Generating Instructions in Virtual Environments

Session 2: GIVE

Alexander Koller

22 October 2009

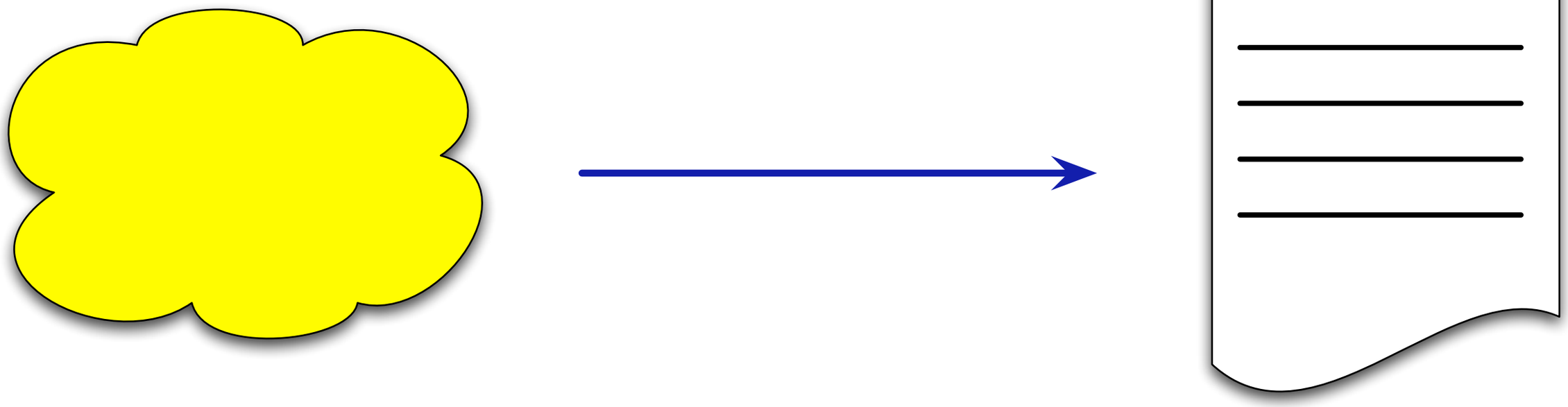
Overview

- Scheduling
- The rest of "NLG in a nutshell"
- Introduction to GIVE

Scheduling

- One time slot where everyone has time:
  - ~~Fri 16-18~~ I veto this.
- Time slots where only one person has conflict:
  - Tue 10-12
  - Wed 14-16
  - Thu 10-12 (??), 16-18
  - Fri any time

Natural language generation



whatever

text

NLG: Example applications

Input Data

Hour of the day	Wind Direction	Wind Speed
0	S	10.0
3	S	12.0
6	N	15.0
9	S	20.0
12	S	15.0
15	S	13.0
18	S	12.0
21	S	10.0
24	S	10.0

Output Text

Generate >>>

S 8-13 backing N 13-18 by early morning, then veering S 18-23 by morning, and easing 8-13 by midnight.

http://www.csd.abdn.ac.uk/research/sumtime/

Reiter et al. (2005)

NLG: Example applications

Isard et al. (2003)

"This exhibit is a lekythos, created during the archaic period. It dates from circa 500 BC. It was painted by Amasis with the red figure technique and it originates from Attica."

"... Unlike the previous exhibit, it originates from Attica."

"Questo reperto è una lekythos."

"Αυτο′ το ε′κθεμα ει′ναι μια λη′κυθος."

NLG: Example applications

S: *I found 28 round trips from Edinburgh to Brussels on October 31st and returning Edinburgh on November 6th. For option 1, I have an outbound flight with KLM departing at 7 am and arriving at 8:49 am non-stop. A return flight is with KLM departing at 10:30 am and arriving at 11:28 am with 1 stopover. Ticket price is £165. Please say "next option", or, "flight details", or, "I'll take it".*

U: Next option

S: *For option 2, I have an outbound flight with KLM departing at 3:30 pm and arriving at 5:19 pm nonstop. A return flight is with KLM departing at 8:25 pm and arriving at 12:28 pm with 1 stopover. Ticket price is £165 please say "next option", or, "flight details", or, "I'll take it."*

U: Next option

S: . . .

naive system

U: Hello. I'd like to book a flight from Edinburgh to Brussels for tomorrow arriving at around 3pm.

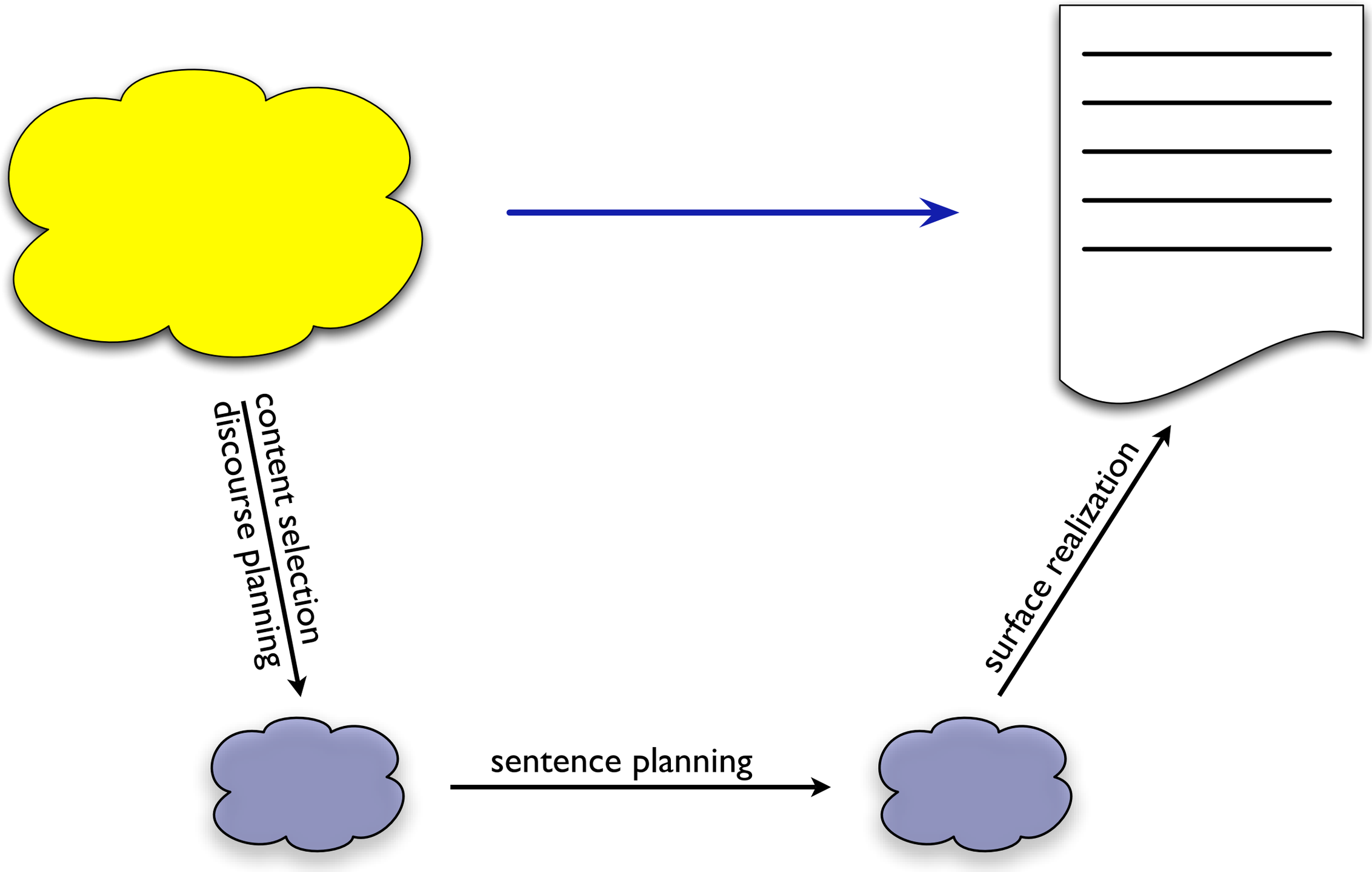
S: *There are four flights with availability in business class. None are on KLM. The only direct flight arrives at 5:30 pm, which is later than you requested. To arrive earlier, you'll have to make a connection.*
If you're willing to travel economy, there is a direct flight on BMI, arriving at 1:10 pm. There's also a KLM flight arriving at 1:50 pm, but it requires a connection in Amsterdam. All other economy flights take longer or arrive after 4 pm.

U: Can you tell me more about the business class flights that arrive earlier?

S: *The two flights with shortest travel time are both on British Airways and cost 308 pounds. The first takes 4 hours and 20 minutes and arrives at 1:20 pm. The other takes 4 hours and 30 minutes and arrives at 3:30pm.*
There's a cheaper flight on Lufthansa. It arrives at 3:10pm and costs just 270 pounds, but it takes 5 hours and 35 minutes. All other business class flights arrive very early or too late.

Demberg & Moore (2006)

How do we do this?



Pipeline model; Reiter (1994)

Content determination

- Content selection:
  - what objects do we want to talk about?
  - what do we want to say about them?
  - if necessary, interpret domain data

- Discourse planning:
  - how to arrange information into a discourse?
  - how to break information up into sentence-sized chunks?

- Both are typically domain-specific.

Content selection

	depart	arrive	airline	business	direct	price
1	9:00	13:20	BA	+	-	308
2	9:27	13:50	KLM	-	-	200
3	10:00	15:10	Lufthansa	+	-	270
4	11:00	15:30	BA	+	-	308
5	11:10	13:10	BMI	-	+	150
6	15:30	17:19	KLM	-	+	165
7	15:40	17:30	BMI	+	+	350

Demberg & Moore (2006)

User model: + direct, + business, + KLM

Input: arrive around 3pm

Discourse planning

1. Flights in business class: 4
   a) KLM: none
   b) describe f7: +direct, arrives 17:30, too late

2. Flights in economy class
   a) describe f5: BMI, arrives 13:10
   b) describe f2: KLM, arrives 13:50, *but* -direct

Surface realization

- Take specification of each sentence and translate it into the actual sentence in the output language.

- Input:
  - specification of sentence (e.g., semantic representation)
  - a grammar or something similar

- Output:
  - one sentence

Surface realization

1. Flights in business class: 4
   a) KLM: none
   b) describe f7: +direct, arrives 17:30, too late

2. Flights in economy class
   a) describe f5: BMI, arrives 13:10
   b) describe f2: KLM, arrives 13:50, *but* -direct

Template-based realization

Hard-coded rule:
"flights in business class: X"
→ "There are X flights in business class".

Flights in business class: 4

"There are four flights in business class."

