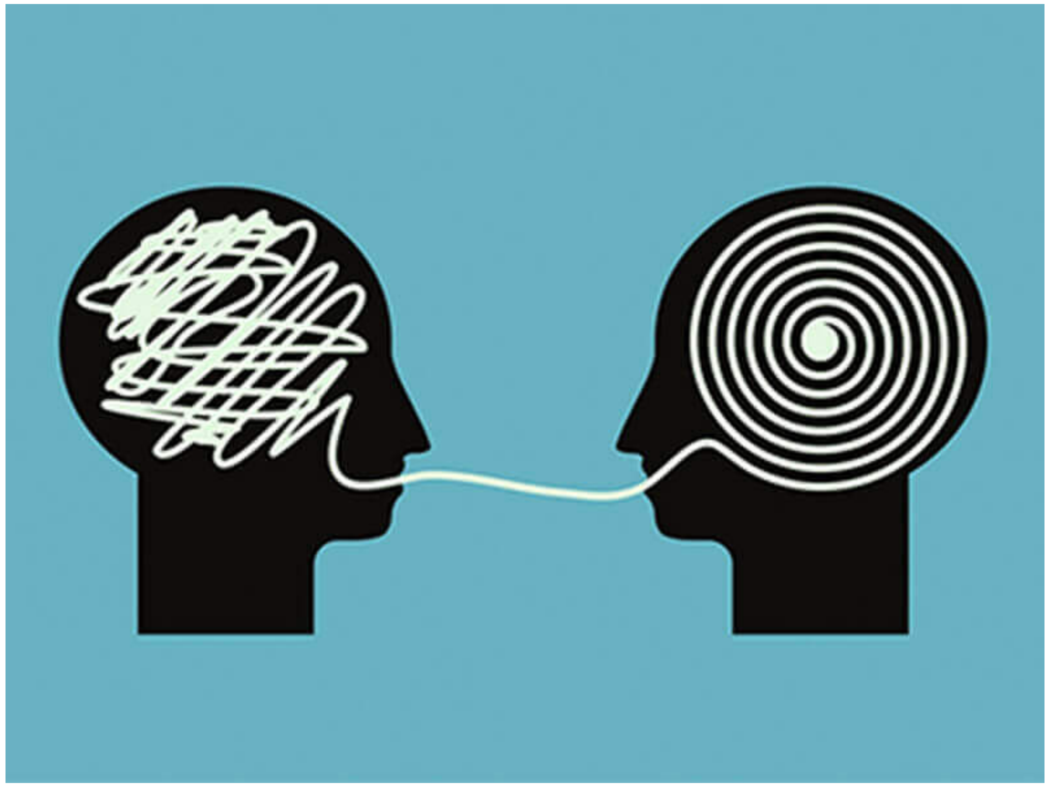
Dialogue Modelling

Raquel Fernández
Institute for Logic, Language and Computation
University of Amsterdam

NLP1 — 27 November 2019

Dialogue

- Using language to dynamically interact and communicate between multiple agents.
- The primary form of language use and language learning!
- The hallmark of human intelligence?

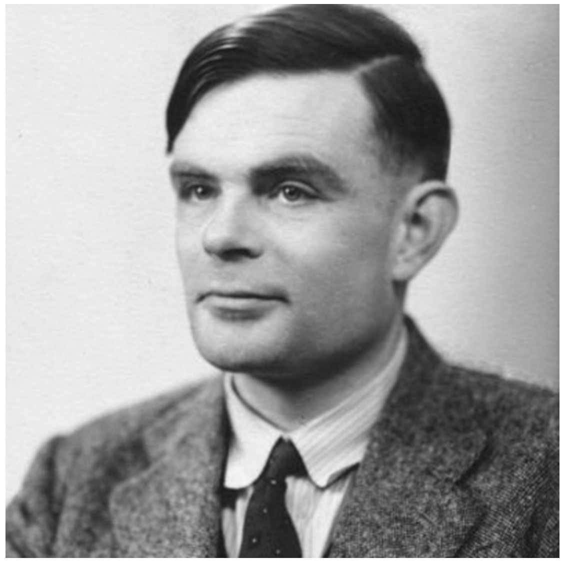


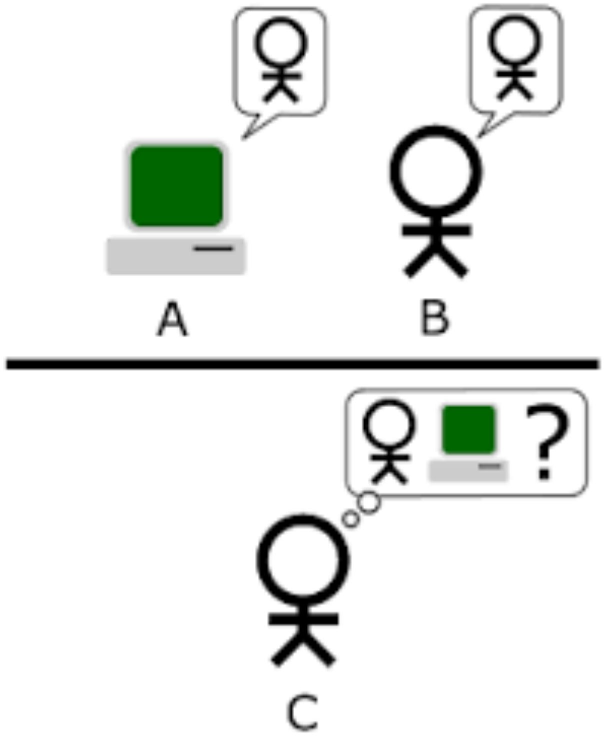
Origins of NLP within AI

Alan Turing, Machine and Intelligence (1950).

The imitation game: can machines think?

- Test this using **dialogue**.





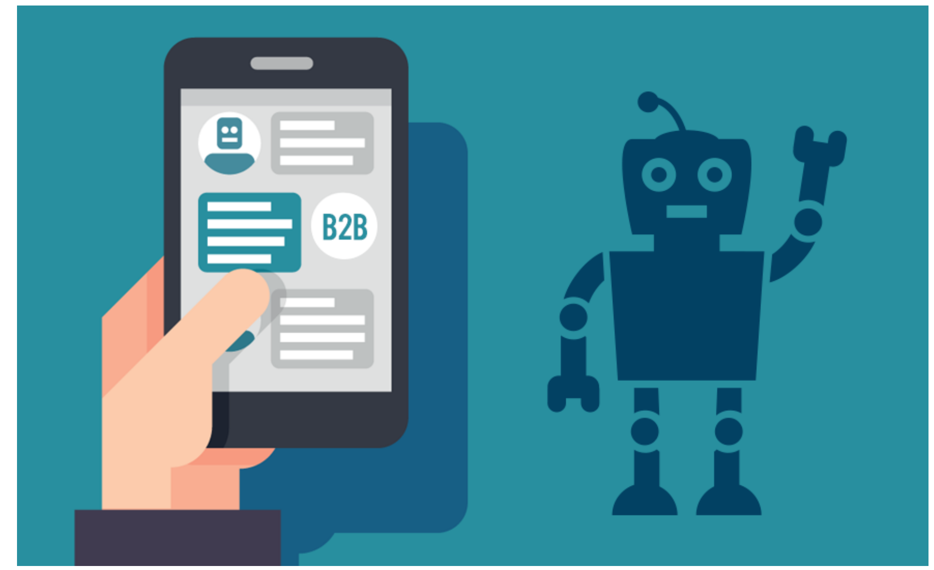
Probing question by C: *Please write me a sonnet on the subject of the Forth Bridge.*

A or B: *Count me out of this one. I never could write poetry.*

- Language in dialogue as the hallmark of human intelligence.

Currently a hot topic

- Human-Computer Interaction
- Chatbots



- Automatic speech recognition and spoken language processing
  Siri (2011), Alexa (2014), Google Assistant (2016)

Challenges of Dialogue

All levels of linguistic analysis (morphology, syntax, semantics, discourse…) are at play — plus more:

- Both ***understanding*** and ***generation***.
- Coordination among dialogue participants:
  - **When** to speak (turn taking)
  - **What** to say (content, function, coherence)
  - **How** to say it (style, adaptation)

Basic units

Dialogues are organised into **turns** and **utterances**.

- Utterances are functional units (not quite like sentences).
- Each turn may contain more than one utterance.

