

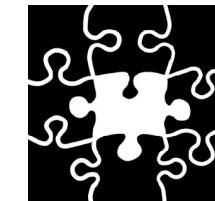
Grounded Dialogue Modelling

An Information-Theoretic Perspective

Information Retrieval 2 — MSc AI
21 September 2022



UNIVERSITY OF AMSTERDAM



Mario Julianelli

Today

- Introduction: Dialogue and Grounded Dialogue
- Grounded Dialogue Datasets
- A Bit of Theory
- Case Study: Reference Games
- Outlook: IR to Model Human Behaviour?

Dialogue

What is it and why do we care?

- **What?** Using language for inter-personal communication and interaction
- **Why?** The primary form of language use and language learning
- **Where?** Face-to-face, on the phone, on Zoom, on Signal, on Reddit, ...
- **How?** Linguistics, psychology, sociology, cognitive science, mathematics, ...



Grounded Dialogue

Interactive language use in context

Dialogue (just like any type of language use) happens in **context**, in an *environment*.

Speakers communicate to change the state of the environment and achieve **goals**.

Communicating is an **action** – through dialogue, an *interaction*.

Grounded Dialogue Modelling

The study of interactive language use in context

Dialogue (just like any type of language use) happens in **context**, in an **environment**.

Speakers communicate to change the state of the environment and achieve **goals**.

Communicating is an **action** – through dialogue, an *interaction*.

- What is the relevant context of an interaction?
- How does the context relate to a speaker's communicative goals?
- What are the decision making strategies that humans follow to choose words and achieve goals in their environment?
- Can we replicate them in a computer system?

Grounded Dialogue Modelling

The study of interactive language use in context

Dialogue (just like any type of language use) happens in **context**, in an **environment**.

Speakers communicate to change the state of the environment and achieve **goals**.

Communicating is an **action** – through dialogue, an *interaction*.

- **What is the relevant context of an interaction?**
- **How does the context relate to a speaker's communicative goals?**
- What are the decision making strategies that humans follow to choose words and achieve goals in their environment?
- Can we replicate them in a computer system?

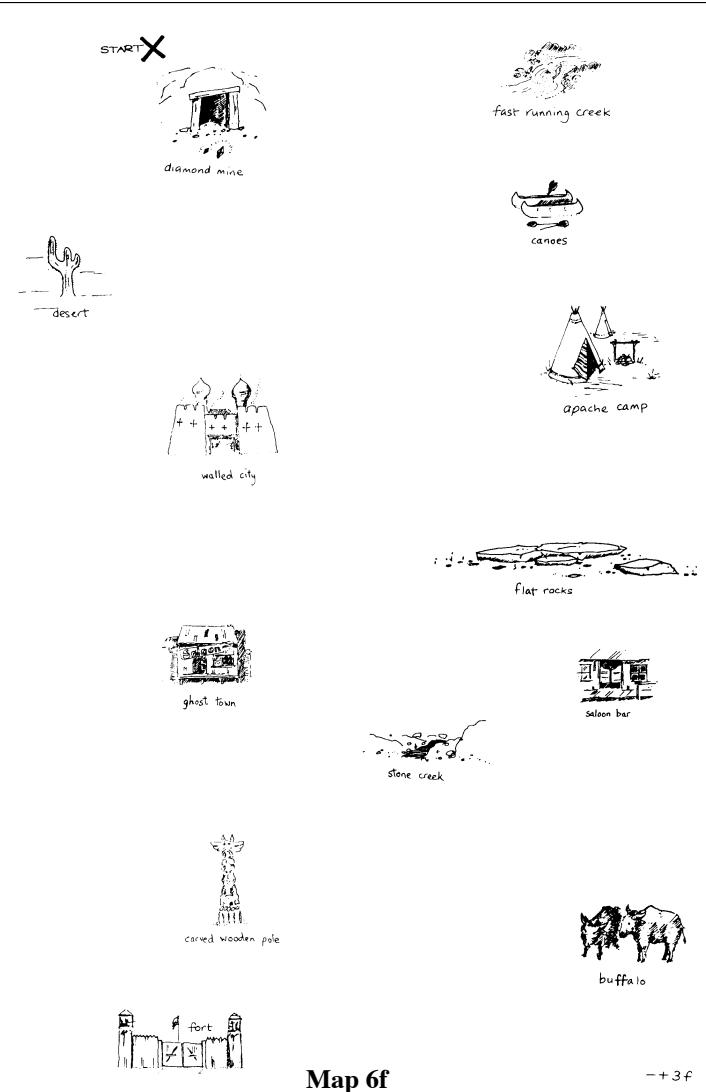
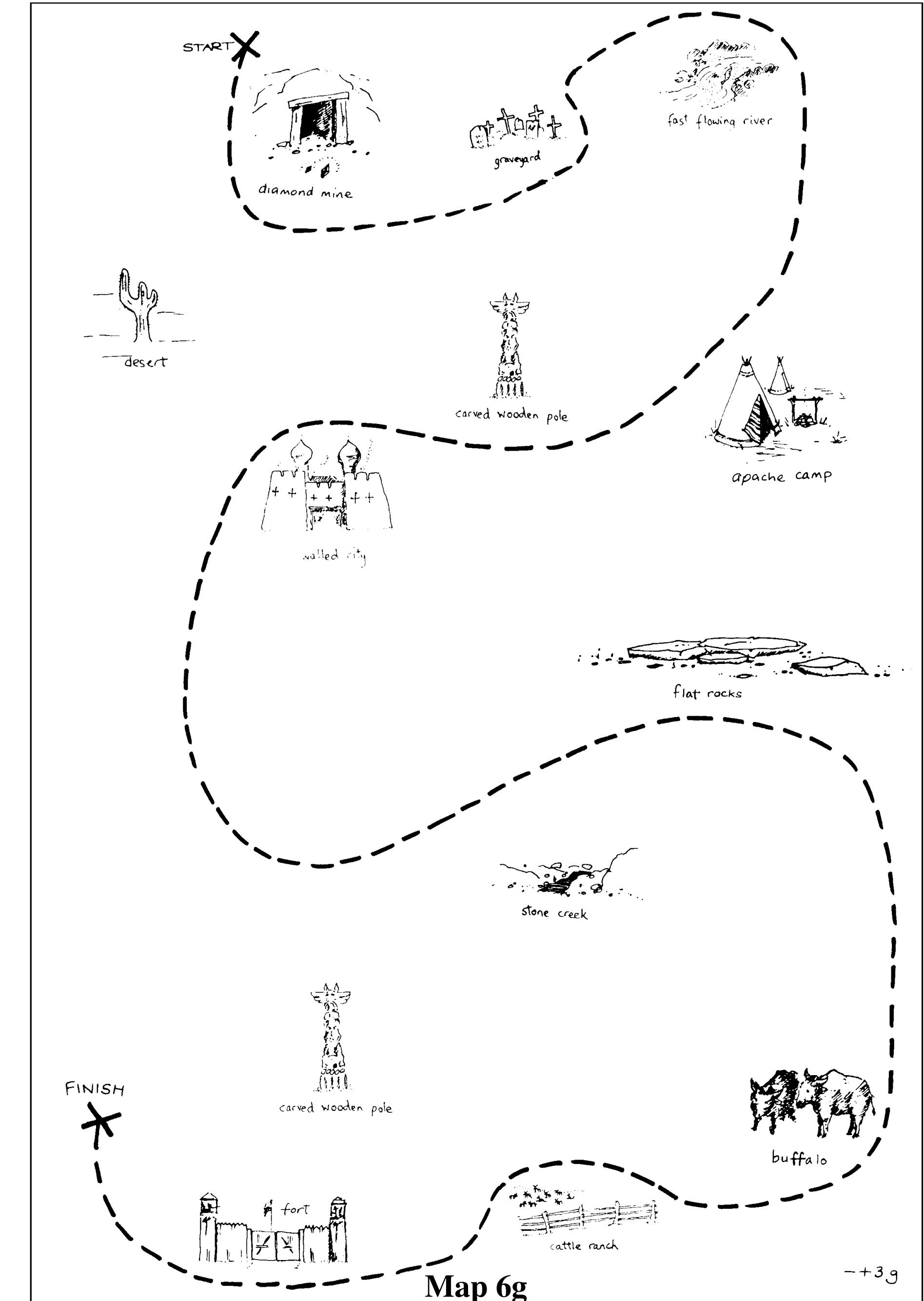
Grounded Dialogue Datasets

Map Task

Anderson et al., 1991. Language & Speech.

Spoken dialogues (transcribed):
instruction **giving** and following to
navigate to a point on a map.

- g start off above the diamond mine
f okay yeah
g now go south from the diamond mine until you are just above the desert
f so that's with the diamond mine on on your right
g that's that's correct uh-huh
g and go below the diamond mine
f mmhmm
g and below the graveyard below the graveyard but above the carved wooden pole
f oh hang on i don't have a graveyard



Links to:
[Data](#)
[Annotations](#)

PhotoBook

Haber et al., 2019. ACL.

Written cooperative *reference game*:
describe images in turn to find
common sets of photographs.

YOU: Do you have a man with two dogs on a bed?

Robin: With a purple wall in the background?

YOU: Yes

Robin: Then yes.

