May we come in?

CJFillmore (ICSI & UCB)

SCSS August 10 2011

Is NLU possible?

- I have no personal stake in whether complete high-quality "language understanding" by machine is possible, but I'm interested in various proposed ways of measuring success, and as a linguist I'm especially interested in the kinds of linguistic facts that any such system would have to recognze.
- The adequacy measures I have in mind are
  - whether a system can select the intended meaning of a word that has two or more meanings;
  - whether paraphrases and contradictions can be recognized;
  - whether actions or states of affairs described in a sentence can be simulated in some non-language medium.

Outline:
1. Text to scenes. 2. MWCI? 3. The lesson

1. I'll begin with a peek at a system (WordsEye) that generates images from texts that describe scenes, pointing out some limitations of such a system.

   (My knowledge of WordsEye is completely second-hand.)

2. Then I'll do a detailed analysis of the question of our title and ask how we know what kind of world its utterance presupposes.

3. Lastly, I'll give a summary of the phenomena and point out relations to some ongoing research.

Outline:

1. Text to scenes 2. MWCI? 3. The lesson

(a) WordsEye (b) Real 3D (c) Deixis

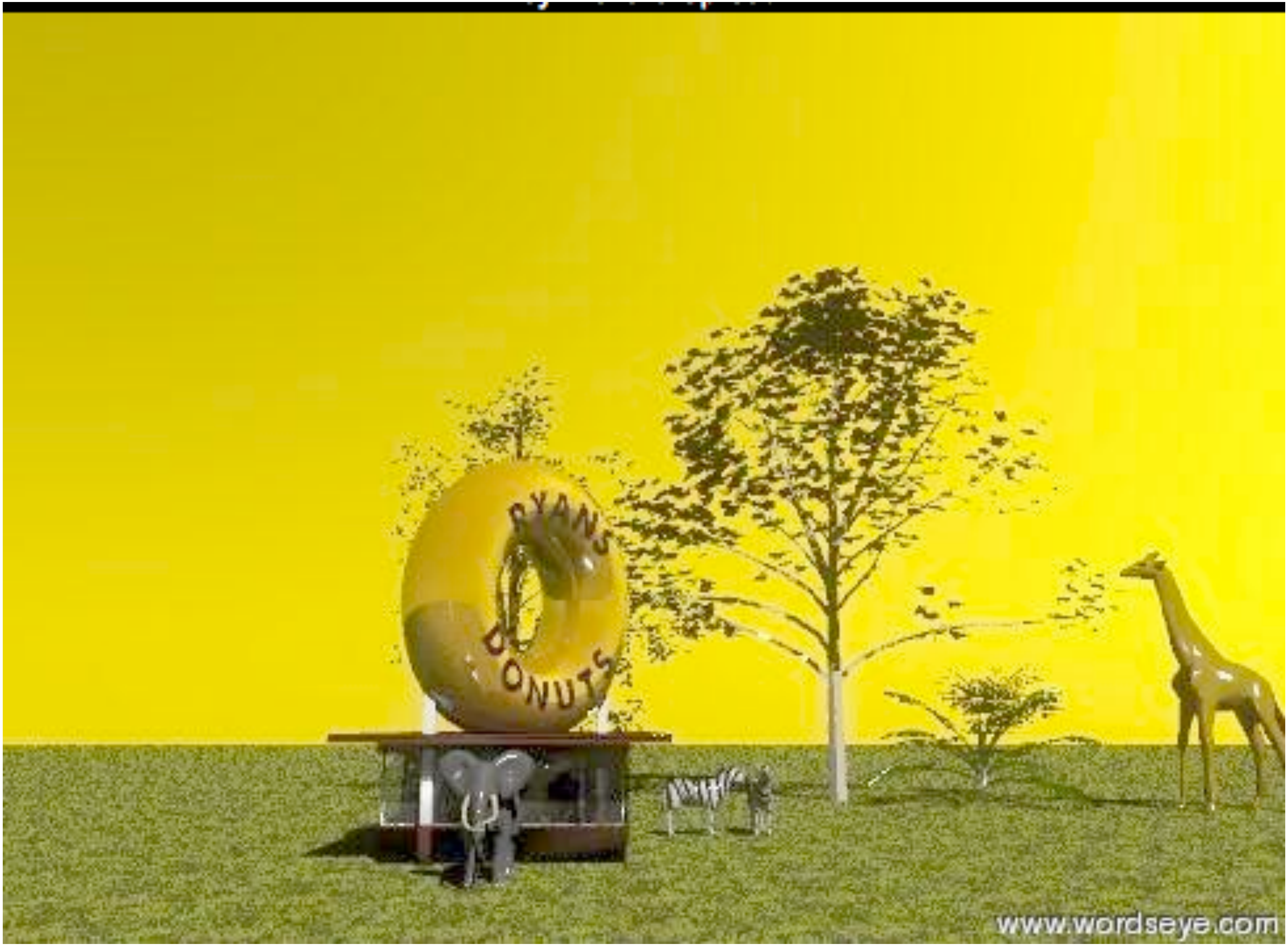
http://lucky.cs.columbia.edu:2001/gallery

WordsEye began as a research project at AT&T Research with Richard Sproat and Bob Coyne, and is now located at CS Columbia. There is now a company (Coyne & Sproat) called Semantic Light.
(They've copyrighted the phrases "Watch your language" and "Type a scene"!)

The Gallery (see URL) contains a number of samples of texts paired with the pictures they generated.

Here are some examples.

1. the donut of the donut shop is shiny.
2. the ground has a grass texture
3. the zebra is to the right of the donut shop.
4. the zebra is facing right.
5. another zebra is to the right of the zebra.
6. the elephant is in front of the donut shop.
7. it is facing left.
8. three large trees are behind the donut shop.
9. the sky is orange.
10. a large giraffe is behind the donut shop.
11. the giraffe is facing left.
12. the elephant is facing right.


RYANS
DONUTS
www.wordseye.com

1. the couch is against the wood wall.
2. the window is on the wall.
3. the window is next to the couch.
4. the door is 2 feet to the right of the window.
5. the man is next to the couch.
6. the animal wall is to the right of the wood wall.
7. the animal wall is in front of the wood wall.
8. the animal wall is facing left.
9. The walls are on the huge floor.
10. The zebra skin coffee table is two feet in front of the couch.
11. The lamp is on the table.
12. The floor is shiny.

www.wordseye.com

This one is taken from Coyne & Sloat, *WordsEye: An Automatic Text-to-Scene Conversion System*