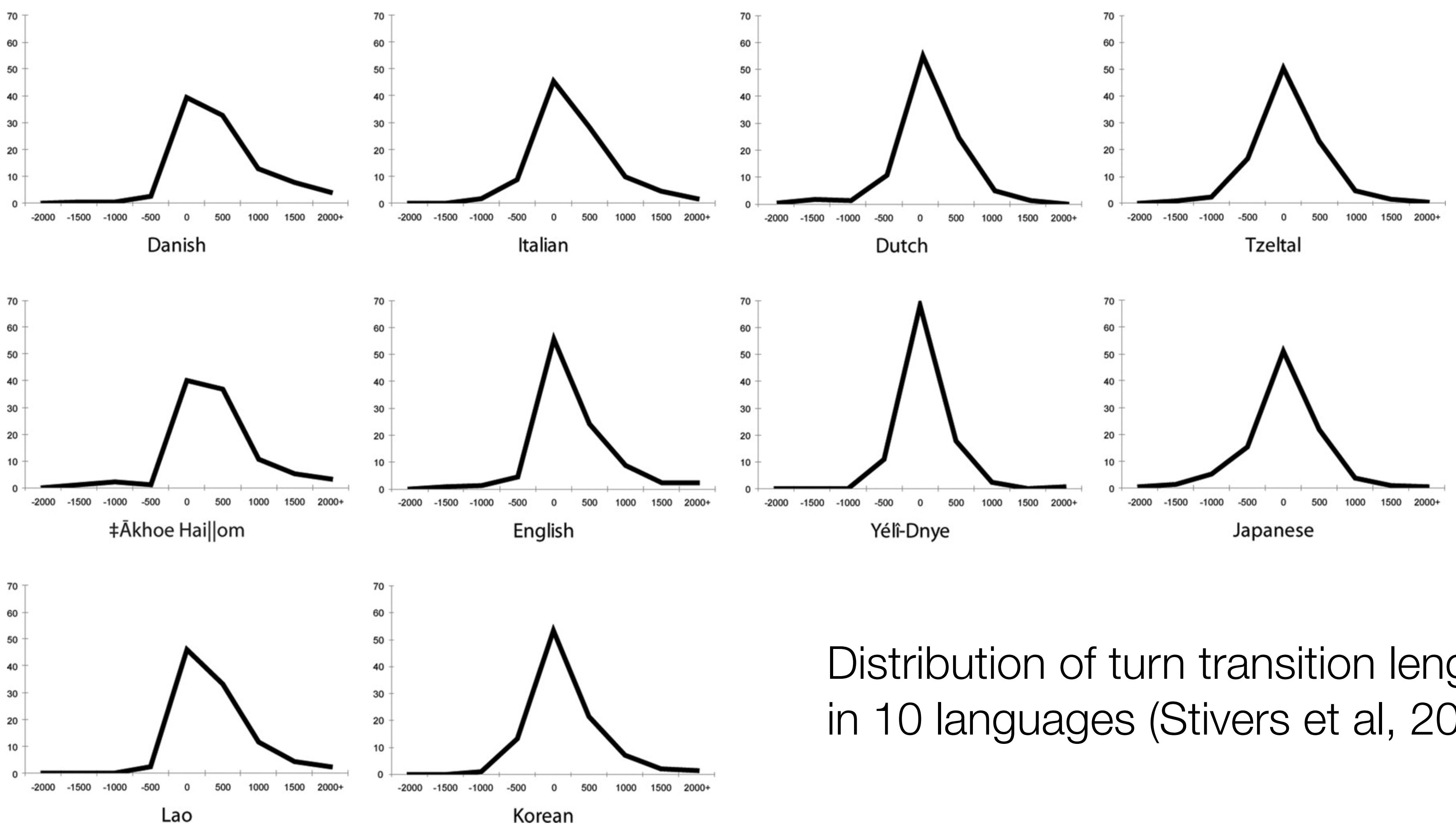
```
B.52 utt1:  Yeah, /
B.52 utt2:  [it's,+ it's] fun getting together with immediate family./
B.52 utt3:  A lot of my cousins are real close /
B.52 utt4:  {C and} we always get together during holidays and
             weddings and stuff like that, /
A.53 utt1:  {F Uh, } those are the ones that are in Texas?  /
B.54 utt1:  # {F Uh, } no, # /
A.55 utt1:  # {C Or } you # go to Indiana on that?  /
B.56 utt1:  the ones in Indiana, /
B.56 utt2:  uh-huh.  /
A.57 utt1:  Uh-huh, /
A.57 utt2:  where in Indiana?  /
B.58 utt1:  Lafayette.  /
```

Transcript fragment from the Switchboard dialogue corpus.

When: turn taking

Turn taking happens very smoothly:

- Overlaps are rare.
- Inter-turn pauses are very short or even absent.
- Strong universal patterns.



Distribution of turn transition length in milliseconds in 10 languages (Stivers et al, 2009)

When: turn taking

Very short inter-turn gaps means:

- Humans do not (always) react to silence to decide when to speak.
- We anticipate the end of the turn and start to plan our utterances before our dialogue partner ends.
- We are good at this prediction — overlaps are rare.

When: turn taking

Very short inter-turn gaps means:

- Humans do not (always) react to silence to decide when to speak.
- We anticipate the end of the turn and start to plan our utterances before our dialogue partner ends.
- We are good at this prediction — overlaps are rare.

Most spoken dialogue systems react to silence or use a push-to-talk strategy.

- A lot of room for improvement: getting timing right is key to develop spoken systems that interact naturally.

Challenges of Dialogue

All levels of linguistic analysis (morphology, syntax, semantics, discourse…) are at play — plus more:

- Both ***understanding*** and ***generation***.
- Coordination among dialogue participants:
  - **When** to speak (turn taking)
  - **What** to say (content, function, coherence)
  - **How** to say it (style, adaptation)

What to say

Modelling what to say next in a conversation is a very difficult problem:

- Understand dialogue **context** (what has been said/agreed).
- Take into account the **goal** of the conversation.
- Produce a **coherent** contribution, given context and goals.

Dialogue acts

Speech act or **dialogue act**: the function of (or the action performed by) an utterance. The intention of the speaker.

- *statement, question, answer, acknowledgement, request, agreement*, ….

Dialogue acts

Speech act or **dialogue act**: the function of (or the action performed by) an utterance. The intention of the speaker.

- *statement, question, answer, acknowledgement, request, agreement*, ….

- Often the dialogue act of an utterance can't be determined by form alone:

  The gun is loaded. *Threat? Warning? Statement?*

Dialogue acts

Speech act or **dialogue act**: the function of (or the action performed by) an utterance. The intention of the speaker.

- *statement, question, answer, acknowledgement, request, agreement*, ….

- Often the dialogue act of an utterance can't be determined by form alone:

  The gun is loaded. *Threat? Warning? Statement?*

- It may require inference (e.g., computing a "conversational implicature"):

  A: Are you going to Paul's party?
  B: I have to work.
  (=> I'm not going — *negative answer*)

Dialogue acts

Dialogue acts contribute to structure dialogues.

- They set up certain expectations: **forward-looking** vs. **backward-looking** acts.

Waiter: What'll you girls have?
Customer: What's the soup of the day?
Waiter: Clam chowder.
Customer: I'll have a bowl of clam chowder.

- **Adjacency pairs**: common sequences of act types.
  - Not strictly adjacent, but most expected dialogue act.
  - Intervening turns perceived as "insertion sequence"

What to say

Modelling what to say has often been addressed with shallow approaches:

- Rule-based chatbots in the early days.
- Data-driven neural chatbots nowadays.
- Current systems (i.e., Alexa) use a combination of both methods.

Rule-based chatbots

A conversation with Eliza (Weizenbaum 1966), the first chatbot:

Men are all alike.

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

Rule-based chatbots

A conversation with Eliza (Weizenbaum 1966), the first chatbot:

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE

Transformation rules based on keywords ranked from specific to general:

```
I know everybody laughed at me
```

"I" is a very general keyword:

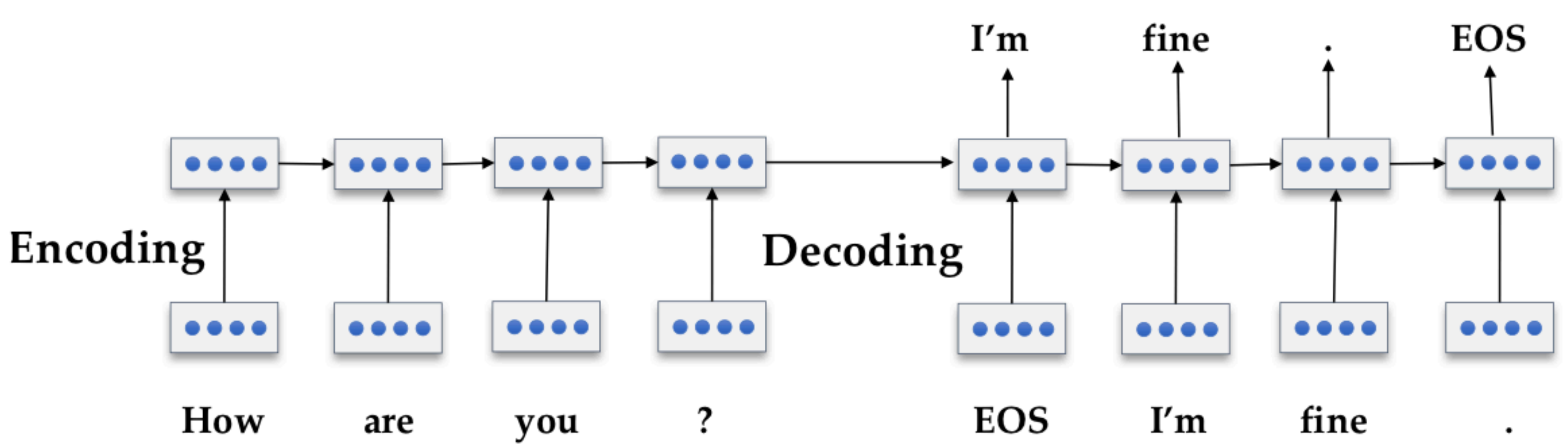
```
I: (I *) -> (You say you 2)
YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU
```

"Everybody" is much more interesting (someone using universals like everybody/always is probably "referring to some quite specific event or person")

```
WHO IN PARTICULAR ARE YOU THINKING OF?
```

Neural chatbots

- Trained on large corpora of conversations.
- **Sequence-to-sequence** architecture.
  - LSTM as encoder of the current turn.
  - LSTM (with attention) as decoder that generates the response with the highest probability (word by word).