Surface realization

1. There are four flights with availability in business class.
   a) KLM: none
   b) describe f7: +direct, arrives 17:30, too late

2. Flights in economy class
   a) describe f5: BMI, arrives 13:10
   b) describe f2: KLM, arrives 13:50, *but* -direct

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) describe f7: +direct, arrives 17:30, too late
2. Flights in economy class
   a) describe f5: BMI, arrives 13:10
   b) describe f2: KLM, arrives 13:50, *but* -direct

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) describe f7: +direct, arrives 17:30, too late

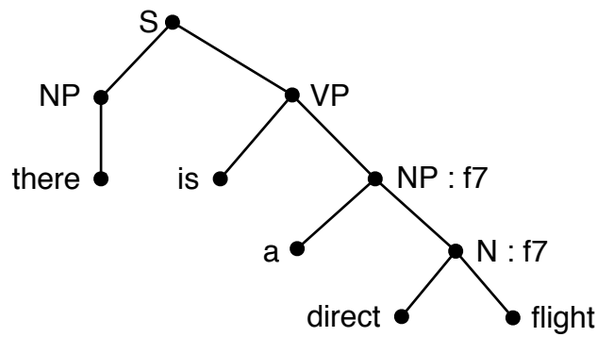
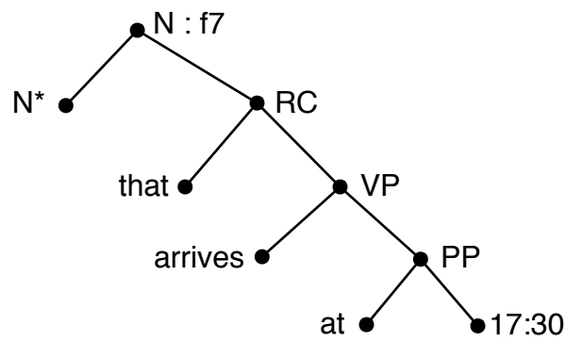
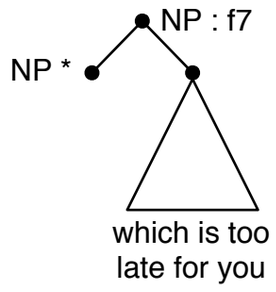
2. Flights in economy class
   a) describe f5: BMI, arrives 13:10
   b) describe f2: KLM, arrives 13:50, *but* -direct

Surface realization

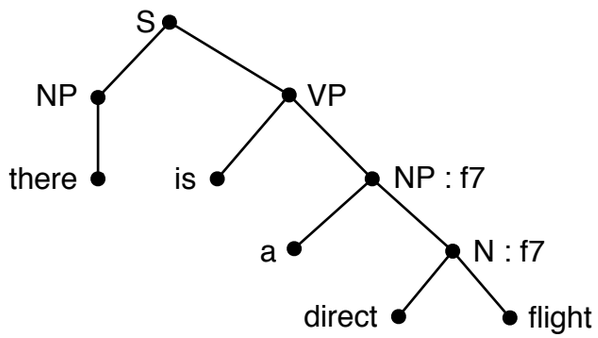
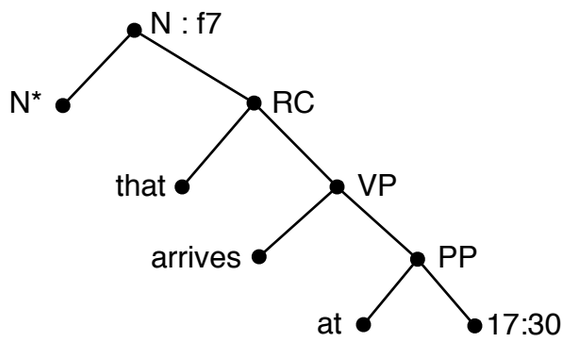
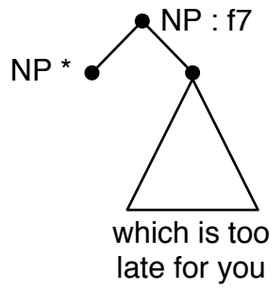
1. There are four flights with availability in business class.
   a) None are on KLM.
   b) describe f7: +direct, arrives 17:30, too late
2. Flights in economy class
   a) describe f5: BMI, arrives 13:10
   b) describe f2: KLM, arrives 13:50, *but* -direct

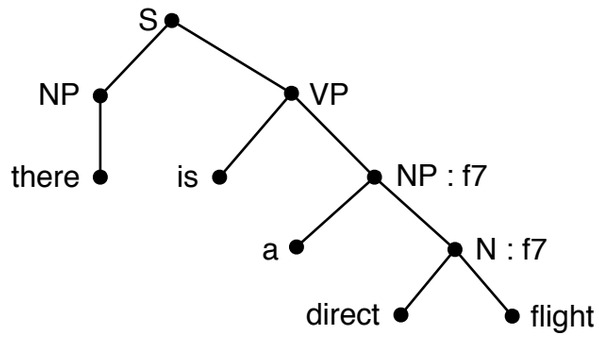
don't want separate template for each combination

Surface realization with TAG

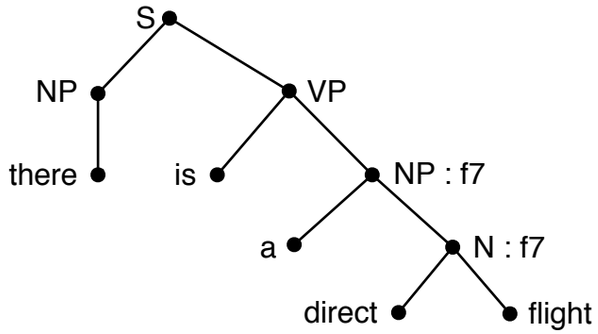
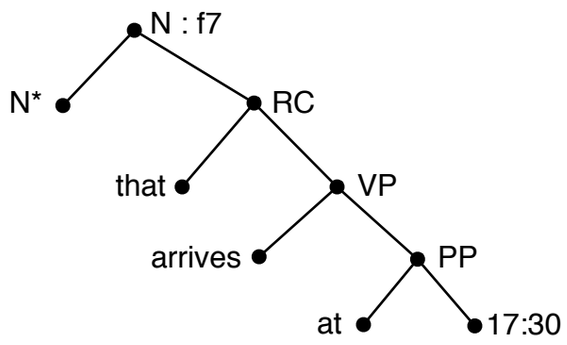
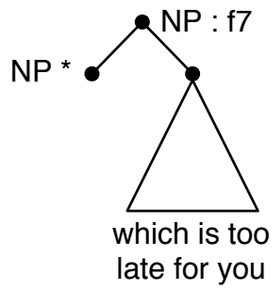
semantics	syntax
direct(f7)	
arrives(f7, 17:30)	
too_late(f7)	

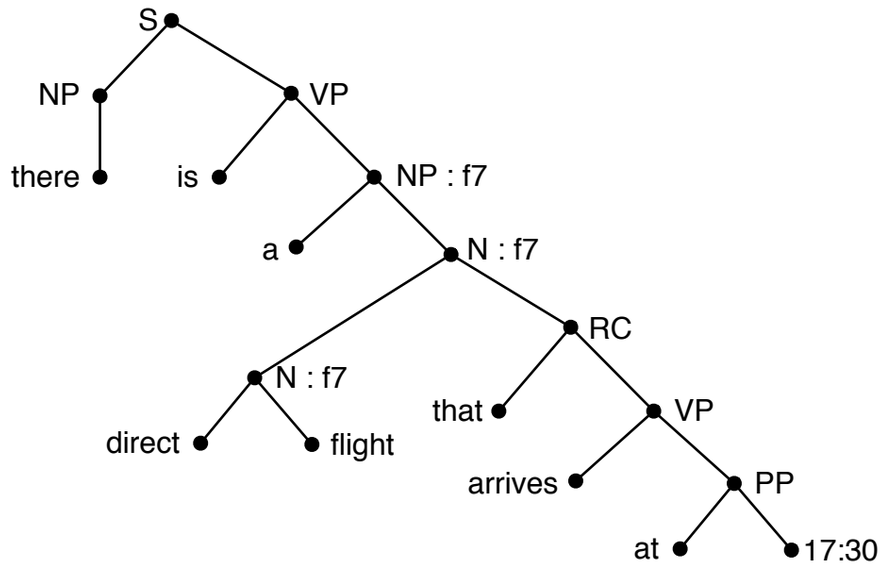
Surface realization with TAG

semantics	syntax
direct(f7)	
arrives(f7, 17:30)	
too_late(f7)	

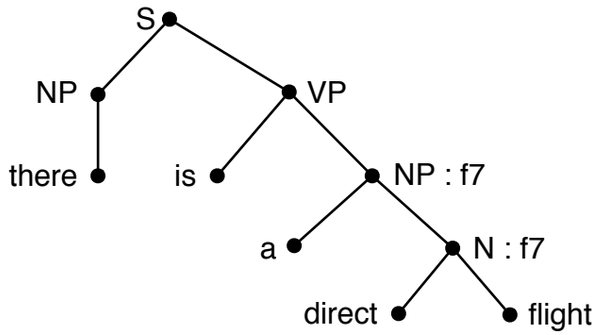
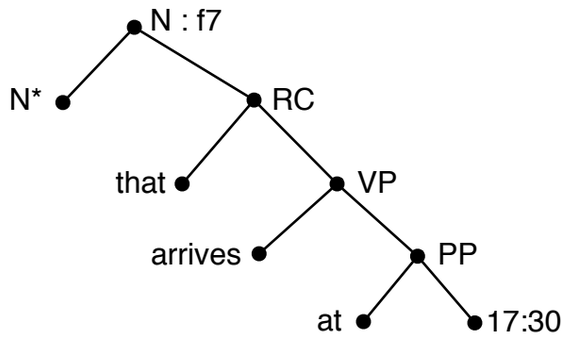
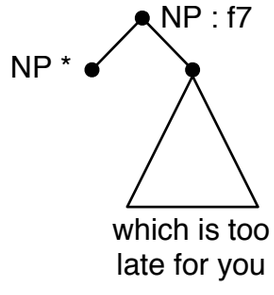


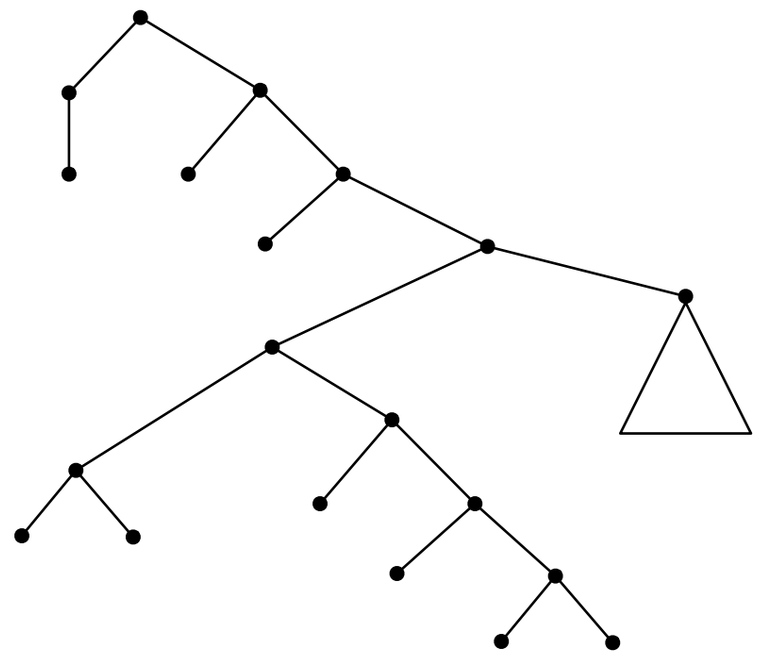
Surface realization with TAG

semantics	syntax
direct(f7)	
arrives(f7, 17:30)	
too_late(f7)	

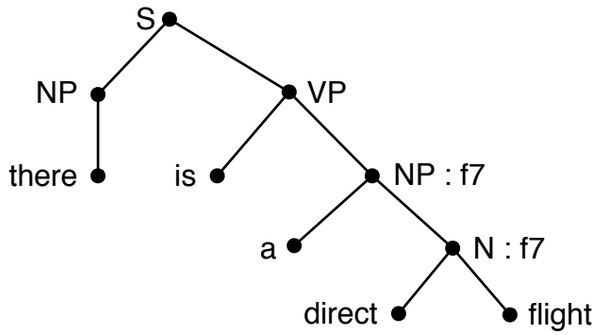
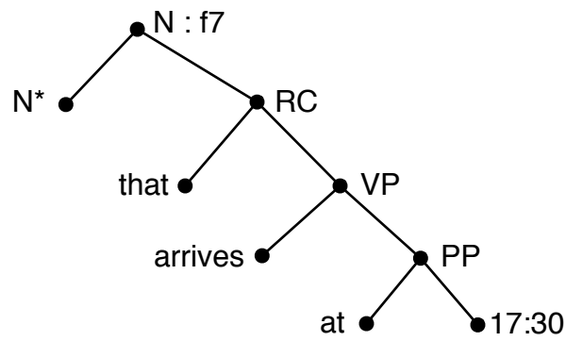
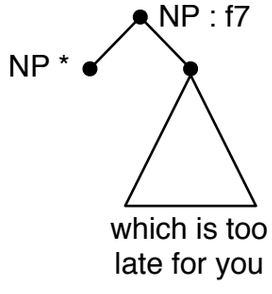