Robin: I have a little boy holding a phone to a teddy bear

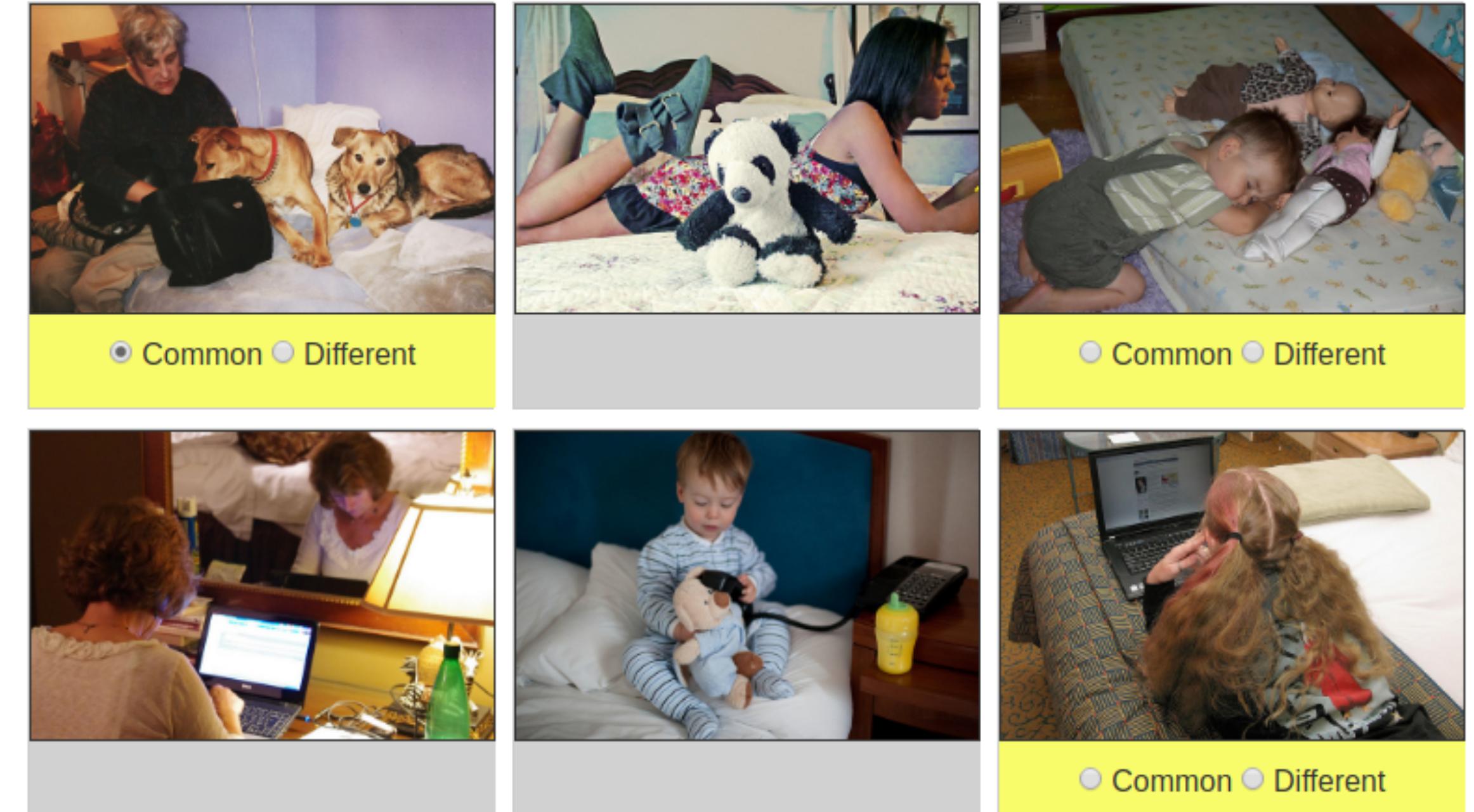
YOU: I have that one as well

My next one is a boy sleeping with dolls|

Send

59 characters remaining.

Page 1 of 5



Submit Selection

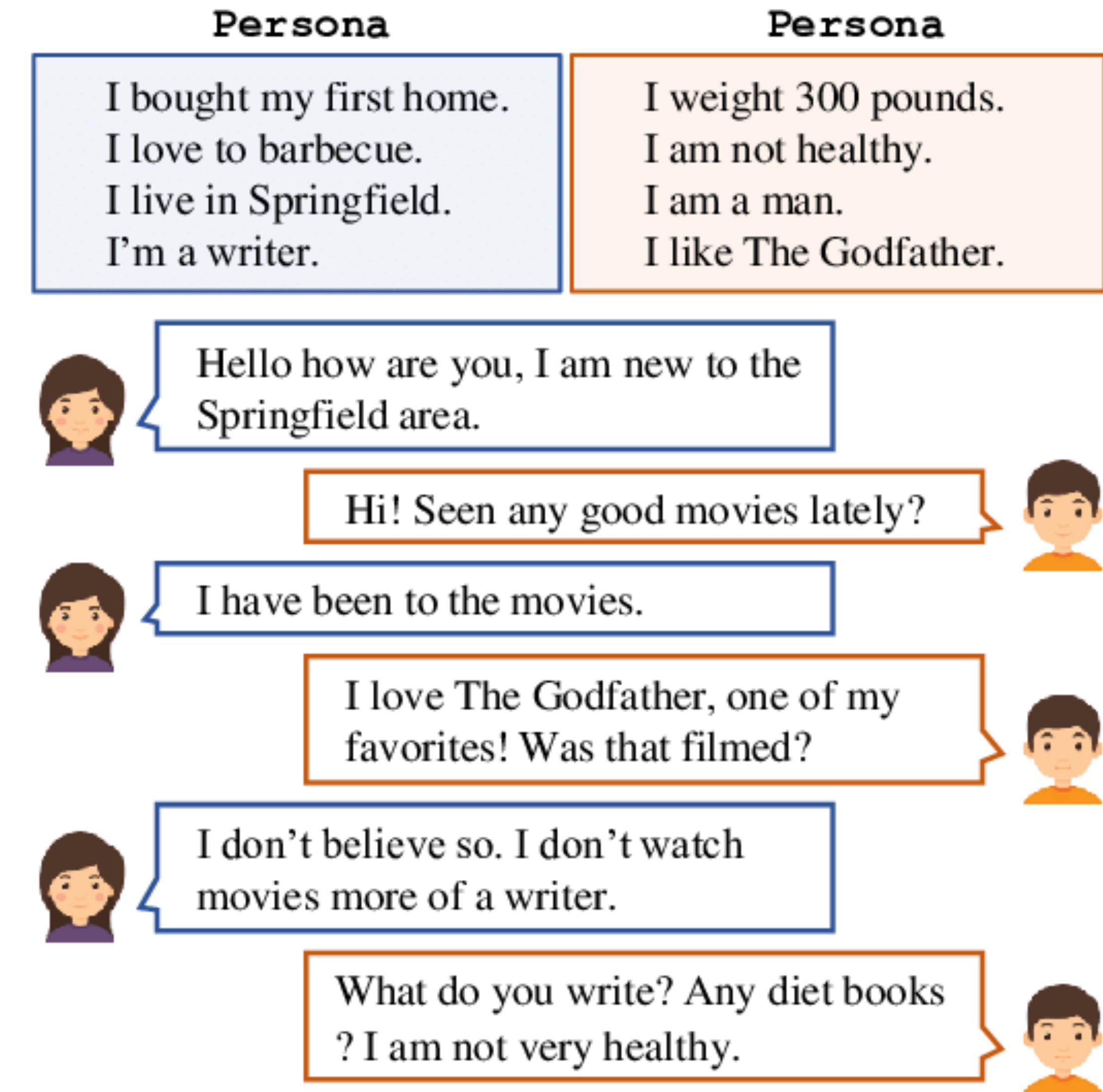
Link to: [PhotoBook website](#)
(data, visualisation, code)

PersonaChat

Zhang et al., 2018. ACL.

Written *chit-chat* dialogue:
given a character description,
chat with another person naturally
and try to get to know each other.

Links to:
[GitHub readme](#)
[ParlAI website](#)



More (Grounded) Dialogue Datasets

A few useful resources

- <https://parl.ai/docs/tasks.html>
- <https://breakend.github.io/DialogDatasets/references.html>
- https://docs.google.com/spreadsheets/d/1N5_5gBKlGR-Origin4jQ6iEqSycyqcoN61JpsHFDQ/htmlview



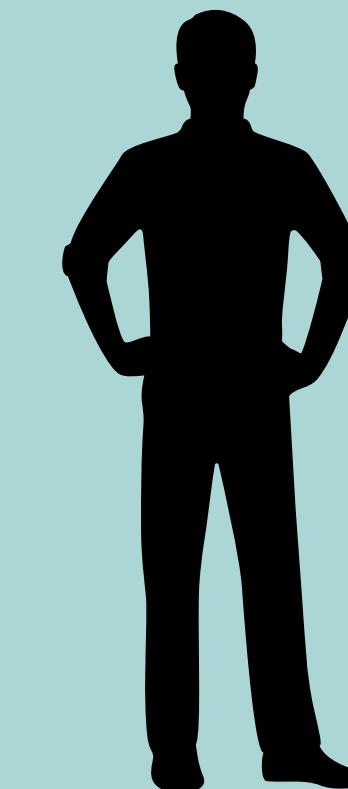
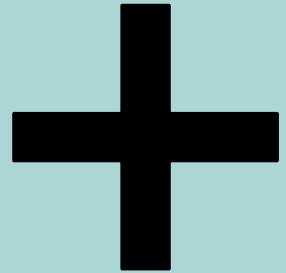
A Bit of Theory

Context

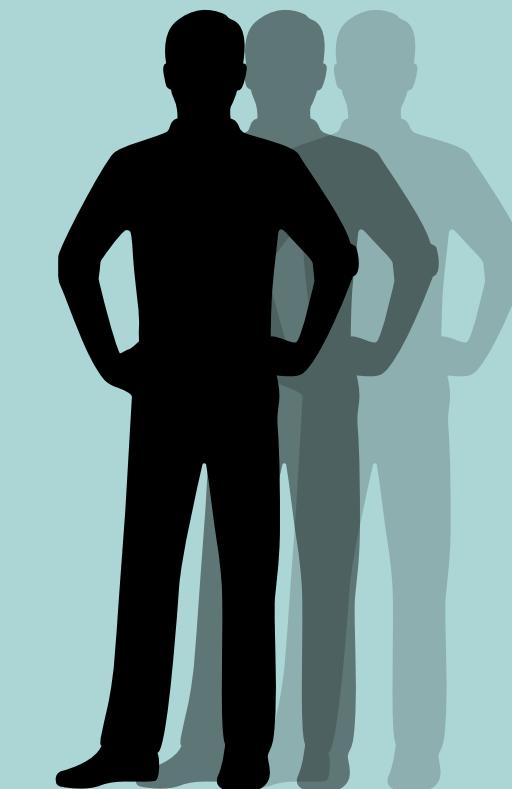
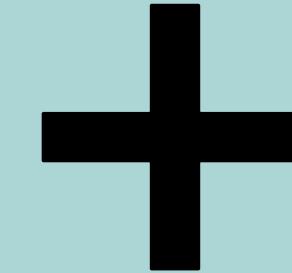


The 'environment' in which the interaction takes place

$w \in W$



Speaker

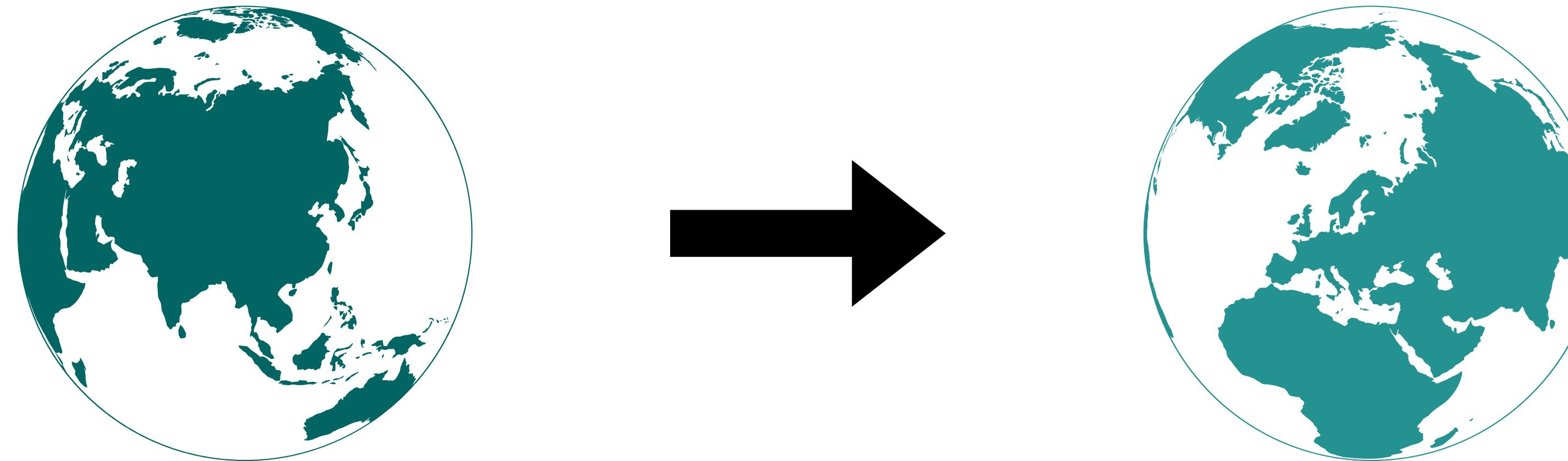


Audience

Communicative Goal



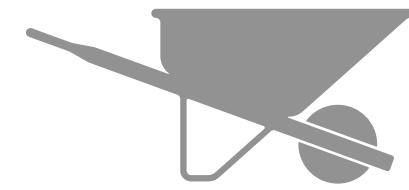
A change of the state of the environment



