Common Ground & Audience Design in Referential Communication

Raquel Fernández

IWCS 2023 Nancy
International Conference on Computational Semantics

European Research Council

UNIVERSITY OF AMSTERDAM

Institute for Logic, Language & Computation

Image description

1. There are several people in chairs and a small child watching one of them play a trumpet
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

(example from the visual & linguistic treebank VLT2K dataset)

Image description

1. There are several people in chairs and a small child watching one of them play a trumpet
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

(example from the visual & linguistic treebank VLT2K dataset)

In image description, several constraints play a role:

Image description

1. There are several people in chairs and a small child watching one of them play a trumpet
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

(example from the visual & linguistic treebank VLT2K dataset)

In image description, several constraints play a role:

- General cooperative principles related to truth and informativeness

Image description

1. There are several people in chairs and a small child watching one of them play a trumpet
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

(example from the visual & linguistic treebank VLT2K dataset)

In image description, several constraints play a role:

- General cooperative principles related to truth and informativeness
- Individual constraints: e.g., lexical availability and visual saliency

Image description

1. There are several people in chairs and a small child watching one of them play a trumpet
2. A man is playing a trumpet in front of a little boy.
3. People sitting on a sofa with a man playing an instrument for entertainment.

(example from the visual & linguistic treebank VLT2K dataset)

In image description, several constraints play a role:

- General cooperative principles related to truth and informativeness
- Individual constraints: e.g., lexical availability and visual saliency
- Social interaction constraints: common ground with dialogue partner

Re-referring in dialogue

- In dialogue, we often refer to the same entities more than once
- Besides the constrains mentioned above, subsequent mentions rely on the common ground established with our dialogue partner

Re-referring in dialogue

- In dialogue, we often refer to the same entities more than once
- Besides the constrains mentioned above, subsequent mentions rely on the common ground established with our dialogue partner

A:	a white fuzzy dog with a wine glass
B:	I see **the wine glass dog**
A:	no I don't have **the wine glass dog**

Re-referring in dialogue

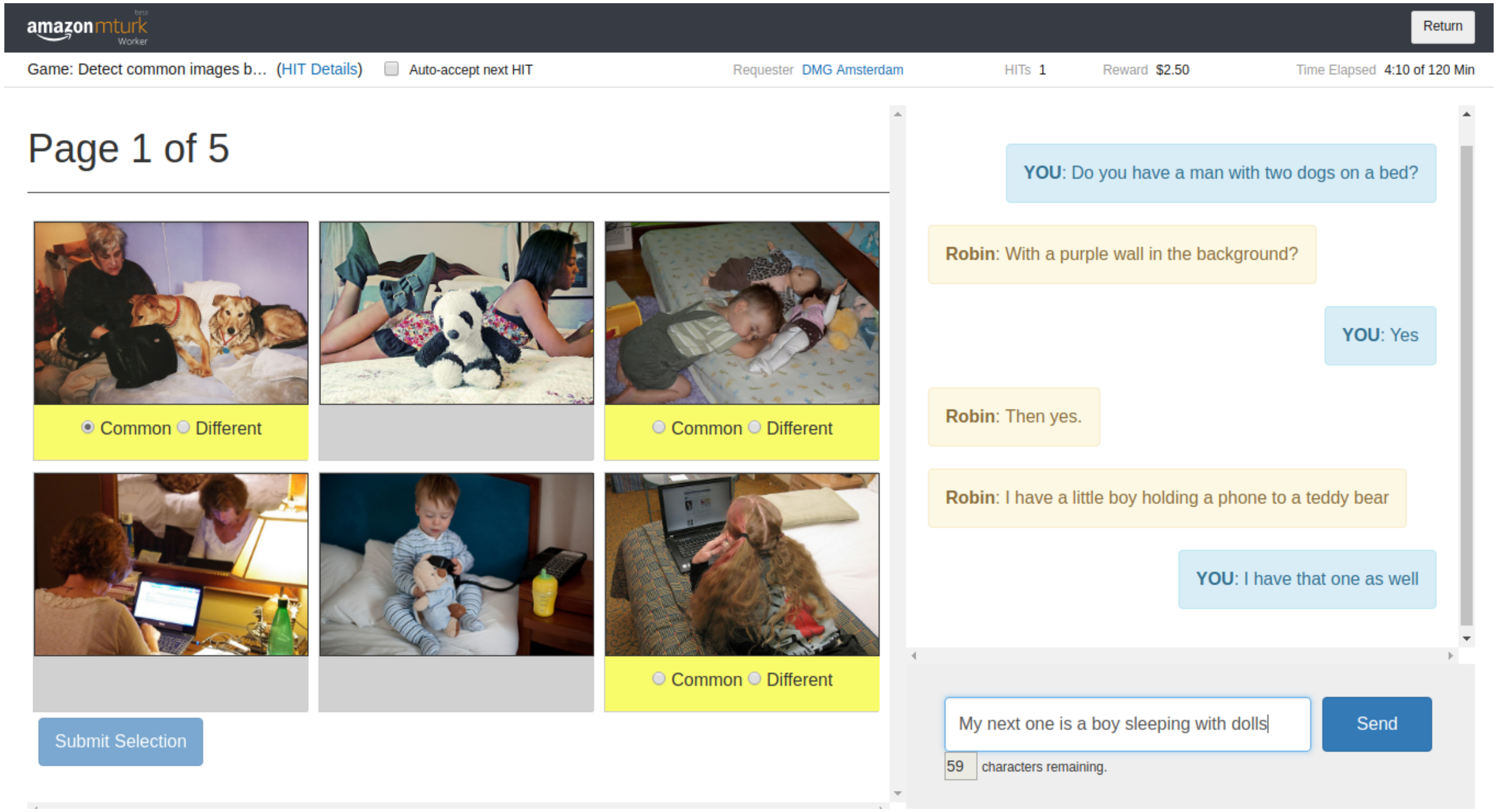
- In dialogue, we often refer to the same entities more than once
- Besides the constrains mentioned above, subsequent mentions rely on the common ground established with our dialogue partner

A:	a white fuzzy dog with a wine glass
B:	I see **the wine glass dog**
A:	no I don’t have **the wine glass dog**

C:	white dog sitting on something red
D:	yes, I have the **dog on the red chair**
C:	white **dog on the red chair**

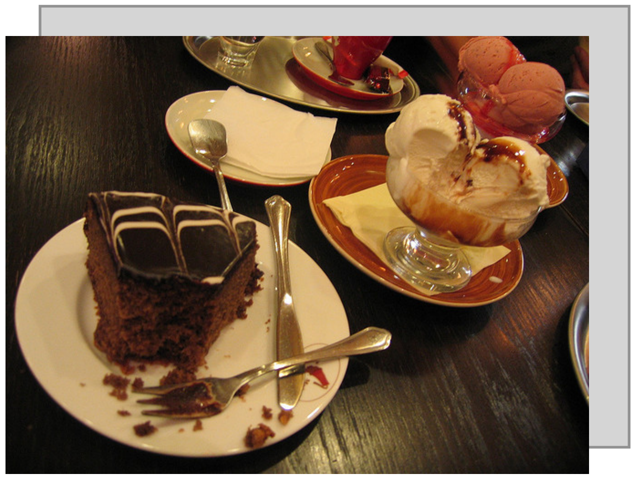
The PhotoBook dataset

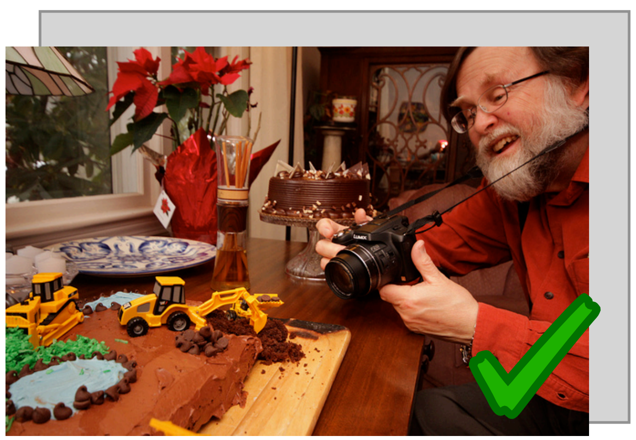
- Two participants see six photos each and need to find out which of three highlighted photos they have in common
- They can chat freely, without predefined roles
- The game consists of five rounds, with the set of images changing at every round and some images reappearing

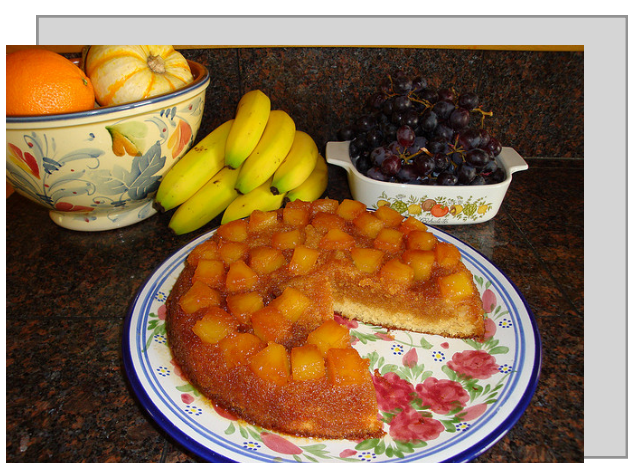


The PhotoBook dataset

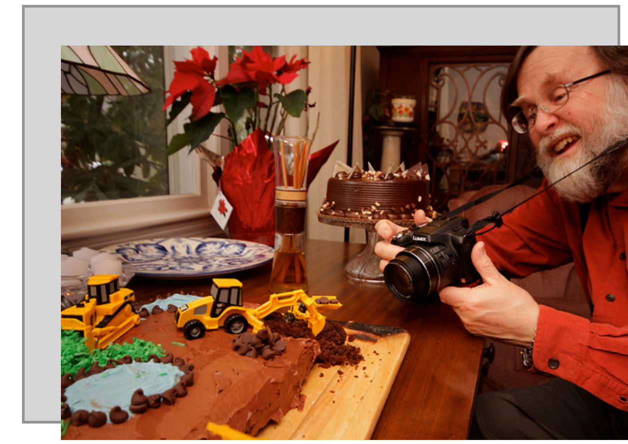
Participant A

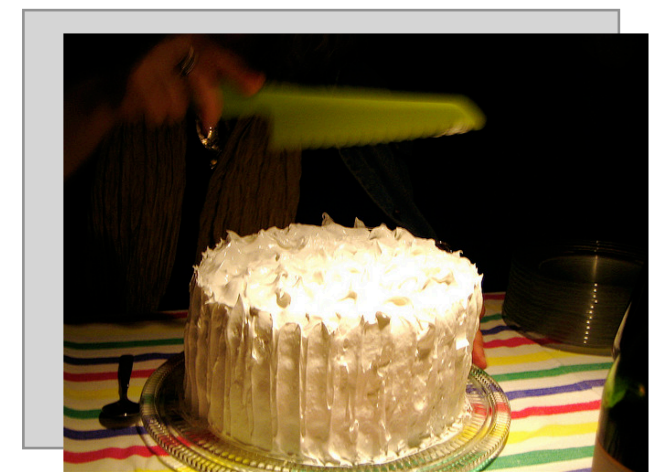


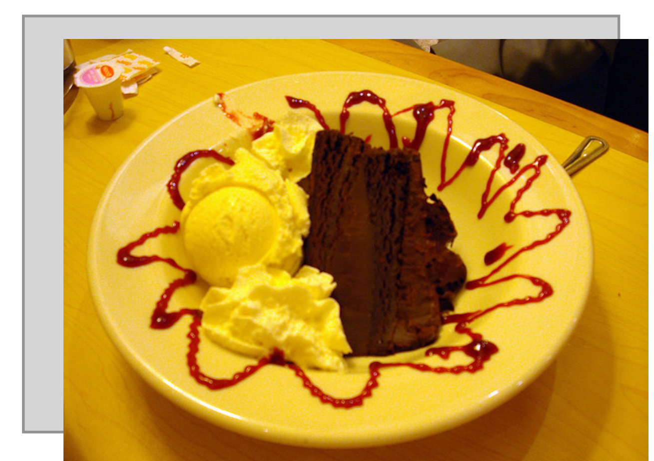




Participant B







Round 1 of 5

A: Hi

B: Hello

B: do you have a white cake on multi colored striped cloth?

A: **I see a guy taking a picture. What about you?**

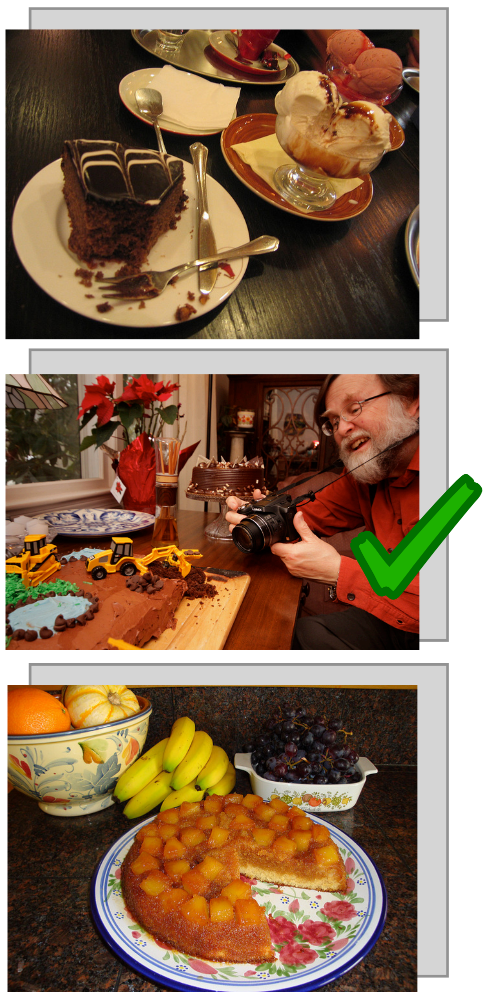
B: is it of a cake with construction trucks on it?

A: Yeah. I don't see the cake you mentioned.

A: `<common img_2>`

The PhotoBook dataset

Round 1
Participant A



A: **I see a guy taking a picture. What about you?**

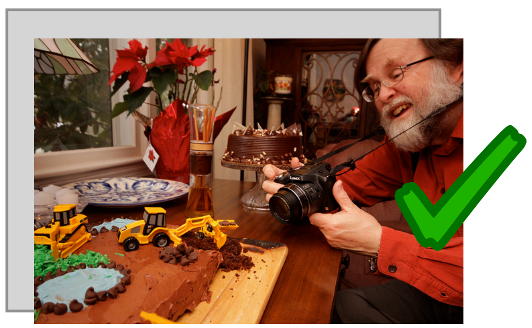
B: guy with camera

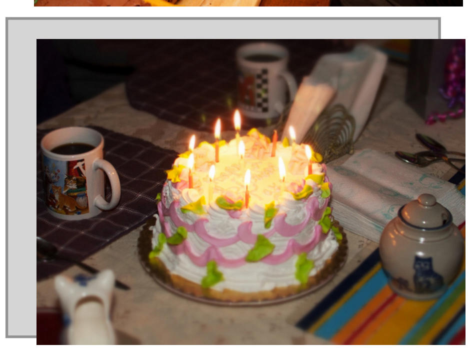
A: I have the guy with camera

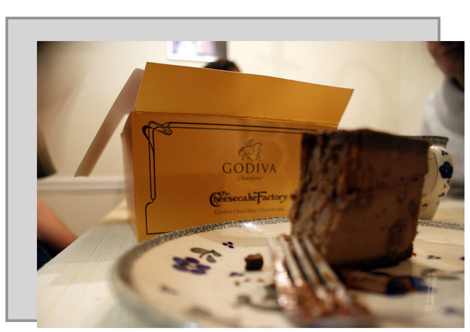
A: the last one is the camera guy

The PhotoBook dataset

Round 2
Participant B







A: I see a guy taking a picture. What about you?

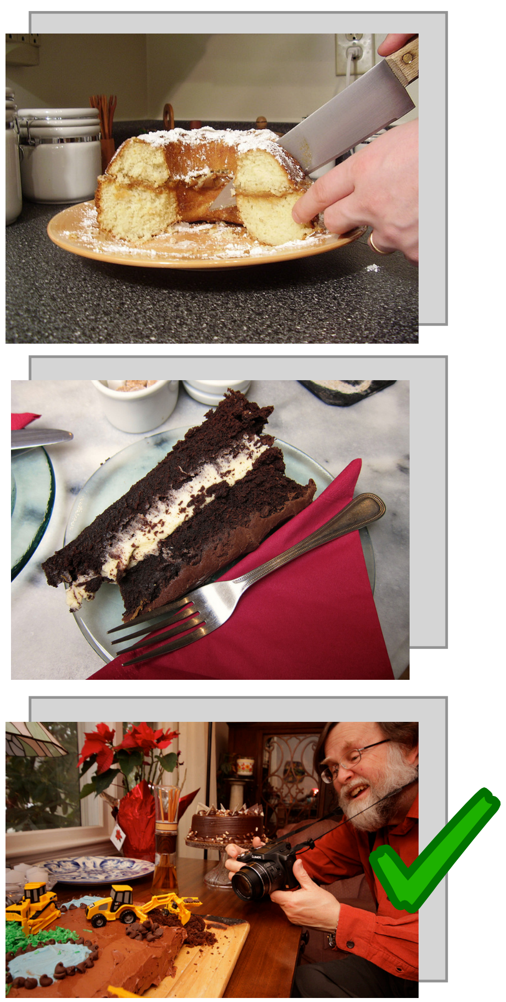
B: guy with camera

A: I have the guy with camera

A: the last one is the camera guy

The PhotoBook dataset

Round 3
Participant A



A: I see a guy taking a picture. What about you?