ACM SIGGRAPH 2001

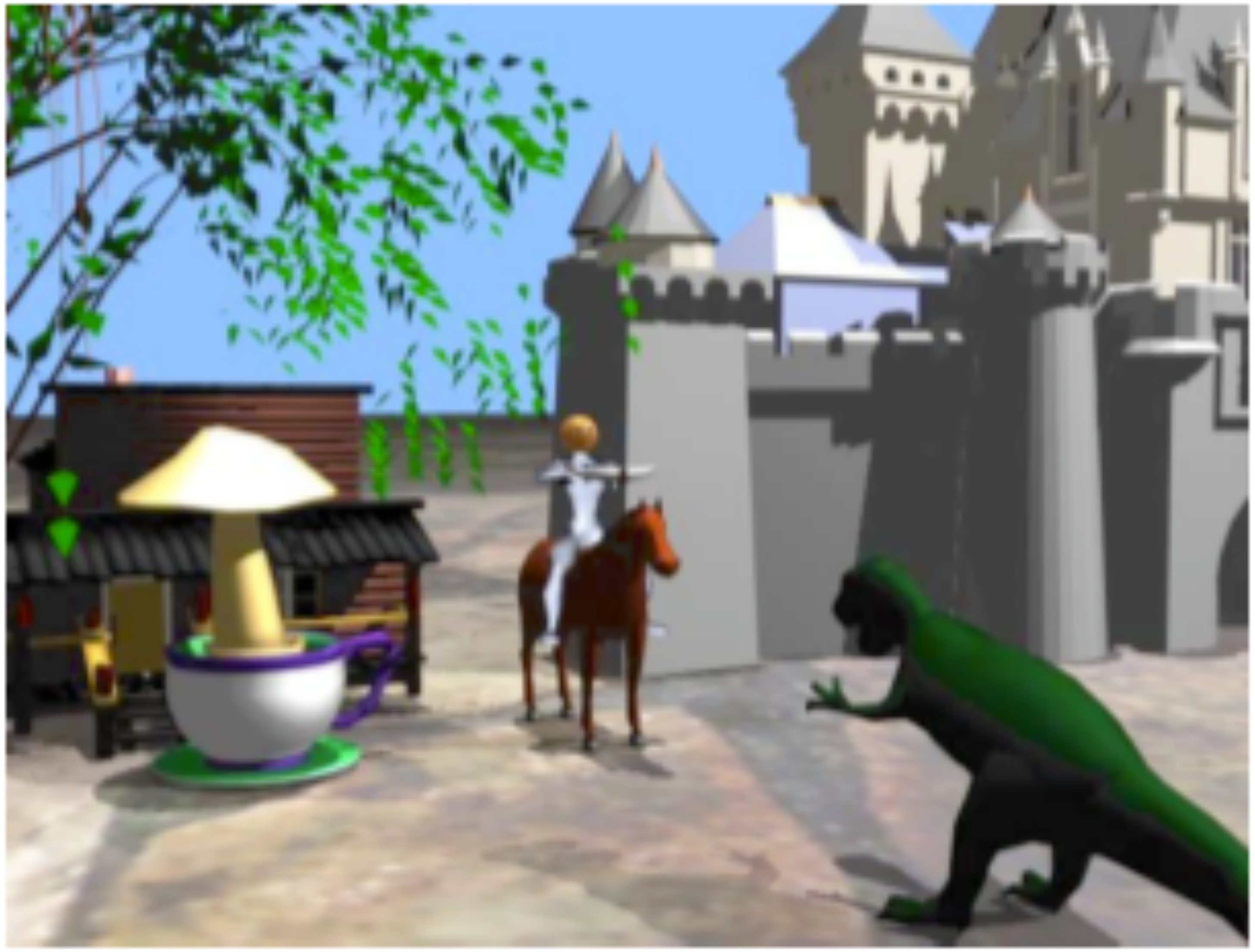


Figure 1: *John uses the crossbow. He rides the horse by the store. The store is under the large willow. The small allosaurus is in front of the horse. The dinosaur faces John. A gigantic teacup is in front of the store. The dinosaur is in front of the horse. The gigantic mushroom is in the teacup. The castle is to the right of the store.*

What have they got?

- A huge collection of images of entities, many rotatable, many colorable, all sizable, humans and some animals are posable.

- Default ways of laying out rooms & landscapes, choices of *renderings* (re light source, shadows, etc.)

- Rules for interpreting locating expressions with *front, behind, under, on, in, left, right*, etc.

- Rules for interactions among entities: facing, pushing, grasping, riding (on horse, bike).

- Grip-tags on graspable places on artifacts.

Outline:

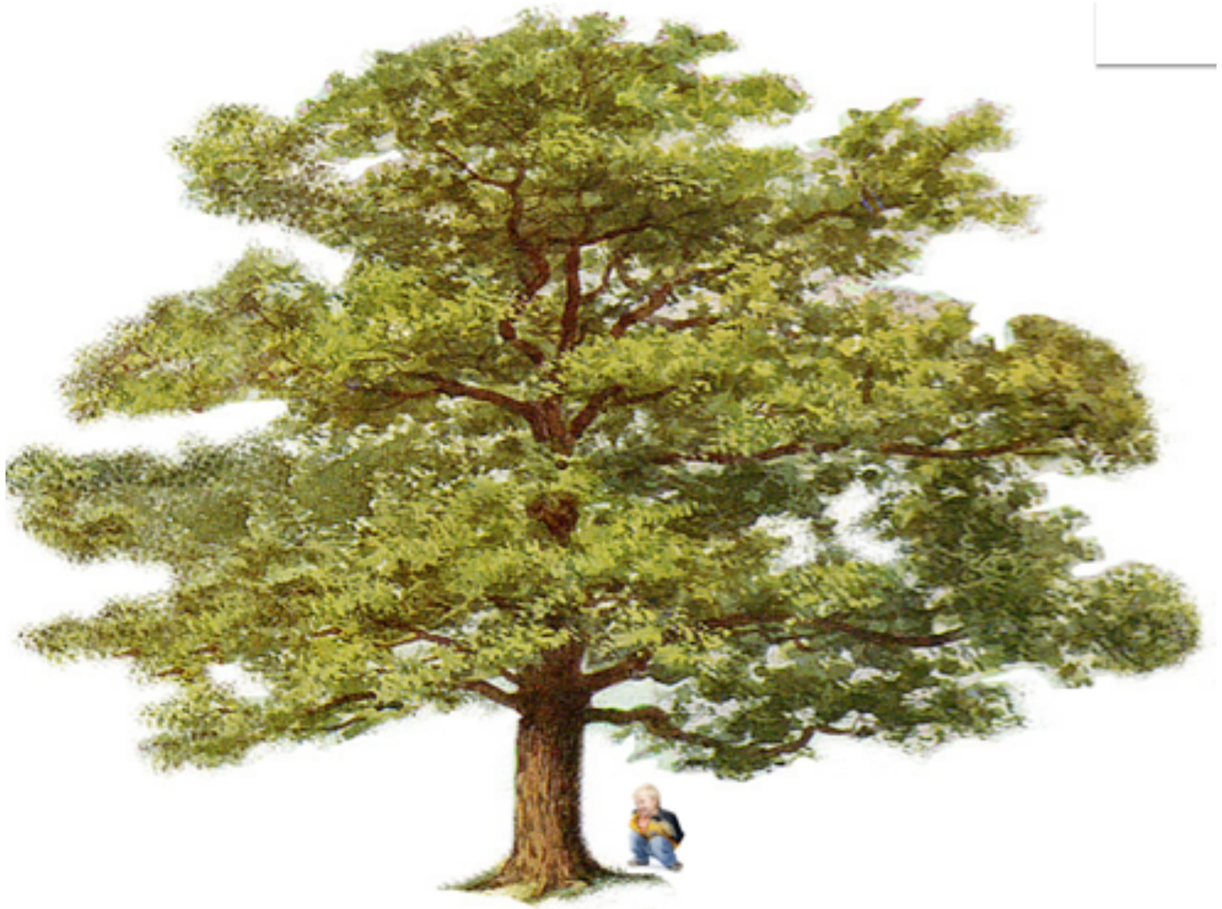
1. Text to scenes 2. MWCI? 3. The lesson

(a) WordsEye (b) Real 3D (c) Deixis

- WordsEye describes itself as allowing people to use text to create 3D images; yet for the most part the "locational" instructions depend on the vantage point of the viewer of the screen.
- Many of the tricks are very clever. ("on", "in", "under")
- Here is my picture (made with images stolen from the web) of "a boy hiding behind a tree."

A boy hiding behind a tree.
Of course you can’t see him because
the image is obscured by the tree trunk.

Another picture of a boy hiding behind a tree. This image has "assigned" to the imagined space another vantage point!



Suppose we created a scene on a table-top,
all of us gathered around it,
following step-by-step instructions.

Start with a clear table.

“In the middle of the village is a huge chestnut tree.”

PLACE CHESTNUT TREE MID-TABLE. DONE.