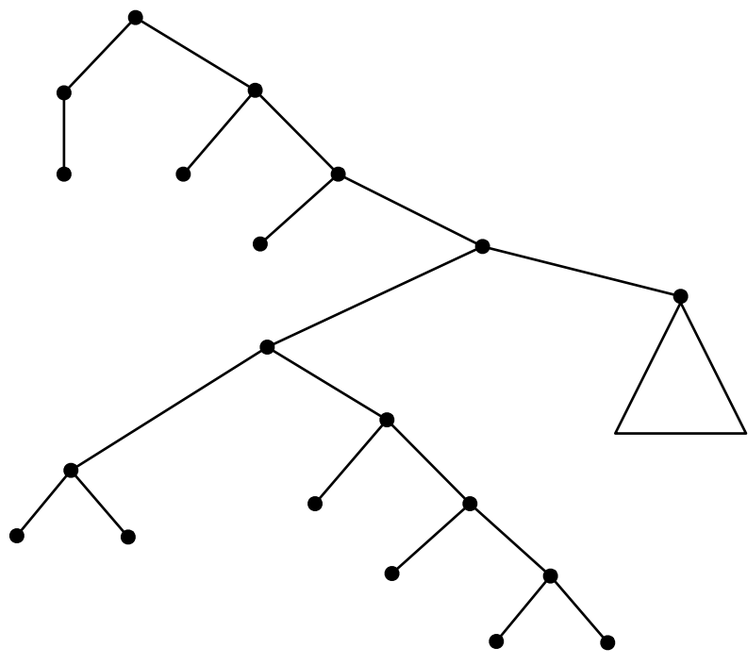
Surface realization with TAG

semantics	syntax
direct(f7)	
arrives(f7, 17:30)	
too_late(f7)	



Surface realization with TAG

semantics	syntax
direct(f7)	
arrives(f7, 17:30)	
too_late(f7)	



"There is a direct flight that arrives at 17:30, which is too late for you."

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) describe f7: +direct, arrives 17:30, too late

2. Flights in economy class
   a) describe f5: BMI, arrives 13:10
   b) describe f2: KLM, arrives 13:50, *but* -direct

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) There is a direct flight that arrives at 17:30, which is too late for you.
2. Flights in economy class
   a) describe f5: BMI, arrives 13:10
   b) describe f2: KLM, arrives 13:50, *but* -direct

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) There is a direct flight that arrives at 17:30, which is too late for you.

2. If you're willing to travel economy,
   a) describe f5: BMI, arrives 13:10
   b) describe f2: KLM, arrives 13:50, *but* -direct

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) There is a direct flight that arrives at 17:30, which is too late for you.

2. If you're willing to travel economy,
   a) There is a direct flight on BMI, arriving at 13:10.
   b) describe f2: KLM, arrives 13:50, *but* -direct

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) There is a direct flight that arrives at 17:30, which is too late for you.

2. If you're willing to travel economy,
   a) There is a direct flight on BMI, arriving at 13:10.
   b) There's also a KLM flight arriving at 13:50, but it requires a connection in Amsterdam.

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) There is a direct flight that arrives at 17:30, which is too late for you.

2. If you're willing to travel economy,
   a) There is a direct flight on BMI, arriving at 13:10.
   b) There's also a KLM flight arriving at 13:50, but it requires a connection in Amsterdam.

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) There is a direct flight that arrives at 17:30, which is too late for you.

2. If you're willing to travel economy,
   a) There is a direct flight on BMI, arriving at 13:10.
   b) There's also a KLM flight arriving at 13:50, but it requires a connection in Amsterdam.

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) There is a direct flight that arrives at 17:30, which is too late for you.

2. If you're willing to travel economy,
   a) There is a direct flight on BMI, arriving at 13:10.
   b) There's also a KLM flight arriving at 13:50, but it requires a connect

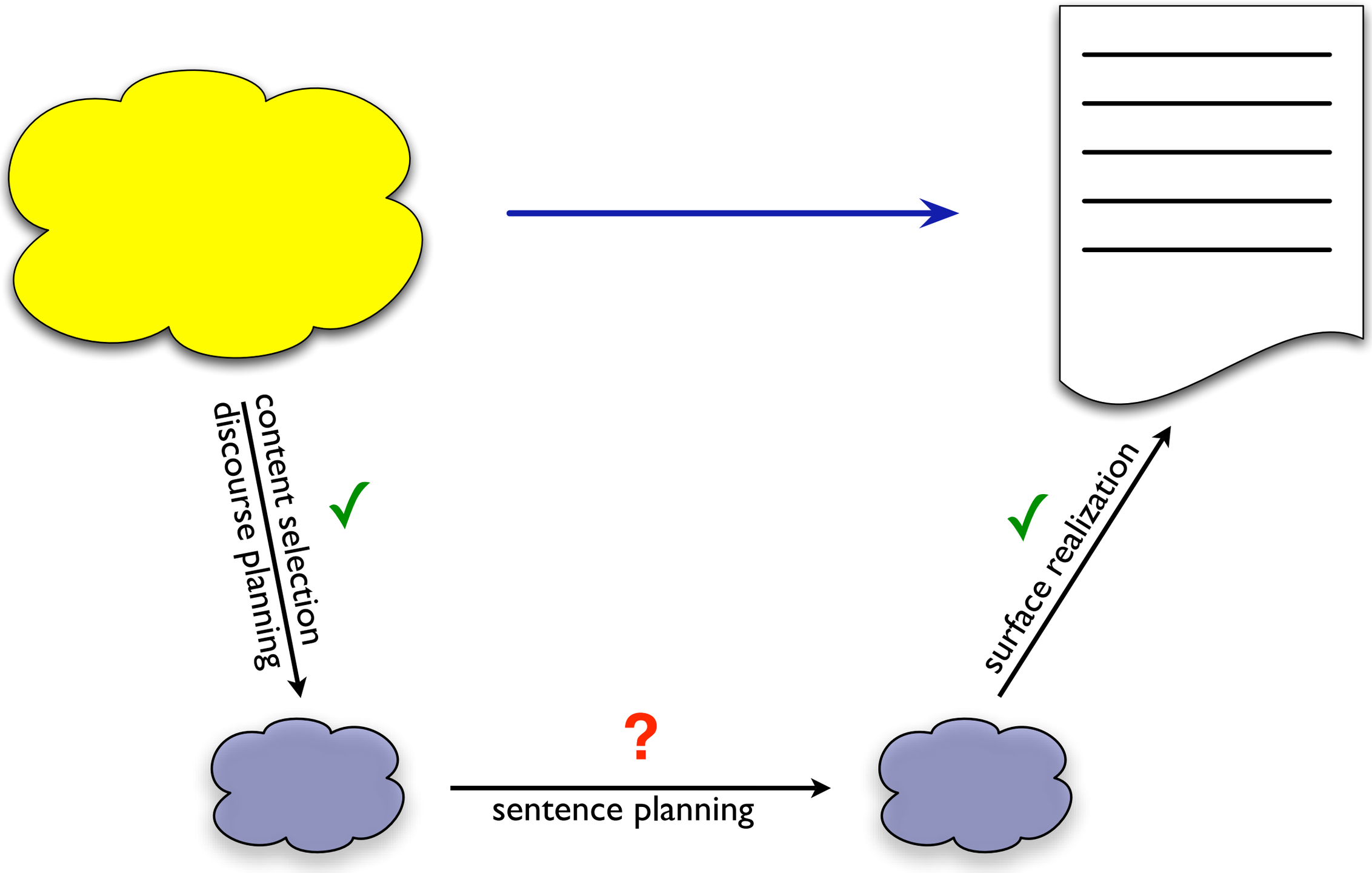
Merge these two sentences: Aggregation.

Surface realization

1. There are four flights with availability in business class.
   a) None are on KLM.
   b) There is a direct flight that arrives at 17:30, which is too late for you.

2. If you're willing to travel economy,
   a) there is a direct flight on BMI, arriving at 13:10.
   b) There's also a KLM flight arriving at 13:50, but it requires a connection in Amsterdam.

Where are we now?



Sentence planning

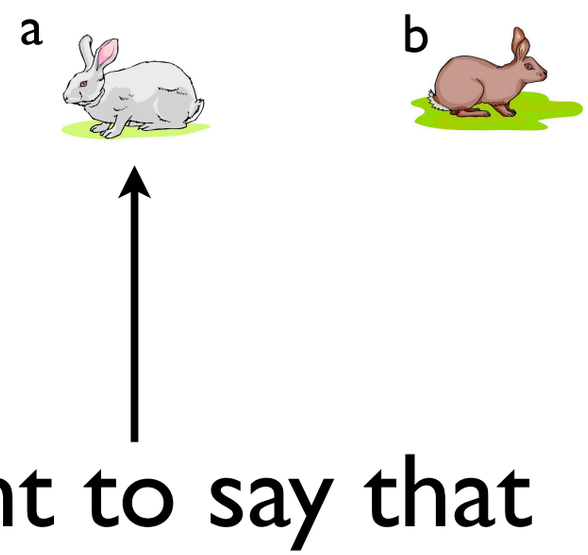
- Output of content determination may be not quite suitable as input of surface realizer.

- Sentence planning: Everything that needs to happen to map CD output to SR input.

- Typically:
  - referring expression generation
  - lexical choice
  - etc.

Referring expressions

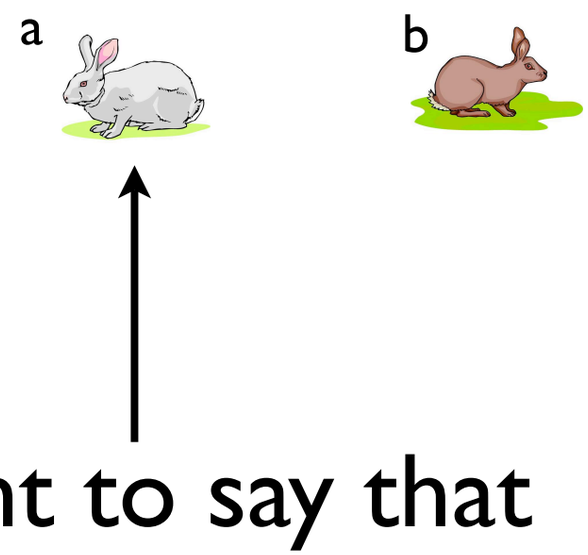
Knowledge base:



We want to say that
this guy sleeps.

Referring expressions

Knowledge base:

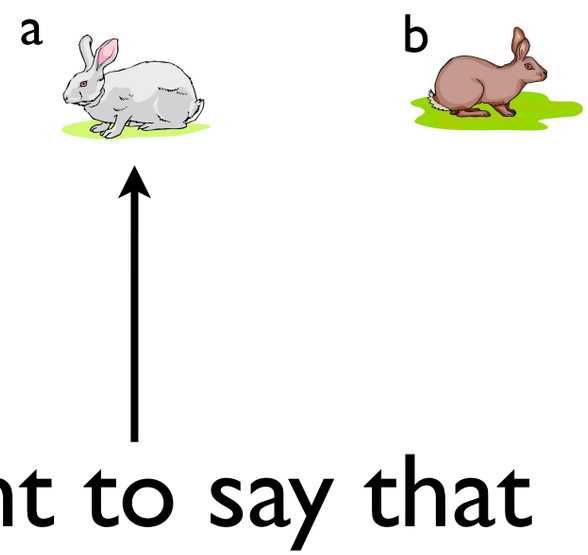


We want to say that
this guy sleeps.