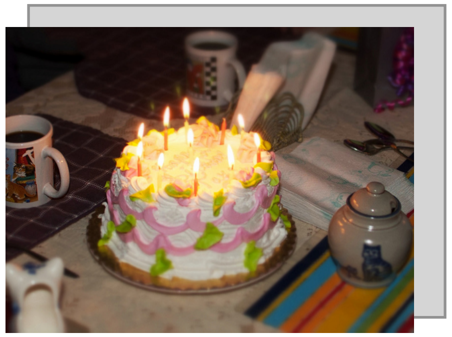
B: guy with camera

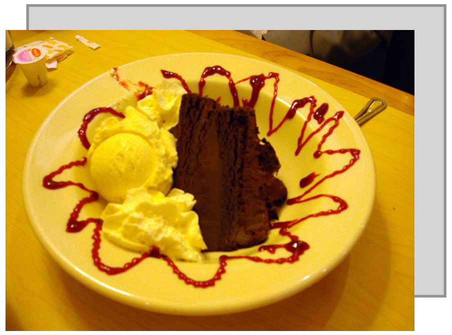
A: I have the guy with camera

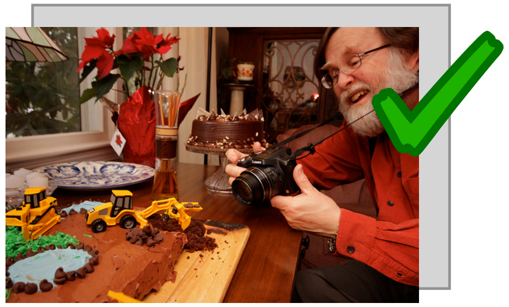
A: the last one is the camera guy

The PhotoBook dataset

Round 5
Participant A







A: I see a guy taking a picture. What about you?

B: guy with camera

A: I have the guy with camera

A: the last one is the camera guy

The PhotoBook dataset

Co-referring chain:
utterances referring to the same target image over a game

A: I see a guy taking a picture. What about you?

B: guy with camera

A: I have the guy with camera

A: the last one is the camera guy

The PhotoBook dataset

Co-referring chain:
utterances referring to the same target image over a game

1. *girl on end of bed with computer, she has pigtails*
2. *Girl with pigtails?*
3. *Pigtail girl?*
4. *Pigtails? lol*

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?*

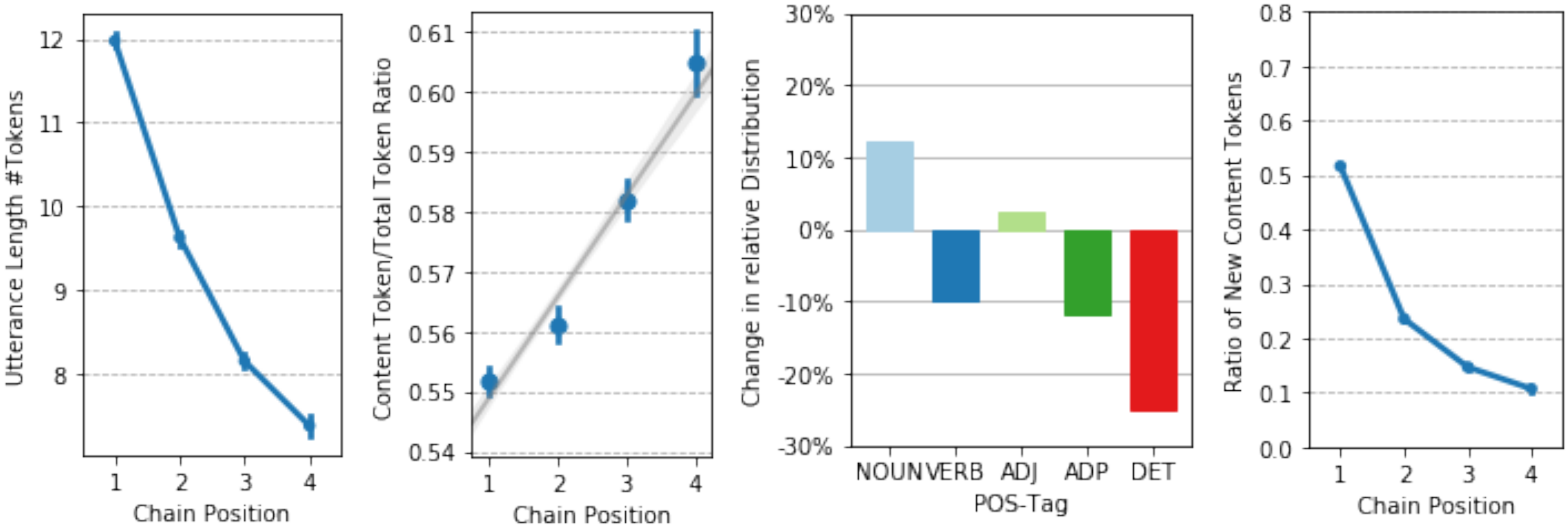
https://dmg-photobook.github.io

2,500 dialogues, 16,525 co-referring chains

Patterns observed in the data

They replicate of previous findings

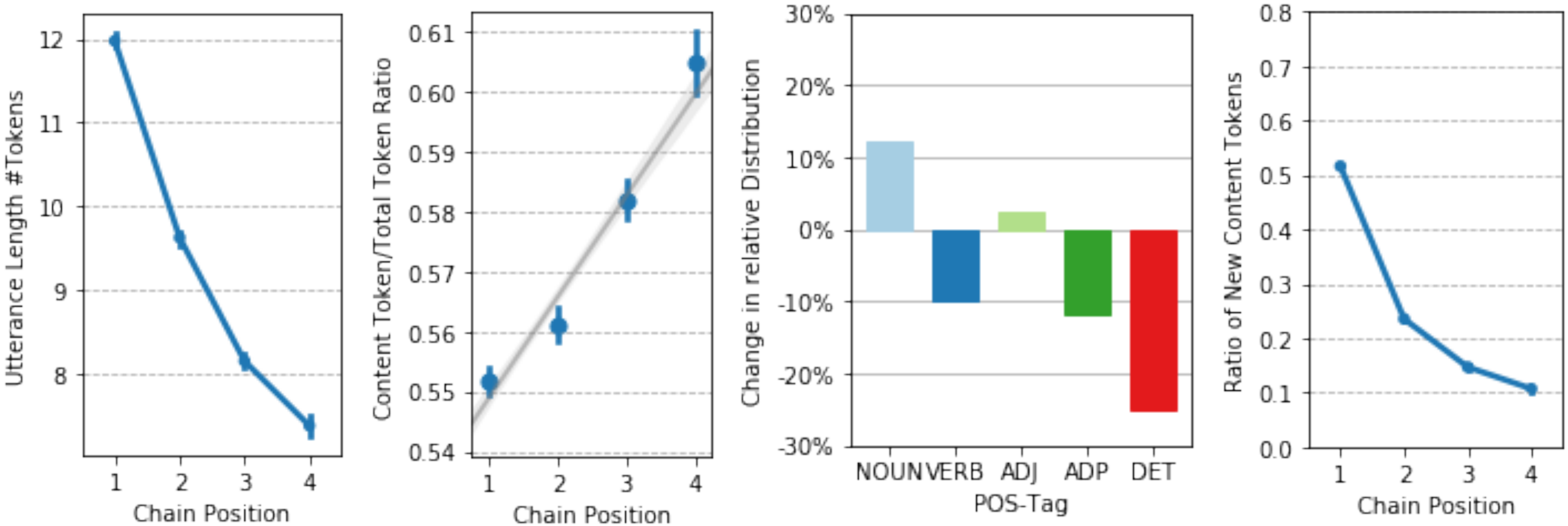
(Krauss&Weinheimer 1964, Clark&Wilkes-Gibbs 1986, Garrod&Anderson 1987, Clark&Brennan 1991)



Patterns observed in the data

They replicate of previous findings

(Krauss&Weinheimer 1964, Clark&Wilkes-Gibbs 1986, Garrod&Anderson 1987, Clark&Brennan 1991)

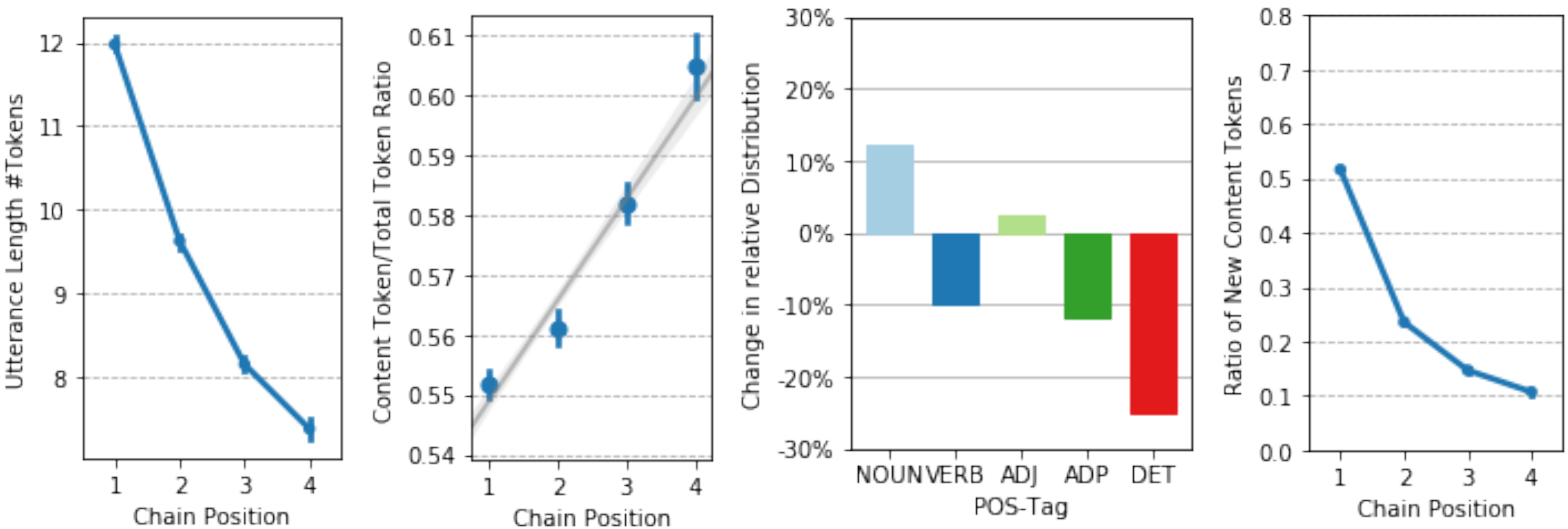


- Referring utterances become shorter.

Patterns observed in the data

They replicate of previous findings

(Krauss&Weinheimer 1964, Clark&Wilkes-Gibbs 1986, Garrod&Anderson 1987, Clark&Brennan 1991)

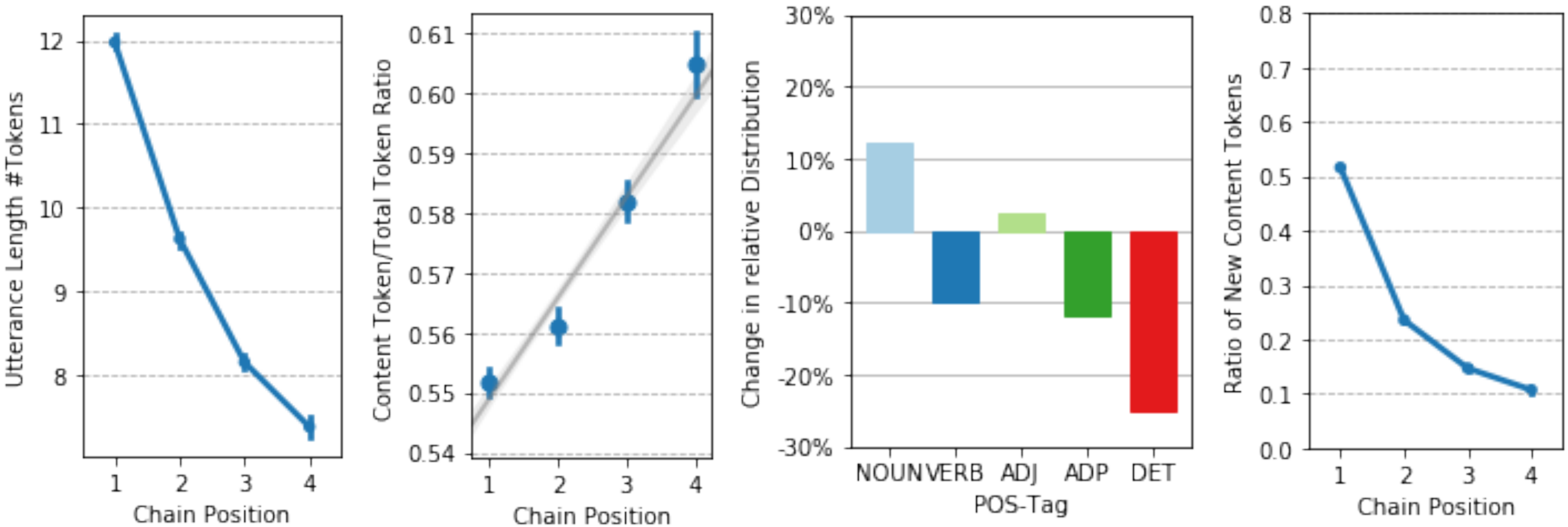


- Referring utterances become shorter.
- Increase of content words ratio: more likely to remain.

Patterns observed in the data

They replicate of previous findings

(Krauss&Weinheimer 1964, Clark&Wilkes-Gibbs 1986, Garrod&Anderson 1987, Clark&Brennan 1991)

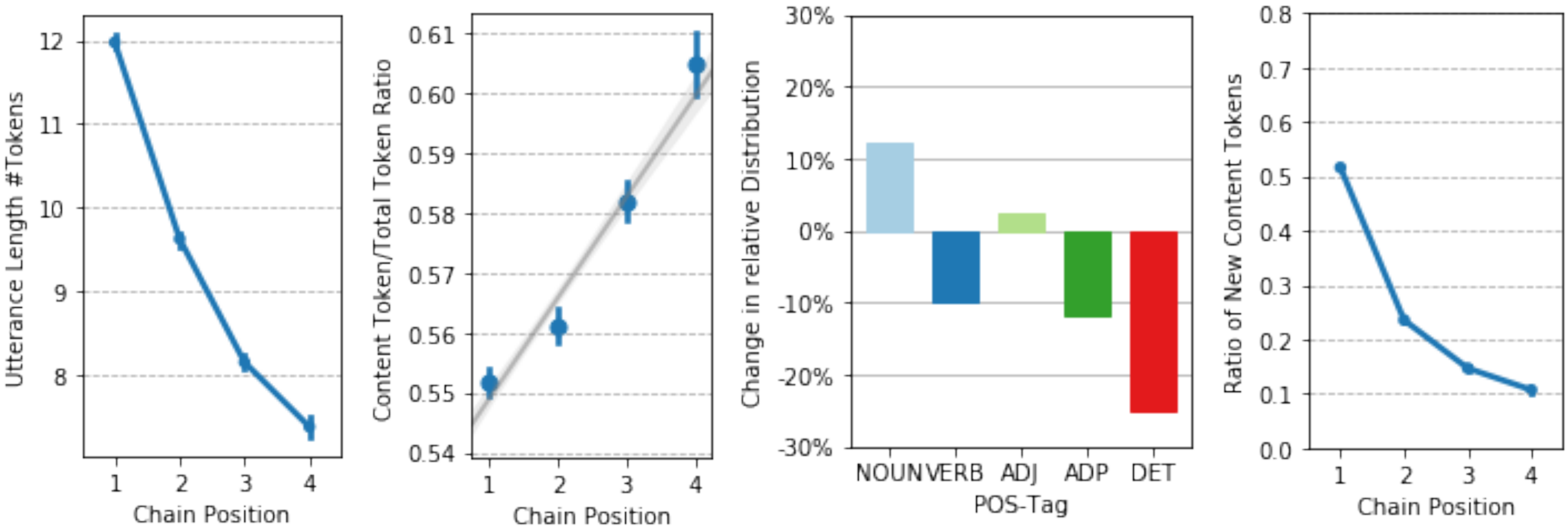


- Referring utterances become shorter.
- Increase of content words ratio: more likely to remain.
- POS distribution: proportion of nouns and adjectives increases.