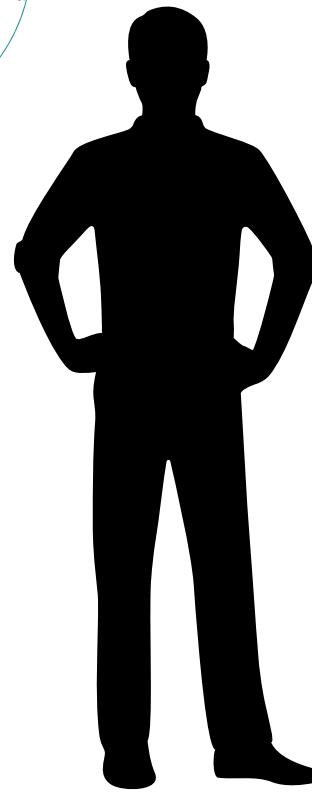
For communication to be successful, the audience must be able to reconstruct the speaker's communicative goal.

Communicative goals shape and constrain the speaker's production choices: different types of utterance correspond to different goals.

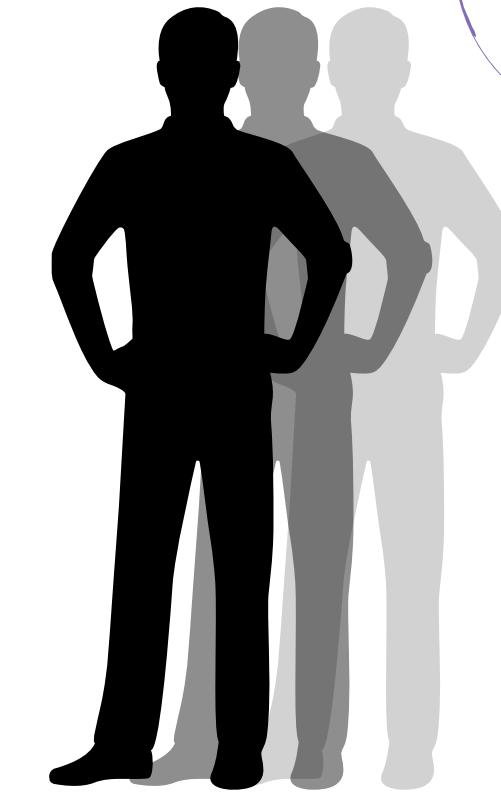
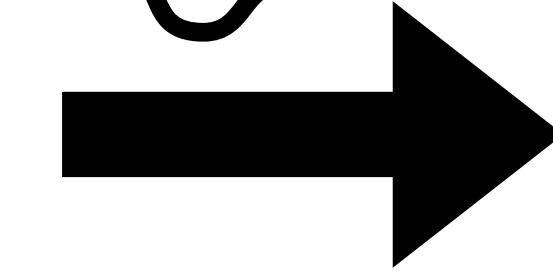
Costs



The cognitive and physical efforts required to communicate



Utterance



Production costs

Speaker executes a bit of behaviour (e.g., speaking or typing) meant to be perceived by the audience in order to convey their communicative intent.

Comprehension costs

Audience attends to and processes the behaviour executed by the speaker in order to reconstruct the speaker's communicative intent.

Speakers estimate these costs and take them into account when they choose words.

Utility

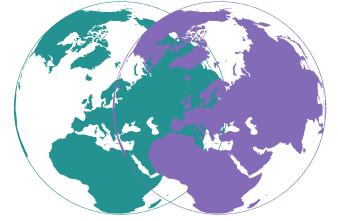


The cognitive, physical, and social effects of a communication act

Inversely proportional to:

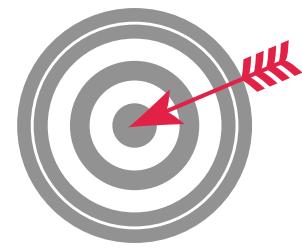


the joint production and comprehension costs (*collaborative effort*)



the distance between the new and the intended state of the world

Directly proportional to:



the positive cognitive, physical, and social effects derived from achieving the intended new state (the communicative goal)

Grounded Dialogue Modelling

The study of interactive language use in context

- What is the relevant context of an interaction?
- How does the context relate to a speaker's communicative goals?
- **What are the decision making strategies that humans follow to choose words and achieve goals in their environment?**
- Can we replicate them in a computer system?

Case Study: Reference Games

PhotoBook

Haber et al., 2019. ACL.

Written cooperative *reference game*:
describe images in turn to find
common sets of photographs.

YOU: Do you have a man with two dogs on a bed?

Robin: With a purple wall in the background?

YOU: Yes

Robin: Then yes.

Robin: I have a little boy holding a phone to a teddy bear

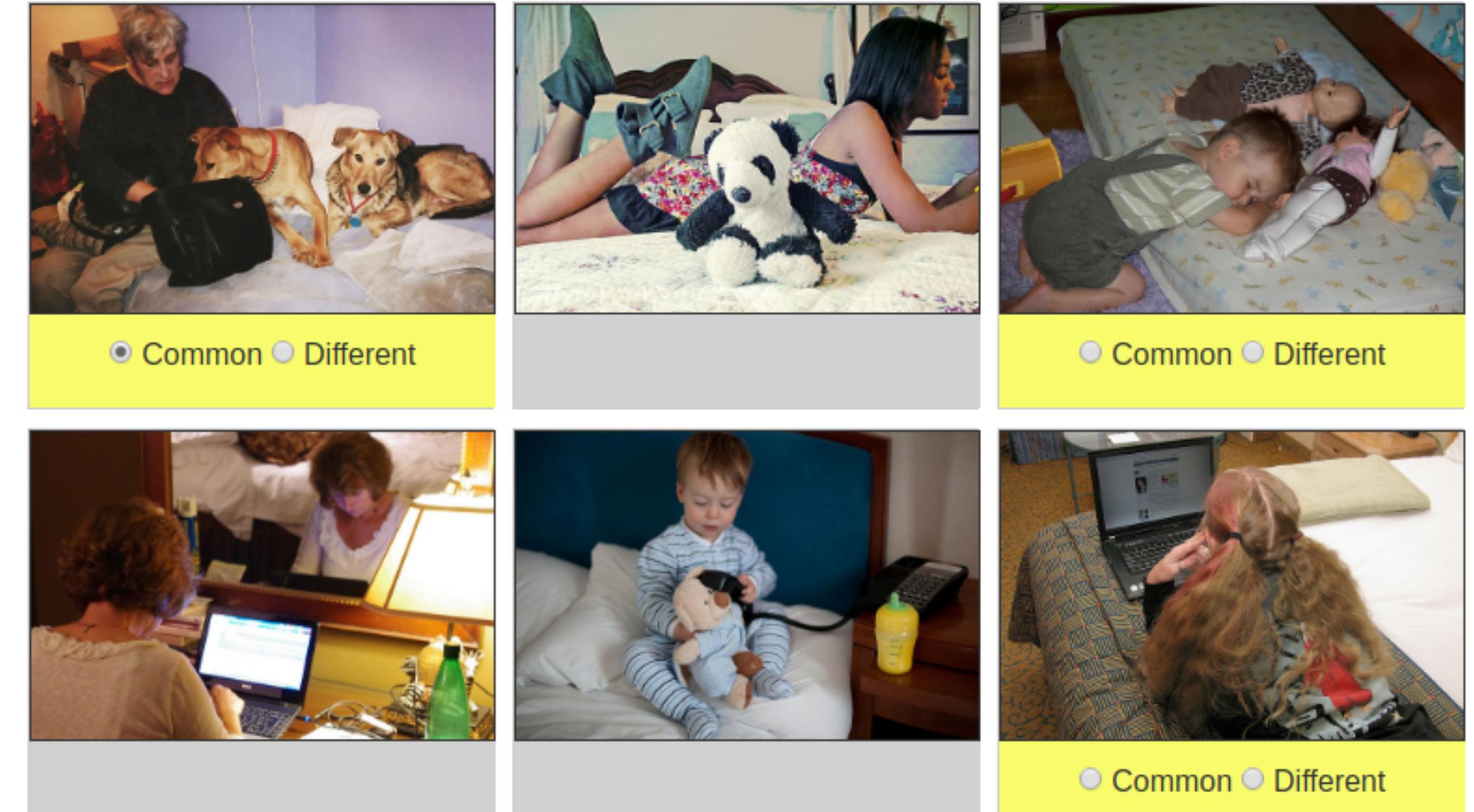
YOU: I have that one as well

My next one is a boy sleeping with dolls|

Send

59 characters remaining.