"The white rabbit sleeps."

Referring expressions

Knowledge base:



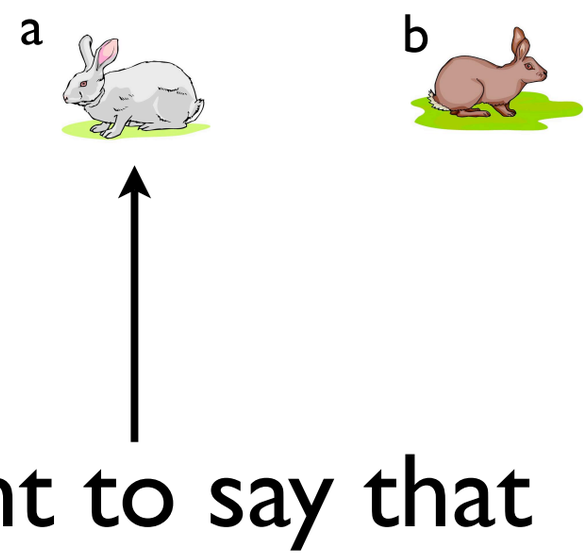
We want to say that
this guy sleeps.

"The white rabbit sleeps."

Is this content determination?

Referring expressions

Knowledge base:



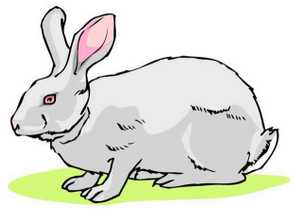
We want to say that
this guy sleeps.

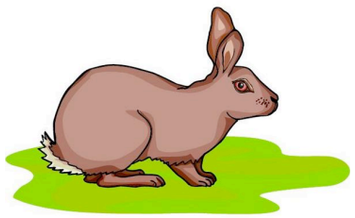
"The white rabbit sleeps."

Is this content determination? Surface realization?

RE generation

Universe:





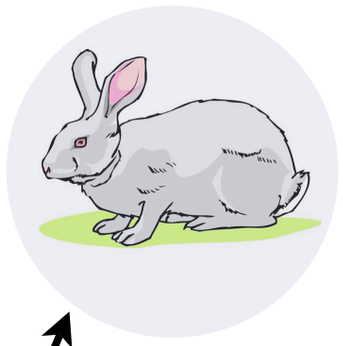
List of properties:
- rabbit
- polar bear
- broccoli
- white

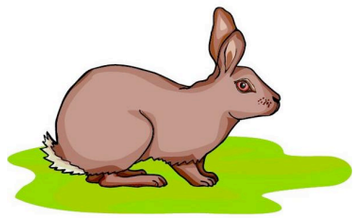
....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





target referent

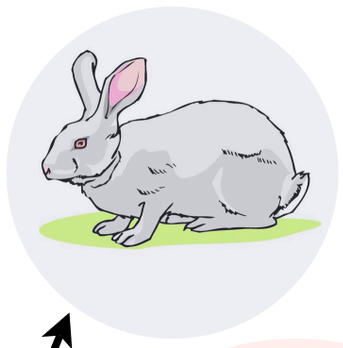
List of properties:
- rabbit
- polar bear
- broccoli
- white

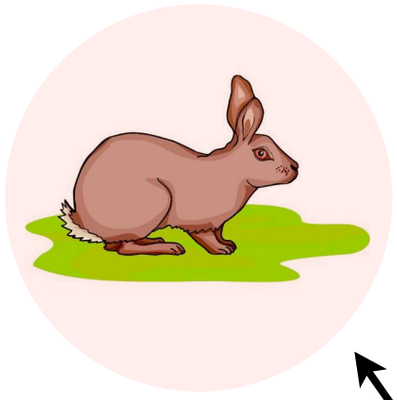
....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





target referent

distractors

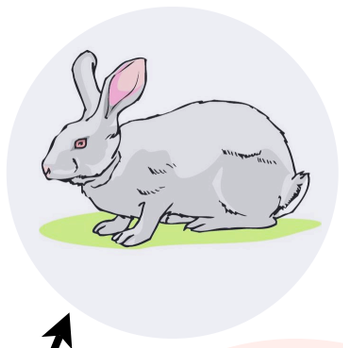
List of properties:
- rabbit
- polar bear
- broccoli
- white

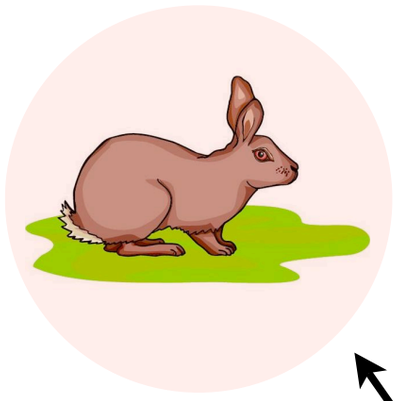
....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





target referent

distractors

List of properties:

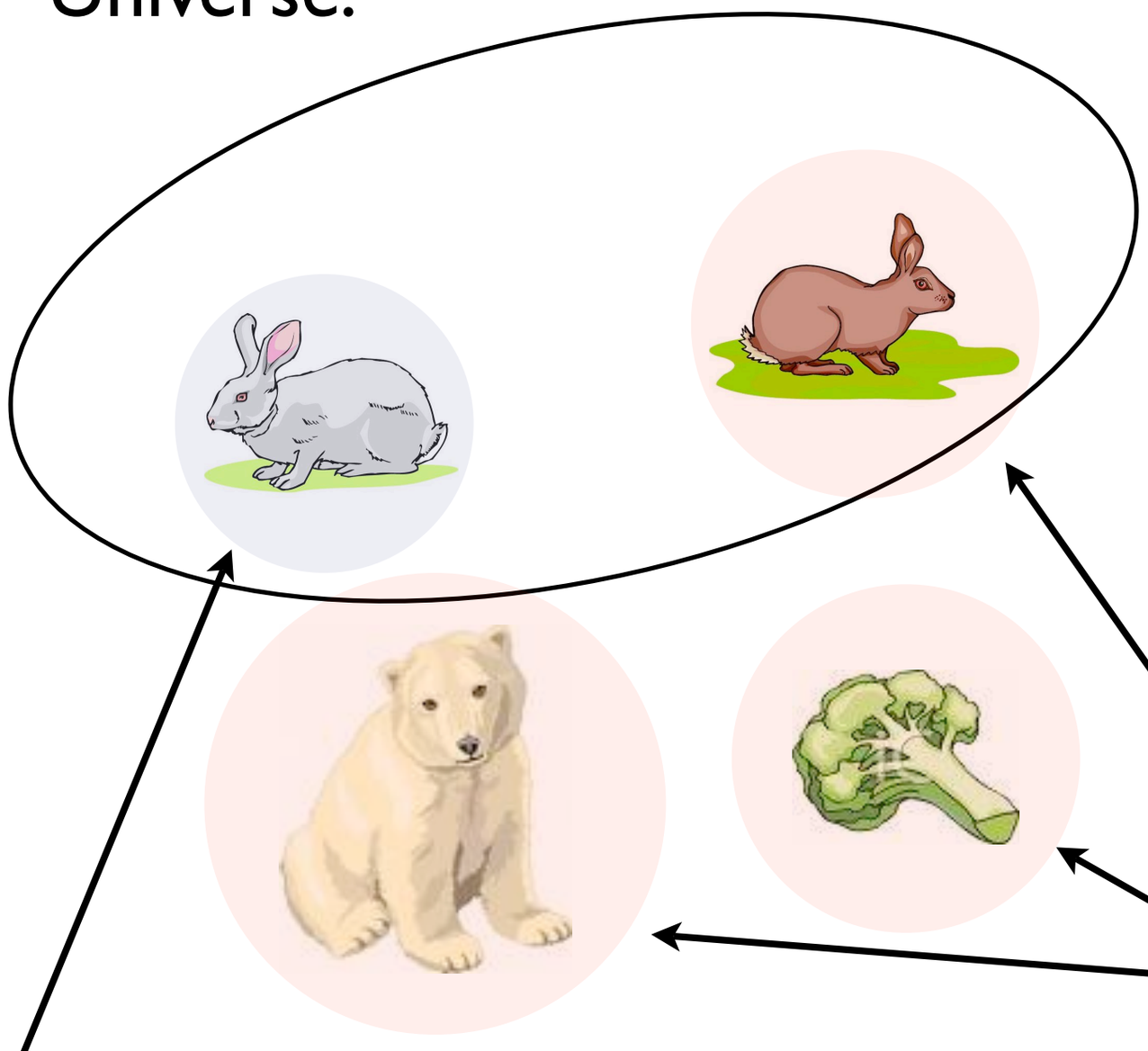
- rabbit
- polar bear
- broccoli
- white

....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:



List of properties:

- rabbit
- polar bear
- broccoli
- white

....

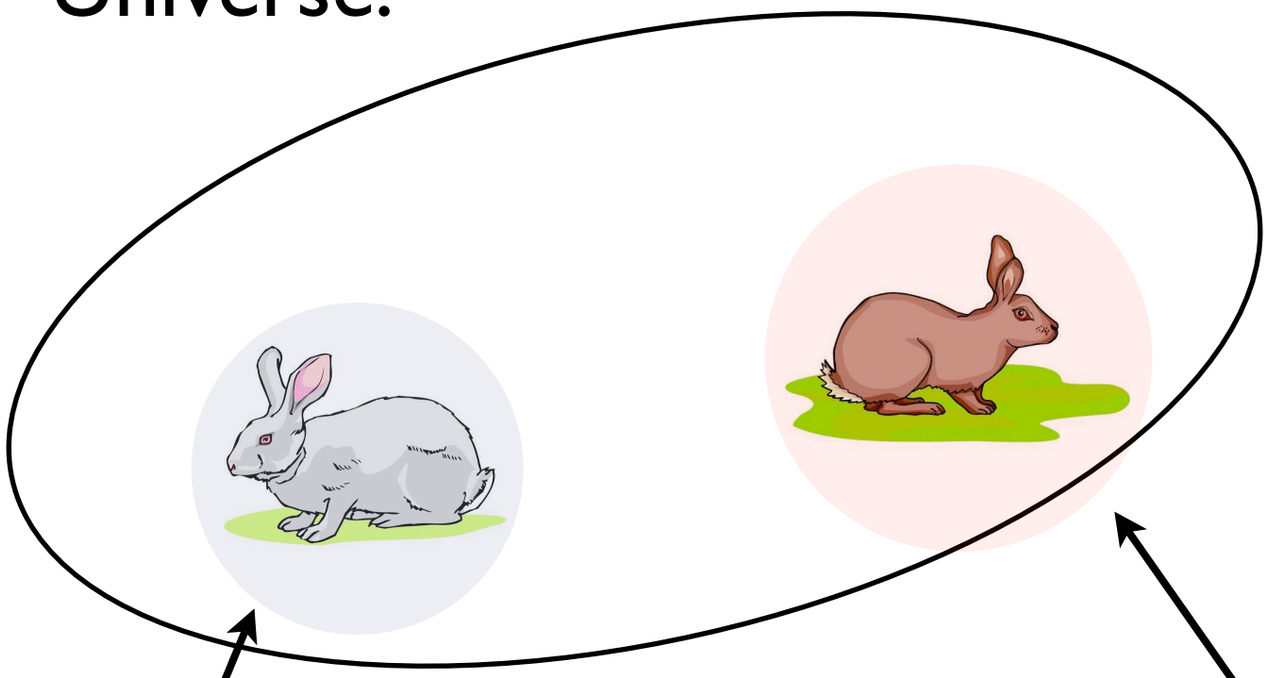
distractors

target referent

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:



List of properties:

- rabbit
- polar bear
- broccoli
- white

....