“A blacksmith shop is to the left of the tree.”

HUH? A TREE DOESN’T HAVE A LEFT SIDE!

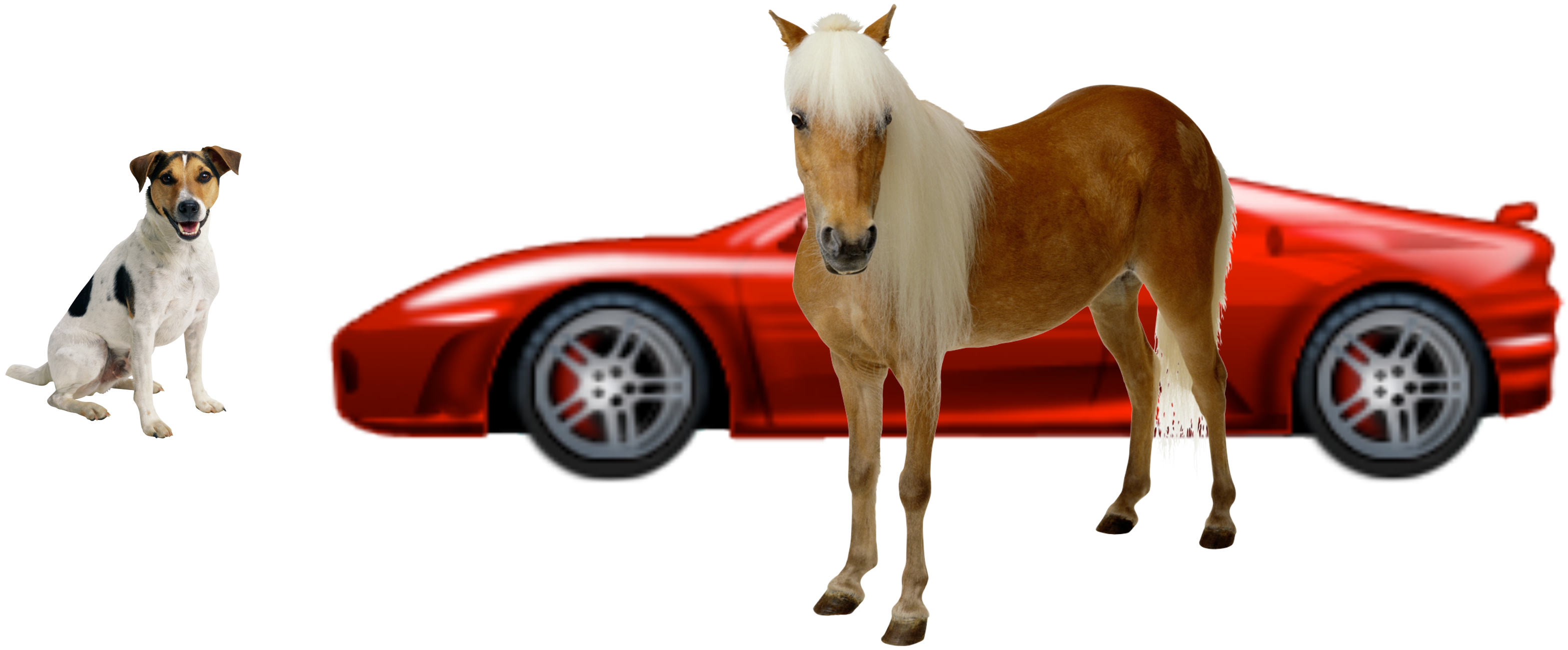
“A small apple tree is behind the chestnut tree.”

A TREE DOESN’T HAVE A FRONT OR BACK EITHER.

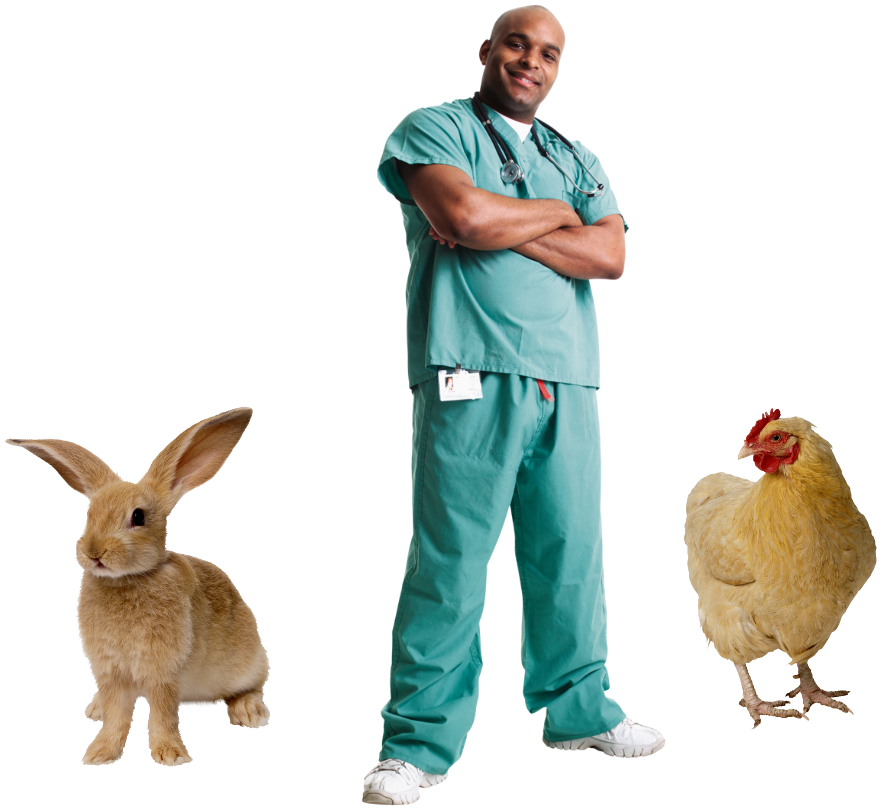
There are three main ways to use language for describing the location of objects.

World-based	**Landmark-based**	**Observer-based**
Using a coordinate system with directions available in the environment: compass directions, uphill, upstream, etc.	Directions determined with reference to asymmetries in the reference object, e.g., human body, vehicle, building, etc.	Coordinate system anchored in landmarks, but directions determined with respect to the speaker's location.
"two miles south of Soda Hall"	"in photographs he always stood by the queen's left side"	"he was standing over there, to the left of the large tree"

Ambiguous case:
(Which animal is in front of the car?)



Confusing case:
(Which animal is to the doctor's left?
Which animal is to the left of the doctor?)



Outline:

1. Text to scenes 2. MWCI? 3. The lesson

(a) WordsEye (b) Real 3D (c) Deixis

- ✤ To get serious about the relation between a speaker and his or her surrounding world, we need to get into the topic of **deixis**.
- ✤ Deixis is the name linguists give to those aspects of language that are sensitive to aspects of the actual speech situation - the speaker's immediate context.
- ✤ The adjective is **deictic**.

Five kinds of deixis.

1. **Person deixis**
   the person speaking ("I"); the person addressed ("you")
2. **Place deixis**
   the location of, or place indicated by, the person speaking ("here")
3. **Time deixis**
   the time of, or a time-span including, the speech event ("now", "today")
4. **Social deixis**
   (not an explicit system in English; shows social status of speaker)
5. **Text deixis**
   indicating places related to points in a reader's ongoing experience with a text ("in the next paragraph")

Person Deixis

- “I”: so-called **first person**
- “you”: so-called **second person**
- but what is “first person plural”?
  multiple speakers, as in a Greek chorus?