Patterns observed in the data

They replicate of previous findings

(Krauss&Weinheimer 1964, Clark&Wilkes-Gibbs 1986, Garrod&Anderson 1987, Clark&Brennan 1991)

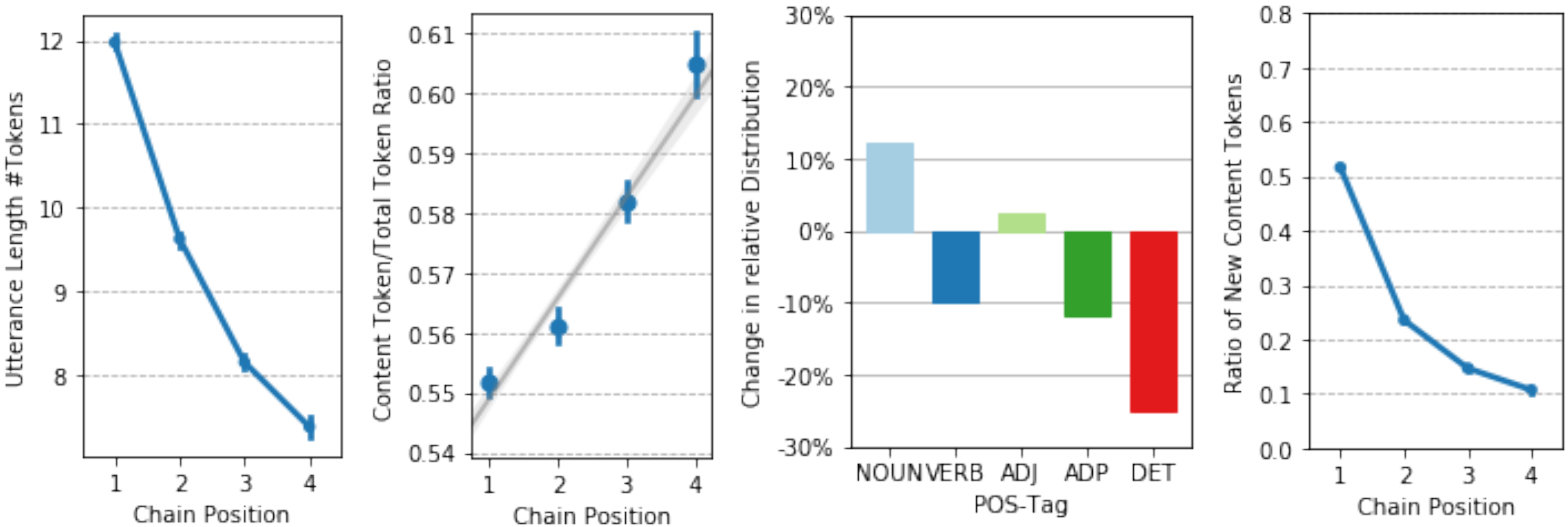


- Referring utterances become shorter.
- Increase of content words ratio: more likely to remain.
- POS distribution: proportion of nouns and adjectives increases.
- Sharp decrease of new content words: lexical entrainment.

Patterns observed in the data

They replicate of previous findings

(Krauss&Weinheimer 1964, Clark&Wilkes-Gibbs 1986, Garrod&Anderson 1987, Clark&Brennan 1991)

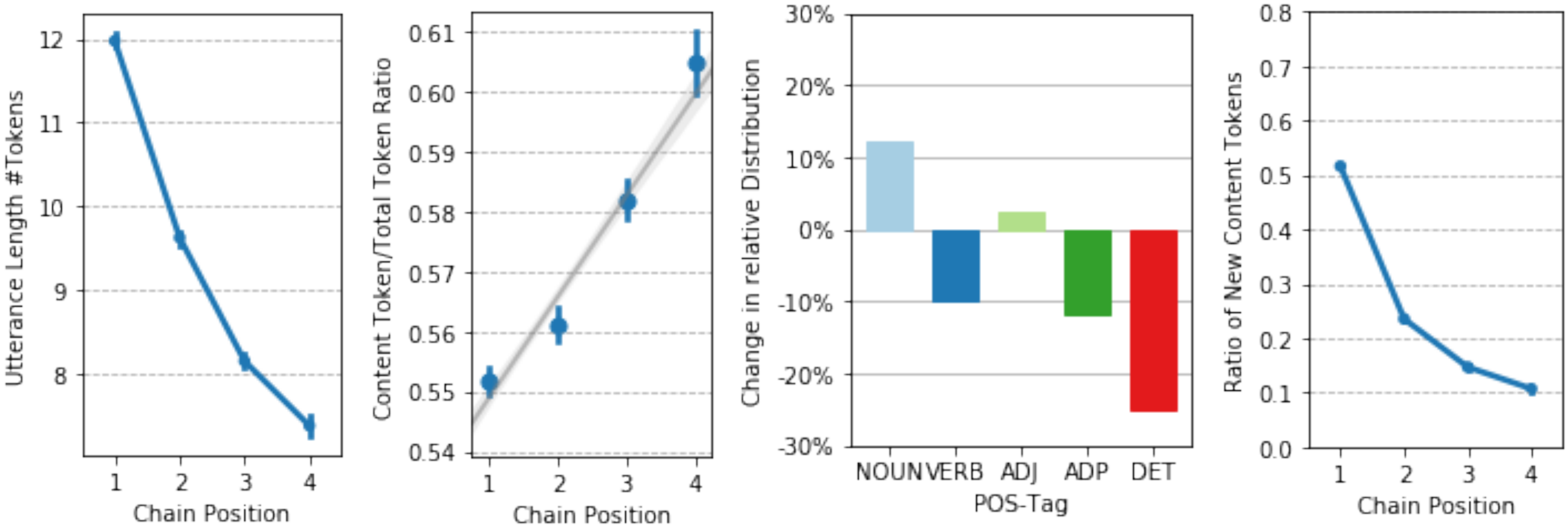


- **What kind of mechanisms would support these patters?**

Patterns observed in the data

They replicate of previous findings

(Krauss&Weinheimer 1964, Clark&Wilkes-Gibbs 1986, Garrod&Anderson 1987, Clark&Brennan 1991)



- **What kind of mechanisms would support these patters?**

Comprehension	**Production**
reference resolution	referring utterance generation

Context dependent resolution

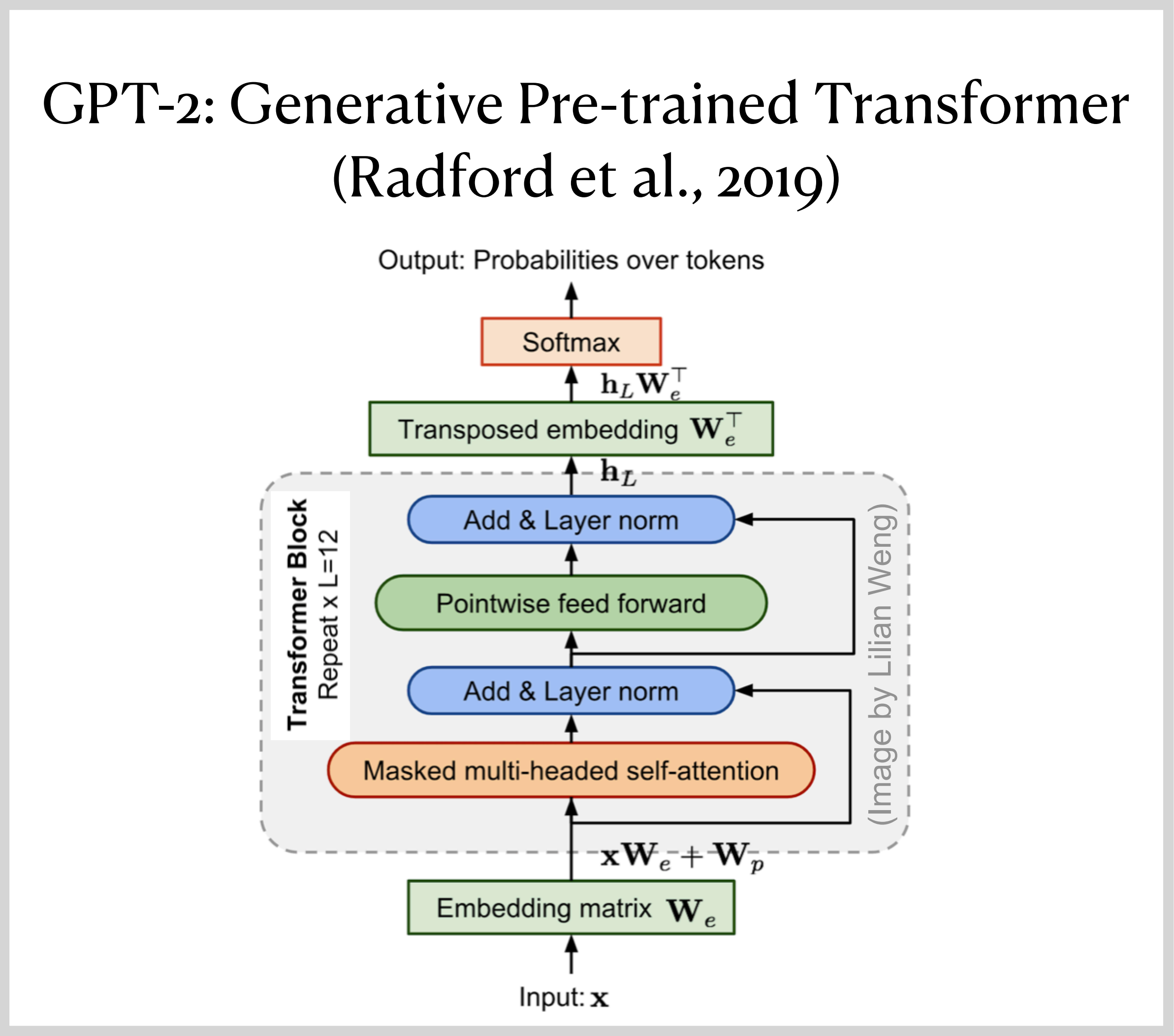
Quantifying effort

- If later references rely on the conversational common ground, they should be more surprising and difficult to **resolve** out of context

Context dependent resolution

Quantifying effort

- If later references rely on the conversational common ground, they should be more surprising and difficult to **resolve** out of context

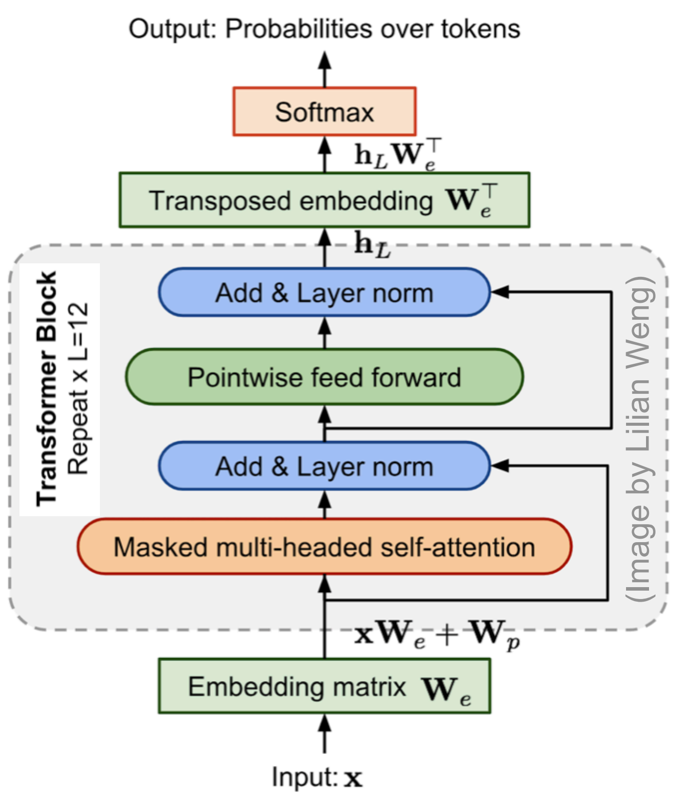


Context dependent resolution

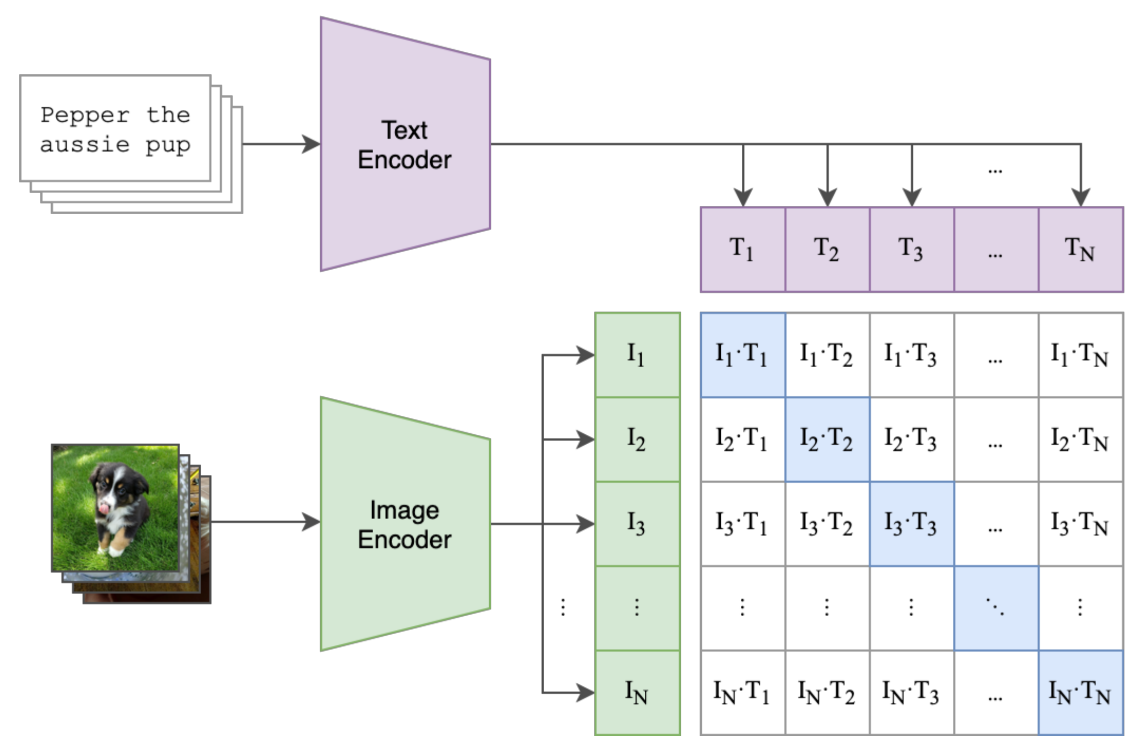
Quantifying effort

- If later references rely on the conversational common ground, they should be more surprising and difficult to **resolve** out of context

GPT-2: Generative Pre-trained Transformer (Radford et al., 2019)



CLIP: Contrastive Language-Image Pre-training via symmetric image-text matching loss (Radford et al., 2021)



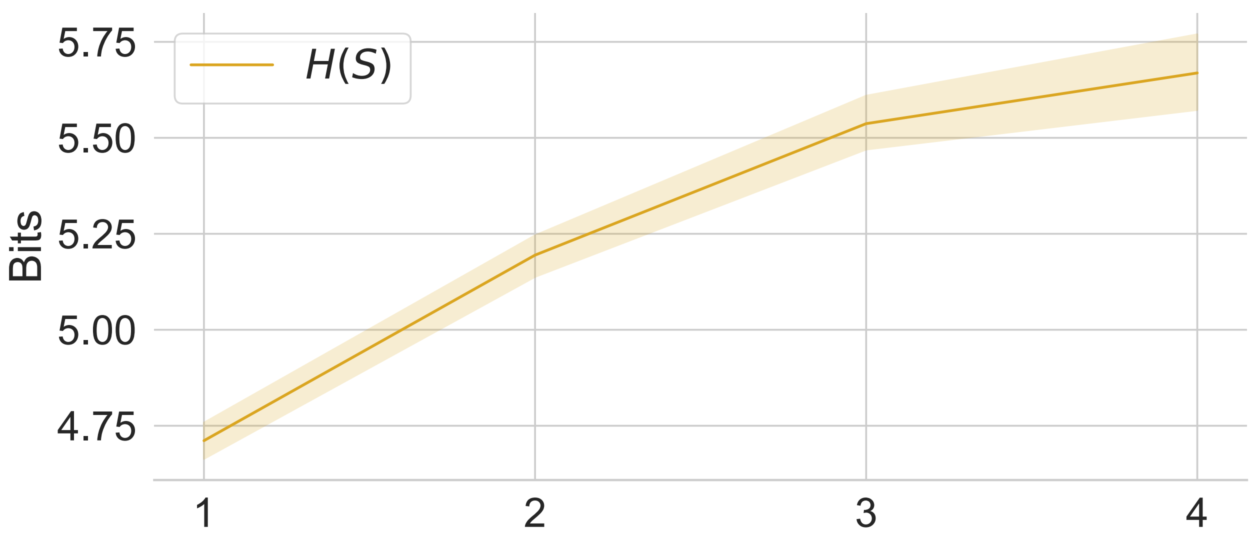
Language

(Giulianelli et al. 2021)

$P(w_i \mid \ldots)$ estimates obtained with GPT-2 fine-tuned on PhotoBook

- Out-of-context surprisal *H(S)*

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



Language
(Giulianelli et al. 2021)

$P(w_i \mid \ldots)$ estimates obtained with GPT-2 fine-tuned on PhotoBook

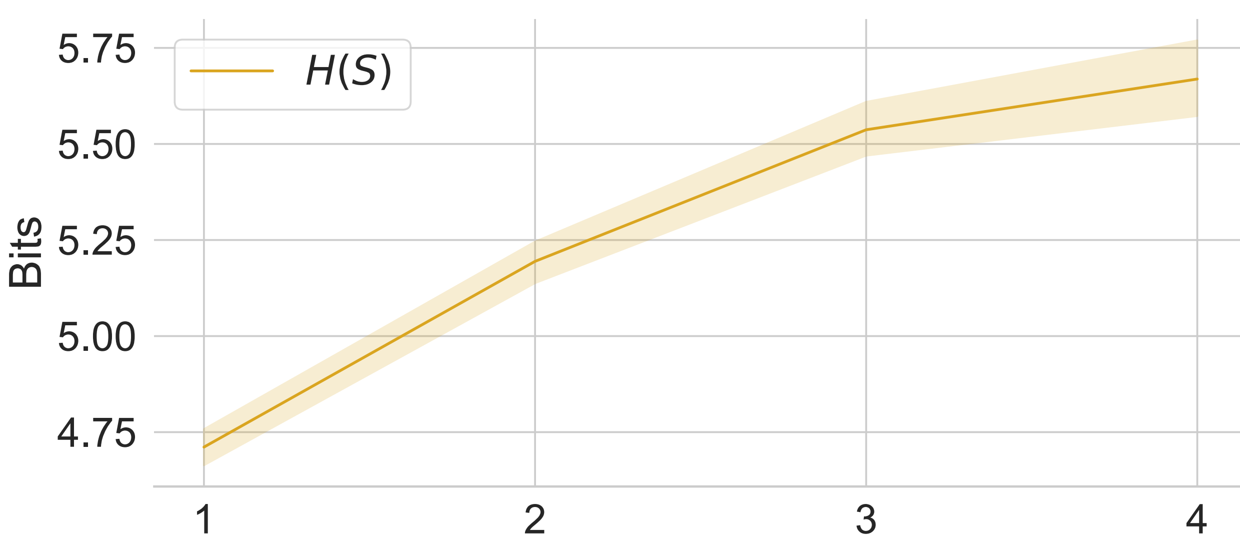
- Out-of-context surprisal *H(S)*

Language & vision
(Takmaz et al. 2022)