Neural chatbots

Examples of generated responses:

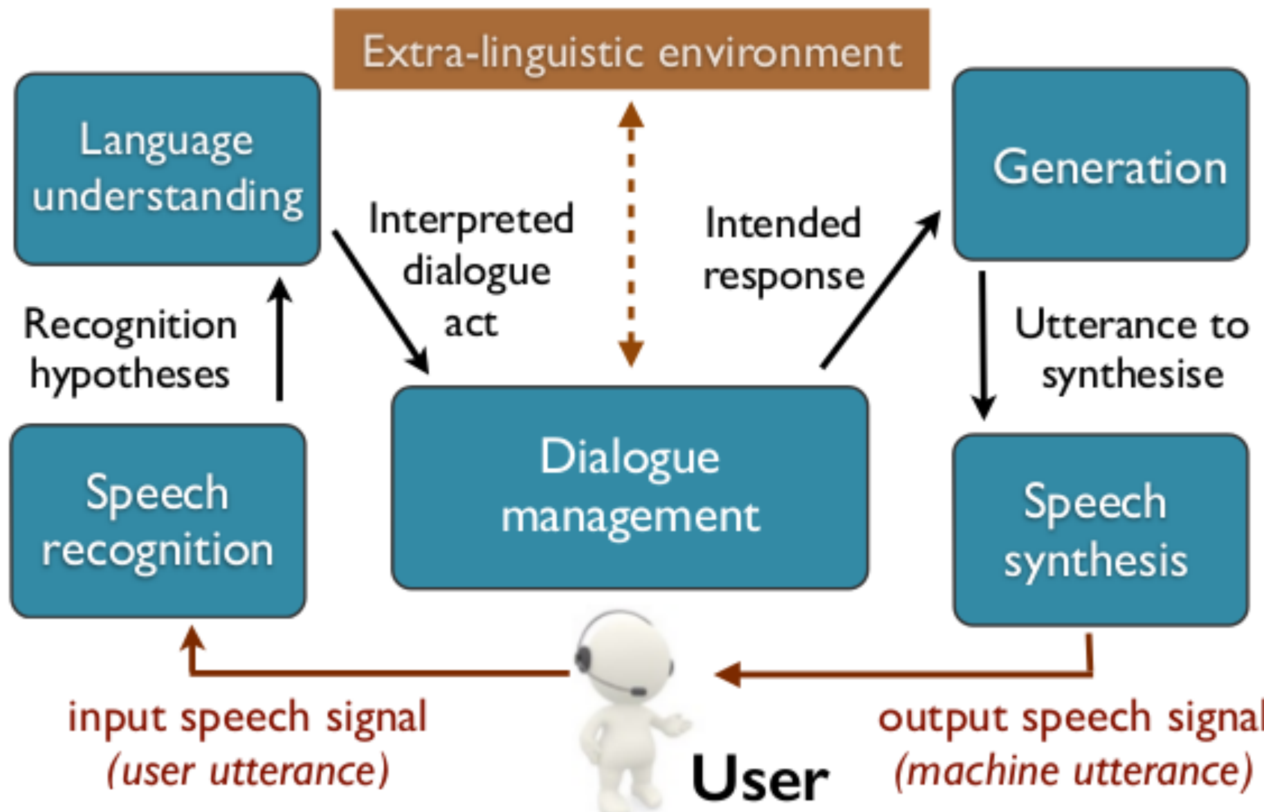
U:	tell me ... how long have you had this falling sickness ?
S:	i'm not a doctor.
U:	How long does it take you guys to learn the drill?
S:	Like 2 weeks ago!!
U:	Where do you live?
S:	I live in England.
U:	In which city do you live?
S:	I live in Los Angeles.

[adapted from Jurafsky & Martin, SLP-3]

Problems:

- No real understanding, lack of coherence / consistency.
- Reproducing general surface patterns the training data.

Modular dialogue systems



(Image by Pierre Lison)

Dialogue act recognition (given an annotated corpus):

- Extract features and train a supervised classifier such as an SVM or a HMM.
- Neural approach: use your favourite RNN to represent an utterance and train it to recognise dialogue acts.

Task-oriented dialogue

Two main types of dialogue:

- Open-domain chit-chat dialogue
- **Task-oriented dialogue**

GUS:	Hello. My name is GUS. I can help you plan a simple trip by air. Where do you want to go ?
Client:	I want to go to San Diego on May 28
GUS:	What time do you want to leave?
Client:	I must be in San Diego before 10 am
GUS:	Would you like P.S.A. flight 102 that arrives at 9.15 am ?
Client:	What is the next flight ?
GUS:	Air California flight 310 that leaves at 8.30am and arrives at 10.00am

Modular architectures are common for task-oriented dialogue.

Task-oriented dialogue

Two main types of dialogue:

- Open-domain chit-chat dialogue.
- **Task-oriented dialogue**
  - Need to keep track of the dialogue state (what has been accomplished, what's missing to achieve the goal, etc)
  - A task restricts the range of relevant dialogue acts.
  - Easier to evaluate: task success.

Task-oriented visual dialogue

Is it a person?	*No*
Is it an item being worn or held?	*Yes*
Is it a snowboard?	*Yes*
Is it the red one?	*No*
Is it the one being held by the person in blue?	*Yes*

Is it a cow?	*Yes*
Is it the big cow in the middle?	*No*
Is the cow on the left?	*No*
On the right ?	*Yes*
First cow near us?	*Yes*

(De Vries et al. 2017)

- Referential task: identify target object.
- Dialogue about visual content — grounded in perception.

Challenges of Dialogue

All levels of linguistic analysis (morphology, syntax, semantics, discourse…) are at play — plus more:

- Both ***understanding*** and ***generation***.
- Coordination among dialogue participants:
  - **When** to speak (turn taking)
  - **What** to say (content, function, coherence)
  - **How** to say it (style, adaptation)

How: style & adaptation

Participants in dialogue coordinate on how to use language.

Dialogue is a form of **joint action**: and instance of two or more agents coordinating to achieve a joint outcome.

Not only in language!



Adaptation

Speakers in dialogue tend to align or adapt to each other at different levels:

- Gestures and postural sway
- Speech rate
- Syntactic structures
- Lexical choice

Adaptation

Speakers in dialogue tend to align or adapt to each other at different levels:

- Gestures and postural sway
- Speech rate
- Syntactic structures
- Lexical choice

Different factors behind this:

- Priming
- Contributes to achieving mutual understanding

Lexical choice

- To coordinate, participants rely on their shared linguistic experience — their **common ground**.
- According to Clark (1996), common ground can be:
  - **Communal**: knowledge shared in virtue of belonging to the same social community.
  - **Personal**: knowledge shared by personally interacting with a a given speaker.
- Speakers anticipate what their dialogue partner knows and plan their utterances accordingly.

Lexical choice

Example of some of our recent work visually grounded dialogue:

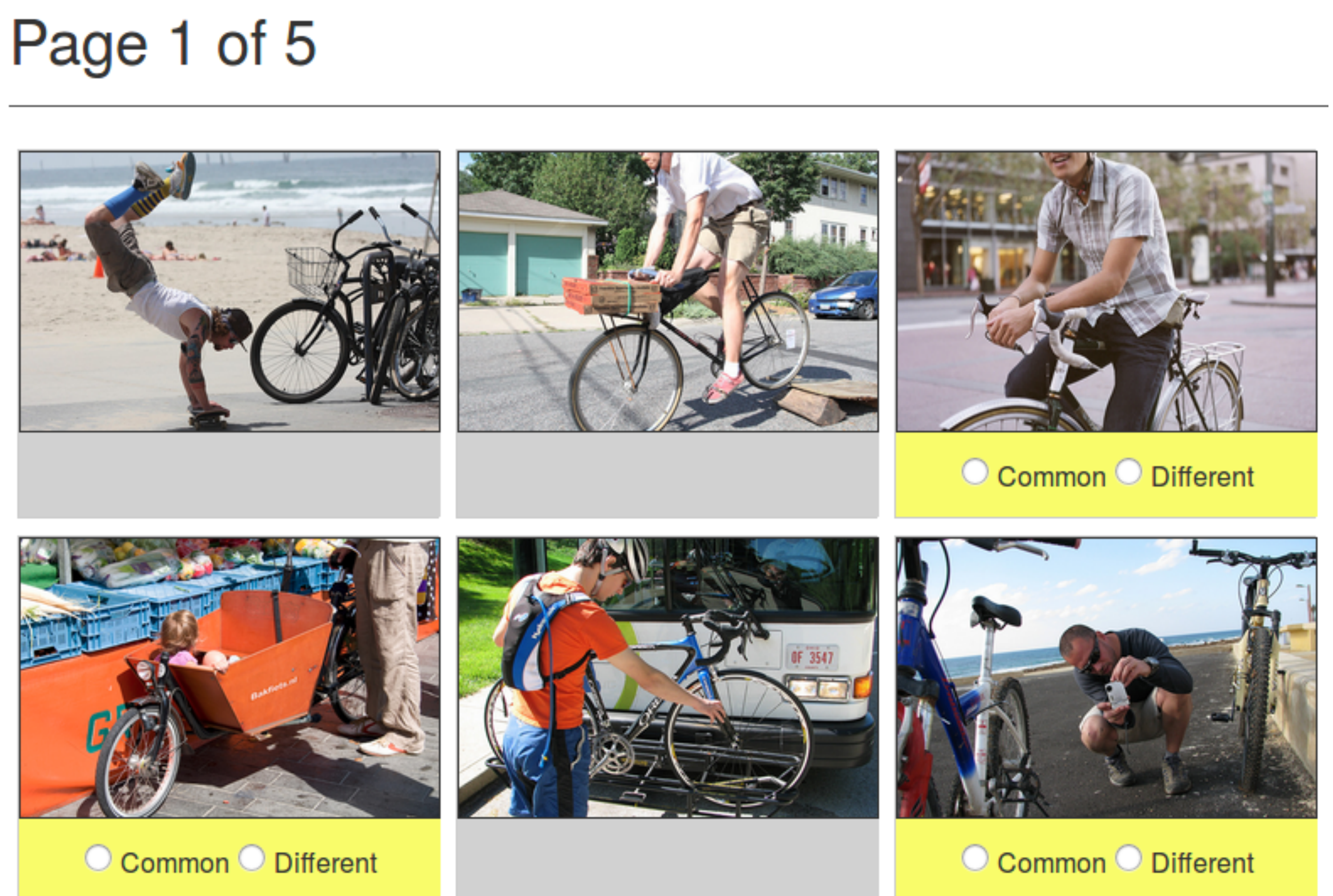
- Alignment of referring expressions
- Exploitation of common ground

Haber et al. The PhotoBook dataset: Building common ground through visually grounded dialogue. ACL 2019.