Place Deixis

- The scope of the word “here” is relative: ‘where I am and you aren’t’ - ‘where we both are’ - ‘this room’ - ‘this continent’ - ‘this galaxy’
- Place deixis also includes **gestural deixis**, locations indicated by the speaker’s gestures, using “here” - “there” - “this” - “that”

Time Deixis

- The most general word is “now”; its scope is relative: the moment I say it (give or take a second or two), the current epoch.
- More specific words indicate time periods related to the period that includes ‘now’: today, yesterday, tomorrow; this week, last week, next week (or month, year, season, semester, . . .)
- The word “ago” means ‘before now’.

Social Deixis

- Some languages have ways of representing social stratification differences between speaker and hearer, or social indicators of the relation between the speaker and the situation, built into important aspects of the grammar.
- In English it shows up in various indirect ways of showing politeness, respect, etc.

Text Deixis

- Some of the same words and patterns used for time are also used for text: “this sentence”, “the next chapter”, etc.
- But there are some that are tied to certain facts about the written language: “this point was treated above”, “as we will see below”, “the former / the latter”, etc.
- An extremely local example: “Click here.”

Outline:
1. Text to scenes 2. MWCI? 3. The lesson
(a) in (b) come (c) we (d) may

- Understanding the question “May we come in?” draws on our knowledge of **person deixis, time deixis, place deixis, and social deixis**.
- When we understand this sentence we have in mind a situation in which one party, on behalf of a group of at least two, requests permission to enter some kind of enclosure; and we have an idea of how the participants are located with respect to each other and the enclosure. The social status difference, in this context, is recognized by asking for permission; politeness toward the hearer is shown by using “may” rather than “can”.
- What did we have to know about English to figure all that out?

Outline:

1. Text to scenes 2. MWCI? 3. The lesson

(a) in (b) come (c) we (d) may

- The word “in” and its partner “into” usually require an accompanying word or phrase to tell us something about an enclosed space of some kind; but when “in” is used by itself, that means that that enclosure is obvious in the context, or has recently been mentioned.

- When “in” and “out” are used in connection with a verb of motion (walk in, swim out, go in) they signal a motion from the outside to the inside (“in”), or from the inside to the outside (“out”), of the indicated enclosure.

Outline:
1. Text to scenes 2. MWCI? 3. The lesson
(a) in (b) come (c) we (d) may

- The verb "come" signals a motion from one place to another, with some constraints on what the destination can be.
- Let's define S as speaker, H as hearer, CT as coding time (the time of the utterance), and AT as arrival time (the time when the traveller arrives at the destination). LS is location of speaker; LH is location of hearer.
- The verb "come" can be used under any of the following four conditions.

1. **Motion is toward LS at CT. (*I-here-now*)**

   “Tell them to come here soon.”
   “Please come and help me.”
   “Come in!”

2. **Motion is toward LH at CT. (*you-there-now*)**

   “I’ll come to your office right away.”
   “My son will come right over to help.”

3. **Motion is toward LS at AT. (*I-there-then*)**

   “When I get to the airport tomorrow,
   can someone come and pick me up?”

4. **Motion is toward LH at AT. (*you-there-then*)**

   “No matter where you are, if you get in trouble,
   call me and I’ll come and help you.”

There are also some other possibilities

5. **Either S or H goes somewhere, and somebody accompanies him or her.**

"I'm going shopping, wanna come along?"
"Hey, where you going? Can I come with you?"

6. **A third-person narrative allows a central character to count as the deictic center for a number of deictic expressions, including "come".**

"Manfred knew that his friends were coming over soon,
and that today he'd have to make the decision."

Outline:

1. Text to scenes 2. MWCI? 3. The lesson

(a) in (b) come (c) we (d) may

- We already mentioned that there's something funny about the expression "first person plural".
- "We" refers to me and somebody else: maybe that includes you, and maybe it doesn't. Many languages distinguish an **exclusive "we"** from an **inclusive "we"** - excluding or including the hearer.
  - (a) *You can't hold us here forever. Let us go!* (excl.)
  - (b) *There's a party tonight. Let's go!* (incl.)
  - (c) *May we come in?* (??)

Outline:

1. Text to scenes 2. MWCI? 3. The lesson

(a) in (b) come (c) we (d) may

The word "may" has (at least) three distinct uses:

(a) **epistemic**, indicating likelihood or probability

"It may rain tonight."
"We may not have enough money for the trip."

(b) **deontic**, indicating permission

"You may leave when you're ready."
"May I go home now?"

(c) **magical**, used in blessings and curses

"May you live forever!"
"May you rot in hell!"

Interpretation

1. English "come", unlike its closest translation in other languages, can be used of motion toward a place associated with the Hearer.
2. In our text, the Speaker is the one wanting to move, so the motion can't be "toward SL"; therefore it has to be "toward HL"; therefore the Hearer has to be inside the enclosure.
3. "We" has to be exclusive, for the same reason.
4. The enclosure is not mentioned, so it has to be apparent in the context: the image we create has an enclosure with two people outside, one of them is speaking to someone on the inside, who is taken to be the "gate-keeper" (his permission is needed).

Let's practice:
one speaker, one hearer.

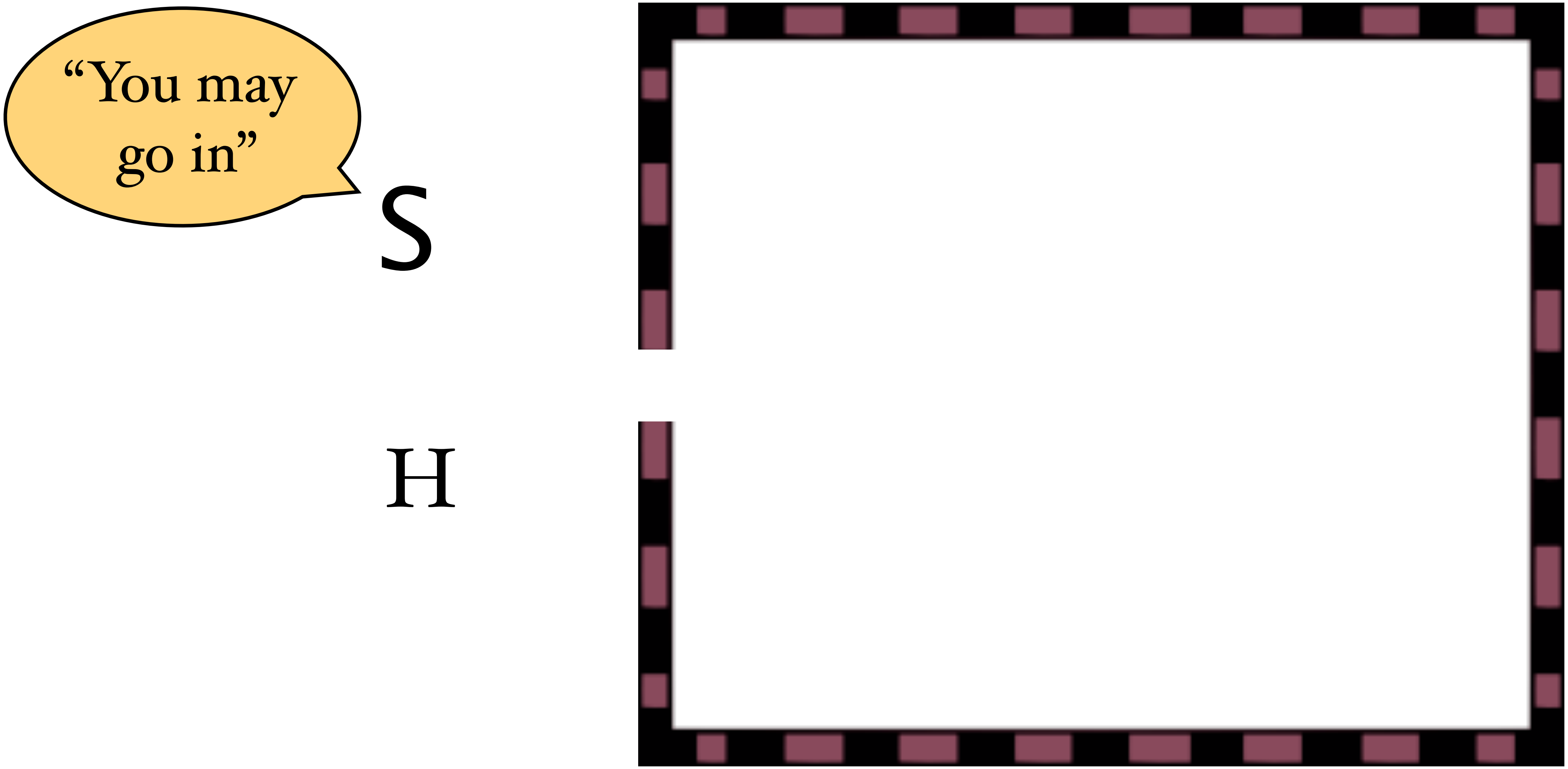
The cast of characters for the following diagrams: the "supplicant" is the one who, like my first dog, wants to be on the other side of the gate.