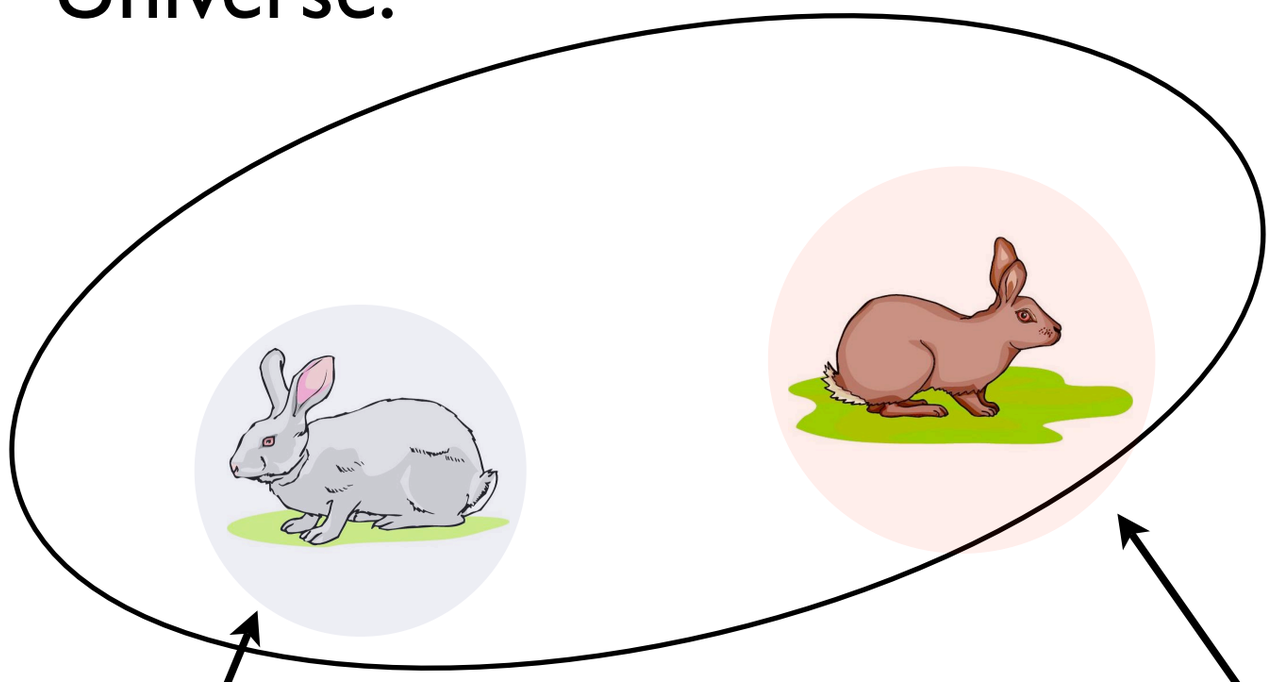
target referent

distractors

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:



List of properties:

- rabbit ✓
- polar bear
- broccoli
- white

....

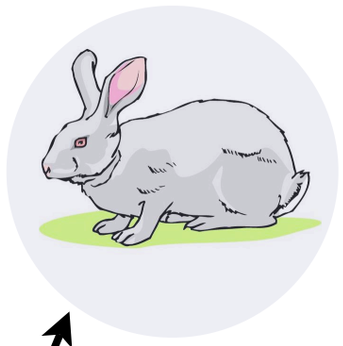
target referent

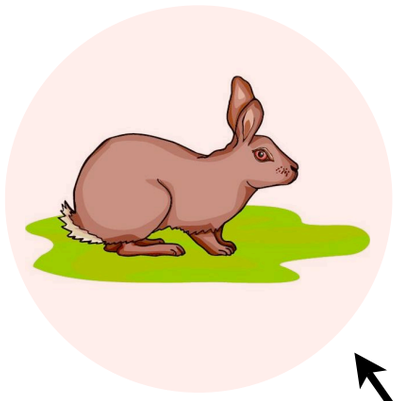
distractors

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





target referent

distractors

List of properties:

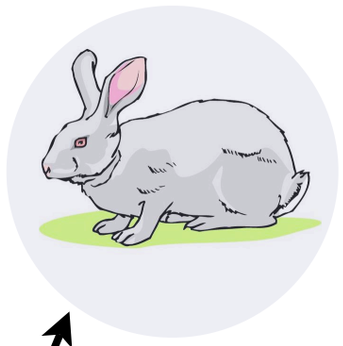
- rabbit ✓
- ▶ - polar bear
- broccoli
- white

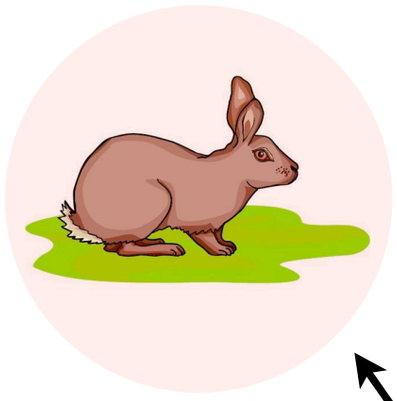
....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





target referent

distractors

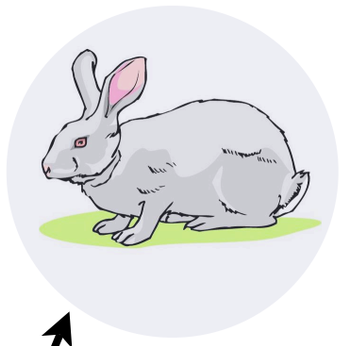
List of properties:
- rabbit ✓
- polar bear
- broccoli
- white

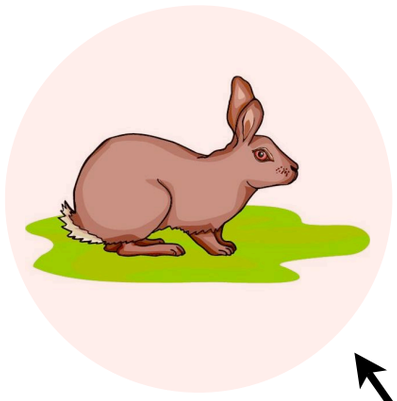
....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





target referent

distractors

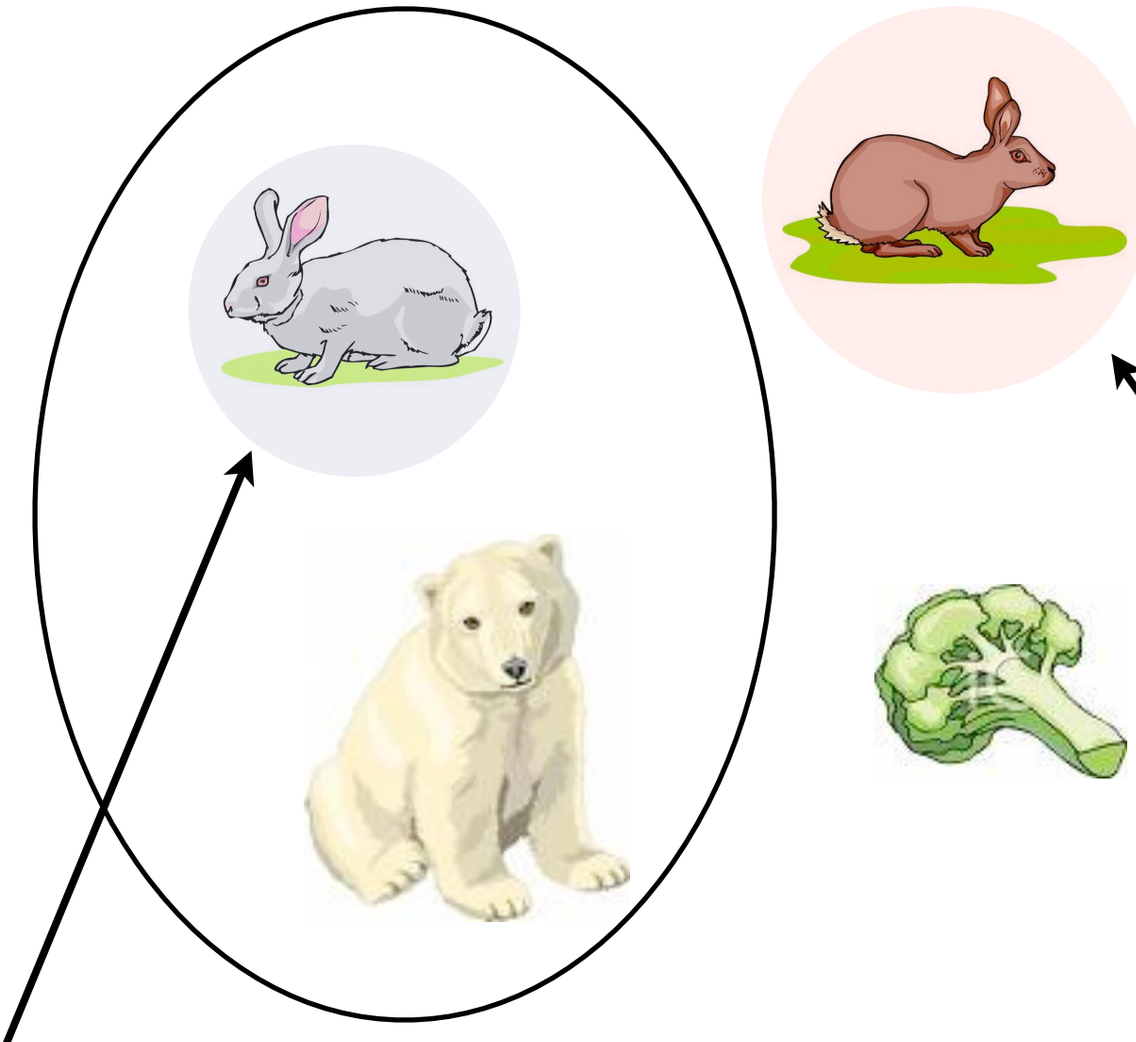
List of properties:
- rabbit ✓
- polar bear
- broccoli
- white

....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:



List of properties:
- rabbit ✓
- polar bear
- broccoli
- white

....