`https://dmg-photobook.github.io`

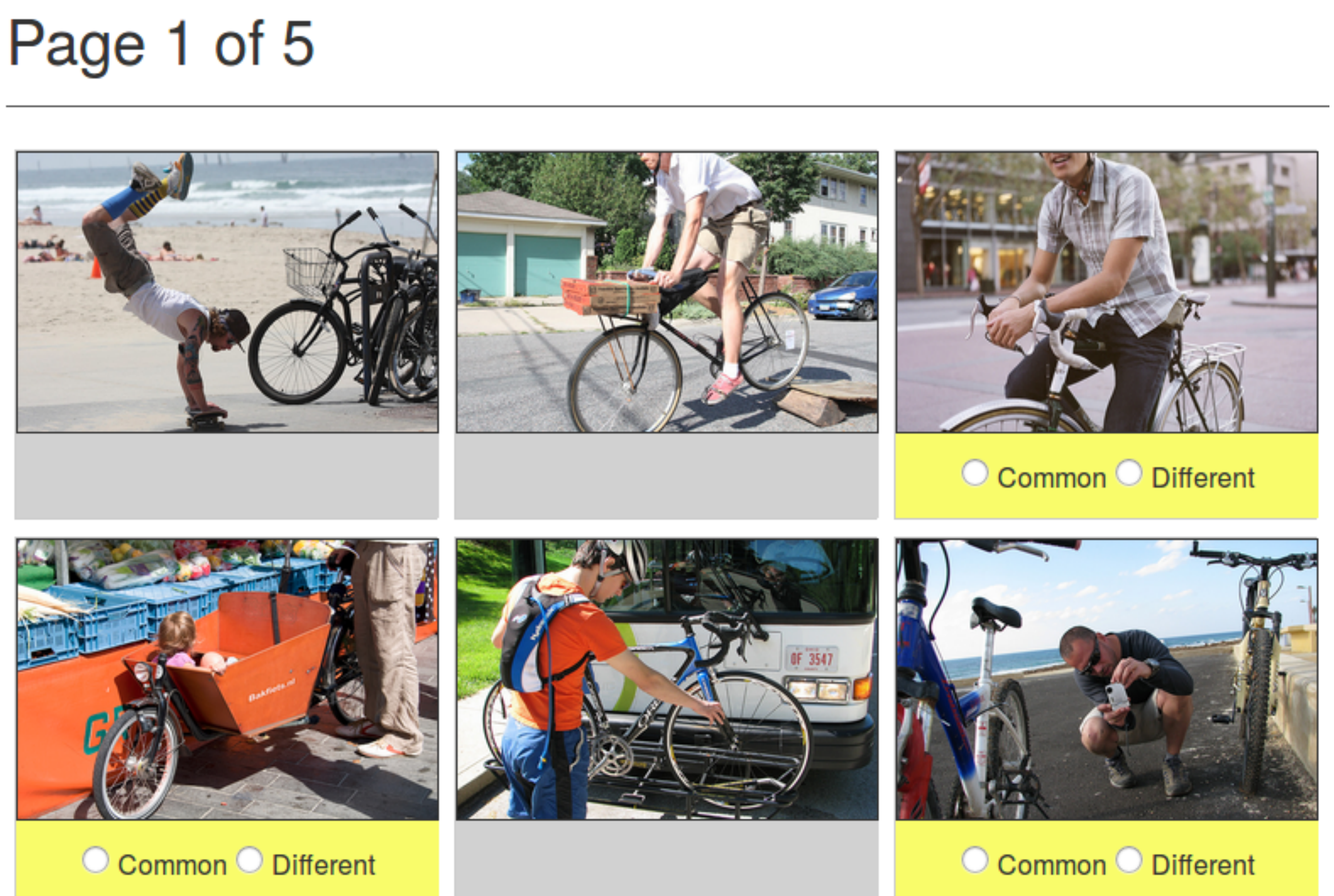
PhotoBook task

Two participants see six photos each, and need to find out which of three highlighted photos they have in common.



PhotoBook task

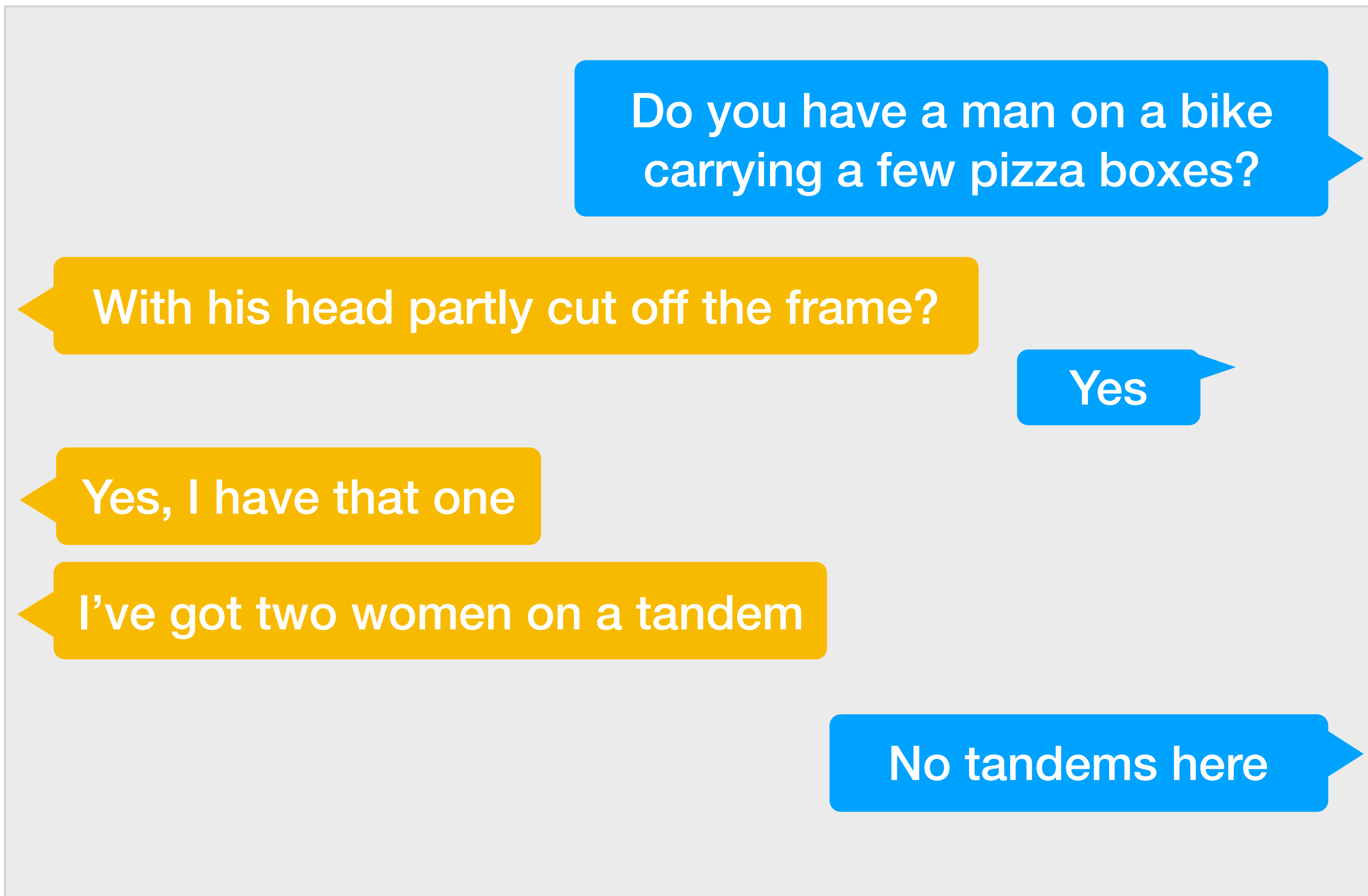
Two participants see six photos each, and need to find out which of three highlighted photos they have in common.



- **Encouraging natural dialogue.** Participants can chat freely and do not have pre-defined roles.

PhotoBook task

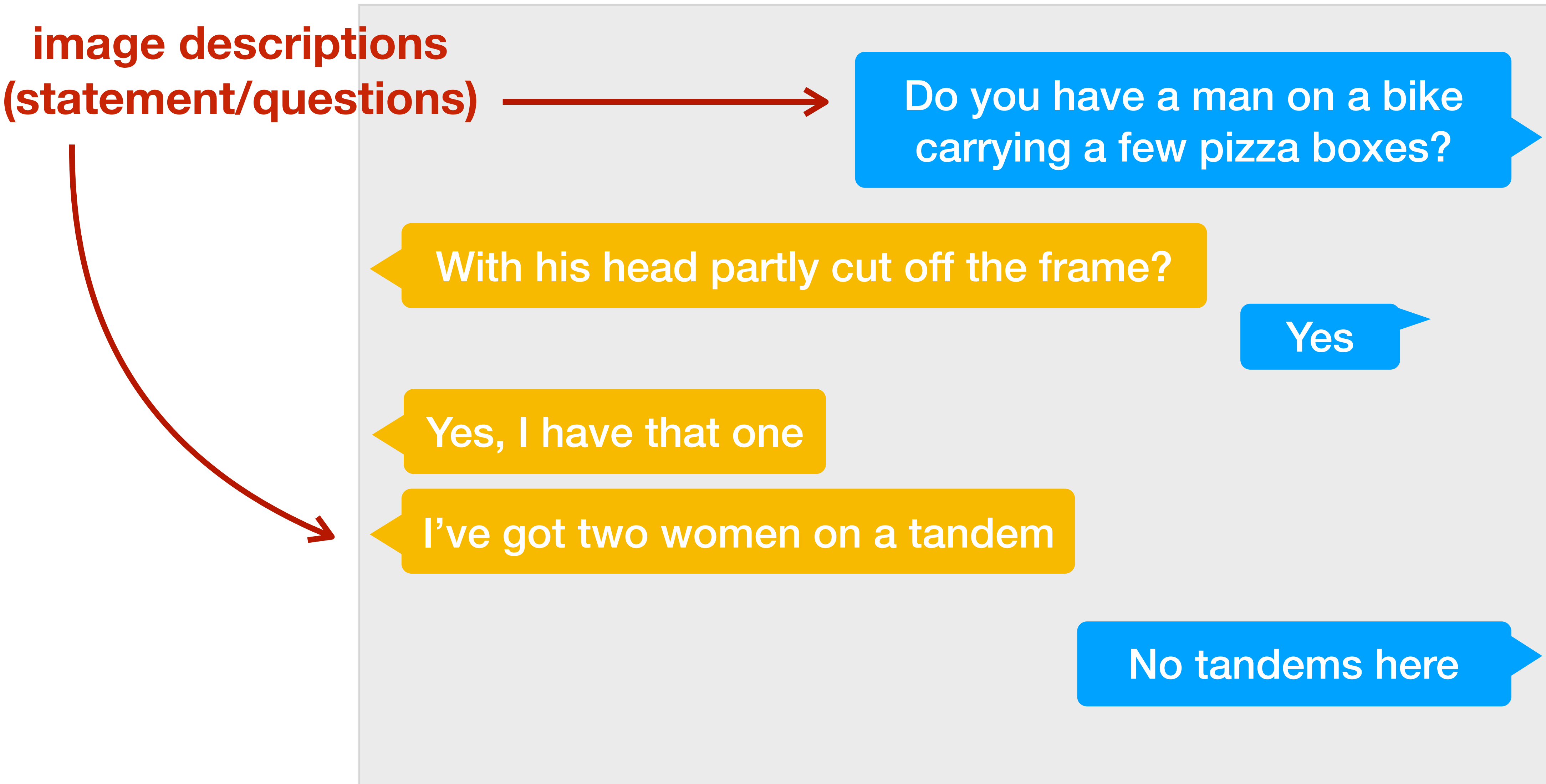
Two participants see six photos each, and need to find out which of three highlighted photos they have in common.