Page 1 of 5

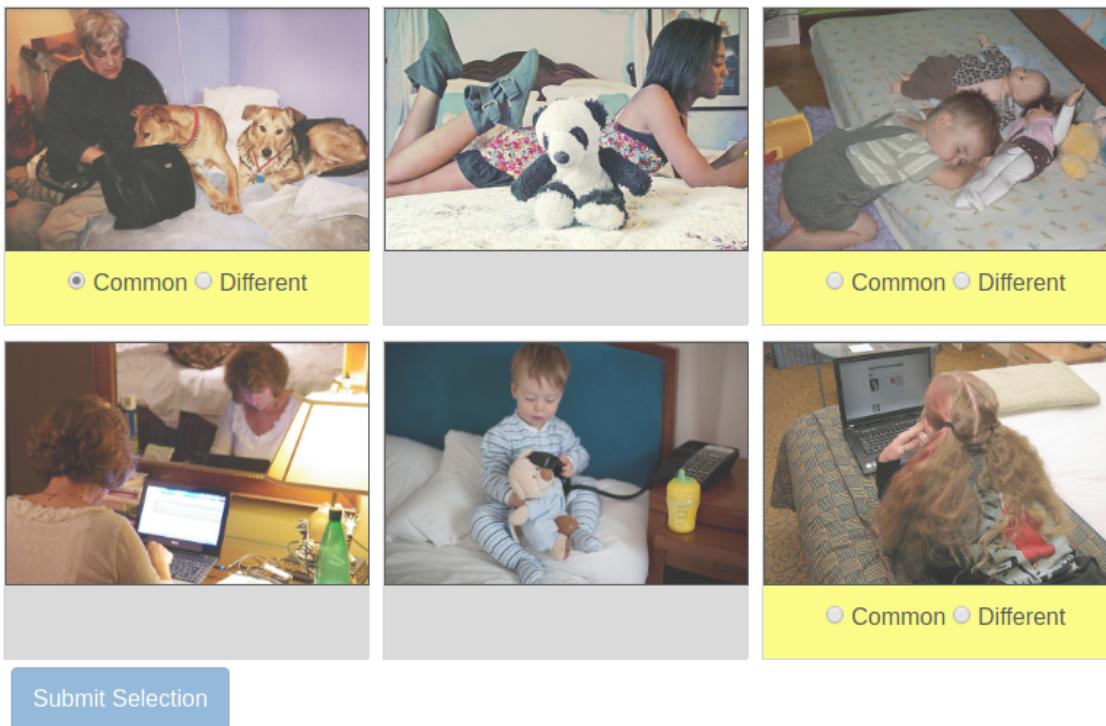


Submit Selection

Link to: [PhotoBook website](#)
(data, visualisation, code)

PhotoBook

Haber et al., 2019. ACL.



YOU: Do you have a man with two dogs on a bed?

Robin: With a purple wall in the background?

YOU: Yes

Robin: Then yes.

Robin: I have a little boy holding a phone to a teddy bear

YOU: I have that one as well

My next one is a boy sleeping with dolls

59 characters remaining.

Send

Reference chains



- 1. *Do you have the girl with the blue umbrella walking by water?*
- 2. *I have the girl with the blue umbrella by the water this time*
- 3. *What about the blue umbrella girl by the water?*
- 4. *Do you have the blue umbrella water girl?*

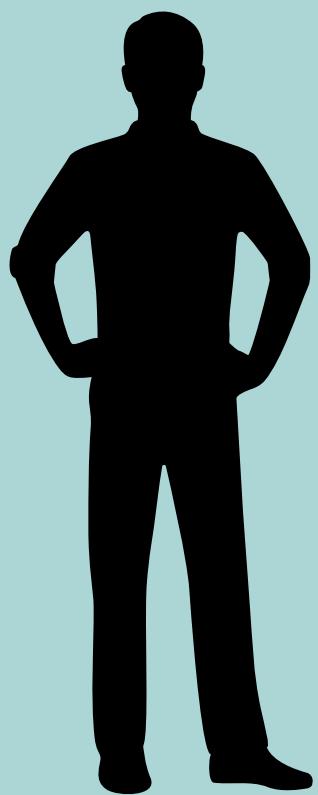
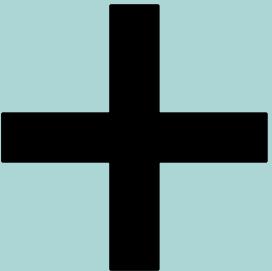
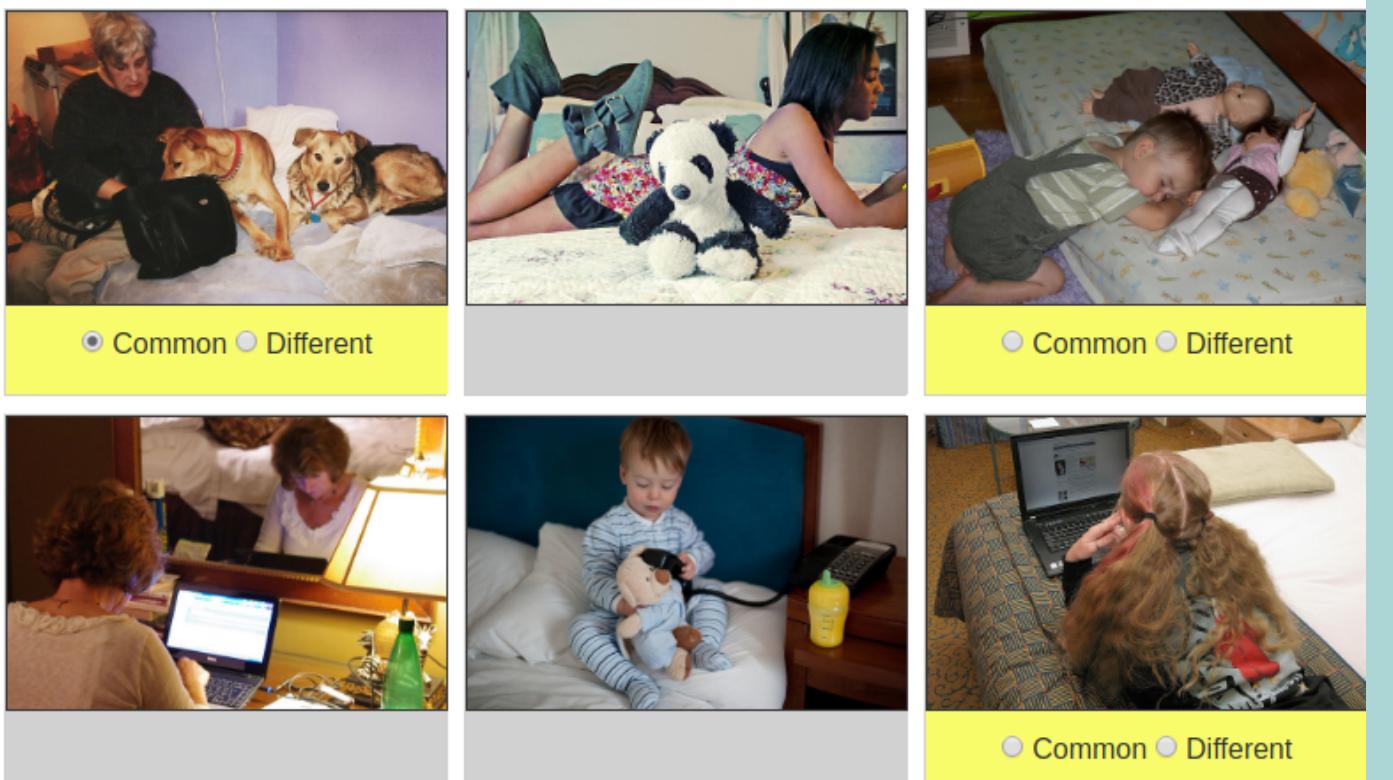
Context



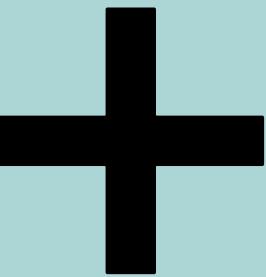
The 'environment' in which the interaction takes place

Visual context

Page 1 of 5

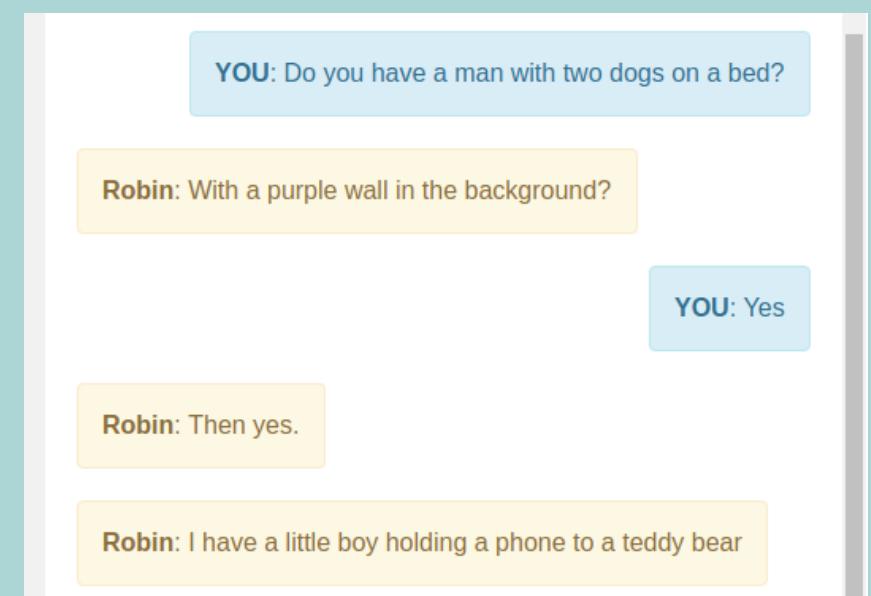


Speaker A



Speaker B

Conversational context

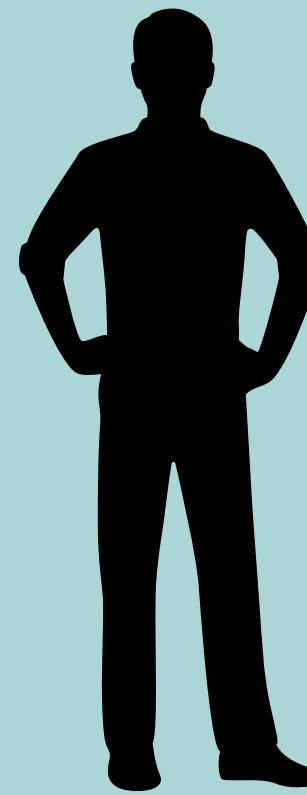
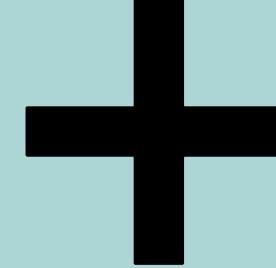


Context

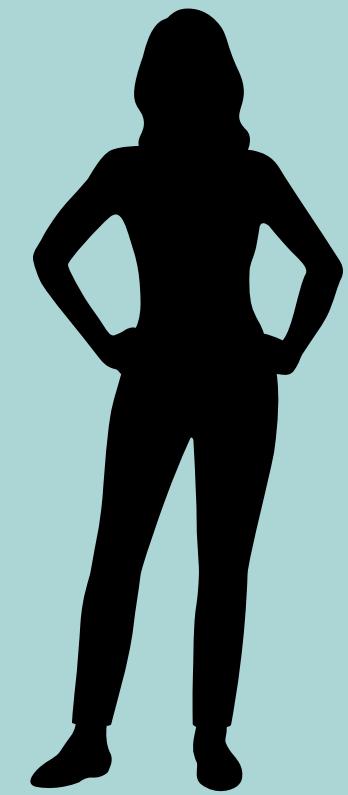
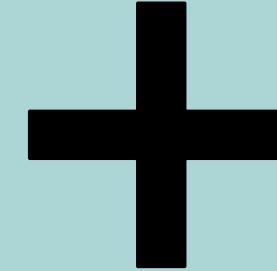


The 'environment' in which the interaction takes place

Visual context



Speaker A



Speaker B

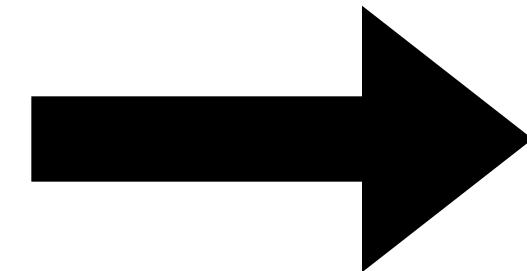
Reference chain

1. Do you have the girl with the blue umbrella walking by water?
2. I have the girl with the blue umbrella by the water this time
3. What about the blue umbrella girl by the water?
4. Do you have the blue umbrella water girl?