S = Speaker as supplicant

H = Hearer as supplicant

S = Speaker as gate-keeper

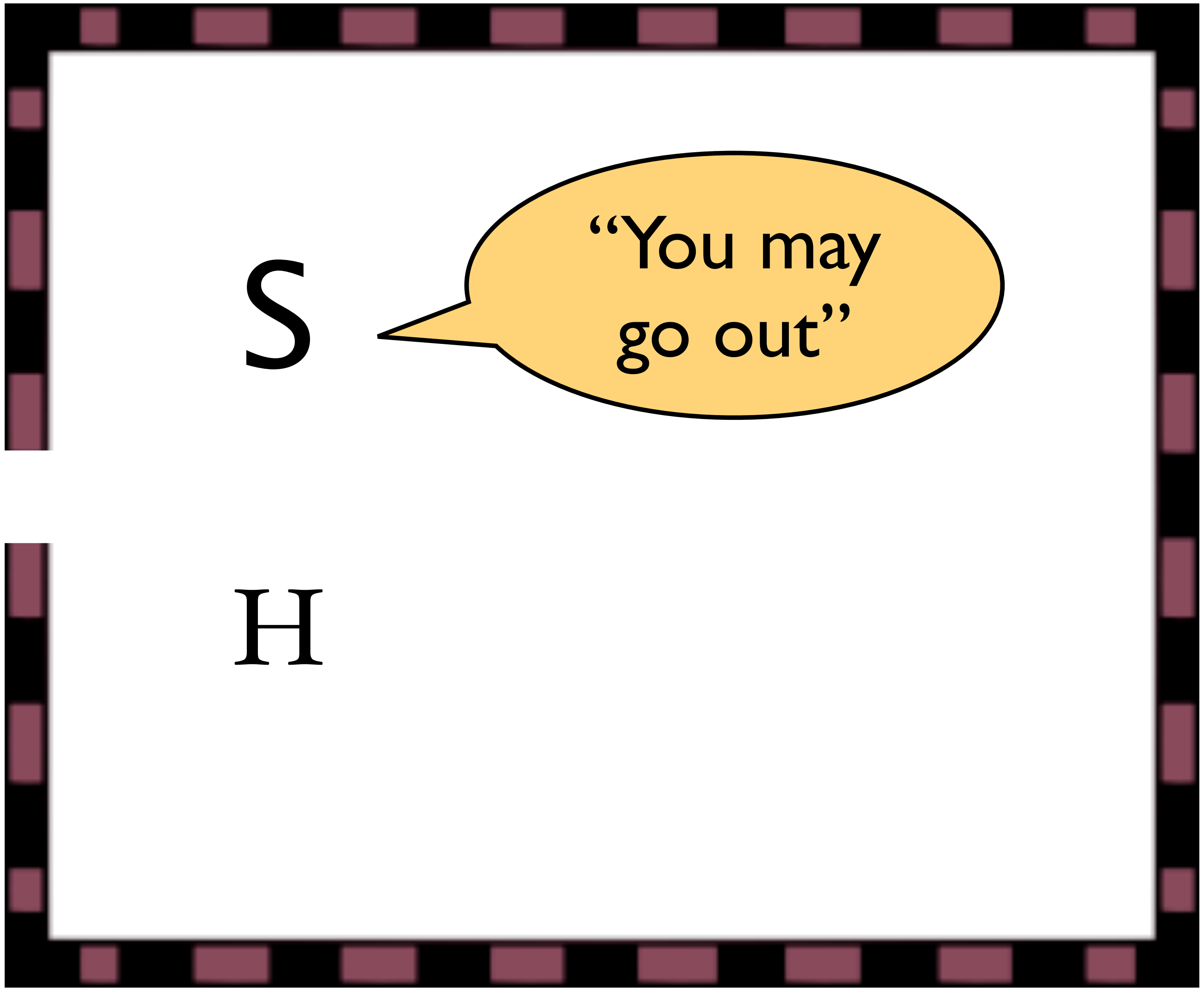