- **Encouraging natural dialogue**. Participants can chat freely and do not have pre-defined roles.

PhotoBook task

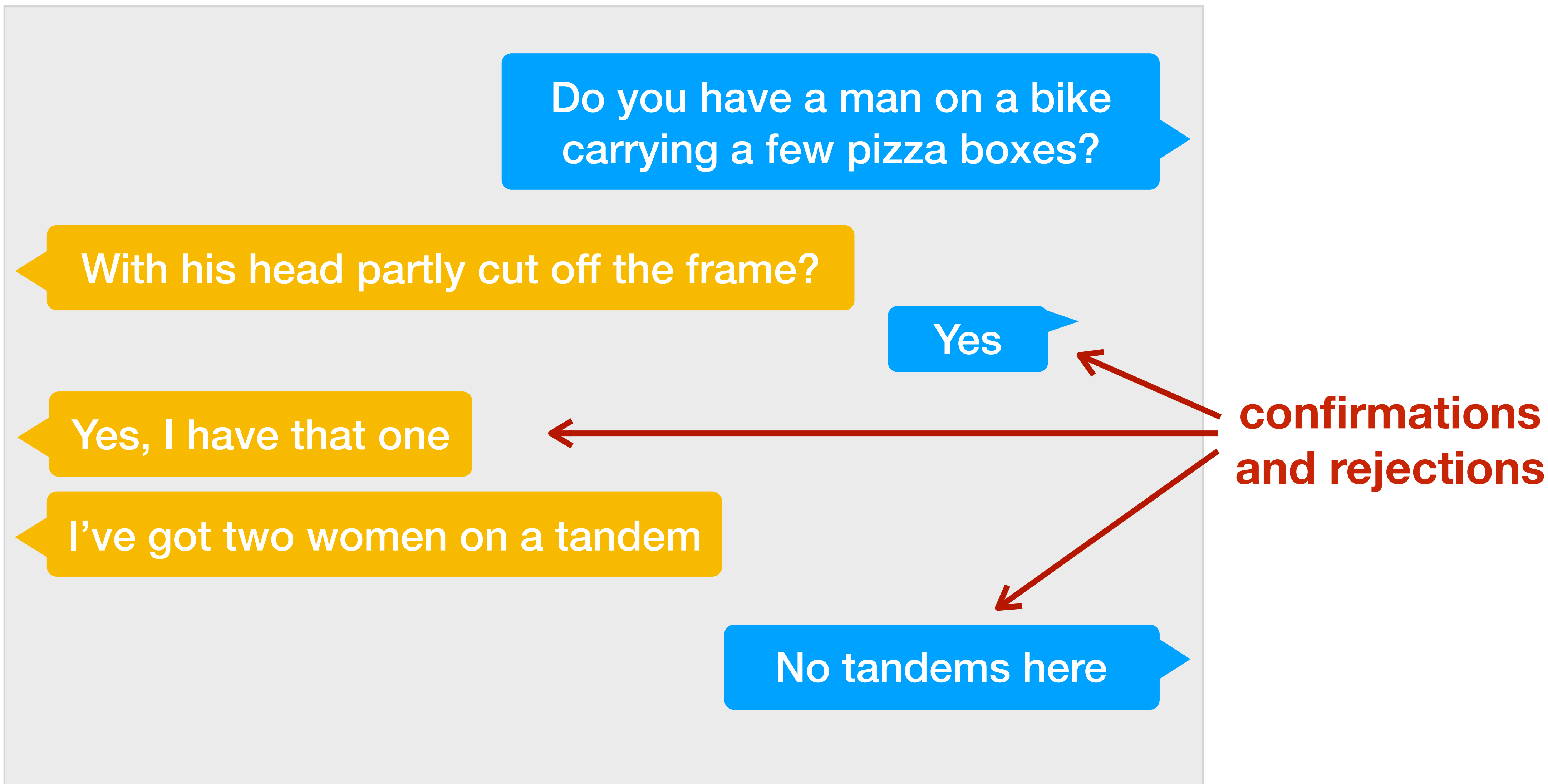
Two participants see six photos each, and need to find out which of three highlighted photos they have in common.



- **Encouraging natural dialogue**. Participants can chat freely and do not have pre-defined roles.

PhotoBook task

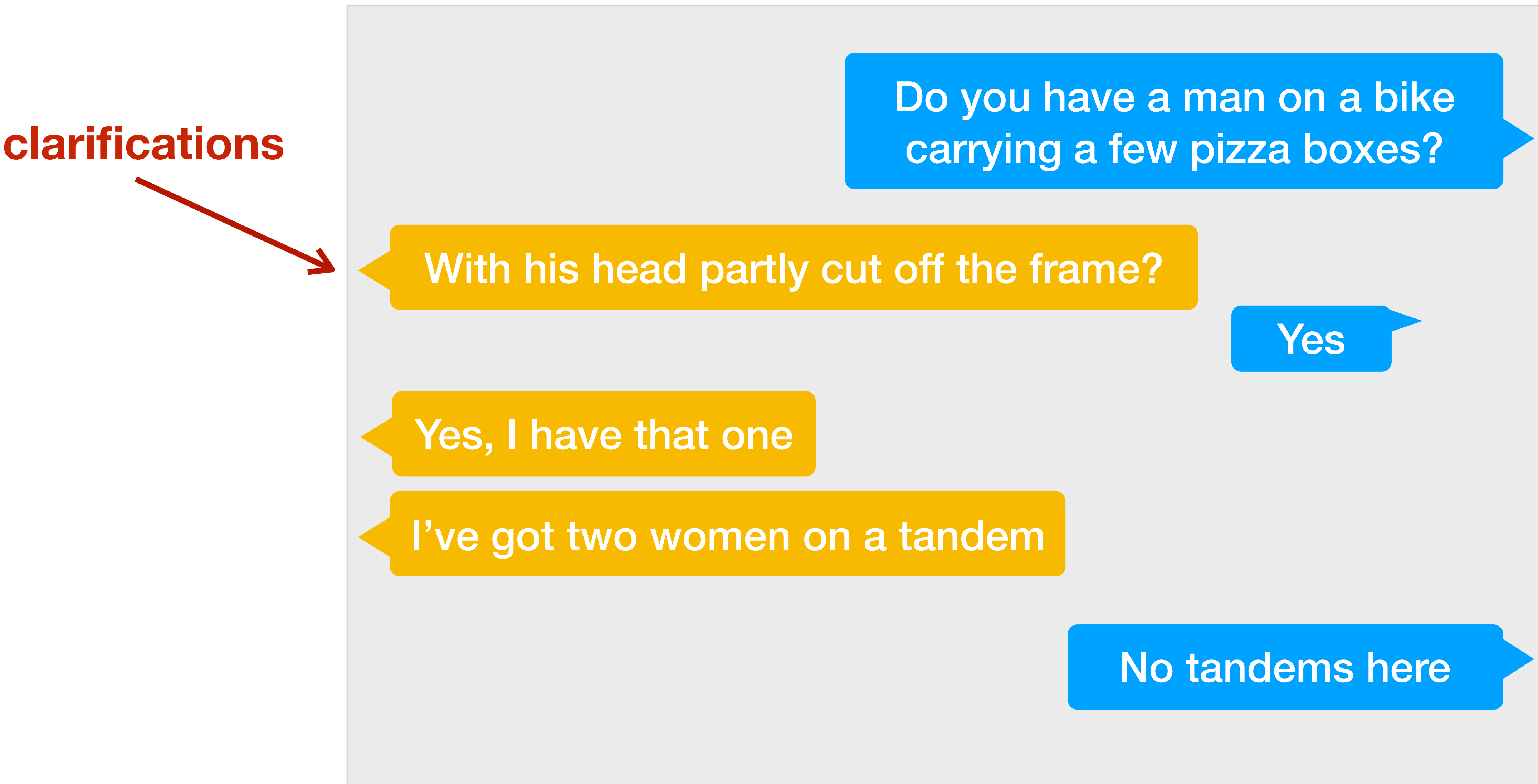
Two participants see six photos each, and need to find out which of three highlighted photos they have in common.



- **Encouraging natural dialogue**. Participants can chat freely and do not have pre-defined roles.