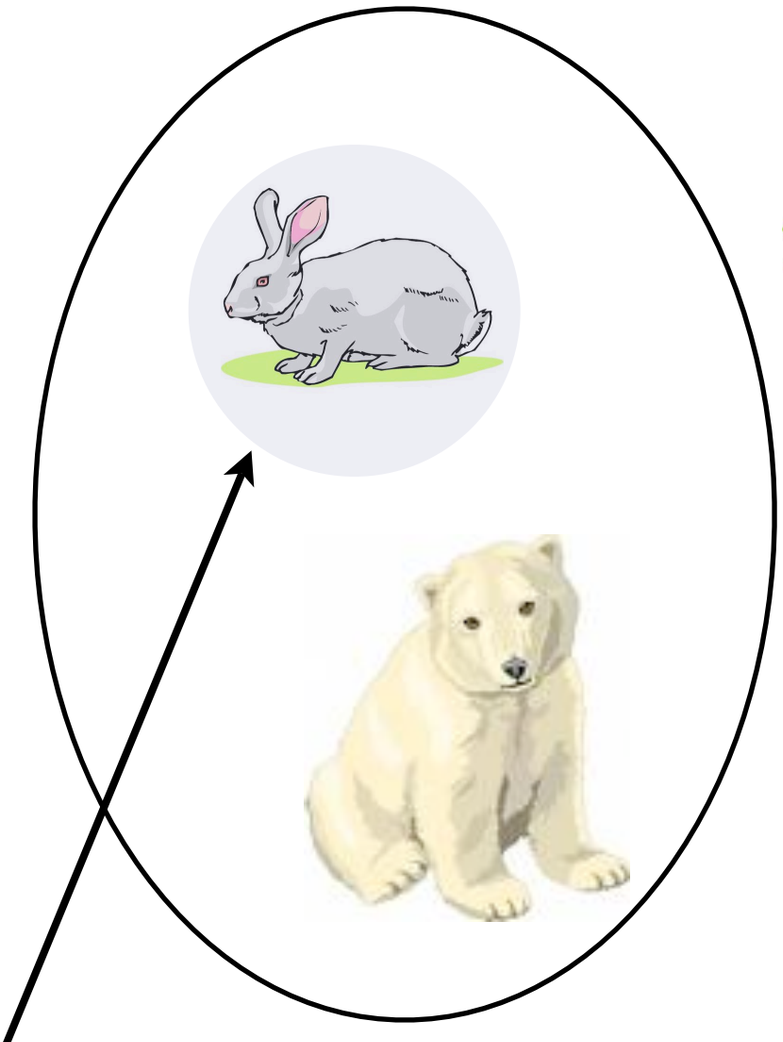
distractors

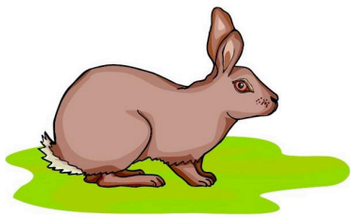
target referent

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





target referent

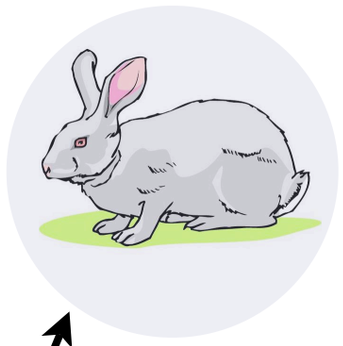
List of properties:
- rabbit ✓
- polar bear
- broccoli
- white

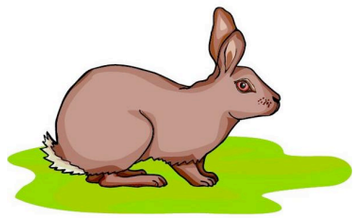
....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





target referent

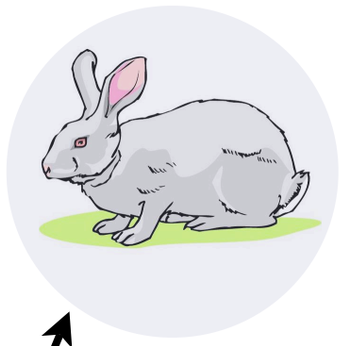
List of properties:
- rabbit ✓
- polar bear
- broccoli
- white

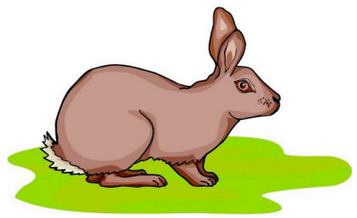
....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





target referent

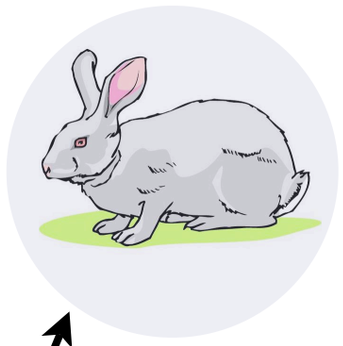
List of properties:
- rabbit ✓
- polar bear
- broccoli
- white ✓

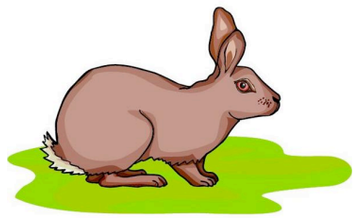
....

Incremental algorithm: Dale & Reiter (1995)

RE generation

Universe:





“the white rabbit”

target referent

List of properties:
- rabbit ✓
- polar bear
- broccoli
- white ✓

....

Incremental algorithm: Dale & Reiter (1995)

RE generation

- Until recently, very active research area.
- 1990s, early 2000s: algorithms for more expressive REs, dominated by logicians
- Recently: focus shift towards cognitive models of "good" REs

Lexical choice

- Lexical choice: Mapping semantic concepts to content words.
- Not as trivial as it sounds at first glance.
- How to realize generic concepts based on what they apply to?
  - The temperature rose.
  - The rain got heavier.
  - The revenue increased.

Lexical choice

- How to map real-world values to words?
  - what RGB values accepted as “red car”?
  - as “red wine”?
  - as “red hair”?
- How to distribute bits of meaning over different words?
  - swim across the lake
  - traverser le lac à la nage

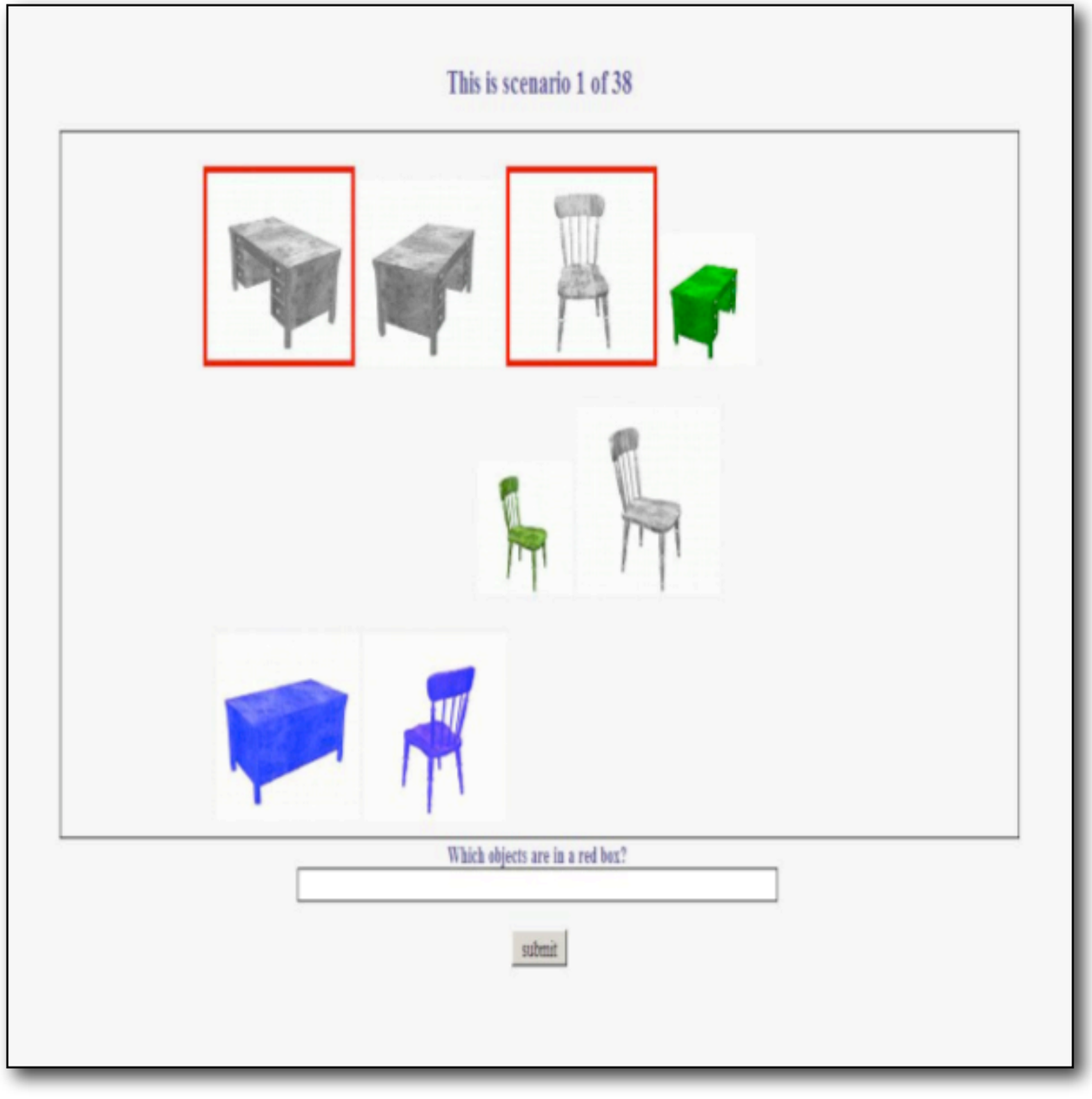
Summary: NLG

- NLG systems: map stuff to text
- Typical components:
  - content determination / discourse planning
  - sentence planning
  - surface realization

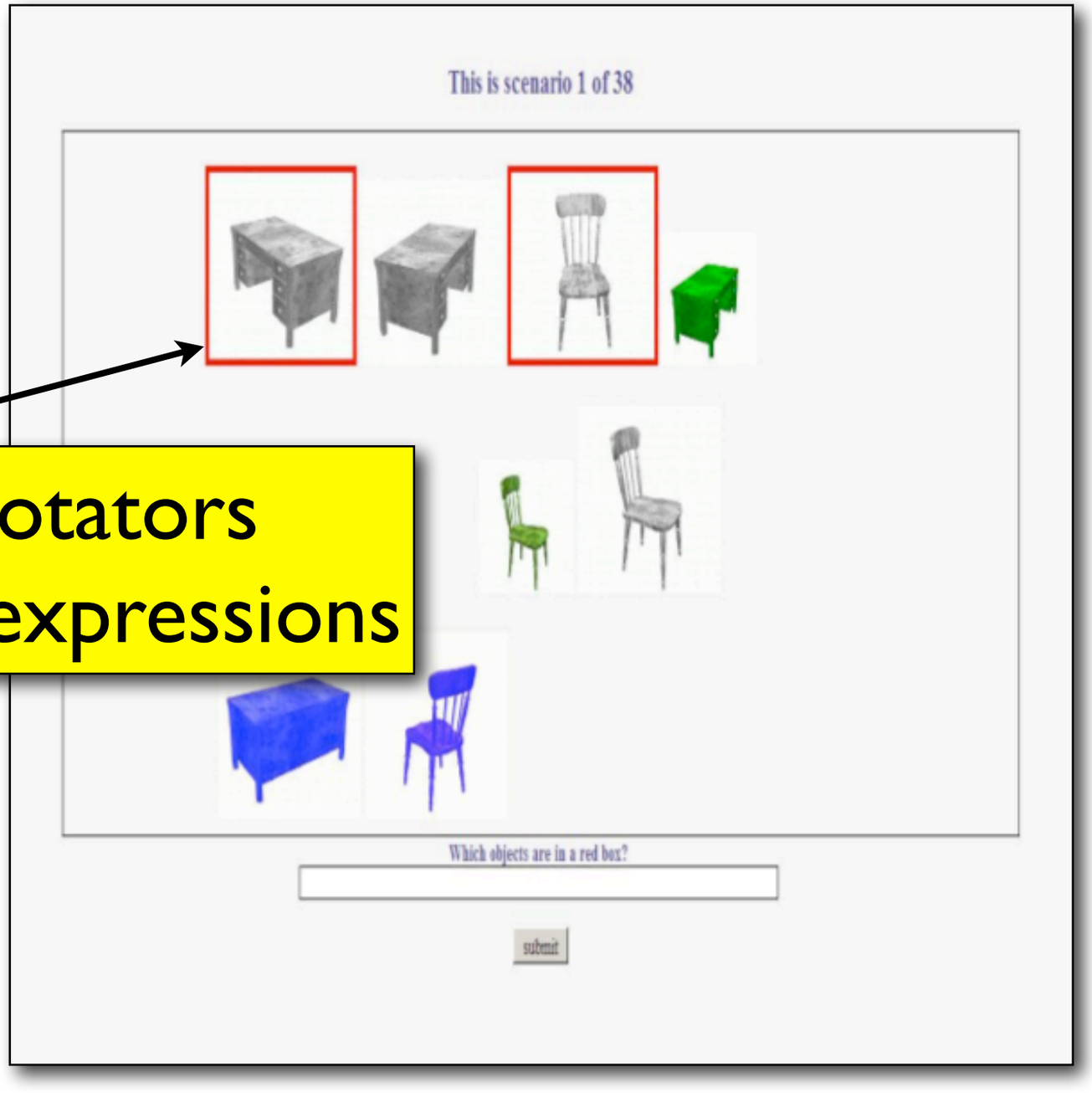
NLG Evaluation

- Evaluating NLG systems is hard.
- However, people find it increasingly important:
  - (DARPA Communicator, Walker et al. 02)
  - Special sessions at INLG conferences since 2006; separate NLG evaluation workshop in 2007
  - ASGRE-07 / REG-08 / TUNA Challenges
  - GREC Challenge
  - “Generation Challenges” umbrella organization