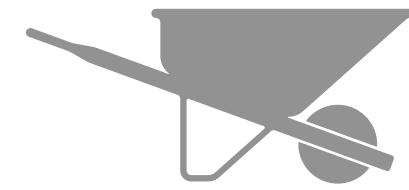
Communicative Goal



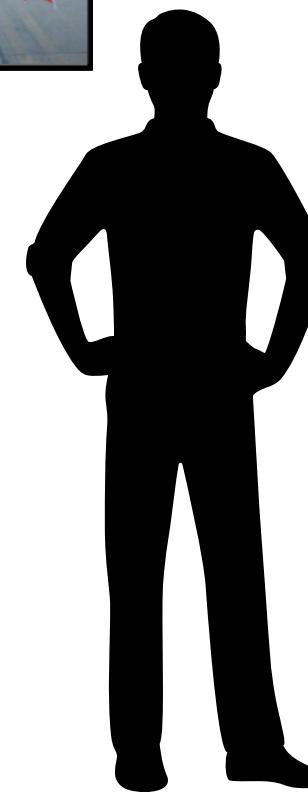
A change of the state of the environment



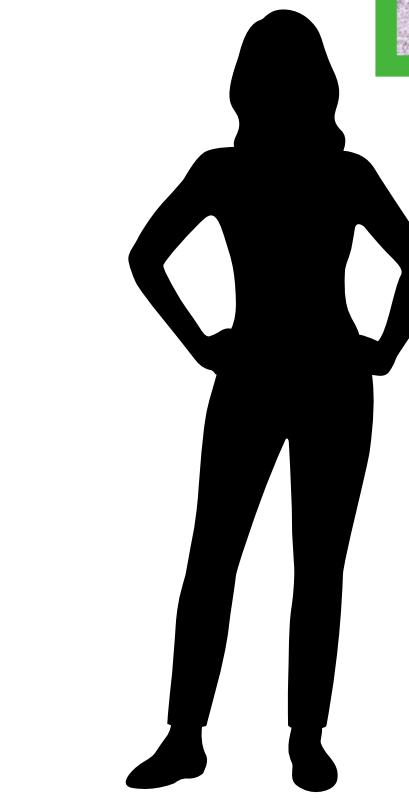
Costs



The cognitive and physical efforts required to communicate



Utterance

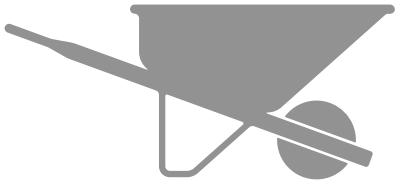


Production costs

Utterance planning, typing, editing, ...