PhotoBook task

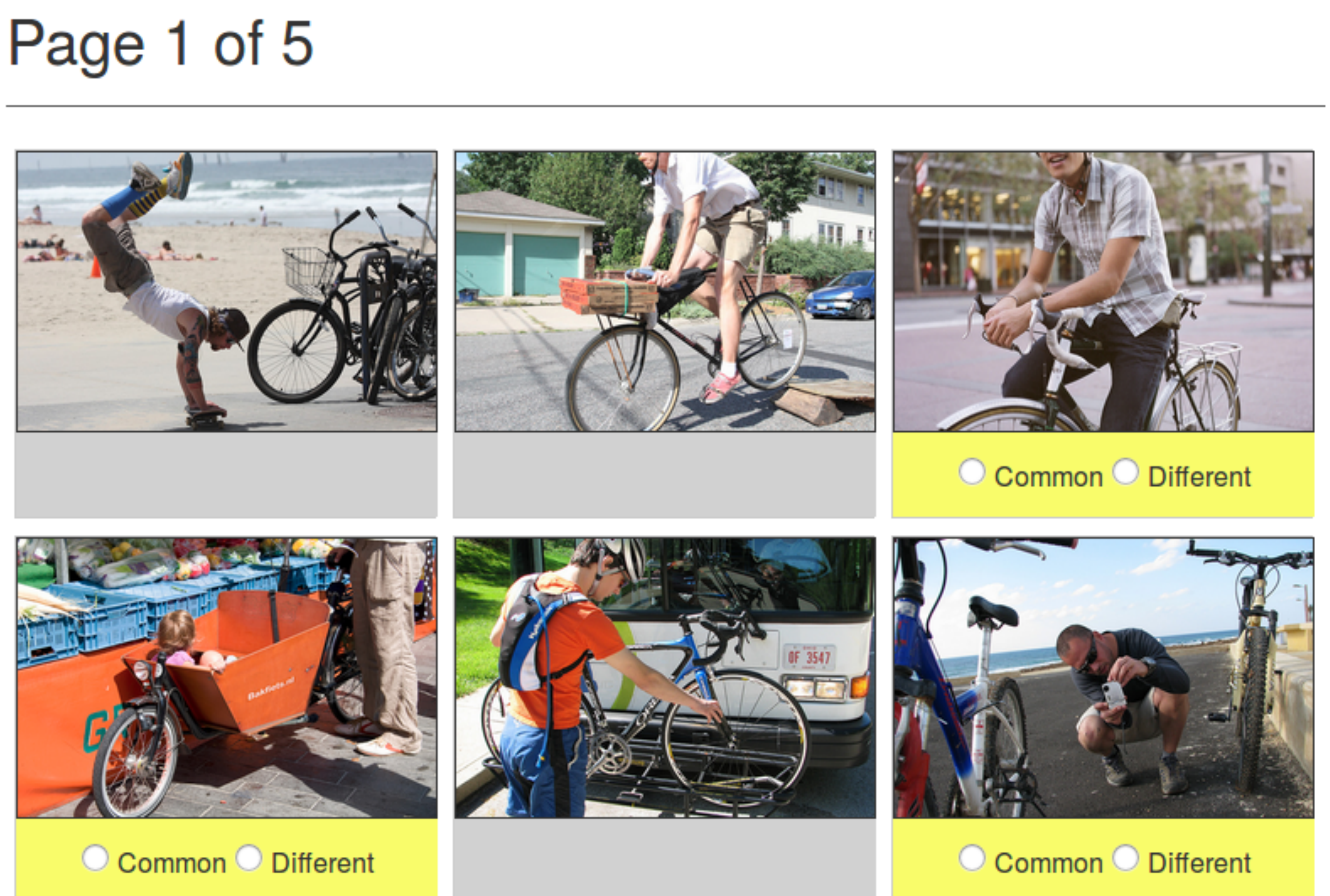
Two participants see six photos each, and need to find out which of three highlighted photos they have in common.



- **Encouraging natural dialogue.** Participants can chat freely and do not have pre-defined roles.

PhotoBook task

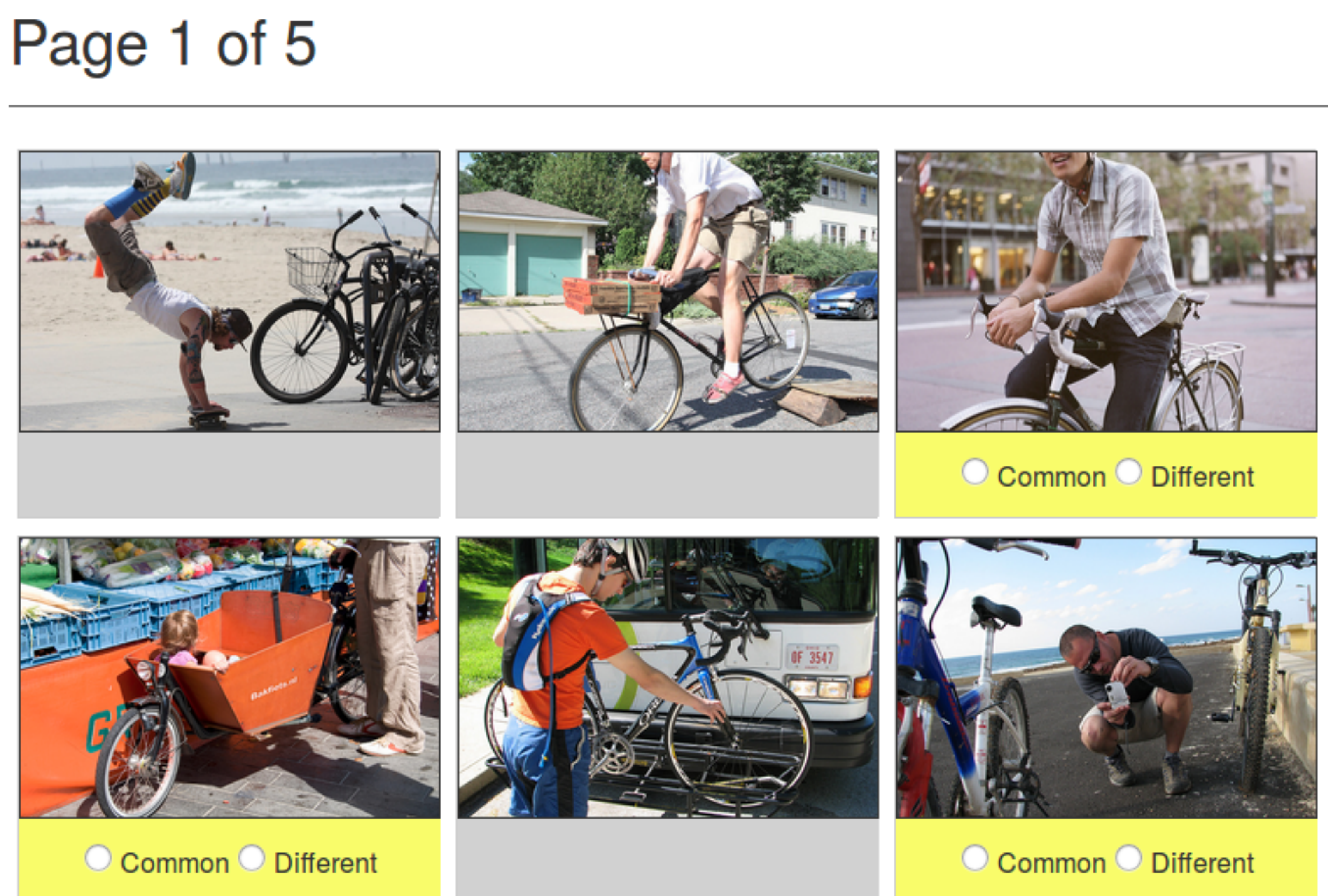
Two participants see six photos each, and need to find out which of three highlighted photos they have in common.



- **Control of the visual context:** Images are similar to each other. They belong to a common domain such "bikes and people".

PhotoBook task

Two participants see six photos each, and need to find out which of three highlighted photos they have in common.



- **Control of the linguistic context:** 5-round game where some images re-occur, inspired by psycholinguistic experiments.

Building common ground

Co-referring descriptions over game rounds | **Referent**

1. ***A:*** *Do you have a boy with a teal coloured shirt with yellow holding a bear with a red shirt?*
2. ***B:*** *Boy with teal shirt and bear with red shirt?*
3. ***A:*** *Teal shirt boy?*

1. ***A:*** *A person that looks like a monk seating on a bench.*
2. *…*
3. *…*
4. ***B:*** *The monk.*

Building common ground

Co-referring descriptions over game rounds | **Referent**

1. ***A:*** *Do you have a boy with a teal coloured shirt with yellow holding a bear with a red shirt?*
2. ***B:*** *Boy with teal shirt and bear with red shirt?*
3. ***A:*** *Teal shirt boy?*

1. ***A:*** *A person that looks like a monk seating on a bench.*
2. *…*
3. *…*
4. ***B:*** *The monk.*

- **First descriptions** are somewhat similar to image captions.

Building common ground

Co-referring descriptions over game rounds | **Referent**

1. ***A:*** *Do you have a boy with a teal coloured shirt with yellow holding a bear with a red shirt?*
2. ***B:*** *Boy with teal shirt and bear with red shirt?*
3. ***A:*** *Teal shirt boy?*

1. ***A:*** *A person that looks like a monk seating on a bench.*
2. *…*
3. *…*
4. ***B:*** *The monk.*

- **First descriptions** are somewhat similar to image captions.
- **Later descriptions** are strongly dependent on the dialogue context.

