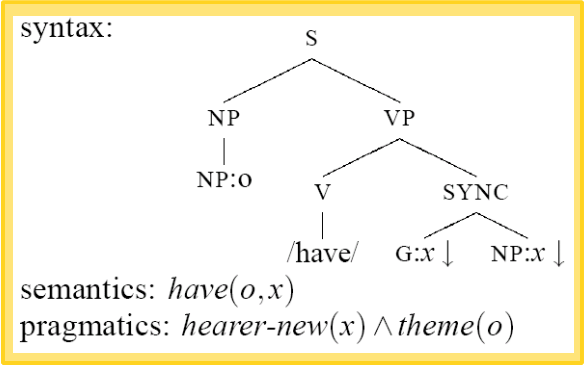
SPUD – integrating gestures

[Cassell, Stone & Yan 2000]

structure for synchronizing gestures with syntactic phrases:



example lexical entry requiring a gesture:

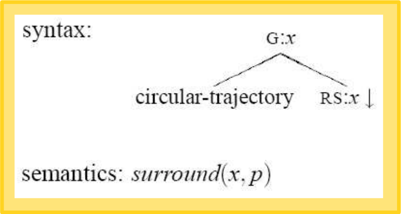


semantics: $have(o, x)$

pragmatics: $hearer\text{-}new(x) \wedge theme(o)$

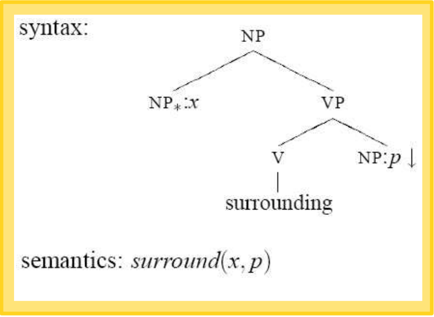
SPUD – lexical entries for gestures

a gesture entry:



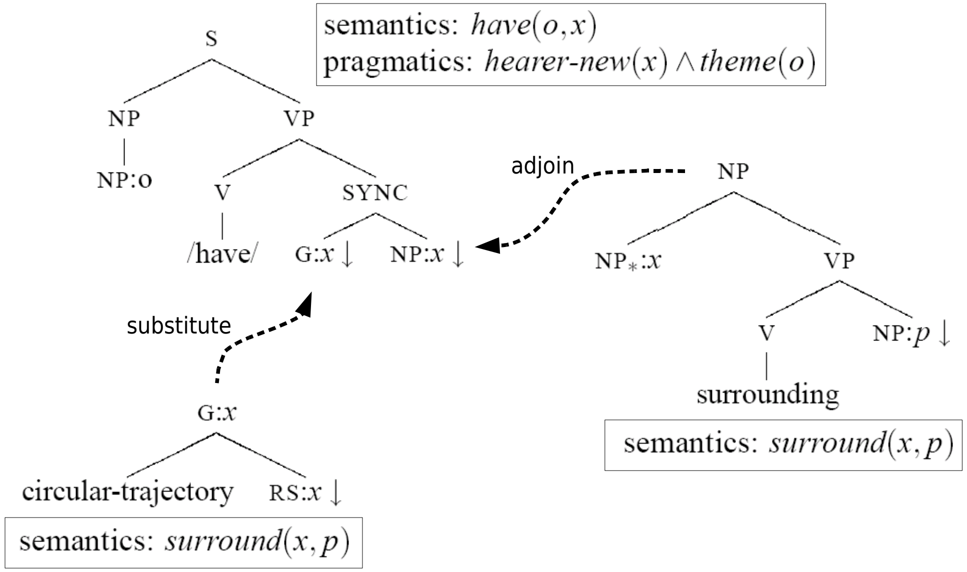
semantics: $surround(x, p)$

A “word” entry with the same semantics. Gestures can be semantically redundant or complementary:



semantics: $surround(x, p)$

SPUD – building a multi-modal utterance specification



semantics: $have(o, x)$
pragmatics: $hearer\text{-}new(x) \wedge theme(o)$

semantics: $surround(x, p)$

semantics: $surround(x, p)$

Today

- Overview of Natural Language Generation (NLG)
- Realizing Multimodal Utterances
- Where do the representations come from?
    - BEAT – a text-to-embodied-speech system
    - a grammar based approach
- Tomorrow: Referring Expression Generation