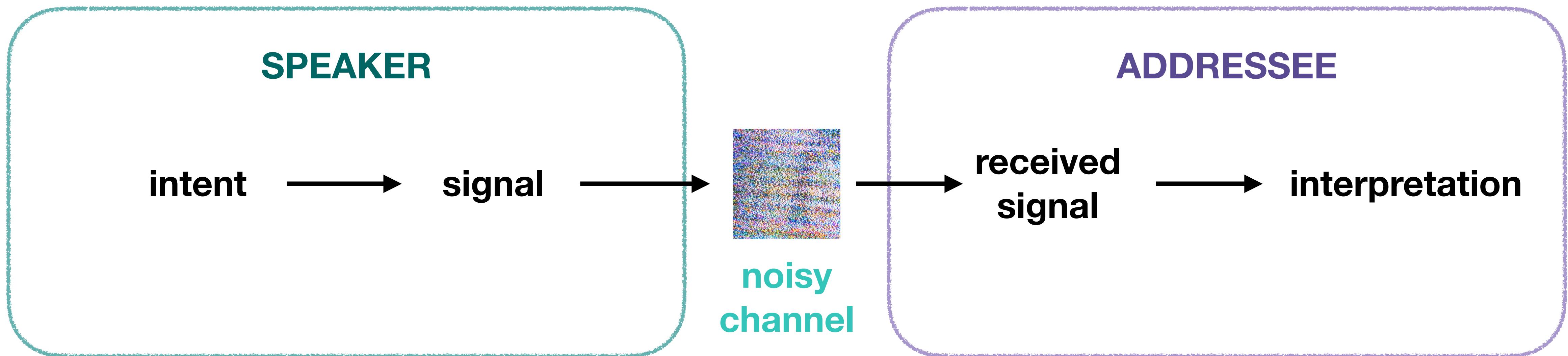
Comprehension costs

Reading, interpretation / reference resolution



Noisy Channel Model of Communication

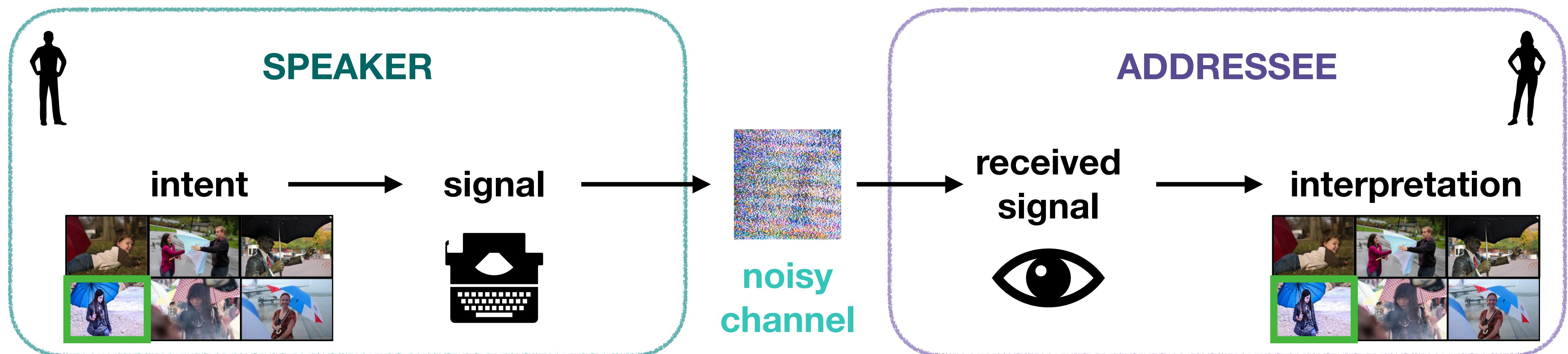
Claude Shannon, 1948

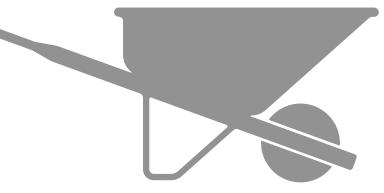




Noisy Channel Model of Communication

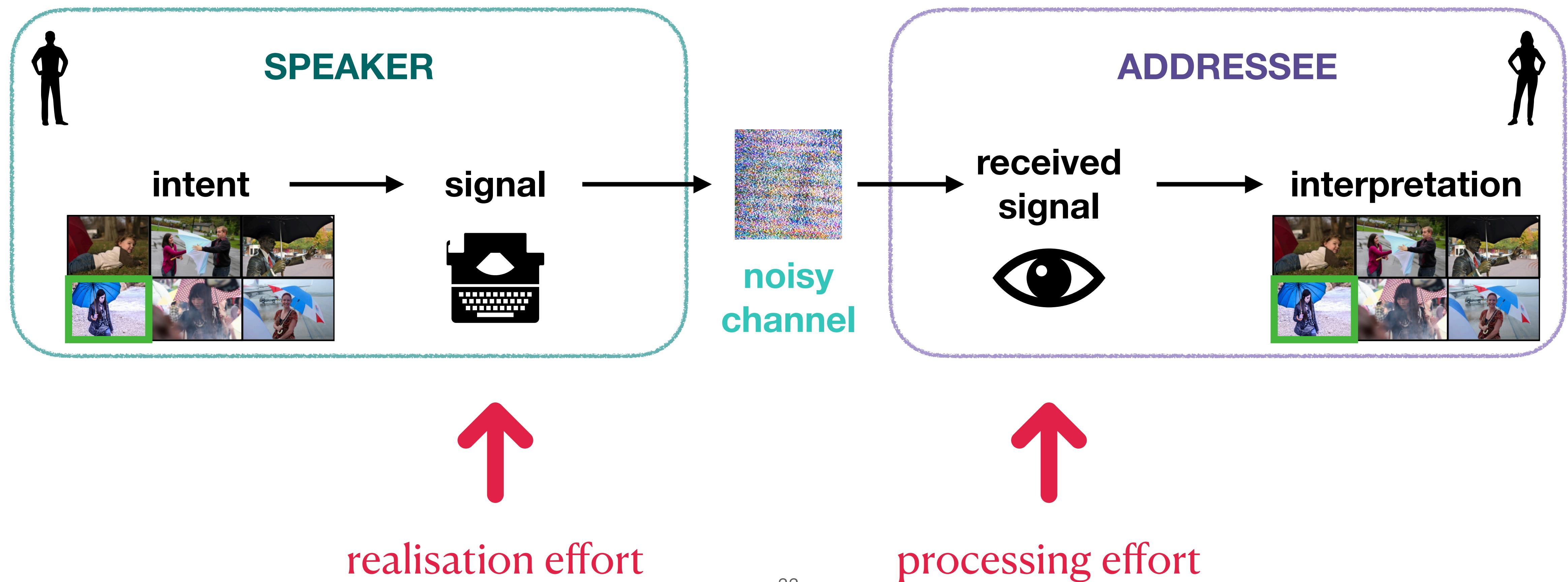
Claude Shannon, 1948

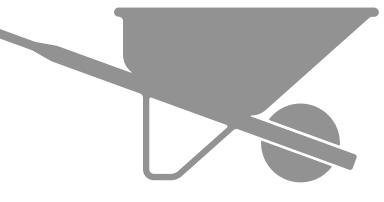




Noisy Channel Model of Communication

Claude Shannon, 1948





Computational Estimates of Processing Effort

via Shannon information content: $-\log P(X)$

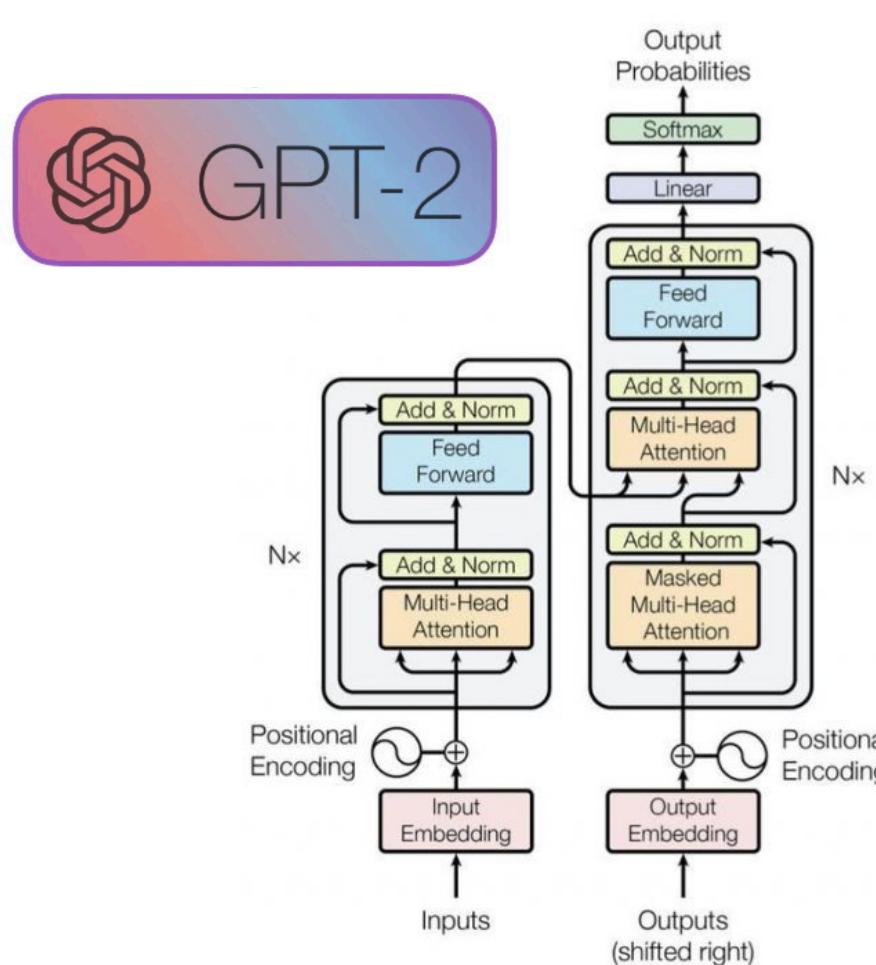
$$H(S) = -\log_2 P(S) = -\frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, \dots, w_{i-1})$$

utterance context (previous words)

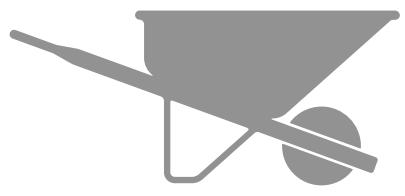


$$H(S | C) = -\log_2 P(S | C) = -\frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i | w_1, \dots, w_{i-1}, C)$$

conversational context

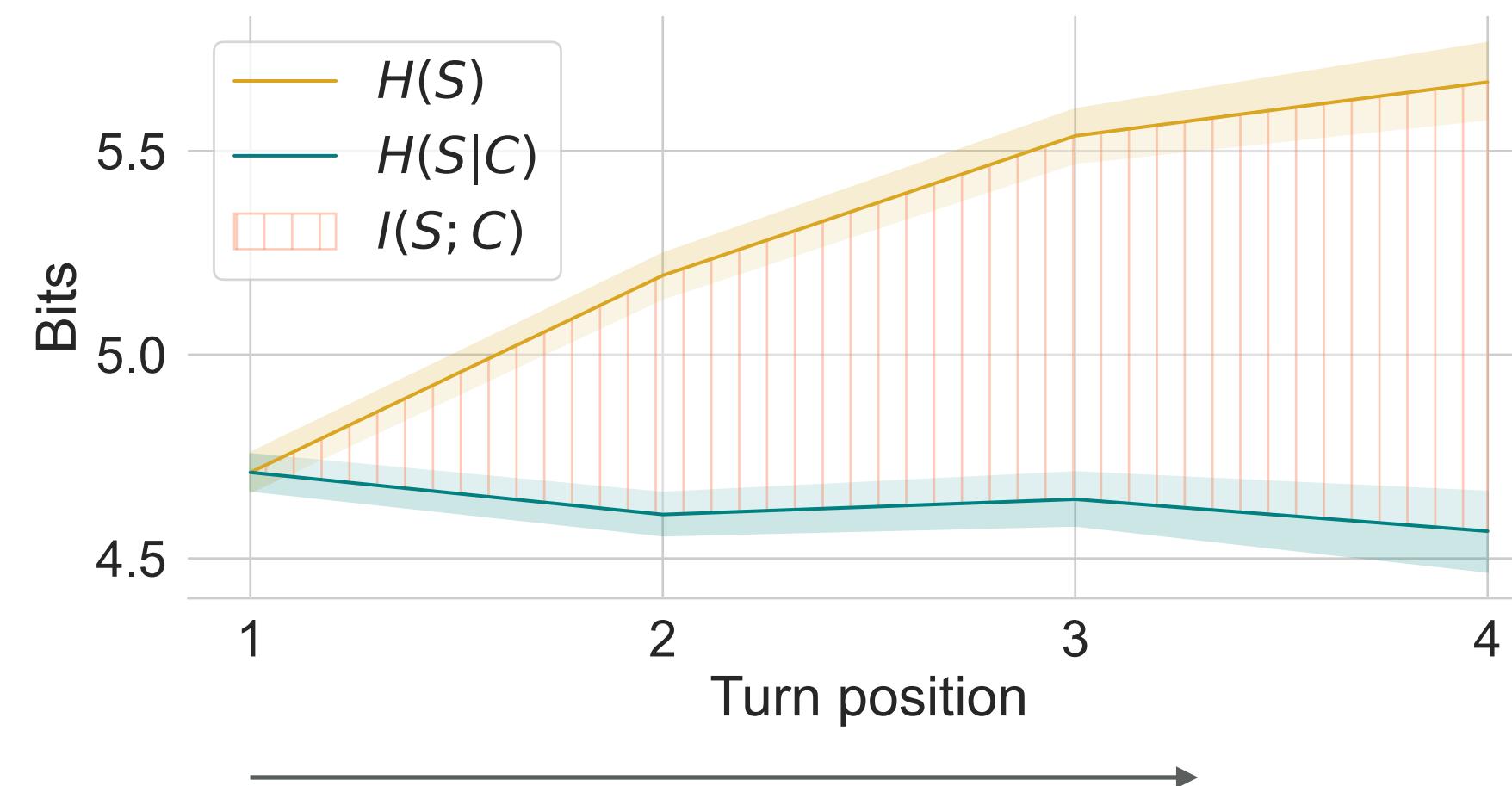


$P(w_i | \dots)$ estimates obtained with **GPT-2** (Radford et al., 2018),
a neural language model which we fine-tune on PhotoBook.



Results: PhotoBook Reference Chains

Speakers reduce collaborative effort

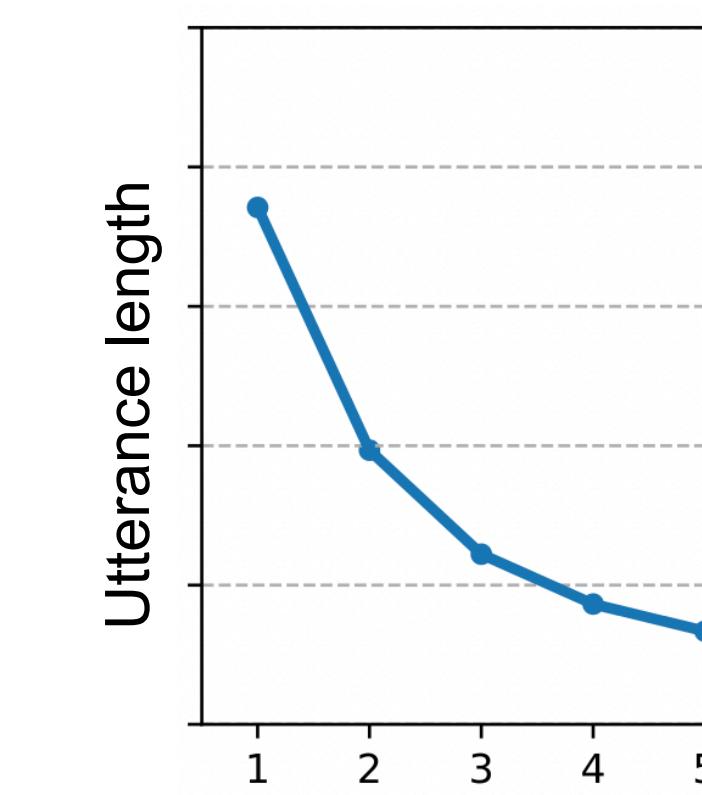


← (Information compression)