Main statistics

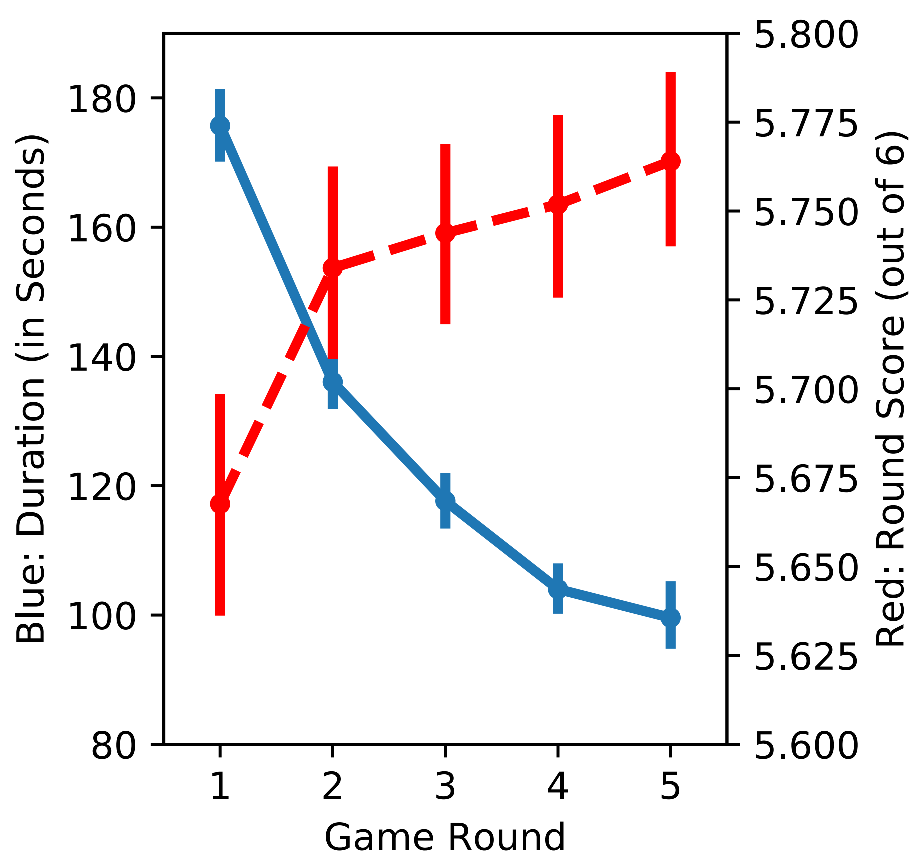
Our data largely confirms observations made by seminal small-scale experiments in psycholinguistics

(Krauss & Weinheimer 1964, Clark & Wilkes-Gibbs 1986, Brennan & Clark 1996, a.o.)

Main statistics

Our data largely confirms observations made by seminal small-scale experiments in psycholinguistics

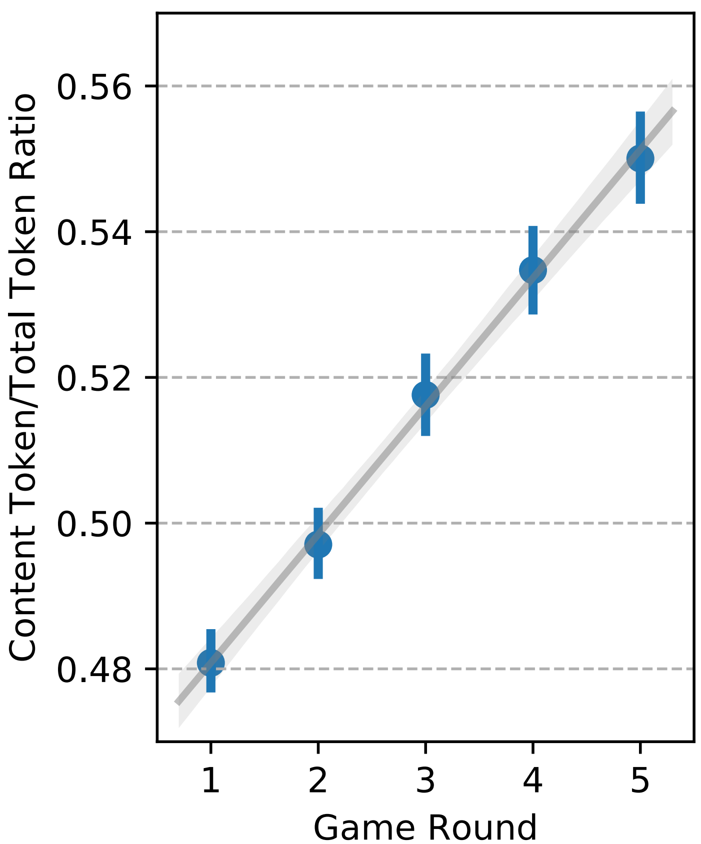
(Krauss & Weinheimer 1964, Clark & Wilkes-Gibbs 1986, Brennan & Clark 1996, a.o.)



Task efficiency

- Number of correct labels increases.
- Completion times get shorter.
- Number of utterances and their length also decreases.

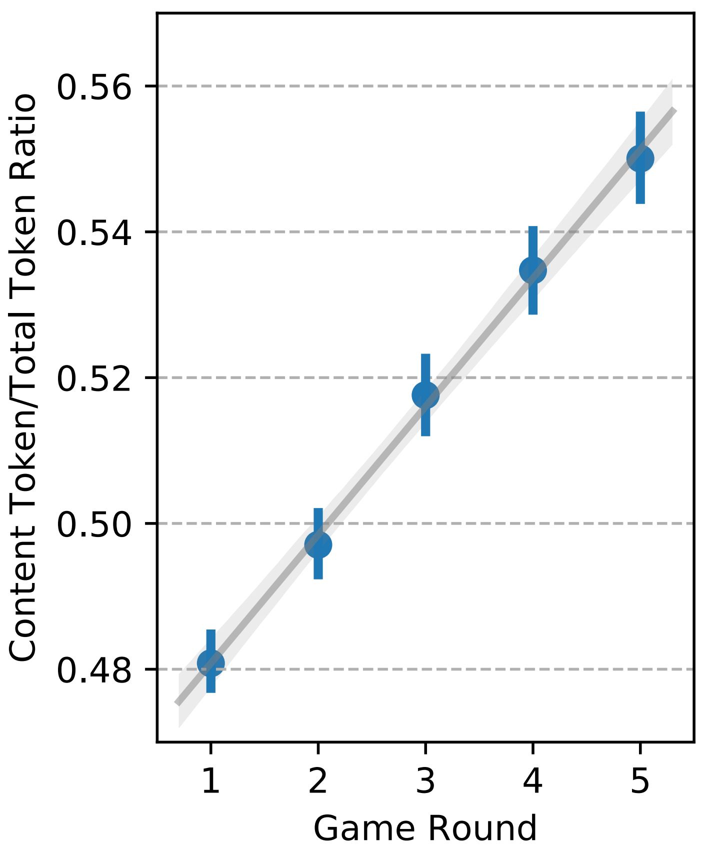
Main statistics

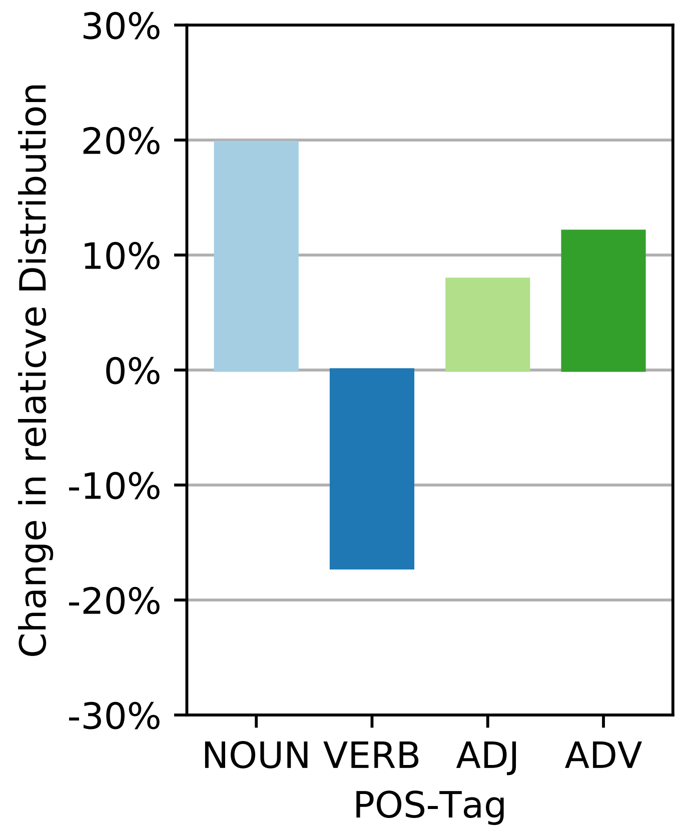


Linguistic properties of utterances

- Increase of content words ratio: shortening, content words remain.

Main statistics

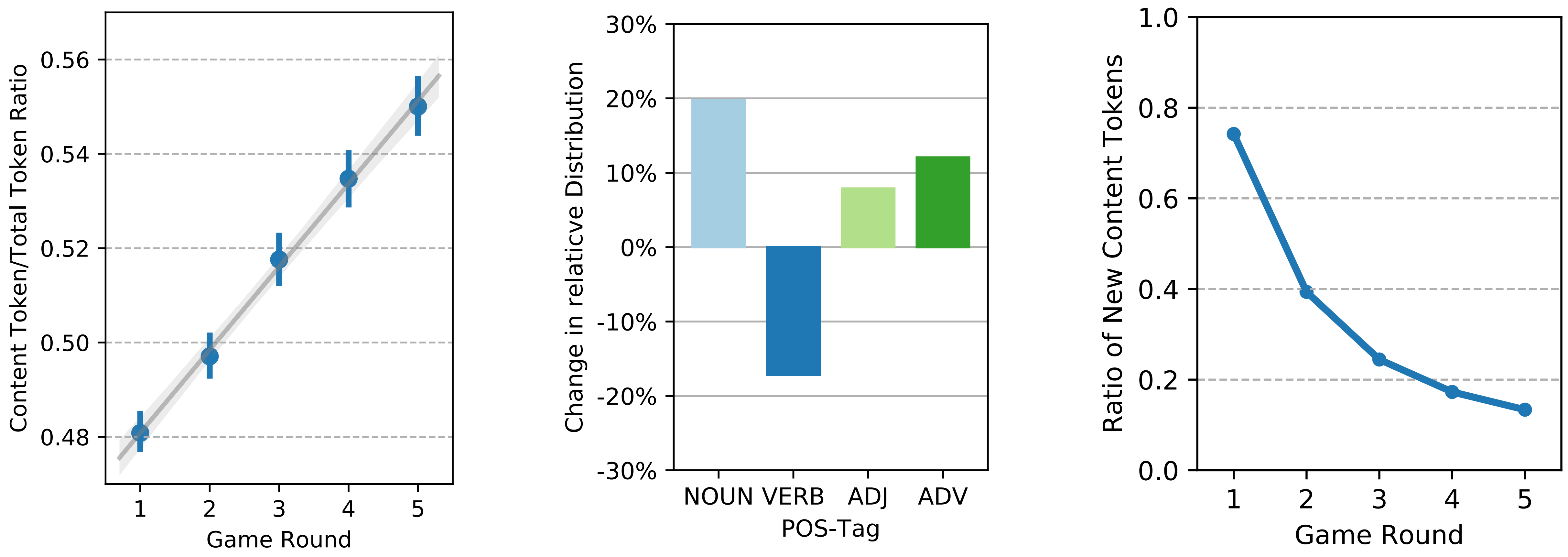




Linguistic properties of utterances

- Increase of content words ratio: shortening, content words remain.
- POS distribution: proportion of nouns and adjectives increases.