H = Hearer as gate-keeper

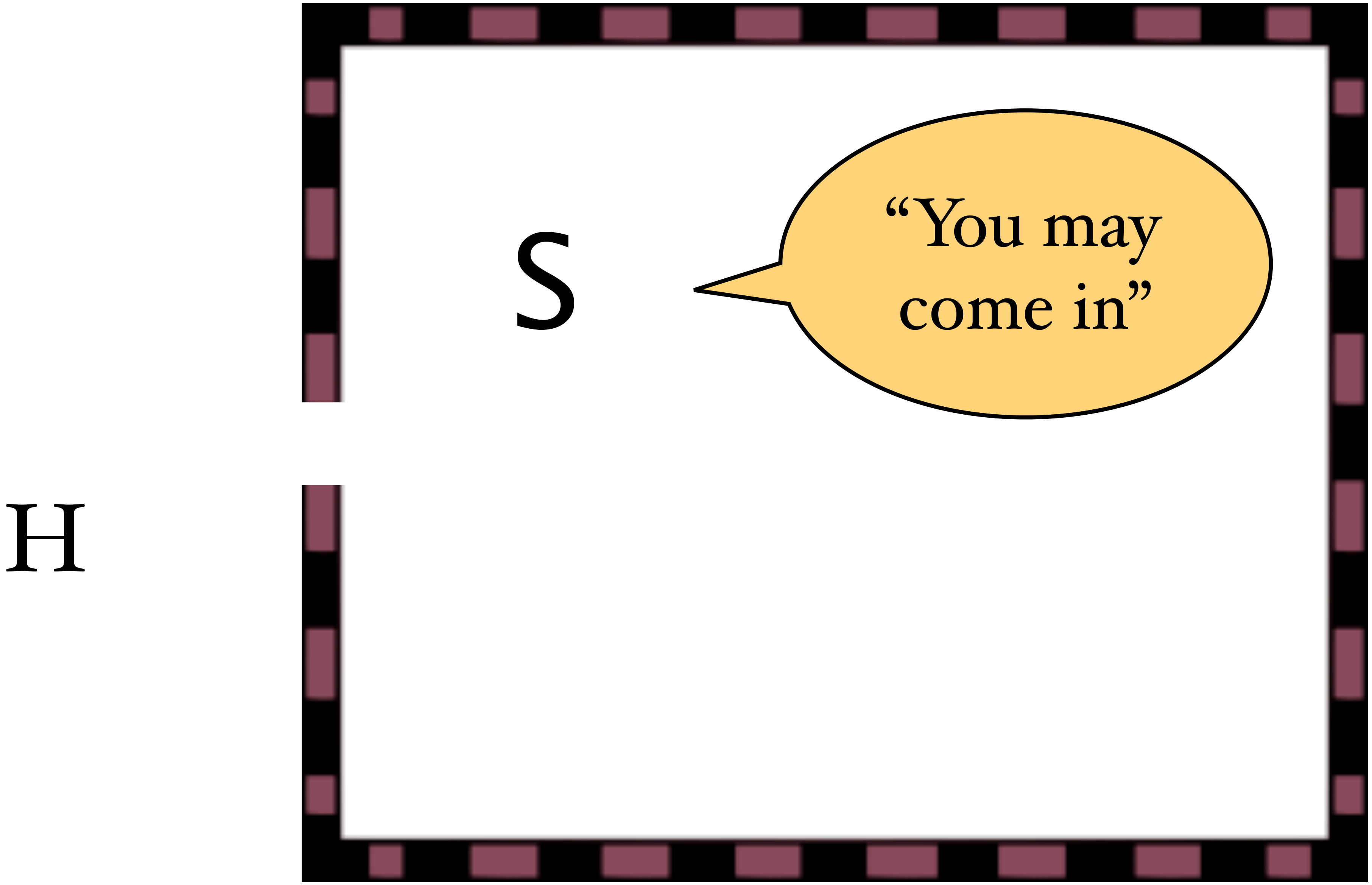


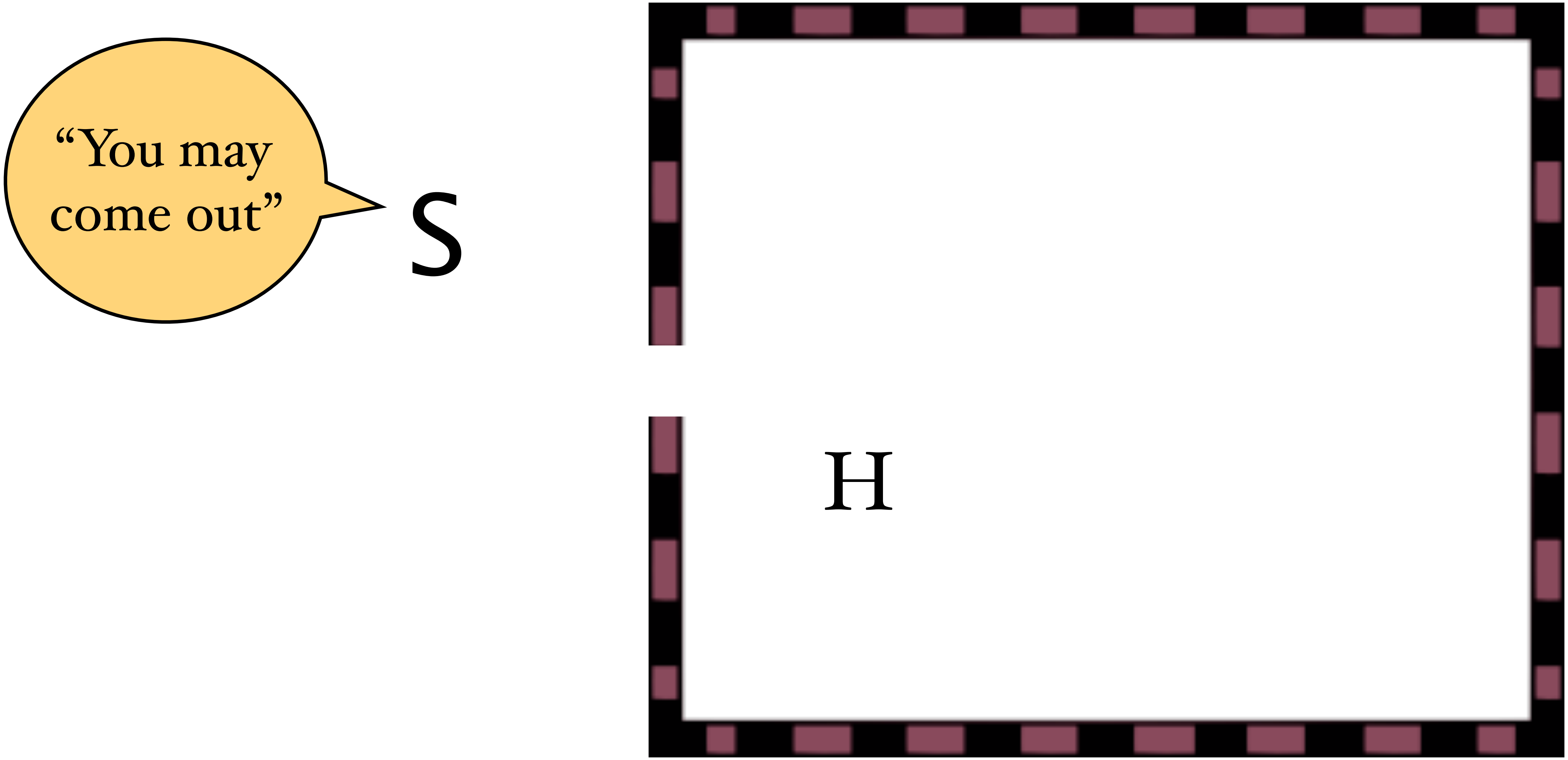
“You may go in”