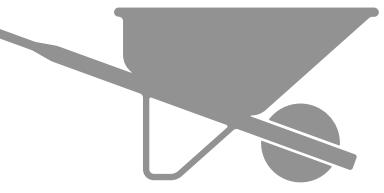
← Reduction of processing effort

$H(S)$ reference chain index i

$H(S|C)$

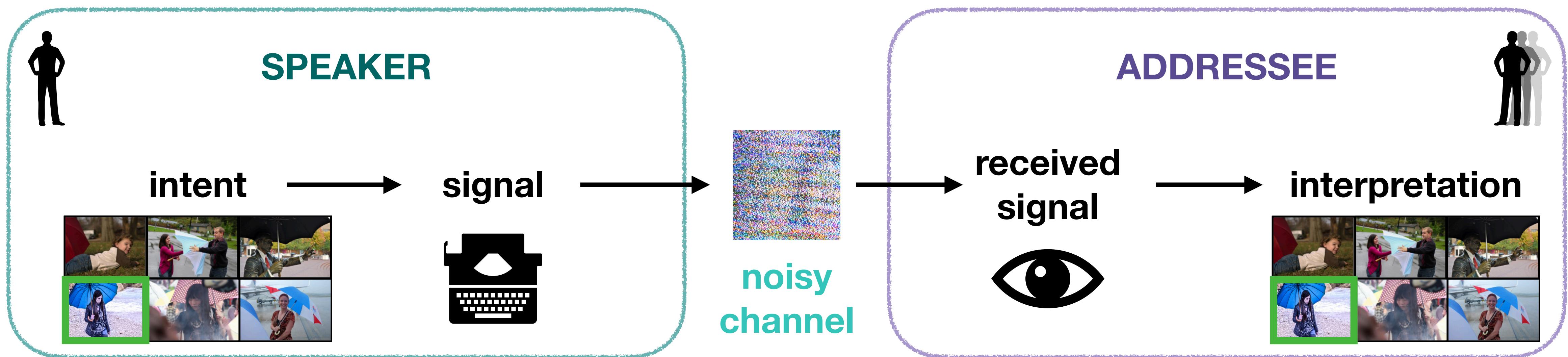


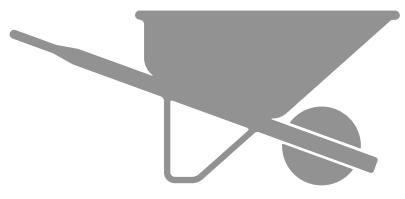
← Reduction of realisation effort



Noisy Channel Model of Communication

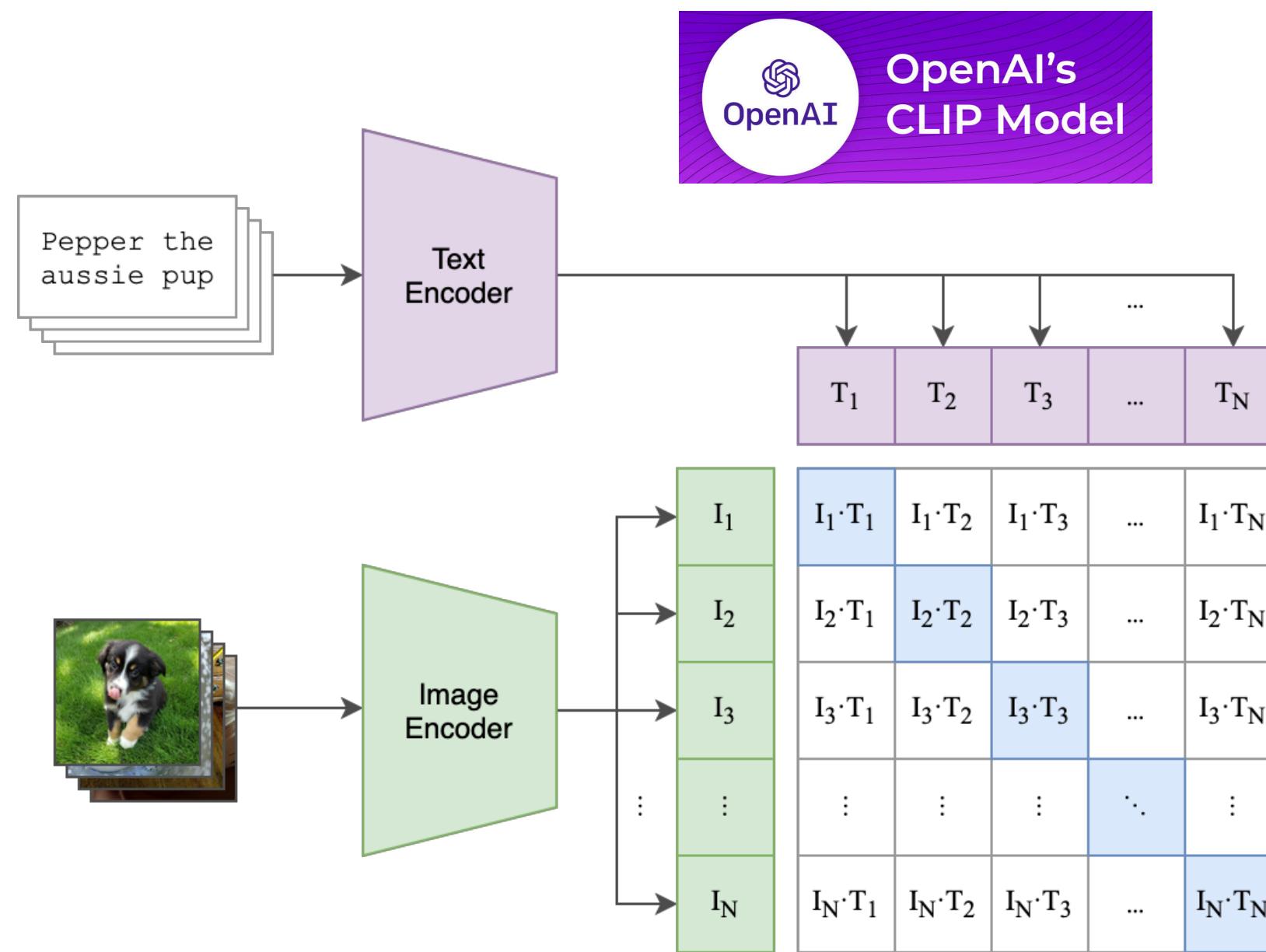
Claude Shannon, 1948





Computational Estimates of Resolution Effort

How well does the utterance describe the target image?



Descriptiveness: $\text{CLIPScore}(\text{image}, \text{utterance})$

High descriptiveness = predictable image-utterance matching



Estimates obtained with **CLIP** (Radford et al., 2021) a neural vision & language model
(*Contrastive Language-Image Pre-training via symmetric image-text matching loss*)

Utility



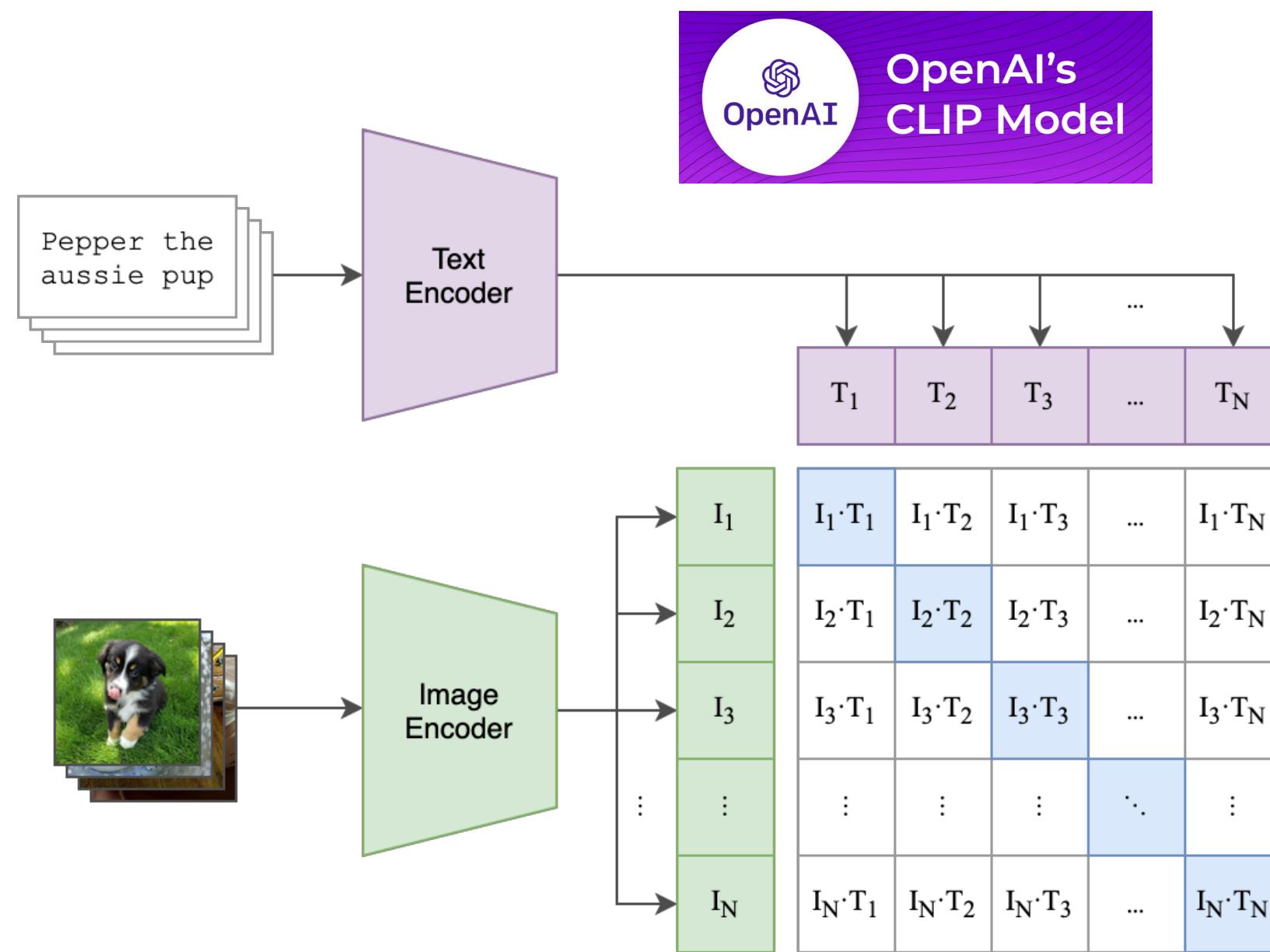
The cognitive, physical, and social effects of a communication act





Computational Estimates of Utility

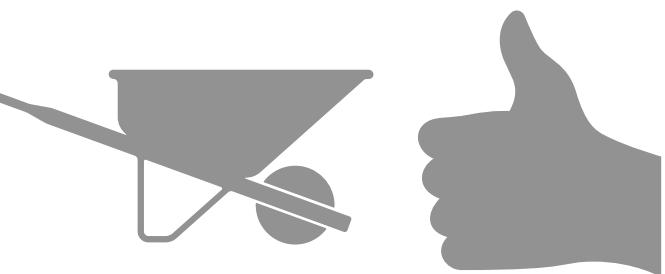
Positive utility: Is the communicative goal achieved?