Given a referring utterance and the images in the context, CLIP yields softmax probabilities

- Accuracy with highest probability image
- Entropy of the distribution

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



Language
(Giulianelli et al. 2021)

$P(w_i \,|\, \ldots)$ estimates obtained with GPT-2 fine-tuned on PhotoBook

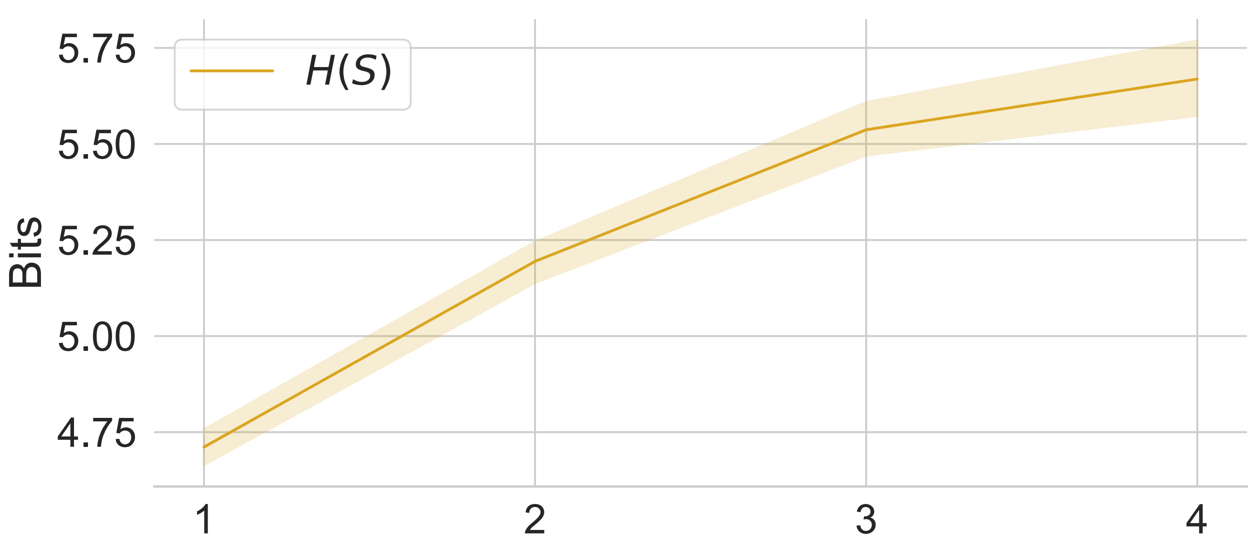
- Out-of-context surprisal *H(S)*

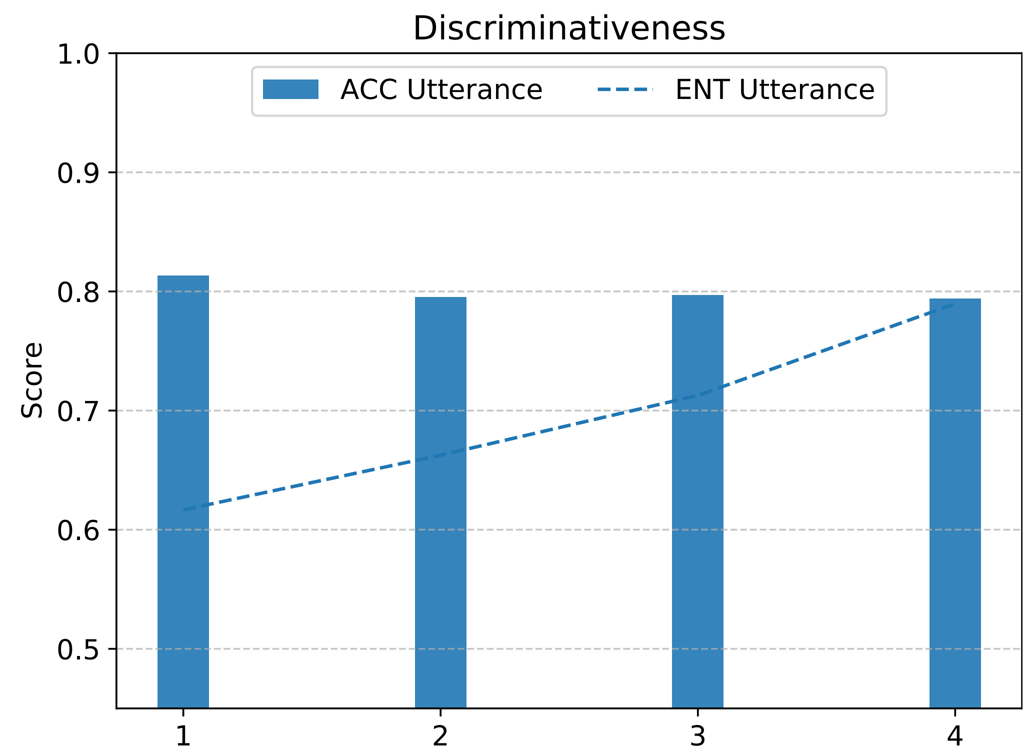
Language & vision
(Takmaz et al. 2022)

Given a referring utterance and the images in the context, CLIP yields softmax probabilities

- Accuracy with highest probability image
- Entropy of the distribution

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



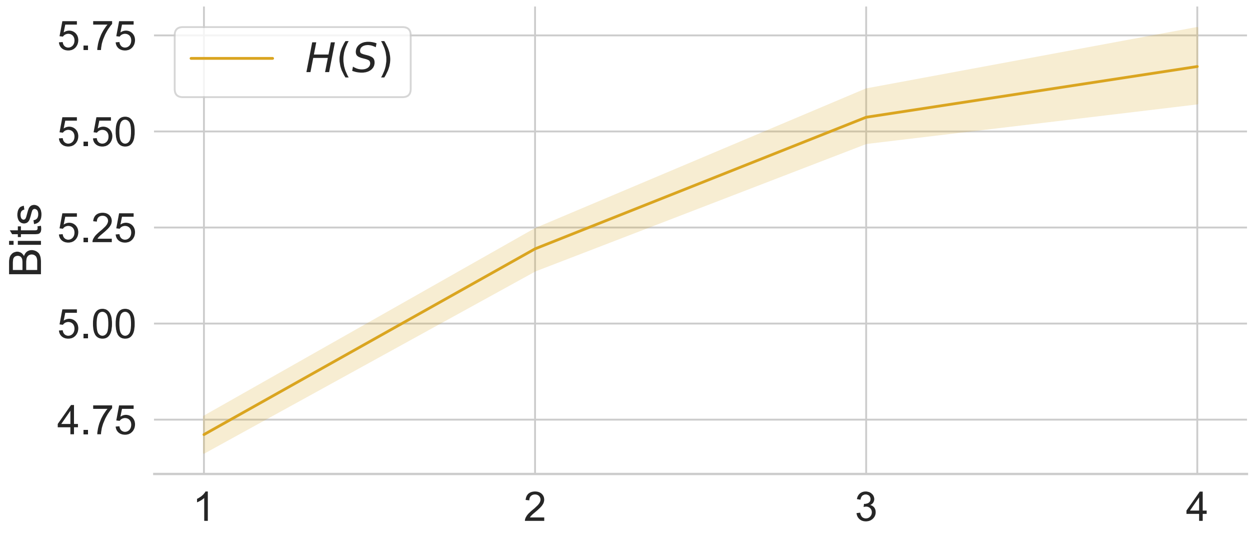


Language
(Giulianelli et al. 2021)

$P(w_i \,|\, \ldots)$ estimates obtained with GPT-2 fine-tuned on PhotoBook

- Out-of-context surprisal *H(S)*

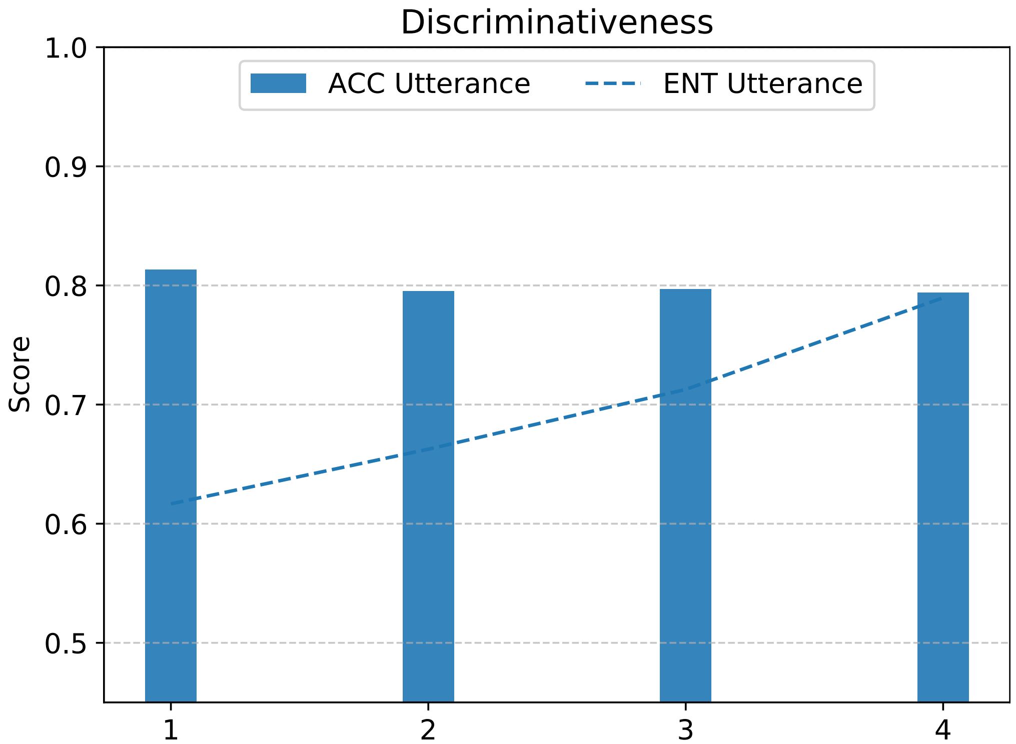
$$H(S) = -\log_2 P(S) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i \,|\, w_1, \ldots, w_{i-1})$$



Language & vision
(Takmaz et al. 2022)

Given a referring utterance and the images in the context, CLIP yields softmax probabilities

- Accuracy with highest probability image
- Entropy of the distribution



Higher surprisal and resolution uncertainty in later references without conversational context

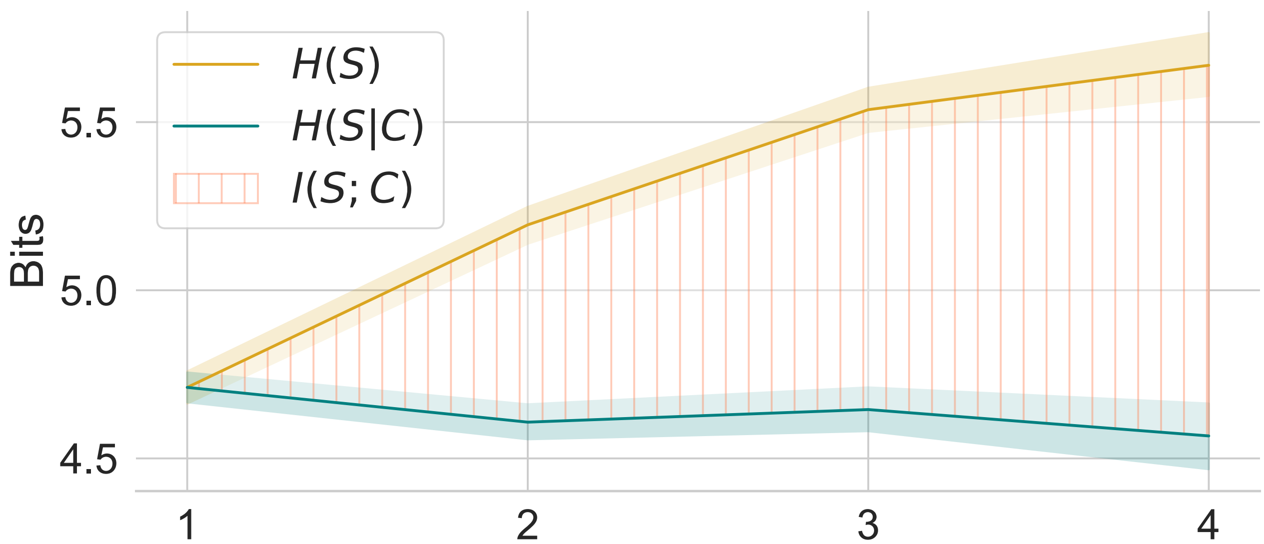
Language

(Giulianelli et al. 2021)

$P(w_i \mid \ldots)$ estimates obtained with GPT-2 fine-tuned on PhotoBook

- Out-of-context surprisal *H(S)*
- In-context surprisal *H(S|C)*

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



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

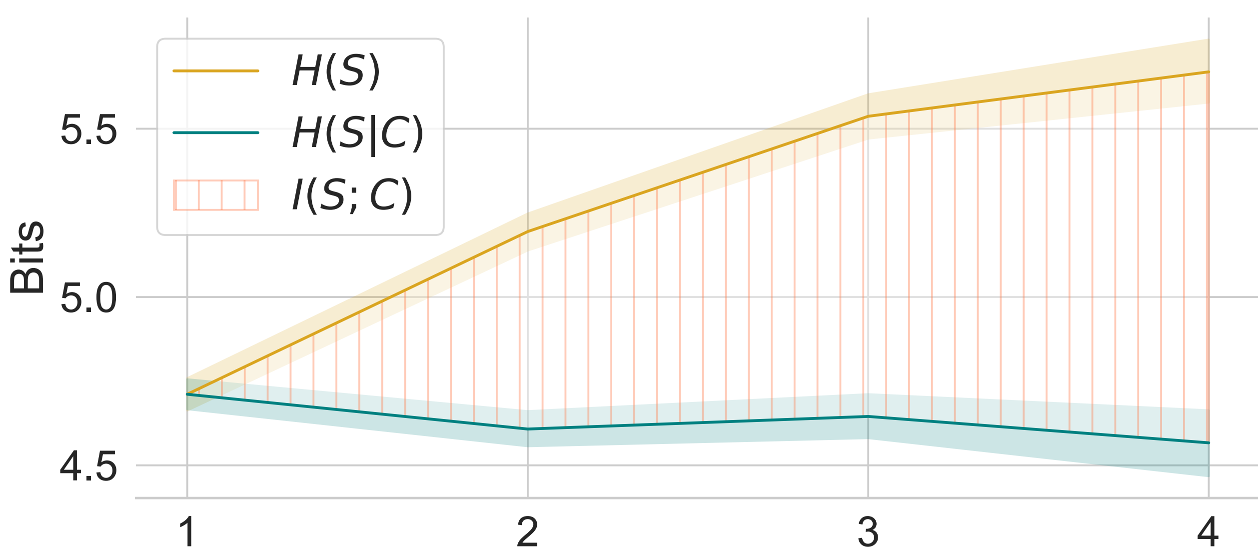
Language

(Giulianelli et al. 2021)

$P(w_i \mid \dots)$ estimates obtained with GPT-2 fine-tuned on PhotoBook

- Out-of-context surprisal *H(S)*
- In-context surprisal *H(S|C)*

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



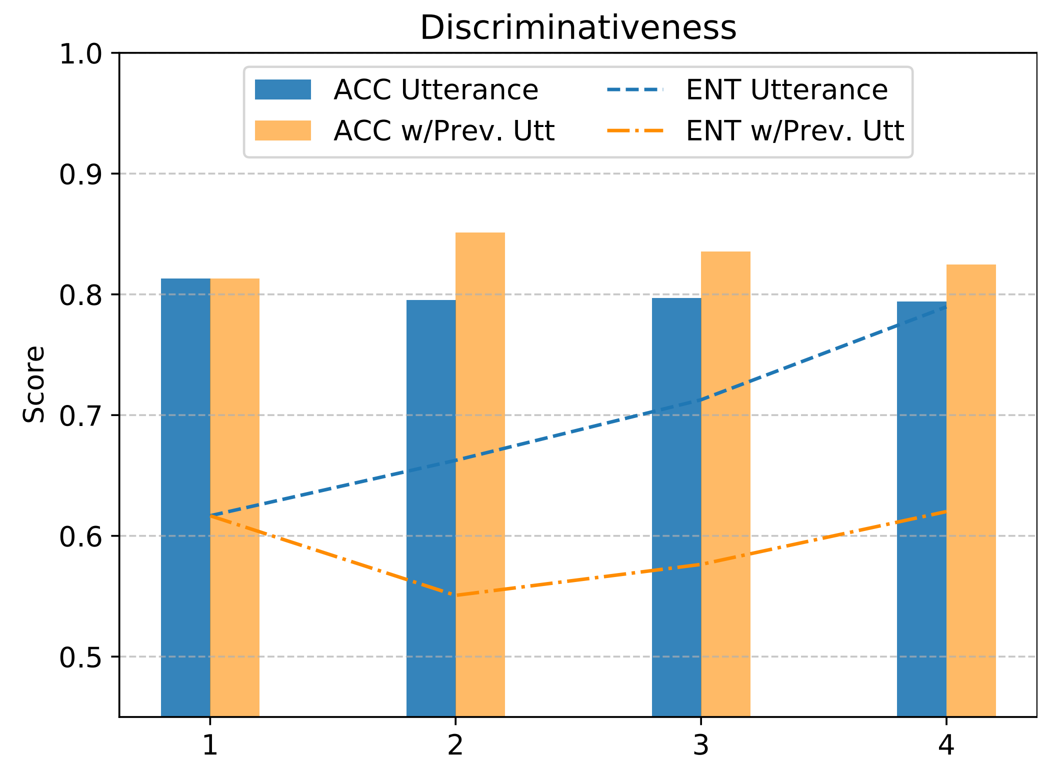
$$H(S \mid C) = -\log_2 P(S \mid C) = -\frac{1}{|S|} \sum_{w_i \in S} \log_2 P(w_i \mid w_1, \dots, w_{i-1}, C)$$