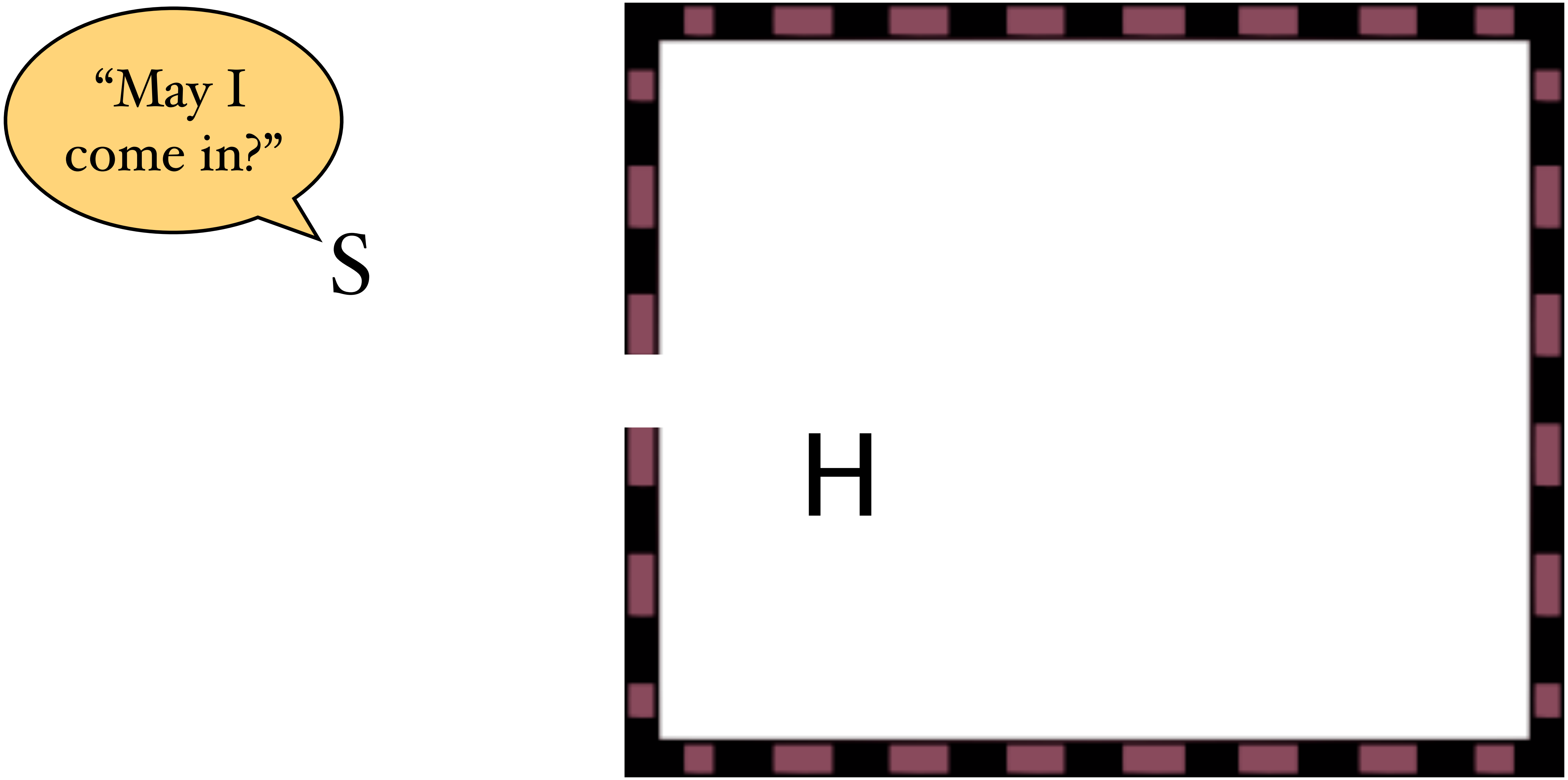
S

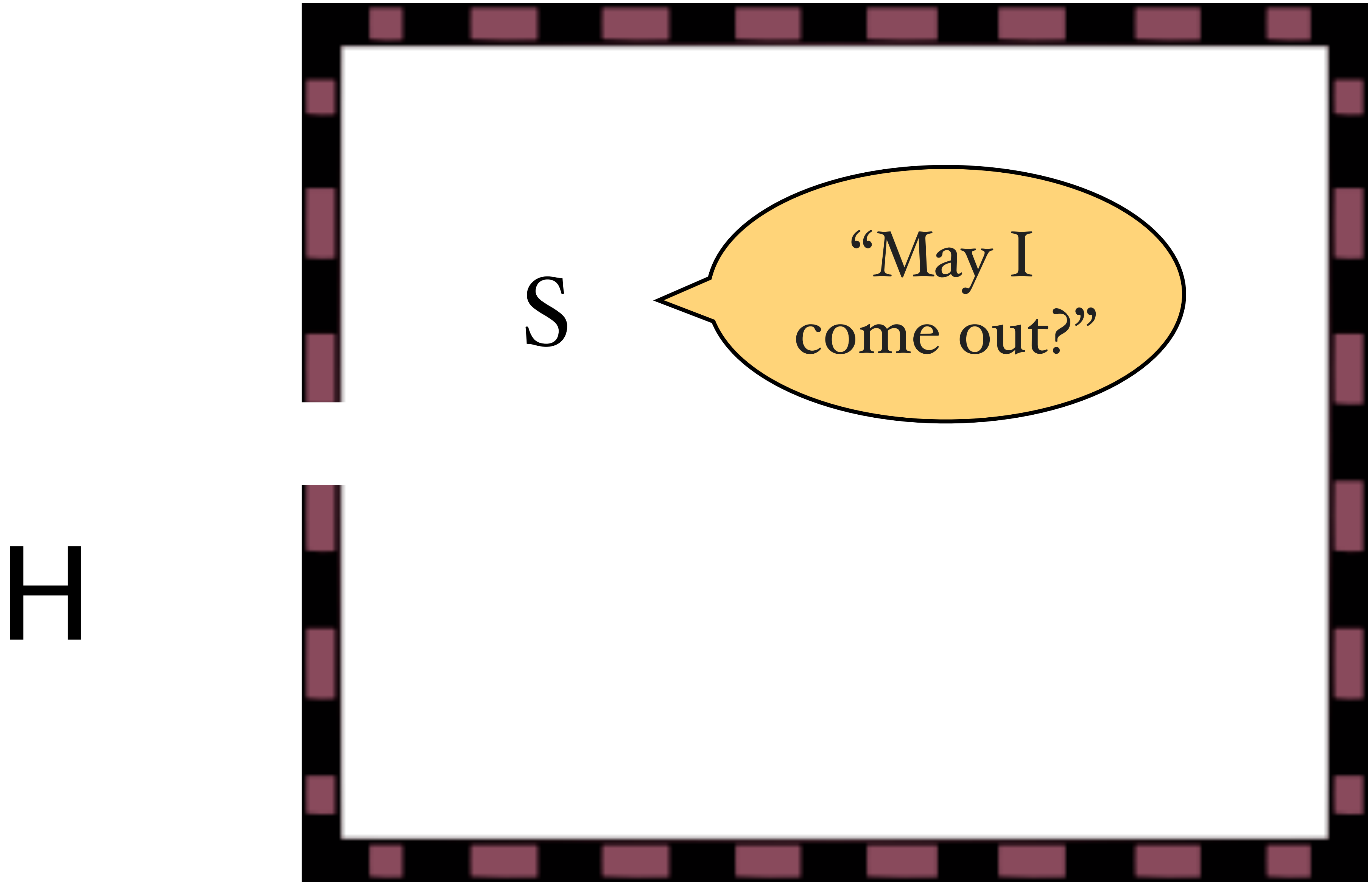
H

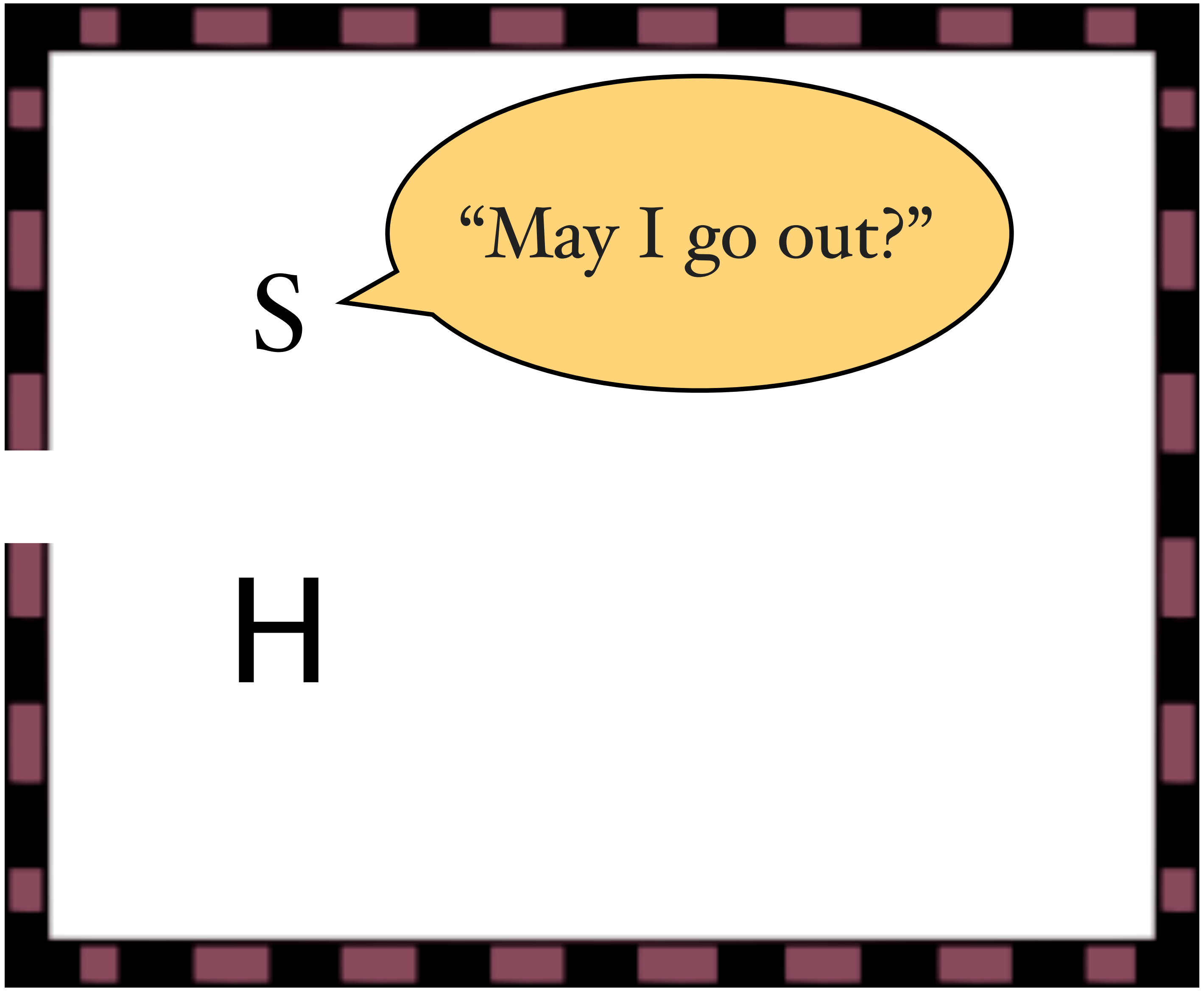

S
“You may go out”
H

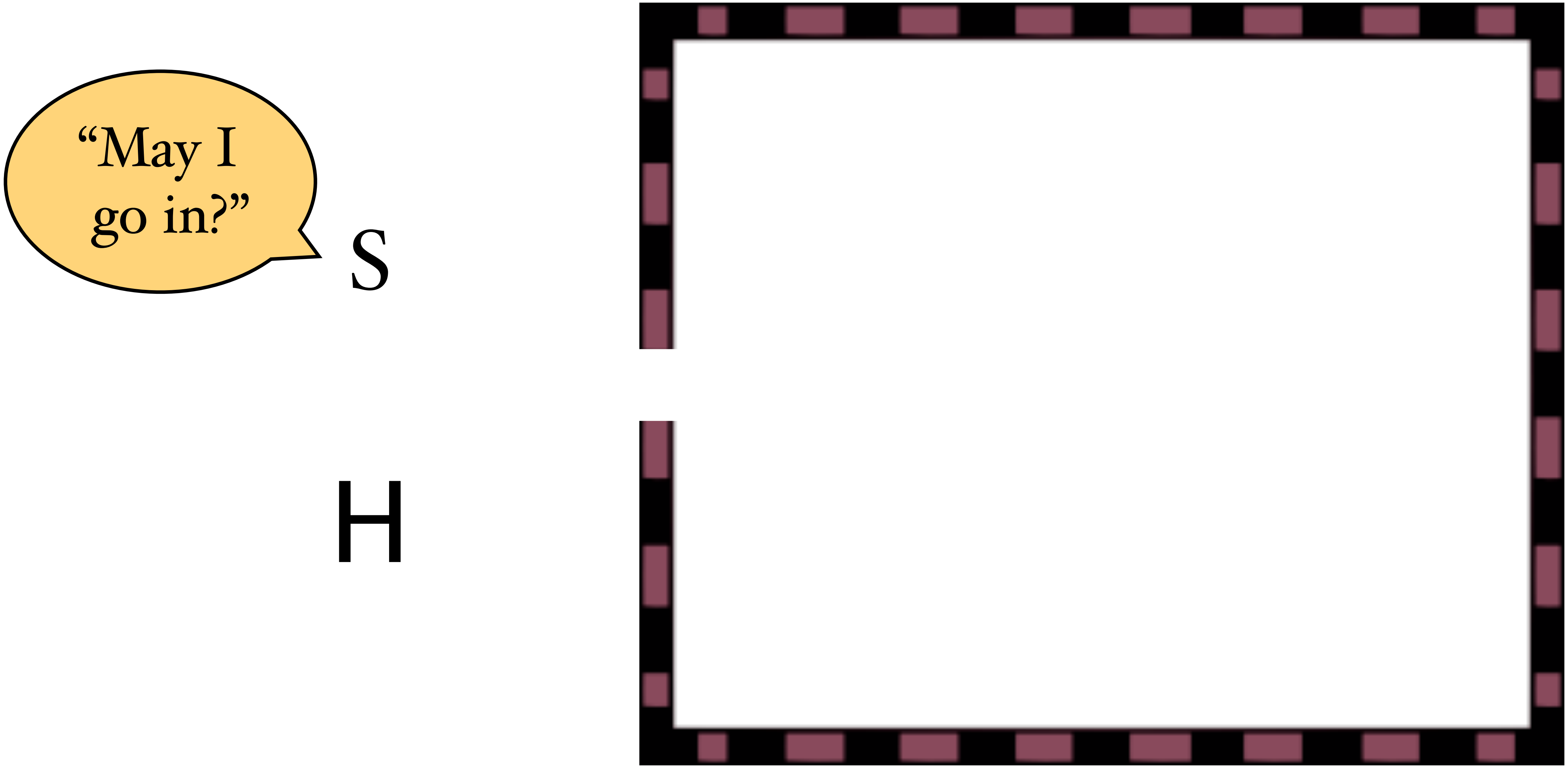

S
"You may come in"
H


“You may come out”
S
H


“May I come in?”
S
H


S
“May I come out?”
H


“May I go out?”
S
H


"May I go in?"
S
H

Outline:
1. Text to scenes. 2. MWCI? 3. The lesson
(a) translation (b) language in the world

- Our theme question is interesting for a number of reasons, among them the facts that many aspects of the English grammar used in it provide a challenge for translation; but also, it is a clear case of an utterance that has to be anchored in some real situation.

- There is also a general interest in dealing with utterances that present constraints on their contextual conditions. We can imagine some agency with a large and detailed map of the relevant terrain receiving great numbers of such reports, especially if they are accompanied by time and location information, and assembling their information into a global picture of what is going on.

Outline:
1. Text to scenes. 2. MWCI? 3. The lesson
(a) translation (b) language in the world

- There are lots of reasons why a structure-matching translation of a sentence like this wouldn't work, unless the target language is very similar to English.
- Not many languages have a *come*-like word that allows motion toward the Hearer.
- Many languages lexicalize both motion and the trajectory of a motion: instead of *come-in* or *go-in,* for example, they would be more likely to use a word which means enter; similarly with meanings like *come-up, go-down, walk-across,* etc., they would use words that mean ascend, descend, pass, traverse, and the like.
- The analysis would have to recognize that in a sentence like this the exclusive "we" is needed, in case the target language uses separate words for the two "first person plural" meanings.

Outline:
1. Text to scenes. 2. MWCI? 3. The lesson
(a) translation (b) language in the world

- Agencies that depend on being able to make decisions on the basis of masses of reports from people in the field will need to be able to take into account the perspectival information that individual reports are likely to contain.
- Citizen reports on near-earth astronomical events, pour in to TV stations, newspapers, police departments: *"I was driving south on highway 83 at approximately 9:45 when suddenly I saw a blue light above me, zooming off to my left."* It may be important to decide rapidly whether these are all the same events or multiple different events. [NASA Fireballs Network]
- My colleagues and I are involved in the linguistic part of a study of reports from military patrols in the battlefield, where a command post may need to assemble the various reports coming in, including GPS readings of the coordinates of the reporter and descriptions of the kind *"after we reached the bridge we noticed an enemy patrol coming toward us from the north..."* They need to know which of these reports identify same or separate events.

All of my examples assumed human participants.
That was too limiting.
It's warm inside, it's cold and miserable outside.
"May we come in? Please?"