Discriminativeness: task success (accuracy)

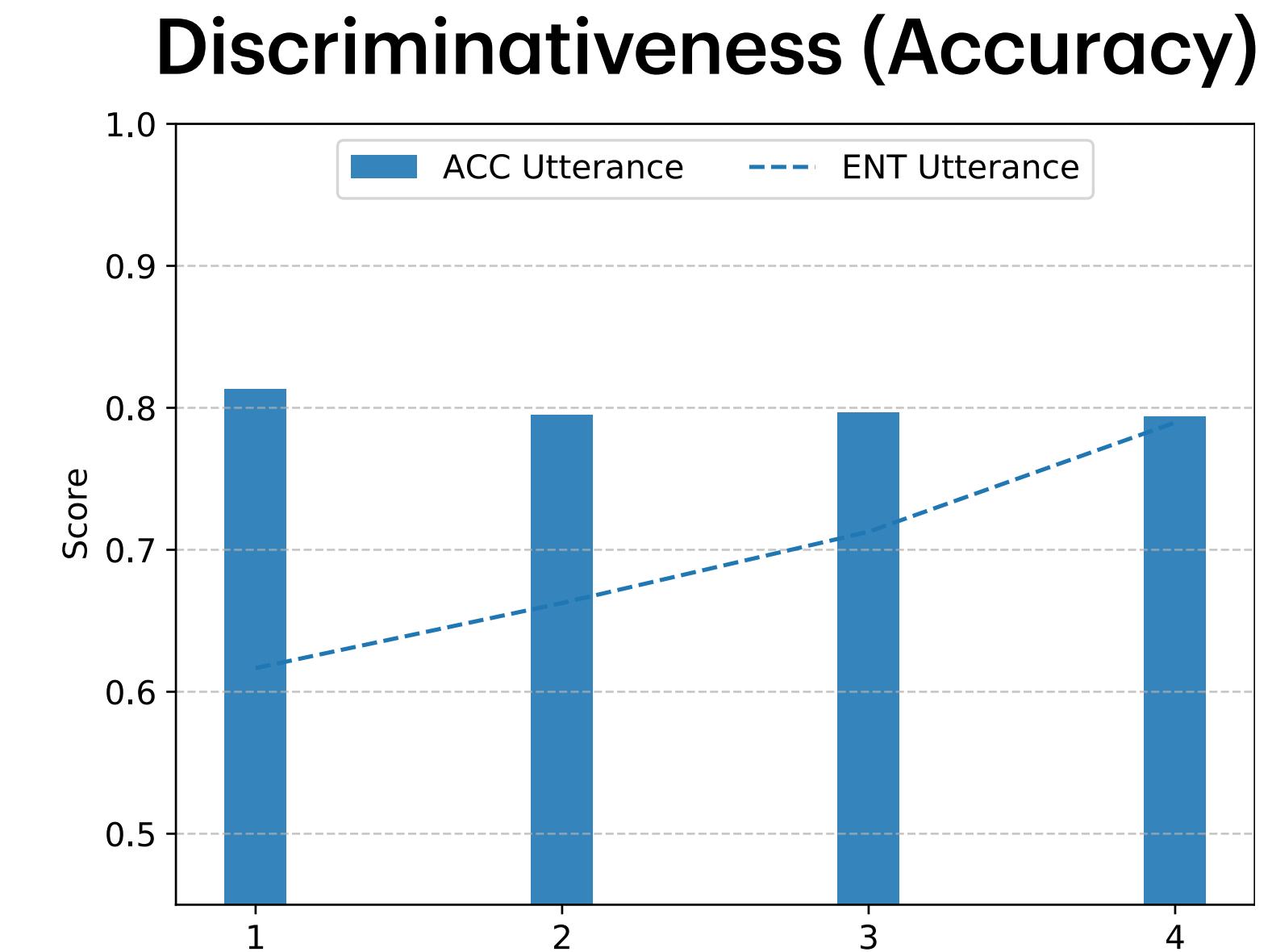
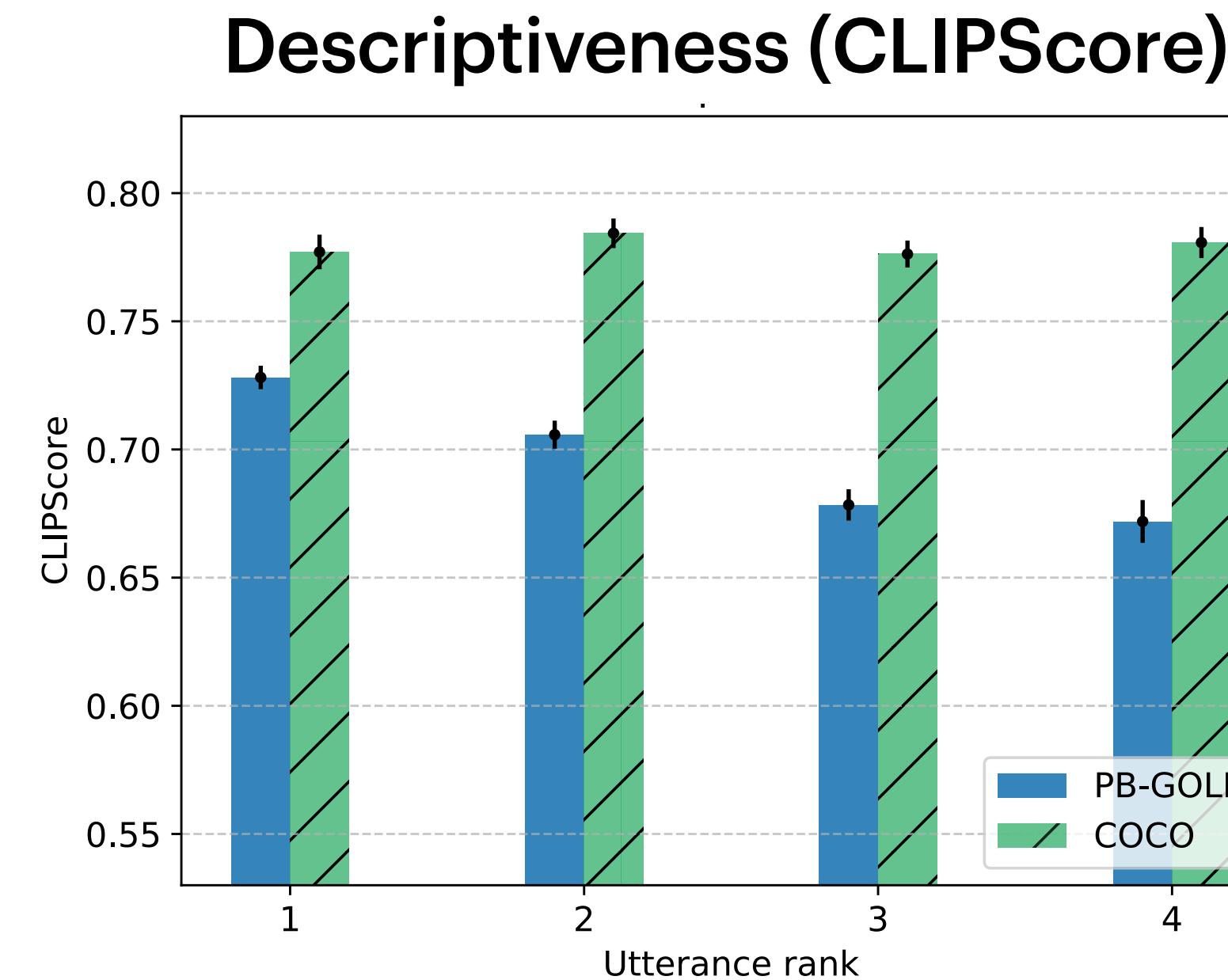
1 if target has the highest probability, otherwise 0

Estimates obtained with **CLIP** (Radford et al., 2021) a neural vision & language model
(*Contrastive Language-Image Pre-training via symmetric image-text matching loss*)



Results: PhotoBook Reference Chains

Speakers reduce collaborative effort while ensuring task success



Descriptiveness decreases over time yet discriminativeness is not significantly affected.

Grounded Dialogue Modelling

The study of interactive language use in context

- What is the relevant context of an interaction?
- How does the context relate to a speaker's communicative goals?
- What are the decision making strategies that humans follow to choose words and achieve goals in their environment?
- **Can we replicate them in a computer system?**

Grounded Dialogue Modelling

The study of interactive language use in context

- What is the relevant context?
- How does the context relate to the dialogue?
- What are the decision making steps to choose words and achieve the goal?
- **Can we replicate them?**

Refer, Reuse, Reduce Generating Subsequent References in Visual and Conversational Contexts

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, Raquel Fernández

Institute for Logic, Language and Computation

University of Amsterdam

{e.takmaz|m.giulianelli|s.pezzelle|
a.j.sinclair|raquel.fernandez}@uva.nl

Abstract

Dialogue participants often refer to entities or situations repeatedly within a conversation, which contributes to its cohesiveness. Subsequent references exploit the common ground accumulated by the interlocutors and hence have several interesting properties, namely, they tend to be shorter and reuse expressions that were effective in previous mentions. In this paper, we tackle the generation of first and subsequent references in visually grounded dialogue. We propose a generation model that produces referring utterances grounded in both the visual and the conversational context. To



Referring utterances extracted from dialogue 1

A: a white fuzzy dog with a wine glass up to his face
~> B: I see the wine glass dog
~> A: no I don't have the wine glass dog

Referring utterances extracted from dialogue 2

C: white dog sitting on something red
D: no I have the dog on the red chair

Dialogue Modelling Group

Research conducted in our research group at the ILLC with:



Ece
Takmaz



Sandro
Pezzelle



Arabella
Sinclair



Raquel
Fernández

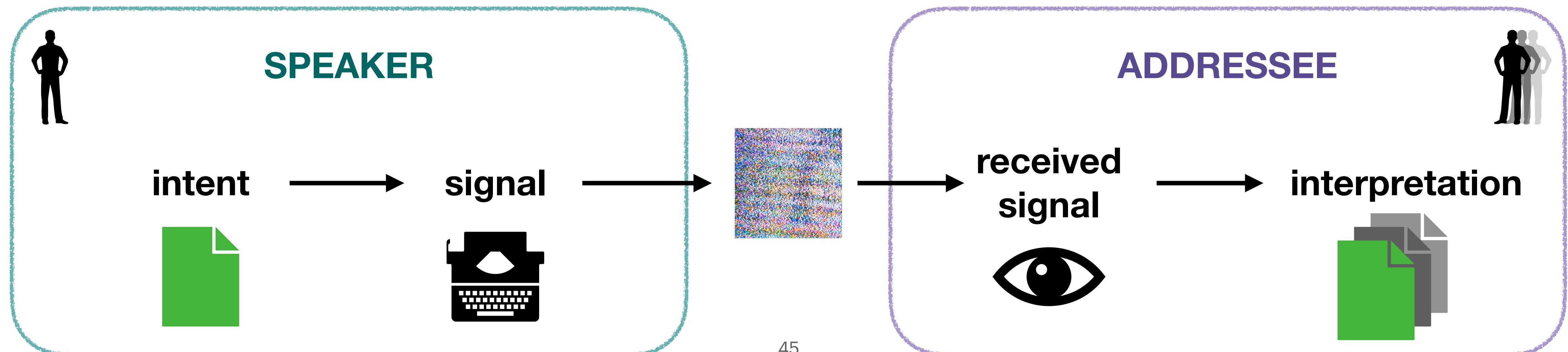
More details & results on other corpora:

- E. Takmaz, M. Julianelli, S. Pezzelle, A. J. Sinclair, R. Fernández. *Refer, Reuse, Reduce. Generating Subsequent References in Visual and Conversational Contexts*. EMNLP 2020.
- M. Julianelli, A. J. Sinclair, R. Fernández. *Is information density uniform in task-oriented dialogues?* EMNLP 2021.
- M. Julianelli & R. Fernández. *Analysing human strategies of information transmission as a function of discourse context*. CoNLL 2021.
- E. Takmaz, S. Pezzelle, R. Fernández. *Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP*. CMCL Workshop, ACL 2022.

Outlook: *IR to Model Human Behaviour?*

Outlook

- Computational modelling of human production strategies using pre-trained language & multimodal models.
- Reference resolution in visual contexts is essentially a retrieval task.
- This idea can be extended to other modes of productions: e.g. summarisation, translation, text simplification, ...



Thanks



For any questions: m.giulianelli@uva.nl