Language & vision

(Takmaz et al. 2022)

Given a referring utterance and the images in the context, CLIP yields softmax probabilities

- Accuracy with highest probability image
- Entropy of the distribution



Language
(Giulianelli et al. 2021)

$P(w_i \mid \dots)$ estimates obtained with GPT-2 fine-tuned on PhotoBook

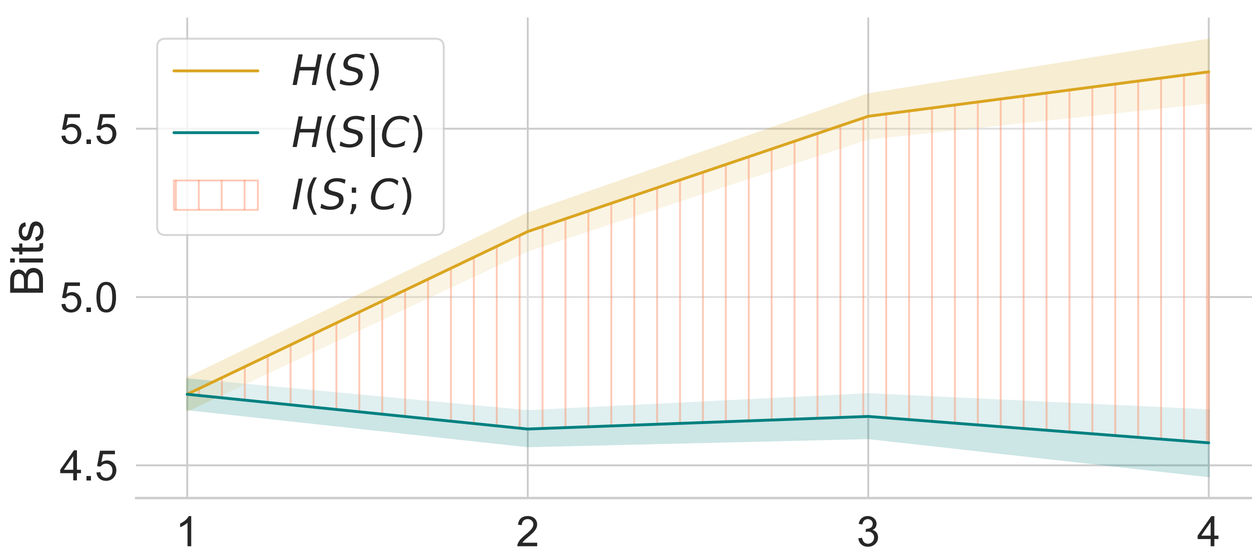
- Out-of-context surprisal *H(S)*
- In-context surprisal *H(S|C)*

Language & vision
(Takmaz et al. 2022)

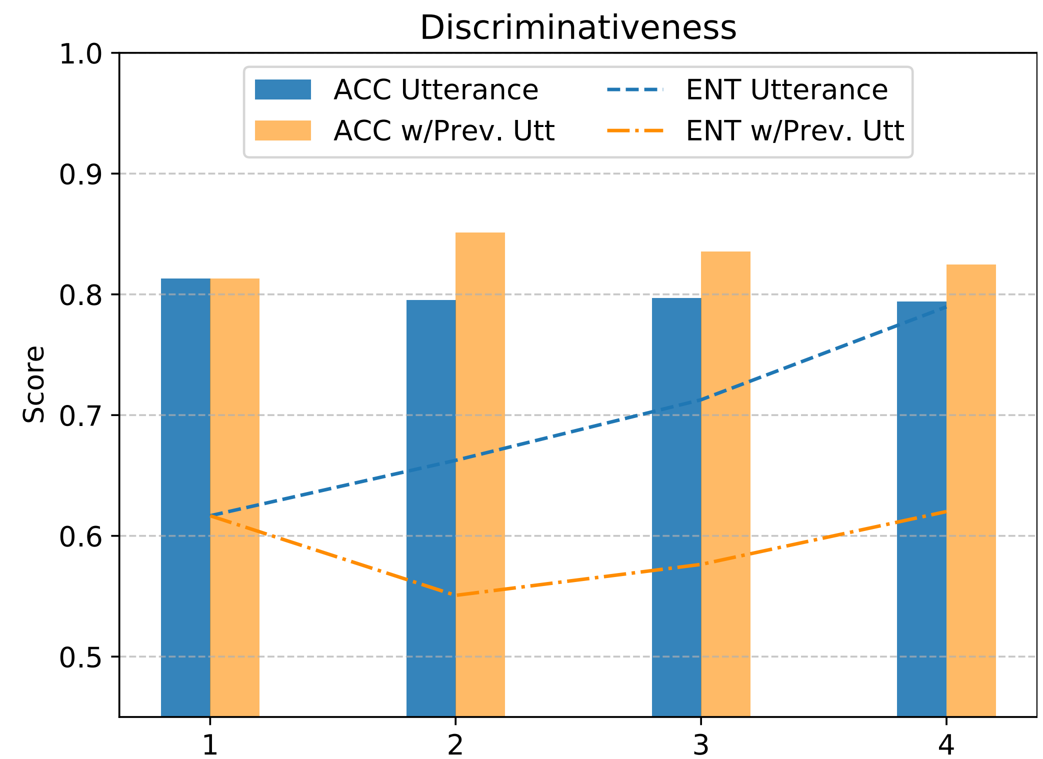
Given a referring utterance and the images in the context, CLIP yields softmax probabilities

- Accuracy with highest probability image
- Entropy of the distribution

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



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



Uniform low surprisal and resolution uncertainty with conversational context

Context dependent generation

- If later references rely on the conversational common ground, context-aware **generation** models will be closer to human patterns

Context dependent generation

- If later references rely on the conversational common ground, context-aware **generation** models will be closer to human patterns

Fine-tuning the model to adapt to the partner (Hawkins et al. CoNLL 2020)	Relying on episodic memory traces to condition generation (Takmaz et al. EMNLP 2020)

Context dependent generation

- If later references rely on the conversational common ground, context-aware **generation** models will be closer to human patterns

Takmaz et al (2020): Different encoder-decoder generation models

Context dependent generation

- If later references rely on the conversational common ground, context-aware **generation** models will be closer to human patterns

Takmaz et al (2020): Different encoder-decoder generation models

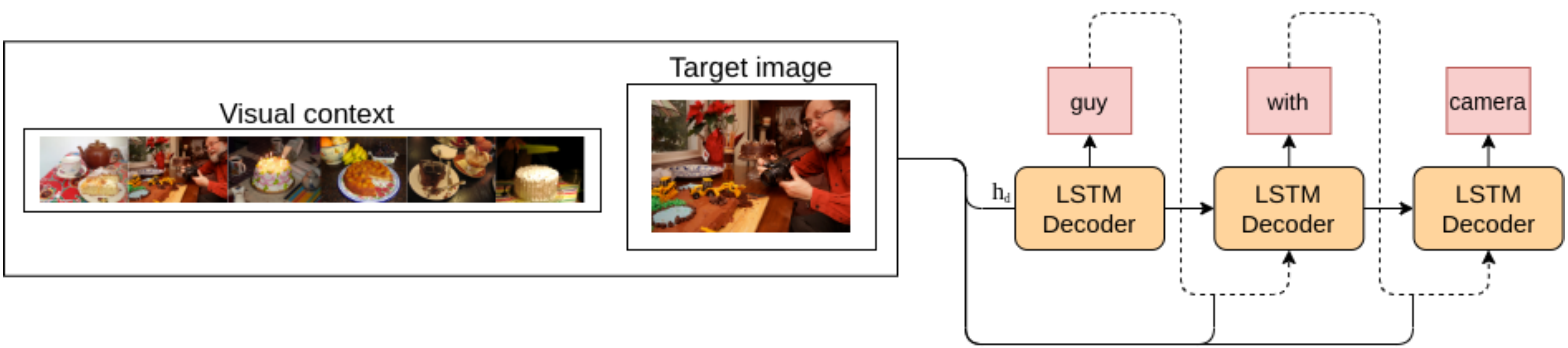
- **Ref:** only the visual context, ignoring the linguistic history

Context dependent generation

- If later references rely on the conversational common ground, context-aware **generation** models will be closer to human patterns

Takmaz et al (2020): Different encoder-decoder generation models

- **Ref:** only the visual context, ignoring the linguistic history

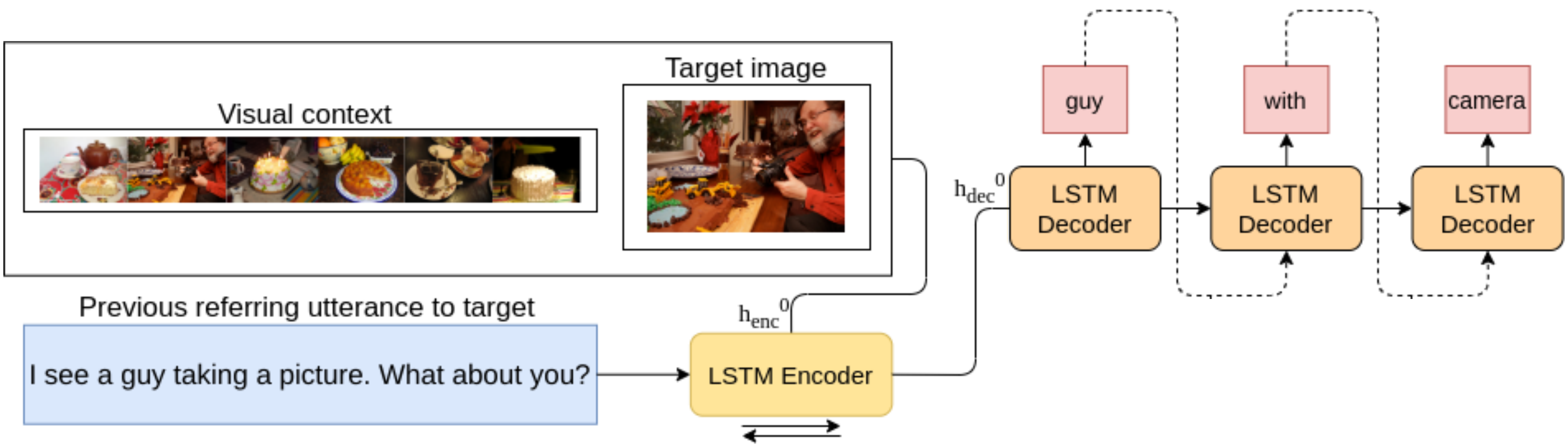


Context dependent generation

- If later references rely on the conversational common ground, context-aware **generation** models will be closer to human patterns

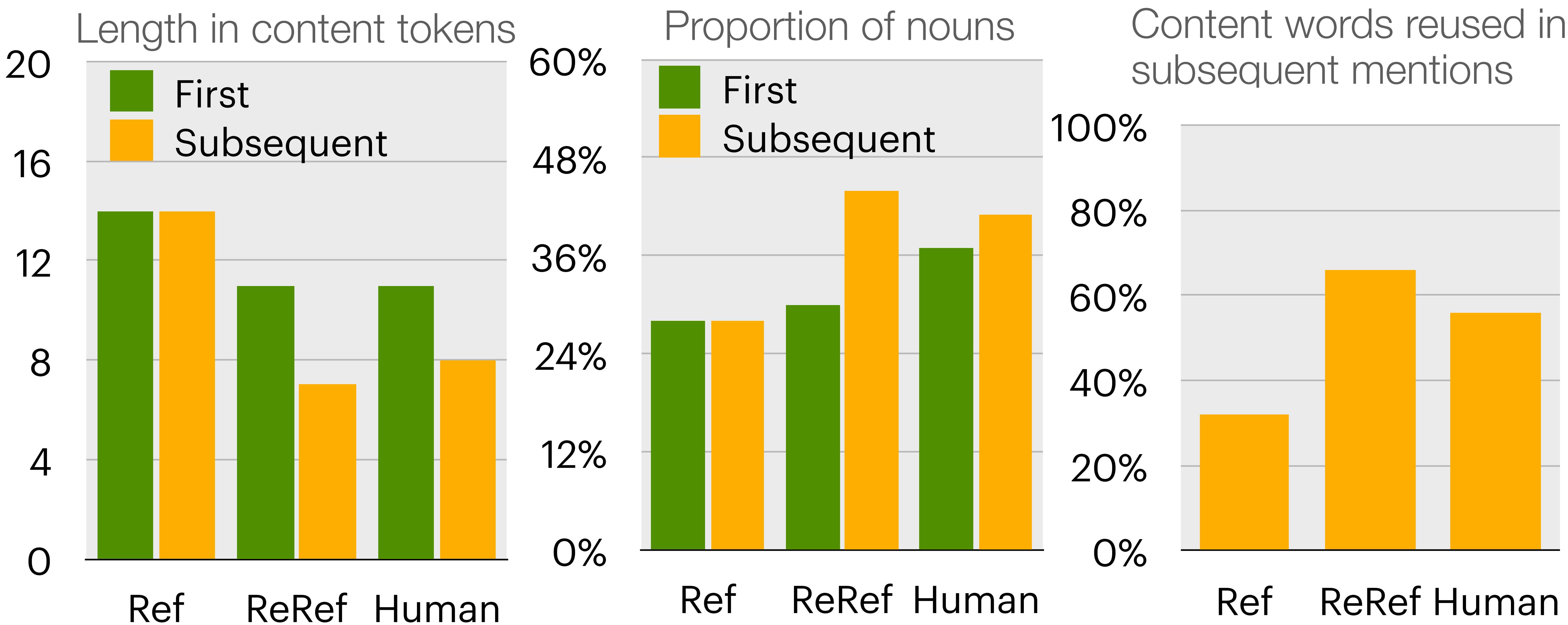
Takmaz et al (2020): Different encoder-decoder generation models

- **Ref:** only the visual context, ignoring the linguistic history
- **ReRef:** takes into account both visual and linguistic context, aware of previous mentions.