REG/TUNA Challenge

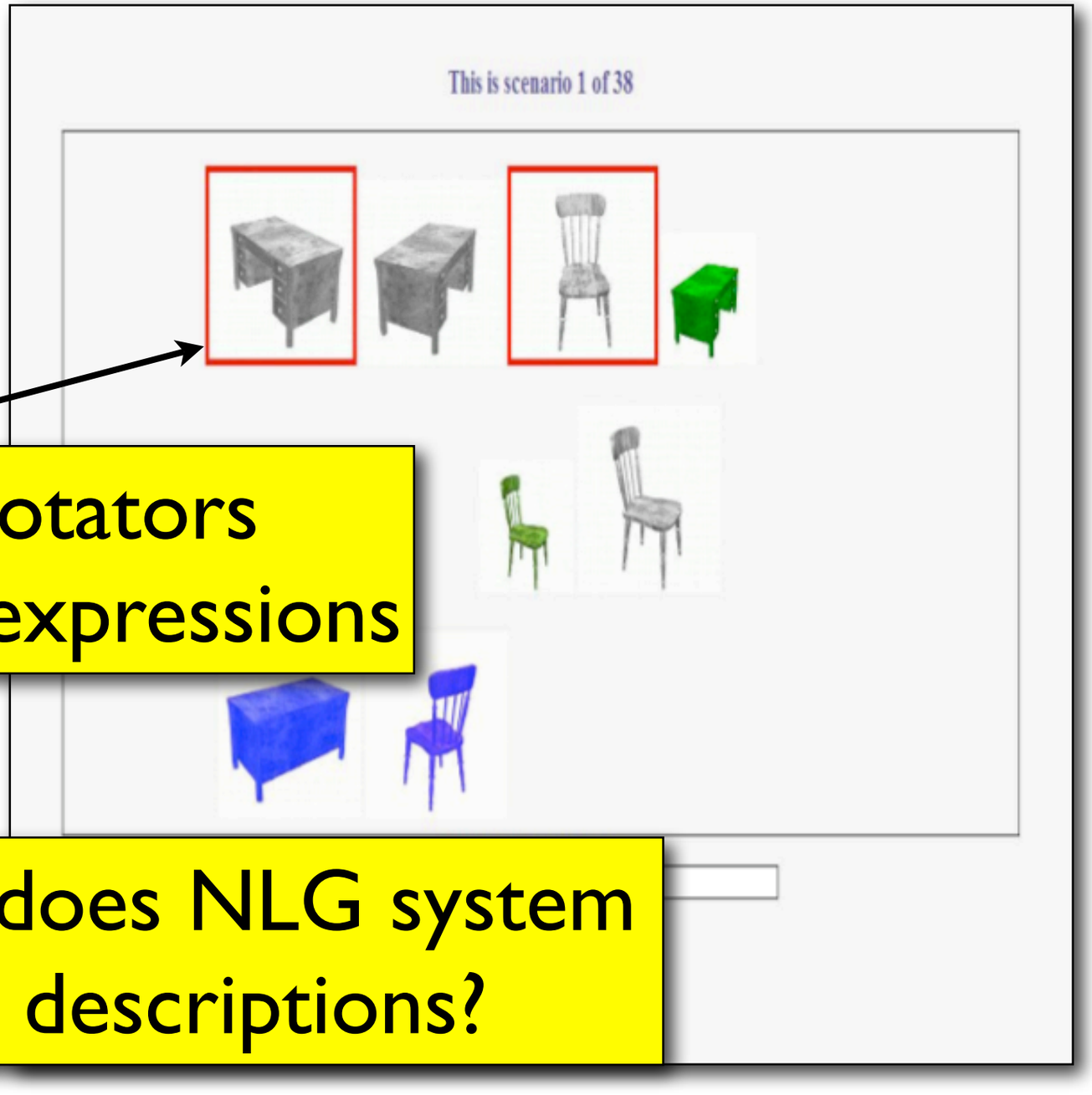


REG/TUNA Challenge



Step 1: Human annotators produce referring expressions

REG/TUNA Challenge



Step 1: Human annotators produce referring expressions

Step 2: How well does NLG system reproduce human descriptions?

Evaluating evaluation metrics

- TUNA 09: Best systems agreed better with humans than humans did with each other.

- Belz & Gatt 08 compared TUNA's "human-likeness" measures against task-based measures (e.g. identification time).

- No correlation between human-likeness and task-based measures.

Problem with NLG Eval

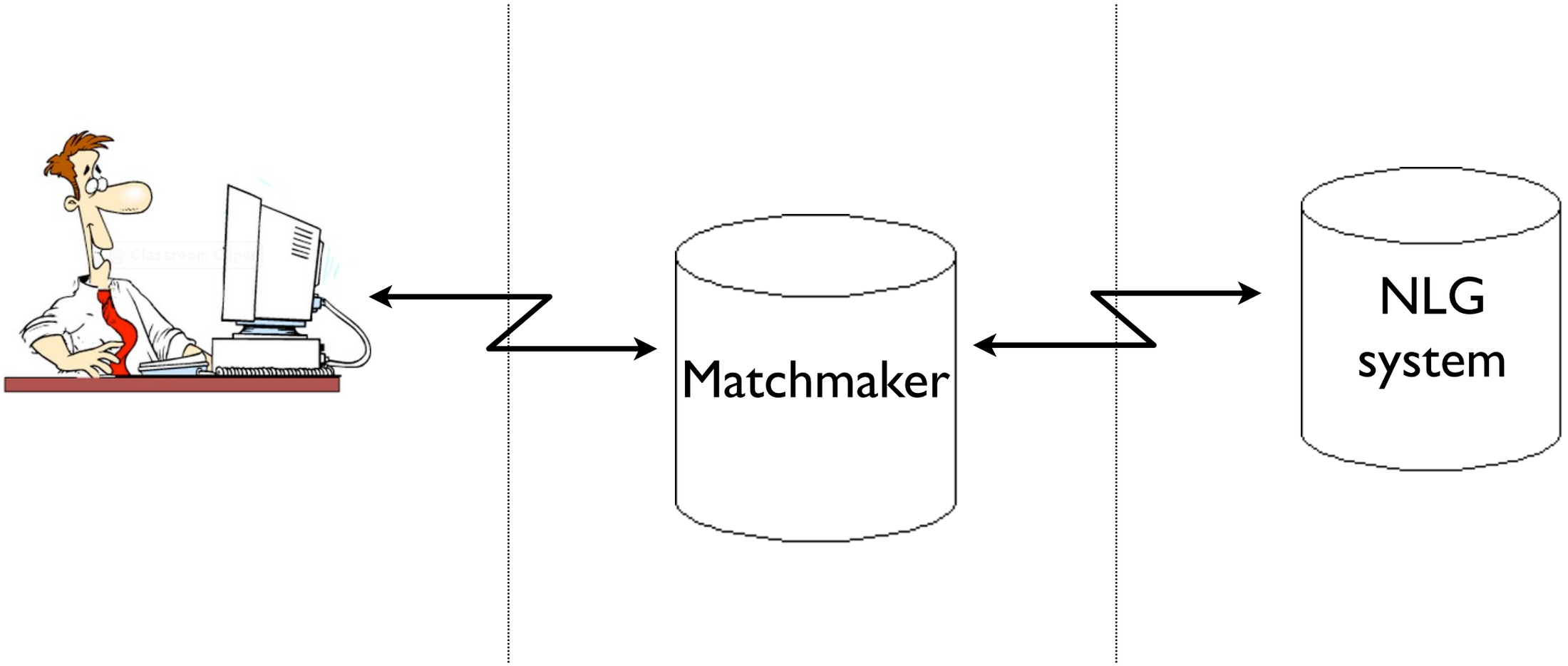
- Two standard options:
  - Evaluate against gold standard: artificial for NLG evaluation because multiple texts may be equally good
  - Evaluate with human annotators or judges: more appropriate, but very expensive and time-consuming

- So how can we do it?

Instruction giving in virtual worlds

- Task: Generate real-time instructions that help user perform some task in a virtual environment.
- Use for end-to-end evaluation of NLG systems.

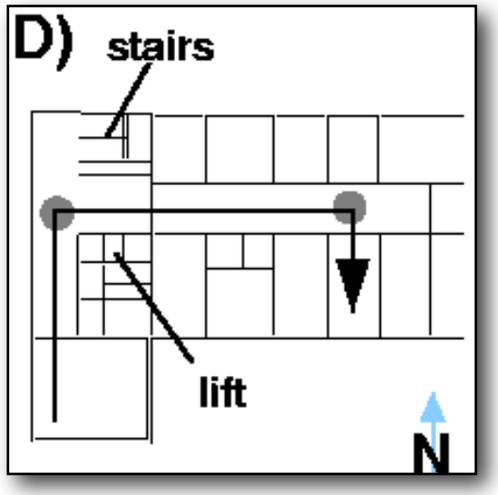
Evaluation



- User and NLG system can be in different places.
- Can perform “web experiments”!

Related Applications

Pedestrian navigation



Task instructions ("Apollo 13")