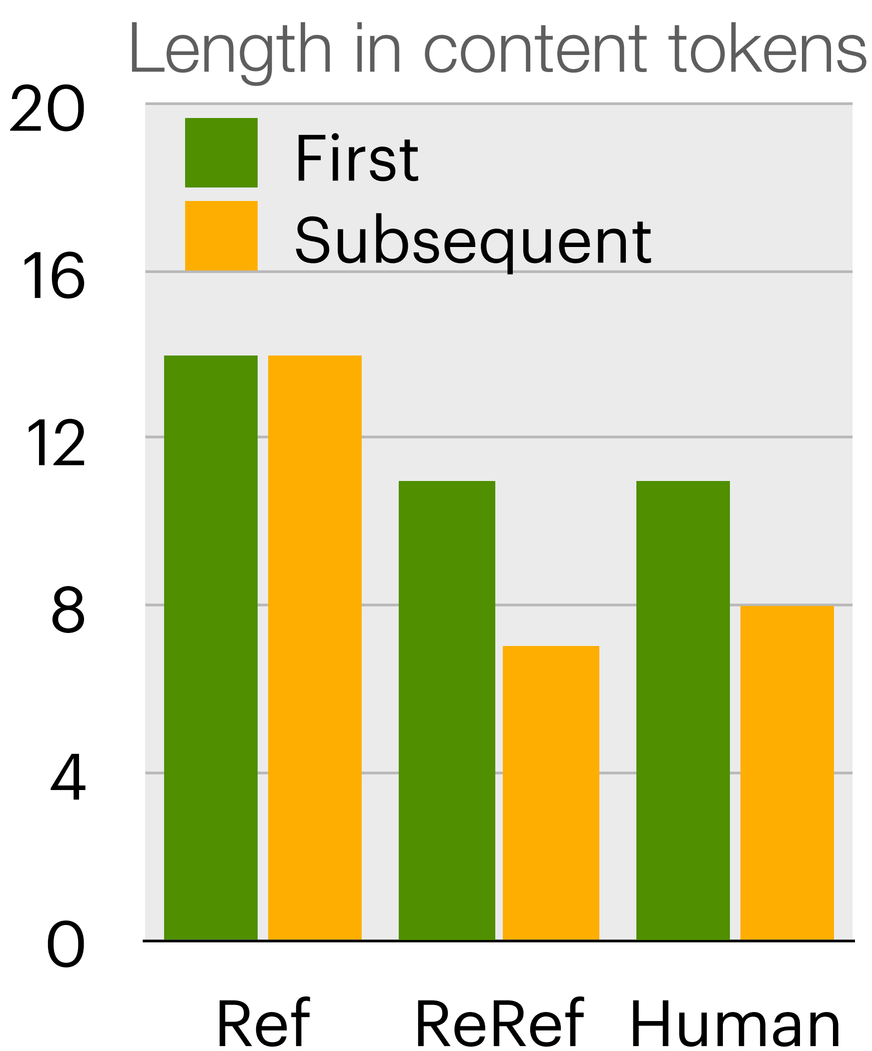


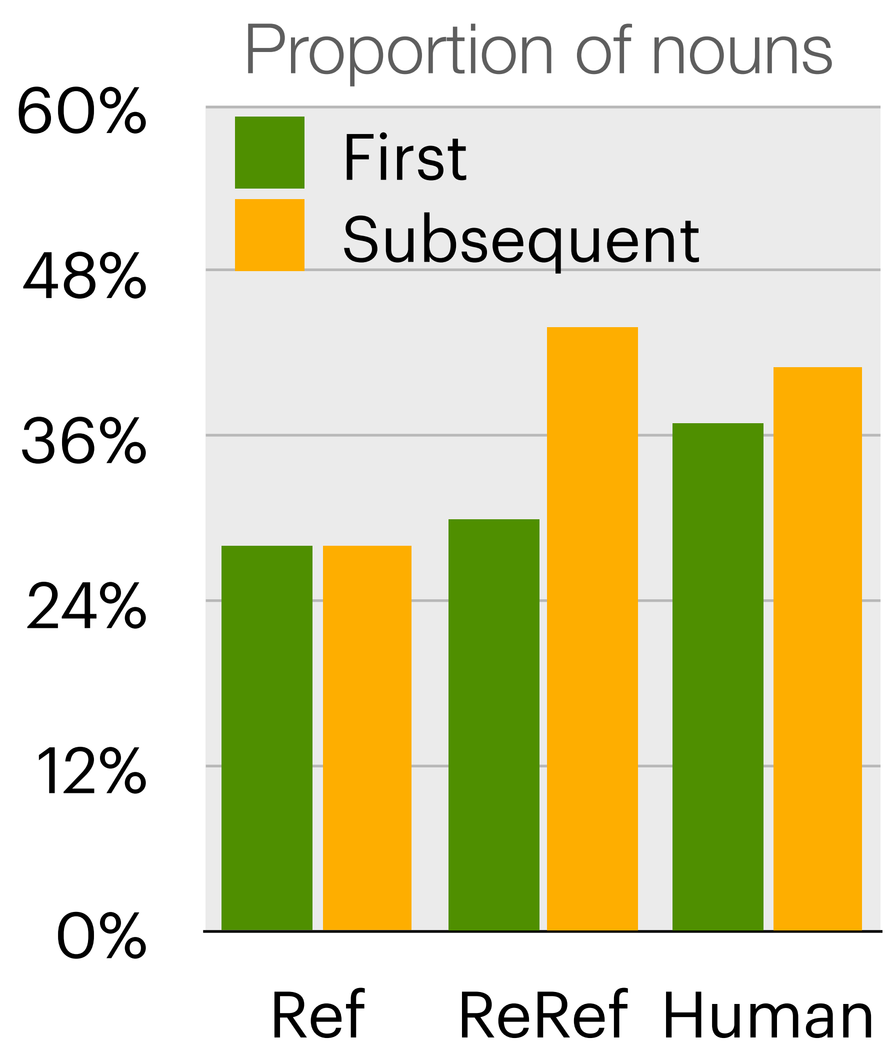
Similarity to human production patterns

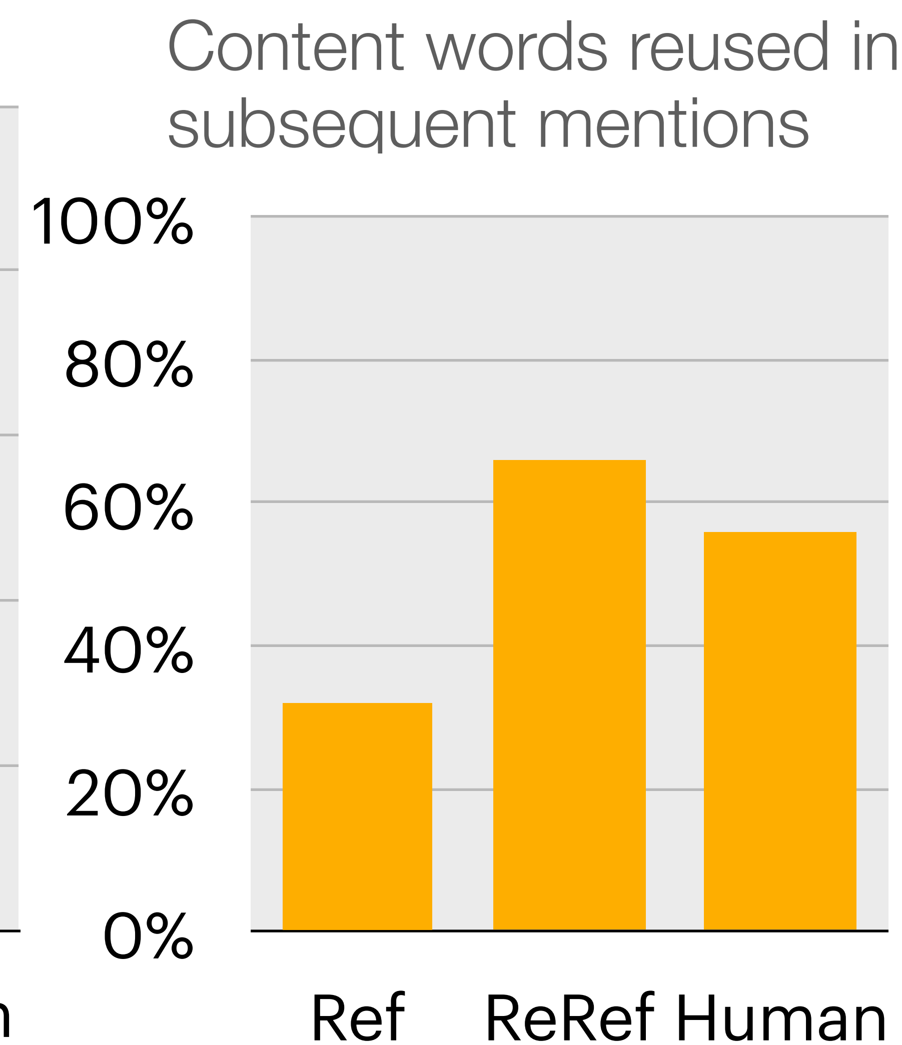
Similarity to human production patterns



Similarity to human production patterns







Previous mention: *a cake with a Godiva package in the background*

- **Human:** *chocolate cake with Godiva package behind it*
- **ReRef:** ***chocolate cake with Godiva in background***
- **Ref:** *do you have a picture of a brown cake on a bed?*

Interim summary

- In conversation, participants converge on referring expressions that they reuse (“conceptual pacts” become part of the context).
- Taking into account this conversational context:
    - Makes resolution less effortful, in line with principles of *uniform information density* and *least collaborative effort*.
    - Helps to generate utterances that are closer to human patterns.

Interim summary

- In conversation, participants converge on referring expressions that they reuse (“conceptual pacts” become part of the context).

- Taking into account this conversational context:
  - Makes resolution less effortful, in line with principles of *uniform information density* and *least collaborative effort*.
  - Helps to generate utterances that are closer to human patterns.

- The process whereby participants collaboratively arrive at “conceptual pacts” assumes shared semantic conventions as the starting point for ad hoc shared conventions….

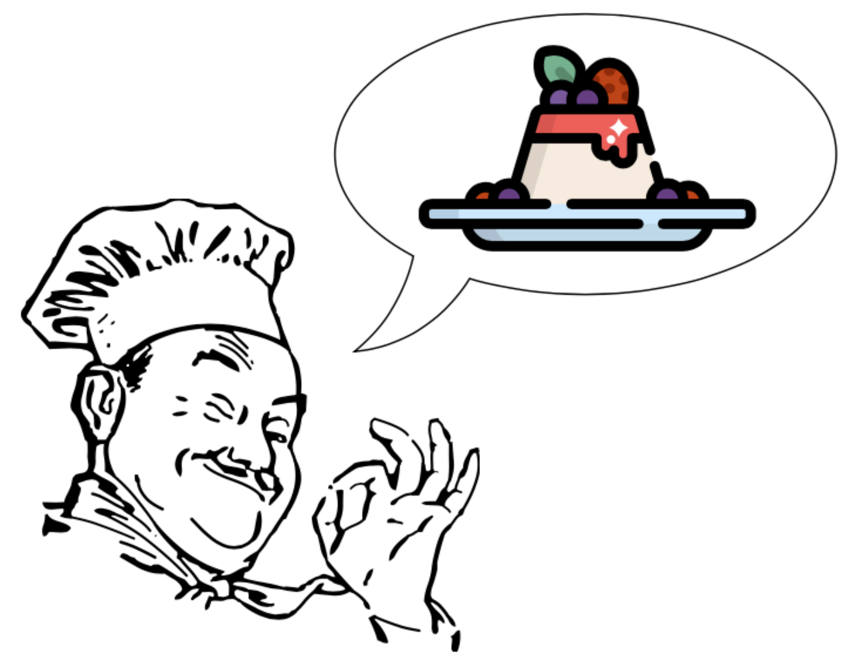
Part 2

- What if the dialogue participants have access to different general conventions and semantic knowledge?

Part 2

- What if the dialogue participants have access to different general conventions and semantic knowledge?

- In other words: How can a cook explain how to make *panna cotta* to someone who has never been in a kitchen?



Part 2

- What if the dialogue participants have access to different general conventions and semantic knowledge?

- In other words: How can a cook explain how to make *panna cotta* to someone who has never been in a kitchen?

- To coordinate not just at the level of dialogue-specific expressions, but also at the level of general semantic knowledge, it is fundamental to be able to represent and reason about others' mental states

Takmaz, Brandizzi et al. *Speaking the Language of Your Listener: Audience-Aware Adaptation via Plug-and-Play Theory of Mind*. Findings of ACL 2023.

Data

- We divide the PhotoBook dialogues into 5 domains with minimum vocabulary overlap

Data

- We divide the PhotoBook dialogues into 5 domains with minimum vocabulary overlap

Appliances

A white fridge with the door open

Food

A bowl of raw veggies next to a grapefruit

Indoor

The living room with lamp on a bookshelf

Outdoor

I have a guy doing a handstand on the beach

Vehicles

A parking lot with cars and motorcycle

Example referring utterance per domain

Data

- We divide the PhotoBook dialogues into 5 domains with minimum vocabulary overlap

Appliances

A white fridge with the door open

Food

A bowl of raw veggies next to a grapefruit

Indoor

The living room with lamp on a bookshelf

Outdoor

I have a guy doing a handstand on the beach

Vehicles

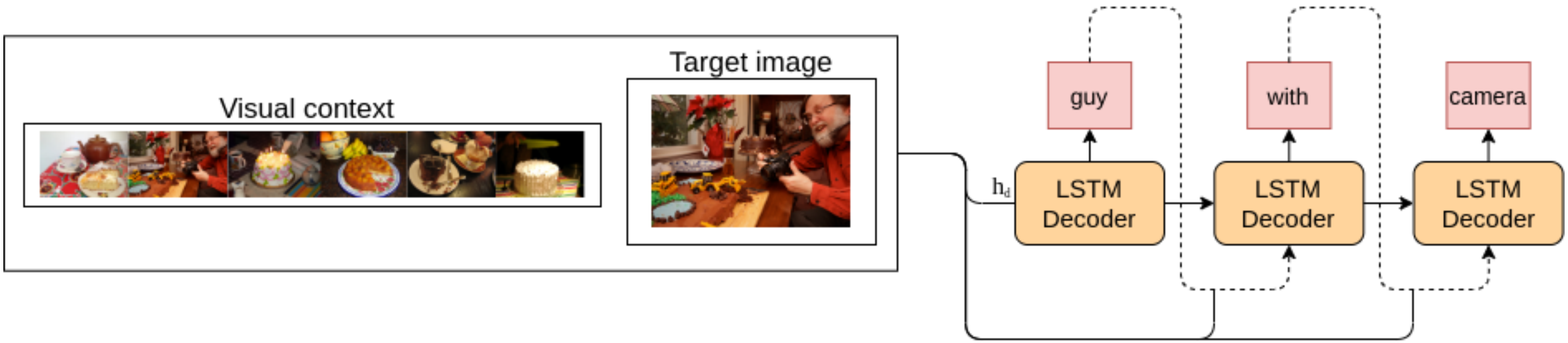
A parking lot with cars and motorcycle

Example referring utterance per domain

Domain	Prop	N	$\|V\|$	Images	Specific	Overlap
Appliances	9.4%	4,310	1,271	36	29.5%	23.2% (*Ind*)
Food	12.4%	5,682	1,646	36	43.3%	22.9% (*App*)
Indoor	26.4%	12,088	2,477	96	44.3%	26.0% (*Out*)
Outdoor	35.9%	16,427	2,858	108	47.0%	26.2% (*Veh*)
Vehicles	15.8%	7,234	1,738	48	36.0%	26.2% (*Out*)
All	100%	45,741	6,038	324	-	-

The speaker

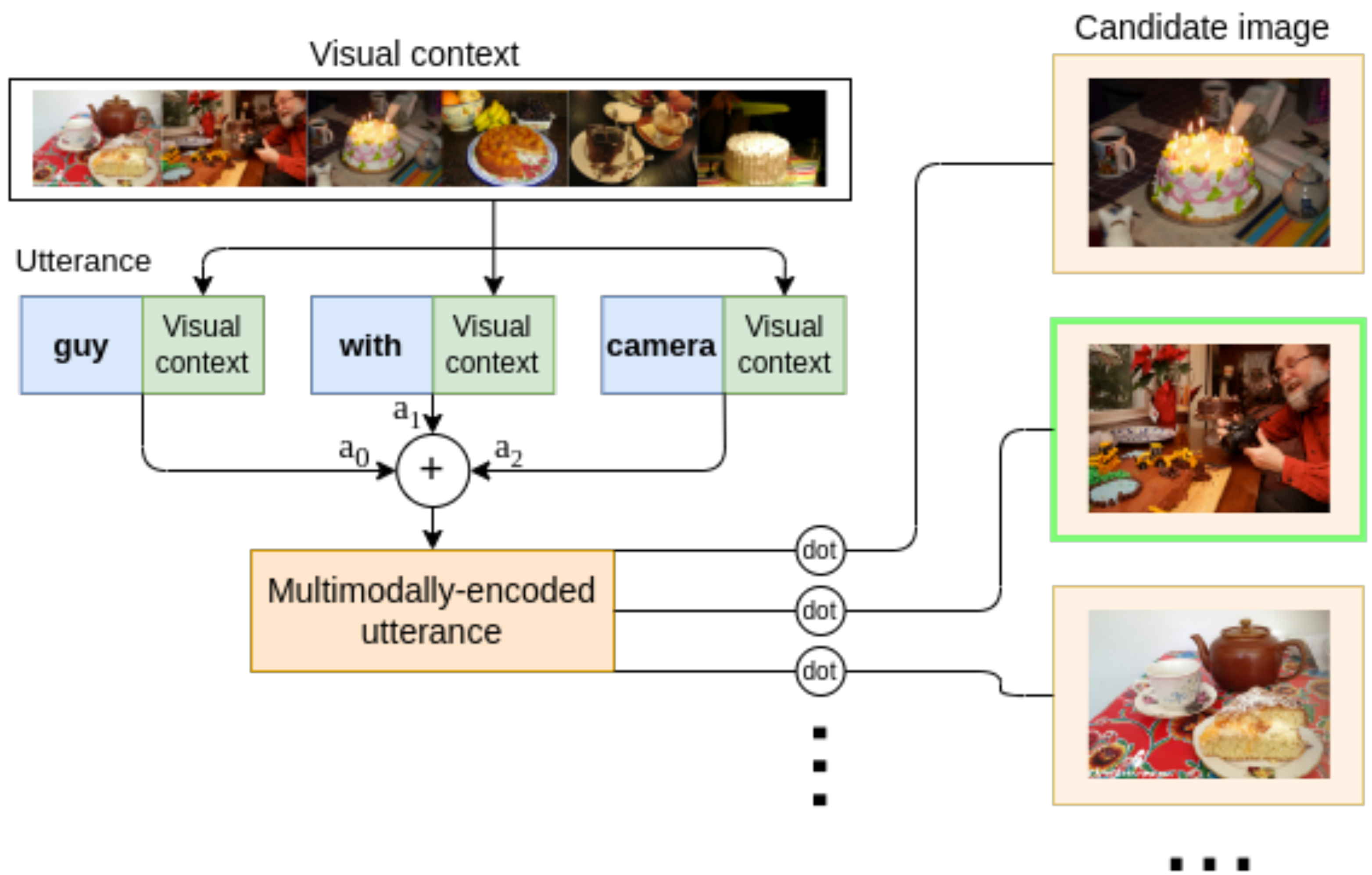
Visually conditioned language model



- **Input**: visual context including target image
- **Goal:** generate a referring utterance for the target
- **Training:** trained on all domains — “proficient speaker”

The listener

Discriminator model



- **Input:** visual context and utterance
- **Goal:** identify the target image the utterance refers to
- **Training:** trained on a single domain — “domain-specific listener”

Resolution performance

without adaptation

With domain-specific listeners, if the speaker does not adapt then communication is unsuccessful:

- High accuracy for in-domain settings (diagonal)
- Near chance accuracy (16%) for out-of-domain cases

Input image domain

Listener domain

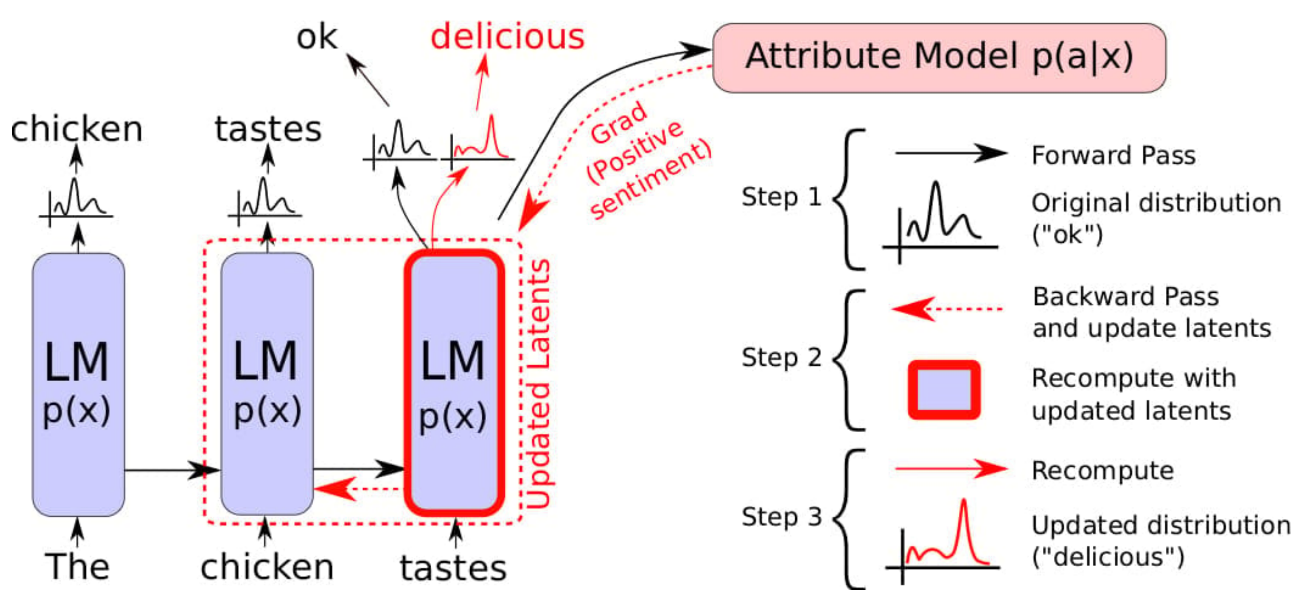
	app	**food**	**indoor**	**outdoor**	**vehi**
appliances	57.61	20.10	19.92	21.27	15.98
food	19.11	54.29	18.60	18.85	18.85
indoor	22.71	19.65	53.62	20.82	16.77
outdoor	15.08	21.46	19.62	52.93	17.69
vehicles	16.36	16.17	17.41	20.13	43.08

Plug-and-Play Theory of Mind

- How can the speaker adapt its utterances to the listener's knowledge?
- On the fly, without fine-tuning the language model permanently?

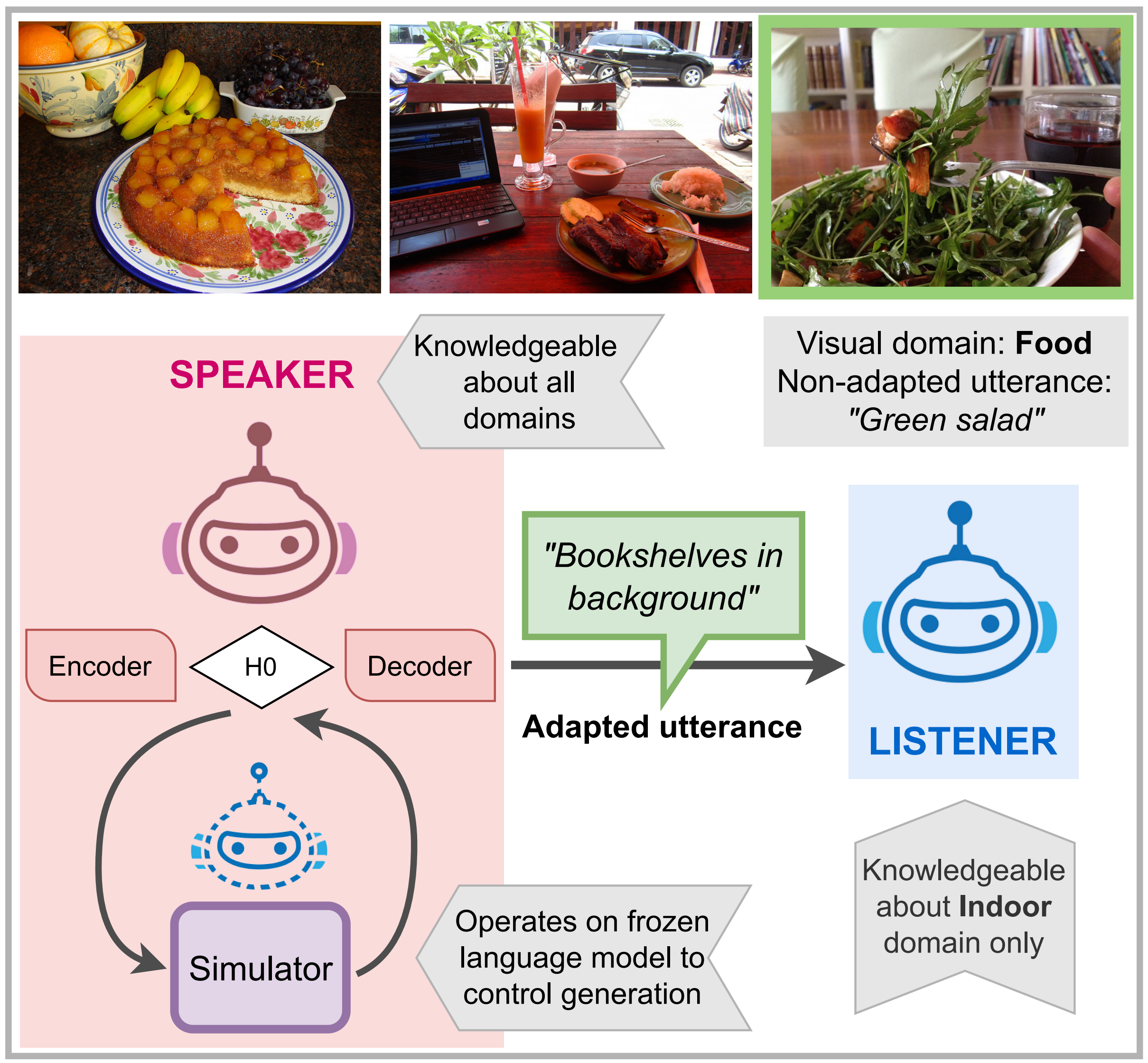
Plug-and-Play Theory of Mind

- How can the speaker adapt its utterances to the listener's knowledge?
- On the fly, without fine-tuning the language model permanently?
- Inspired by work on controlled text generation (Dathathri et al., 2020), we explore a "plug-and-play" approach



(Dathathri et al., 2020)

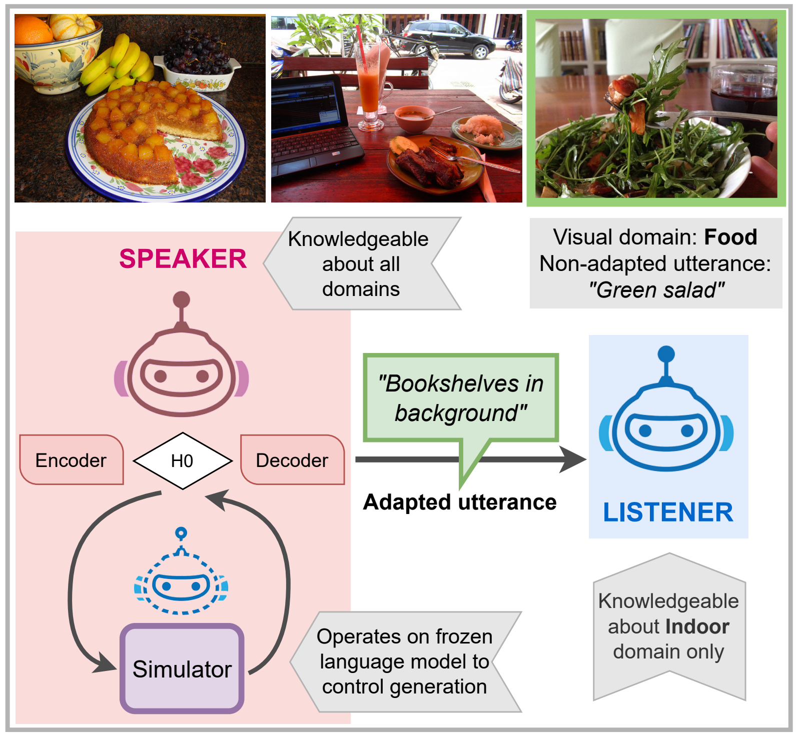
Plug-and-Play Theory of Mind



- Knowledge asymmetry
- The speaker tailors its utterance about a food image for a listener who does not know about food
- The speaker's simulator module guides this adaptation via self-monitoring loop

Plug-and-Play Theory of Mind

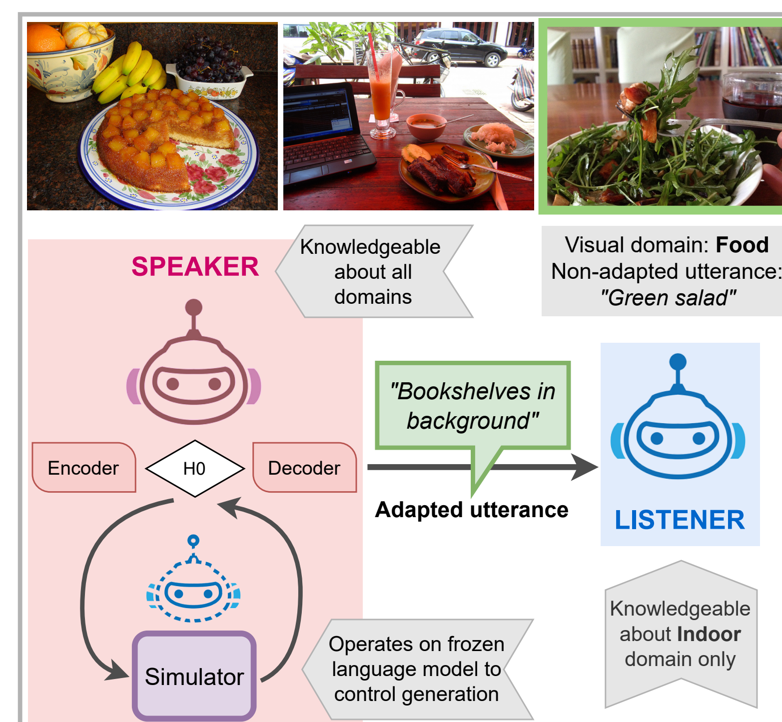
- The simulator is trained to predict the behaviour of a domain-specific listener, given a "planned" utterance and visual context
  - Simplification: *the speaker knows the type of listener a priori*



Plug-and-Play Theory of Mind

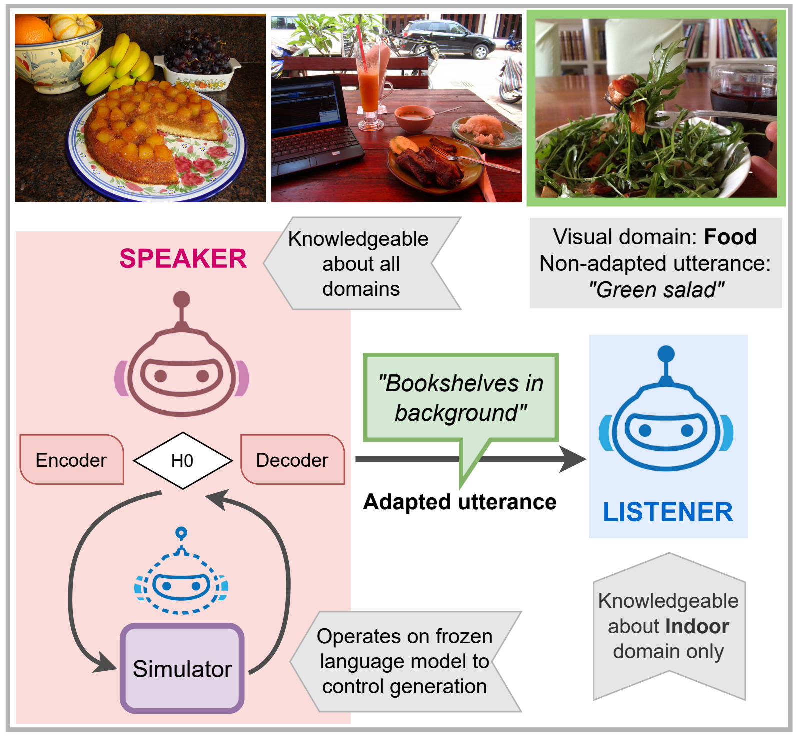
- The simulator is trained to predict the behaviour of a domain-specific listener, given a "planned" utterance and visual context
    - Simplification: *the speaker knows the type of listener a priori*

- It refines the speaker's utterance plan iteratively (≈ self-monitoring) until it considers it sufficiently discriminative for the listener
    - "*Would the listener be able to resolve this utterance?*" If the prediction is negative, this triggers an update to the speaker's decoder initial state, and the utterance gets updated



Plug-and-Play Theory of Mind

- The simulator is trained to predict the behaviour of a domain-specific listener, given a “planned” utterance and visual context
  - Simplification: *the speaker knows the type of listener a priori*
- It refines the speaker’s utterance plan iteratively (≈ self-monitoring) until it considers it sufficiently discriminative for the listener
  - “*Would the listener be able to resolve this utterance?*” If the prediction is negative, this triggers an update to the speaker’s decoder initial state, and the utterance gets updated
- Finally, the referring utterance is overtly passed on to the listener (who may or may not be able to resolve it — the simulator is not perfect!)



Does it work?

	Baseline	**Audience-aware**
OOD	19.06 ± 0.47	26.74 ± 1.48
IND	52.30 ± 1.10	71.77 ± 2.16

averages across domains

- Audience-aware adaptation leads to significant increases in accuracy
- Including more than 7% in scenarios where the image domain is not known to the listener (OOD)

Does it work?

	Baseline	Audience-aware
OOD	19.06 ± 0.47	26.74 ± 1.48
IND	52.30 ± 1.10	71.77 ± 2.16

averages across domains

- Audience-aware adaptation leads to significant increases in accuracy
- Including more than 7% in scenarios where the image domain is not known to the listener (OOD)

How does it work?

Qualitative examples

Target

Outdoor

Listener domain

Food

PhotoBook participant: *I have the pink food truck again ... white shirt lady*

Generated not adapted: *girl at black phone, red truck, brown hair, pink*

Generated adapted: *pink donuts*

Qualitative examples

Target

Food

Listener domain

Indoor

PhotoBook participant: *green salad with a person holding up a portion with fork?*

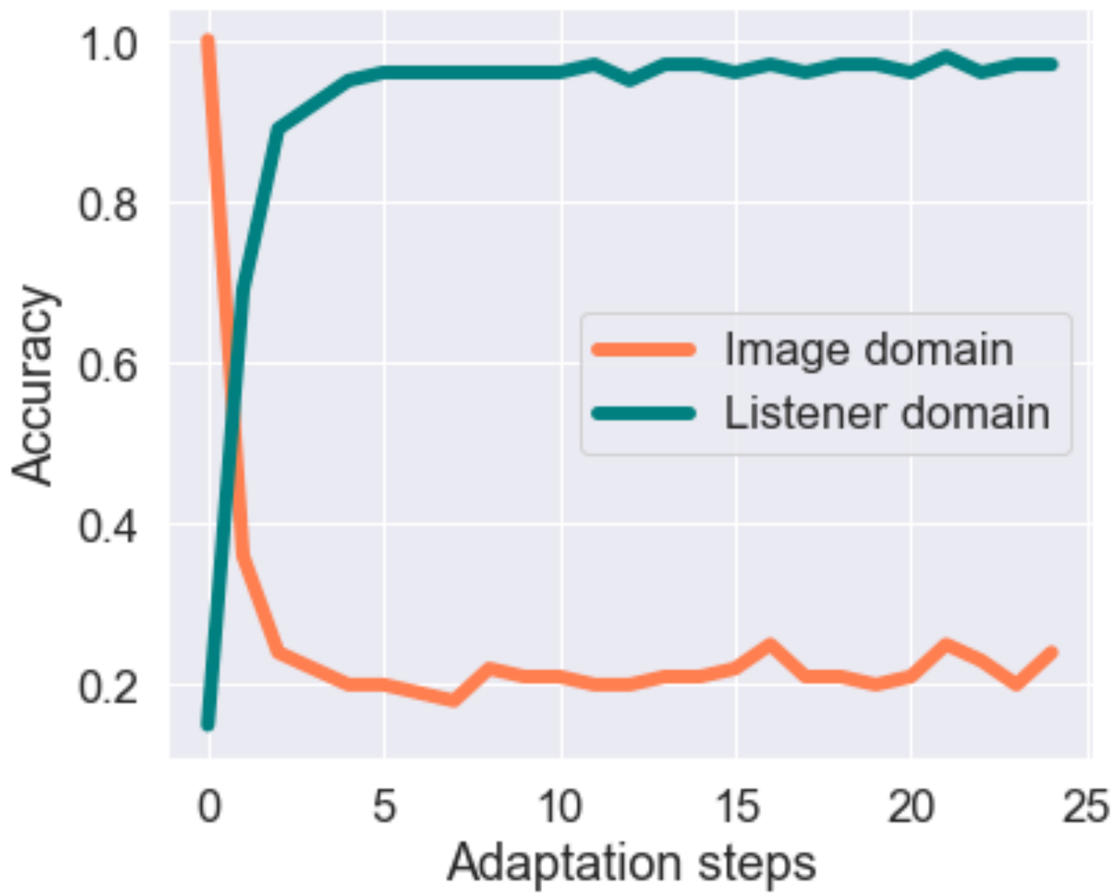
Generated not adapted: *I have one more maybe round you think that has a lime green shaped greens, a salad?*

Generated adapted: *must bookshelves in the salad?*

Probing for domain adaptation

- We expect h_0 to carry information about the **target image domain**, because it is the result of encoding such image.