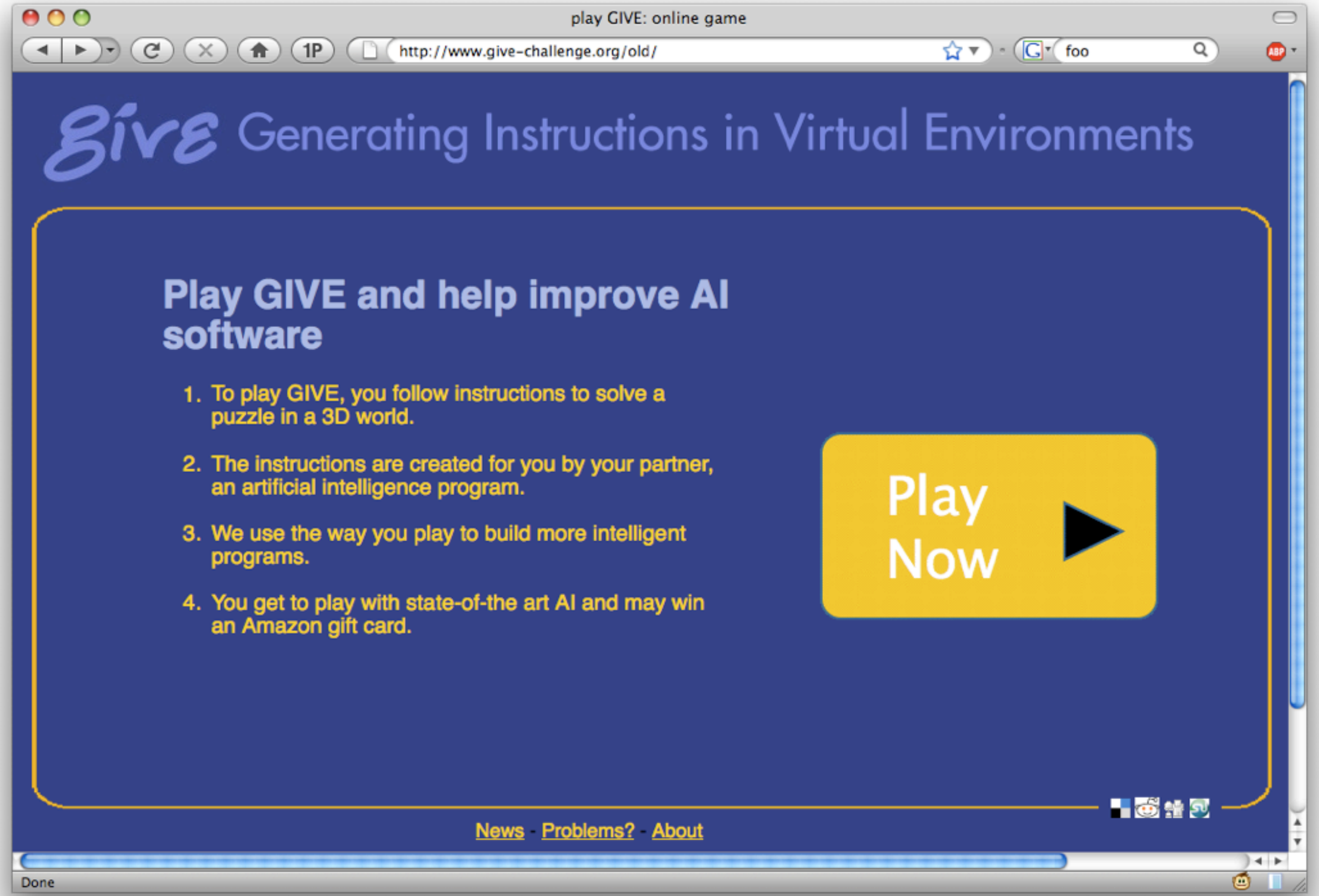
"In vitro" human-robot interaction

GIVE-1

- For the first installment of the challenge:
  - pilot experiment character
  - discrete virtual worlds

- Timeline:
  - announced in March 2008
  - distributed software to participants in May 2008
  - Internet-based evaluation Nov 2008 to Feb 2009
  - data analysis and report writing until March 2009
  - results presented at ENLG in Athens, March 2009

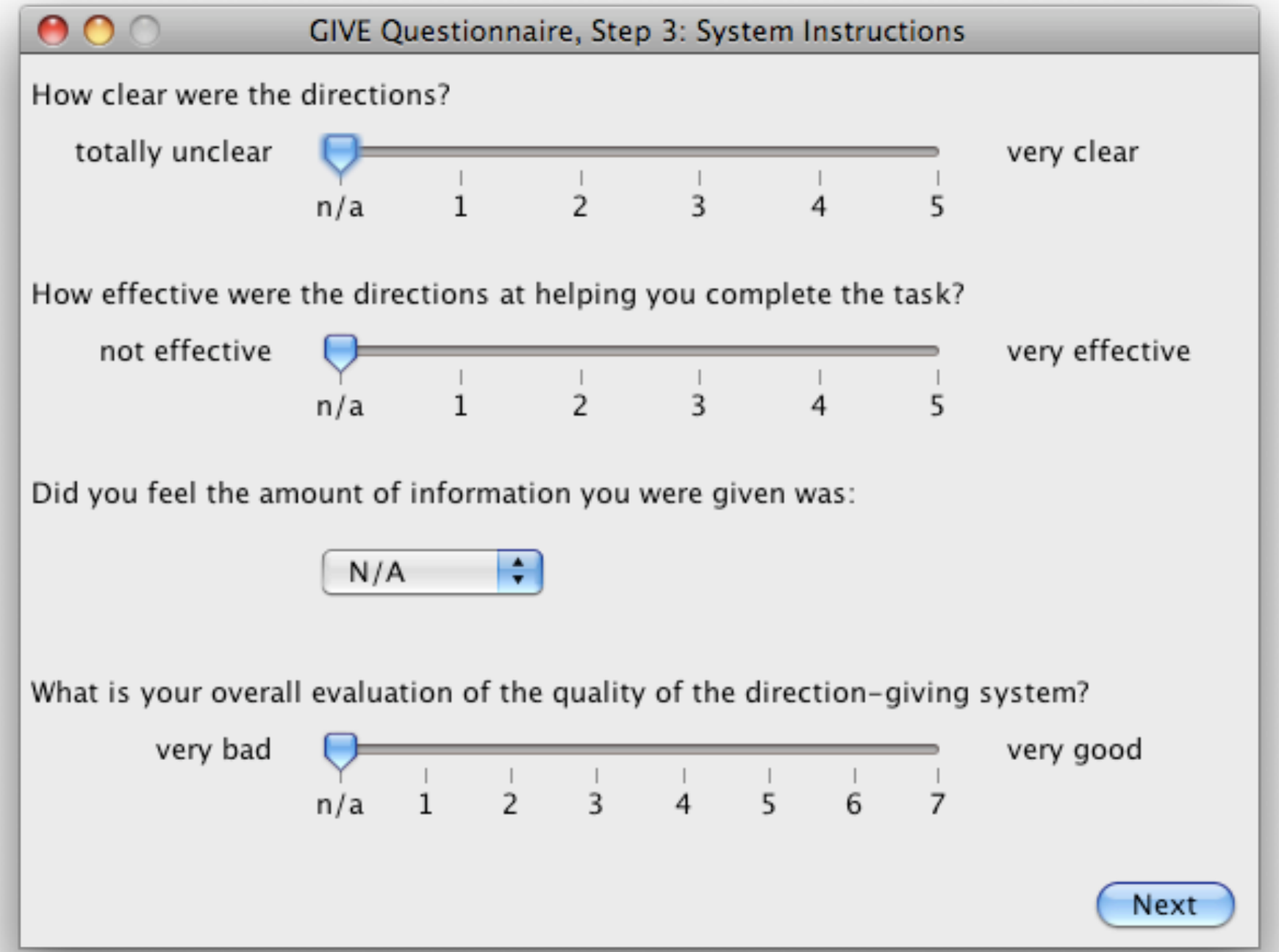
GIVE website



Game client


GIVE 3D Client
Move forward 2 steps, and then press the blue button.
Walk
Turn
Esc
quit
H
What?

Questionnaire



Participating Systems

- Proof-of-concept system: Compute domain plan, realize plan actions one by one.
- Austin: Optimized version of this system (improved paths; some aggregation).
- Madrid: Emphasis on inferring and exploiting "hidden" aspects of world, such as rooms, corners, etc.

Participating Systems (2)

- Union College: Emphasis on navigation instructions, switches between landmark-based and path-based modes.

- Twente: Emphasis on adaptation to user's ability to understand instructions.

- Twente Warm/Cold system: only says "warmer", "colder", etc.; intended to maximize entertainment.

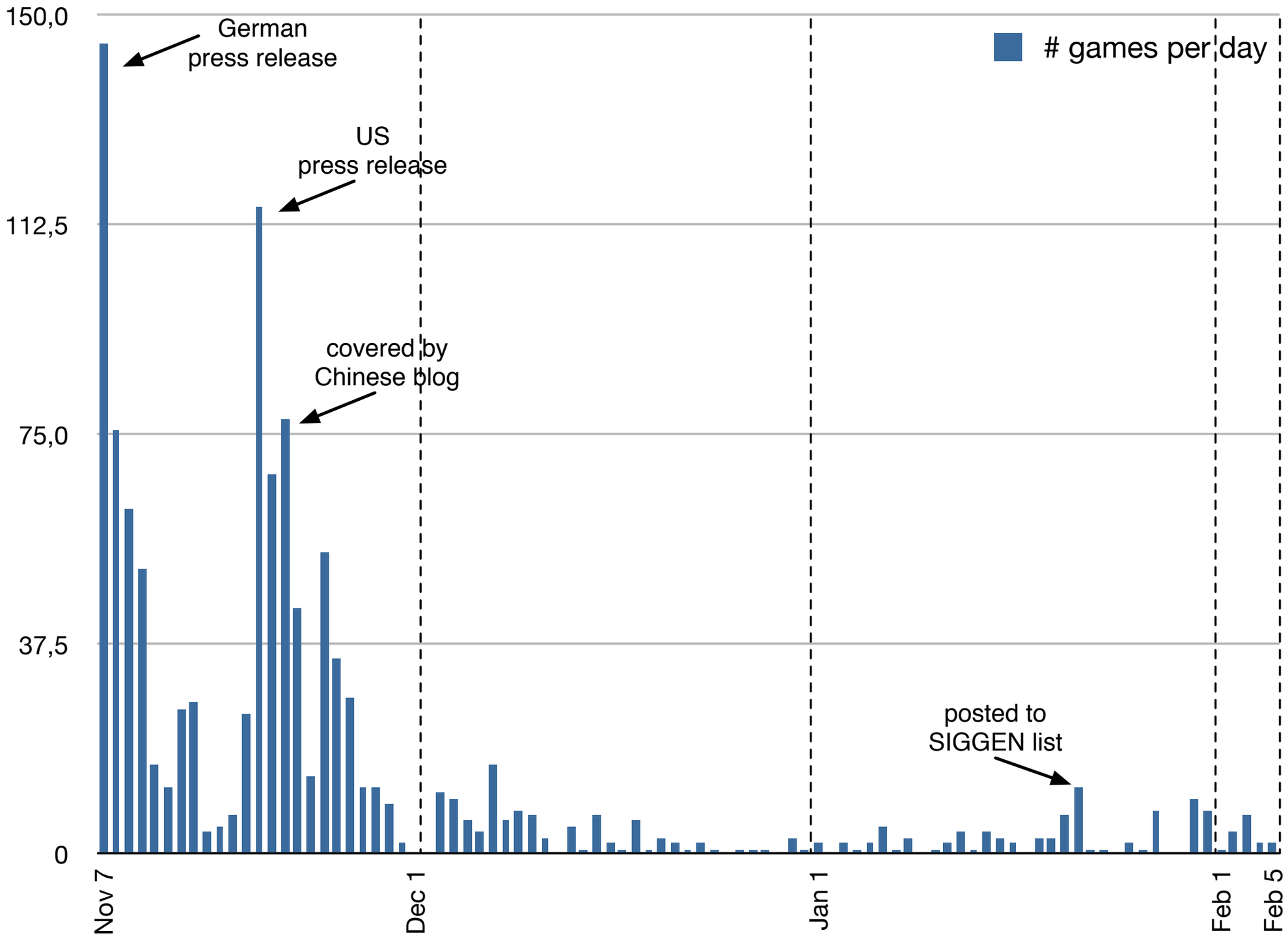
Results

(Connections in 24-hour window on 20 Nov 08)

Results


Greenland
Iceland
Canada
México
Venezuela
Colombia
Perú
Brasil
Brazil
Bolivia
Chile
Argentina
Suomi
Finland
Norge
Norway
Россия
Russia
Kazakhstan
Mongolia
Україна
Ukraine
Tanzania
Angola
Namibia
Botswana
Madagascar
Indian
Ocean
Australia
Papua New
Guinea
New
Zealand
South
Pacific
Ocean
South
Atlantic
Ocean
North Pacific
Ocean
Collected 1143 valid games
over 3-month period
(Connections in 24-hour window on 20 Nov 08)
Map data ©2008 Europa Technologies - Terms of Use

Results: Timeline



Results: Objective measures

Differences are significant if two systems don't share a letter.

Lower letters are better.

task
succ

Results: Subjective measures

- overall on 1-7 scale;
- timing, informativity "just right" vs. not;
- all others on 1-5 scale.

Summary: GIVE

- GIVE-1 was largest evaluation effort for NLG systems in terms of users, ever.
- Evaluated 5 systems, which emphasized different aspects. Significant differences, consistent with lab experiment.
- Simple systems work surprisingly well.

GIVE-2

- Mostly like GIVE-1, but:
  - continuous worlds
  - improved evaluation measures

- Development phase started in August.

- Evaluation phase is Feb - Apr 10.

- Presentation of results in July 10 at INLG.