- Indeed, using a diagnostic classifier we can predict the image domain from the h_0 with 100% accuracy.

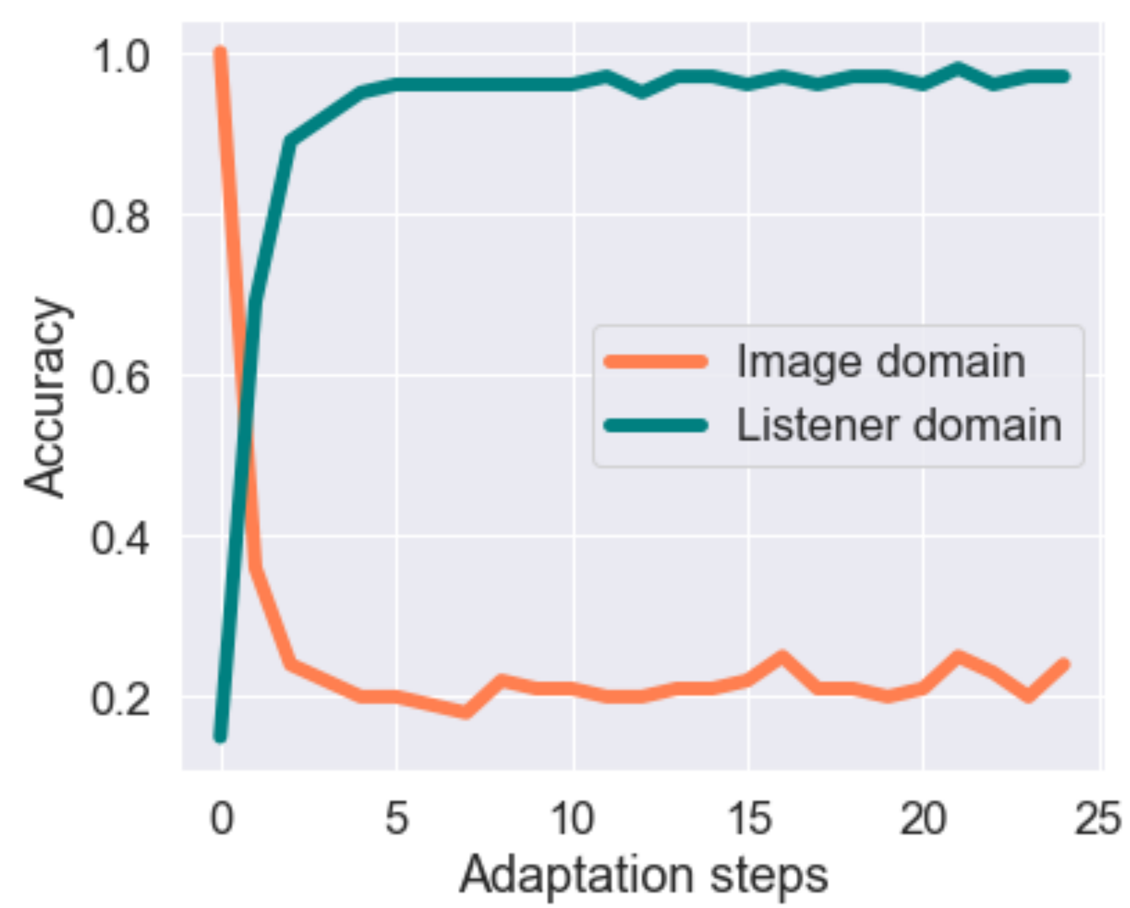


Probing for domain adaptation

- We expect h_0 to carry information about the **target image domain**, because it is the result of encoding such image.

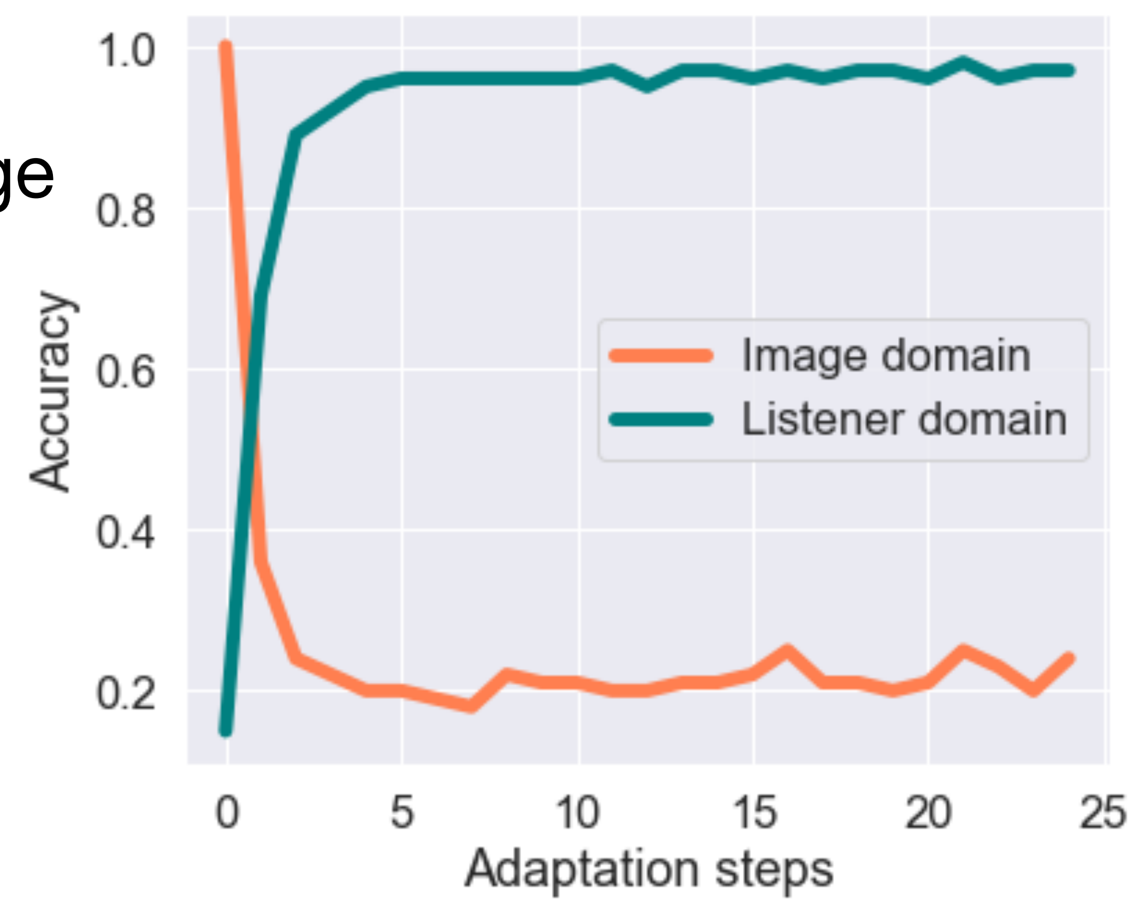
- Indeed, using a diagnostic classifier we can predict the image domain from the h_0 with 100% accuracy.

- Does h_0 carry information about the **listener's domain**? Not before adaptation.



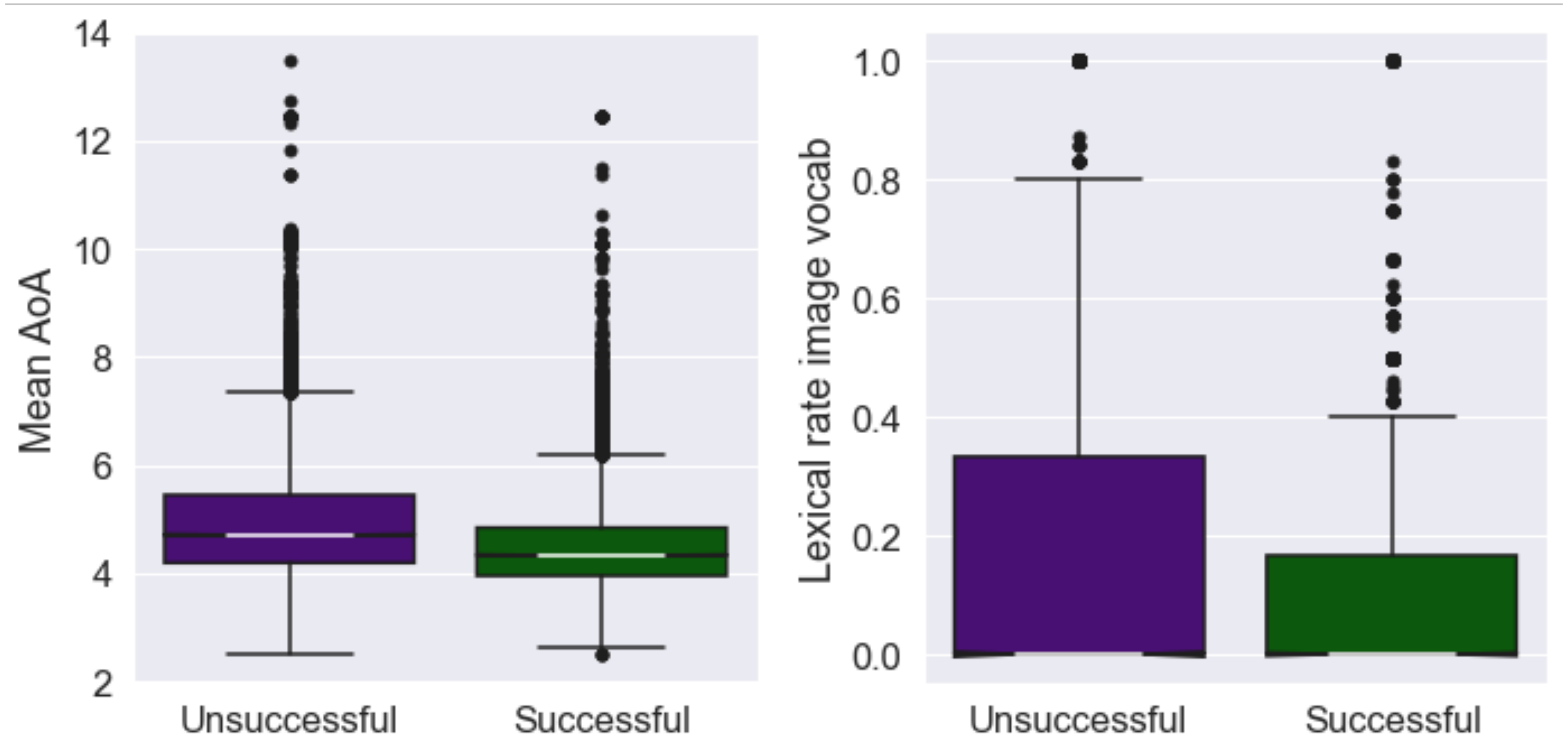
Probing for domain adaptation

- We expect h_0 to carry information about the **target image domain**, because it is the result of encoding such image.
- Indeed, using a diagnostic classifier we can predict the image domain from the h_0 with 100% accuracy.
- Does h_0 carry information about the **listener's domain**? Not before adaptation.
- With adaptation, the encoding of the image domain deteriorates, while the listener's domain becomes highly predictably.



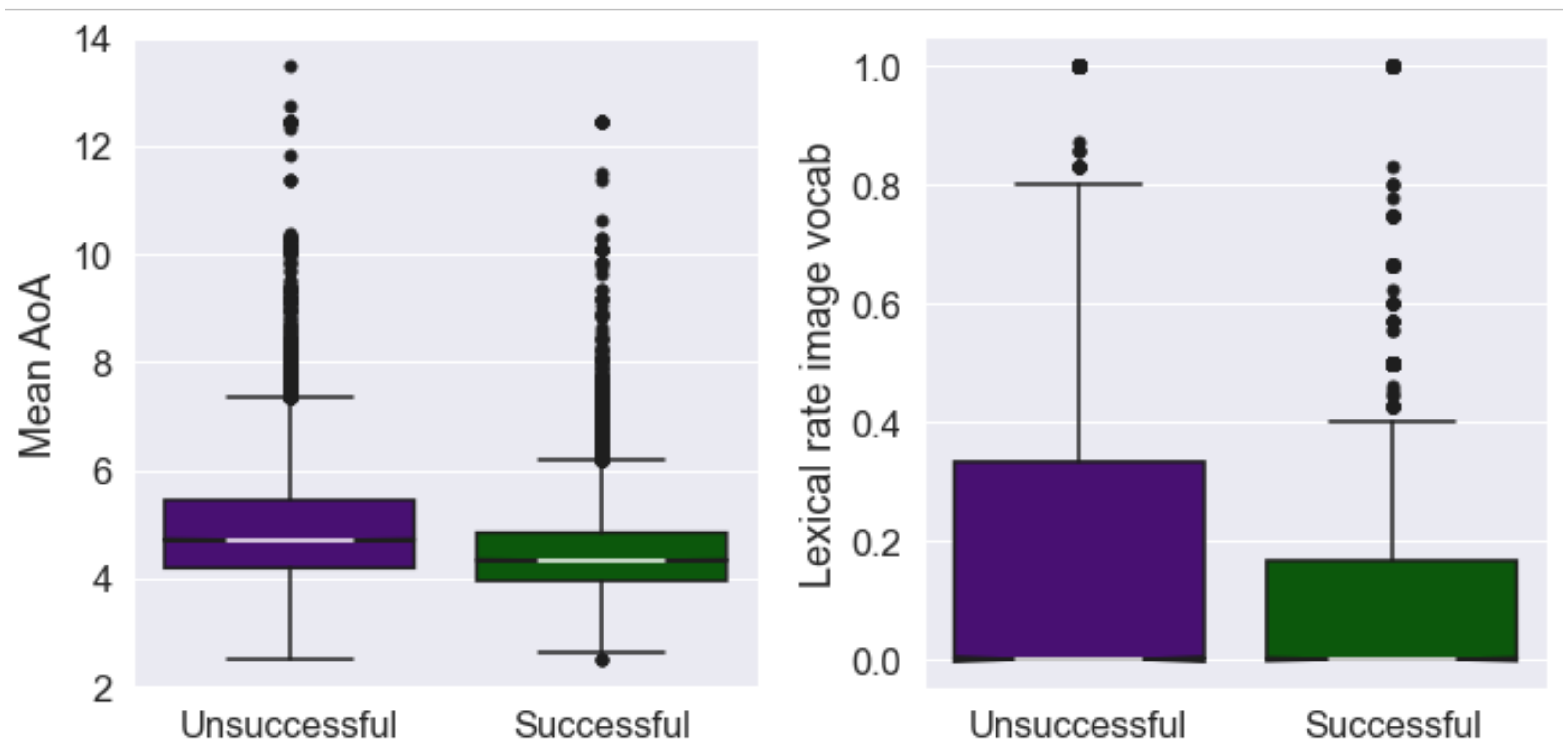
Properties of the adapted utterances

- Adapted utterances: when the speaker *believed* the utterance would lead to communicative success



Properties of the adapted utterances

- Adapted utterances: when the speaker *believed* the utterance would lead to communicative success



- More successful adapted utterances contain:
  - words with **lower age of acquisition**
  - lower rate of lexical choice from the target image vocabulary and **higher rate of words from the listener vocabulary**

Summing Up

- We decide what to say and how to say it on the basis of what we share with our dialogue partner.
- It is an open question how such accommodation can be modelled in computational agents.

Summing Up

- We decide what to say and how to say it on the basis of what we share with our dialogue partner.

- It is an open question how such accommodation can be modelled in computational agents.

- This talk: A few proposals for analysing context dependence of repeated references, generating dialogue-aware references, adapting utterances via theory of mind simulation.

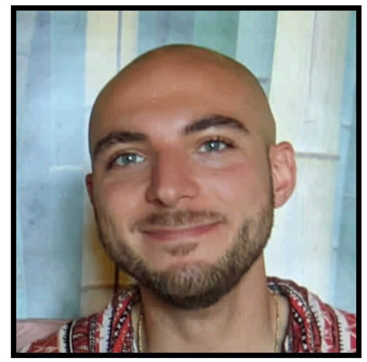
Summing Up

- We decide what to say and how to say it on the basis of what we share with our dialogue partner.
- It is an open question how such accommodation can be modelled in computational agents.
- This talk: A few proposals for analysing context dependence of repeated references, generating dialogue-aware references, adapting utterances via theory of mind simulation.
- Many limitations remain: for example, our agent models (speaker, simulator, listener) are pertained and remain frozen. This has advantages but dramatically oversimplifies the dynamics of interaction.

Thanks

Niccolò Brandizzi

Ece Takmaz



Mario Giulianelli

Sandro Pezzelle

Arabella Sinclair

Janosch Haber

- Haber et al. *The PhotoBook dataset: Building common ground through visually grounded dialogue*. ACL 2019.
- Takmaz et al. Refer, reuse, reduce: Generating subsequent references in visual and conversational contexts. EMNLP 2020.
- Giulianelli et al. *Is information density uniform in task-oriented dialogues?* EMNLP 2021.
- Giulianelli & Fernández. *Analysing human strategies of information transmission as a function of discourse context.* CoNLL 2021.
- Takmaz et al. *Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP*. CMCL Workshop, 2022.
- Takmaz*, Brandizzi* et al. *Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind*. Findings of ACL 2023.