Main statistics



Linguistic properties of utterances

- Increase of content words ratio: shortening, content words remain.
- POS distribution: proportion of nouns and adjectives increases.
- Sharp decrease of new content words: *lexical entrainment*.

Reference resolution

Co-referring descriptions over game rounds	Referent
1. ***A:*** *Do you have a boy with a teal coloured shirt with yellow holding a bear with a red shirt?* 2. ***B:*** *Boy with teal shirt and bear with red shirt?* 3. ***A:*** *Teal shirt boy?*	

Reference resolution

Co-referring descriptions over game rounds

1. ***A:*** *Do you have a boy with a teal coloured shirt with yellow holding a bear with a red shirt?*
2. ***B:*** *Boy with teal shirt and bear with red shirt?*
3. ***A:*** *Teal shirt boy?*

Referent

If later descriptions rely on conversational common ground, they should be more difficult to resolve without dialogue history.

Reference resolution

Co-referring descriptions over game rounds	**Referent**
1. ***A:*** *Do you have a boy with a teal coloured shirt with yellow holding a bear with a red shirt?* 2. ***B:*** *Boy with teal shirt and bear with red shirt?* 3. ***A:*** *Teal shirt boy?*	

If later descriptions rely on conversational common ground, they should be more difficult to resolve without dialogue history.

We develop two baseline reference resolution models:
No-History vs. **History**

Reference chain extraction

We exploit labelling actions to extract co-referring **dialogue segments** over game rounds.

***A:** Do you have a boy with a teal coloured shirt with yellow holding a bear with a red shirt?*
***B:** The bear wears a shirt?*
***A:** Yes, and glasses.*
***B:** I don't think I have that one.*
`A marks #340332 as different`

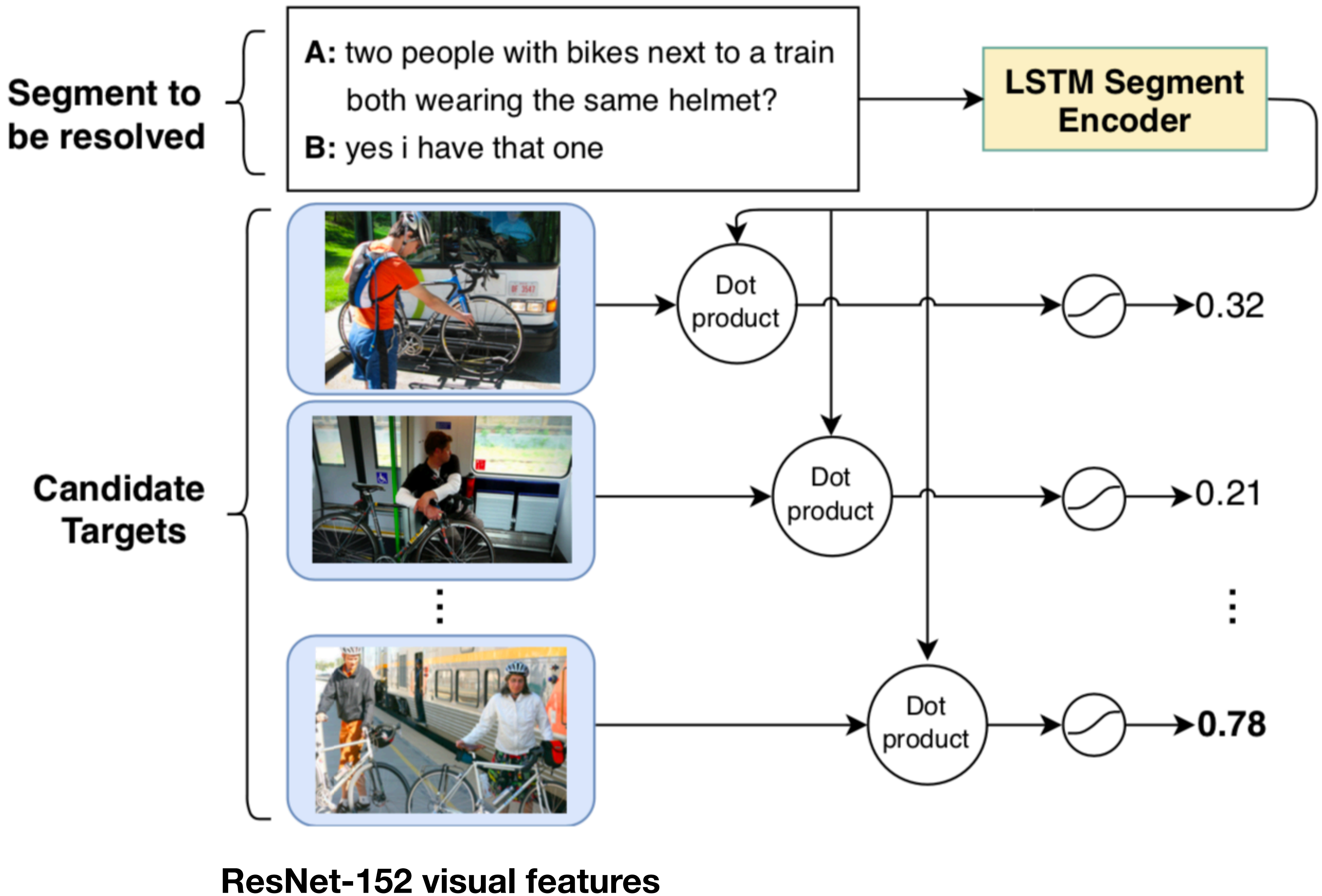
***B:** Boy with teal shirt and bear with red shirt?*
***A:** Yes, I have it.*
`B marks #340332 as common`
`A marks #340332 as common`

***A:** Teal shirt boy?*
***B:** Not this time.*
`A marks #340332 as different`

#340332

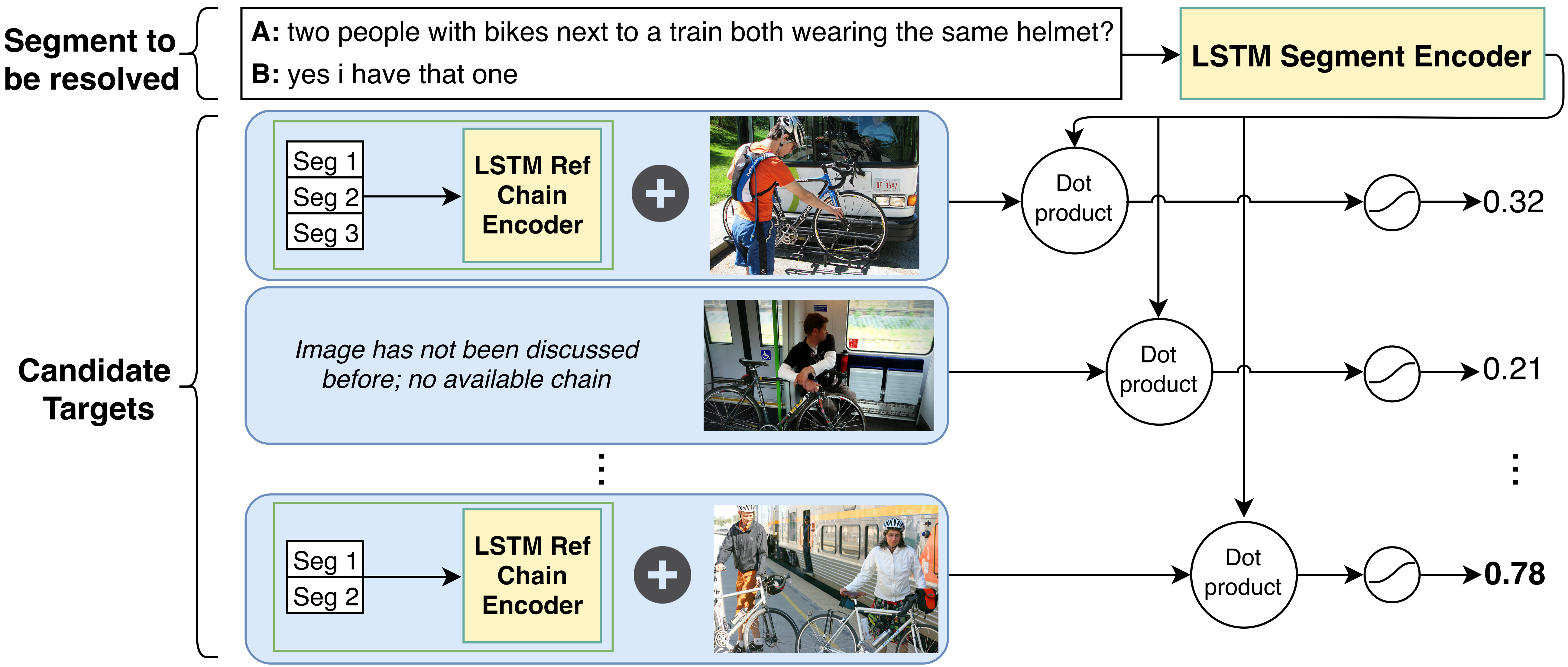
Baseline models

No-History condition



Baseline models

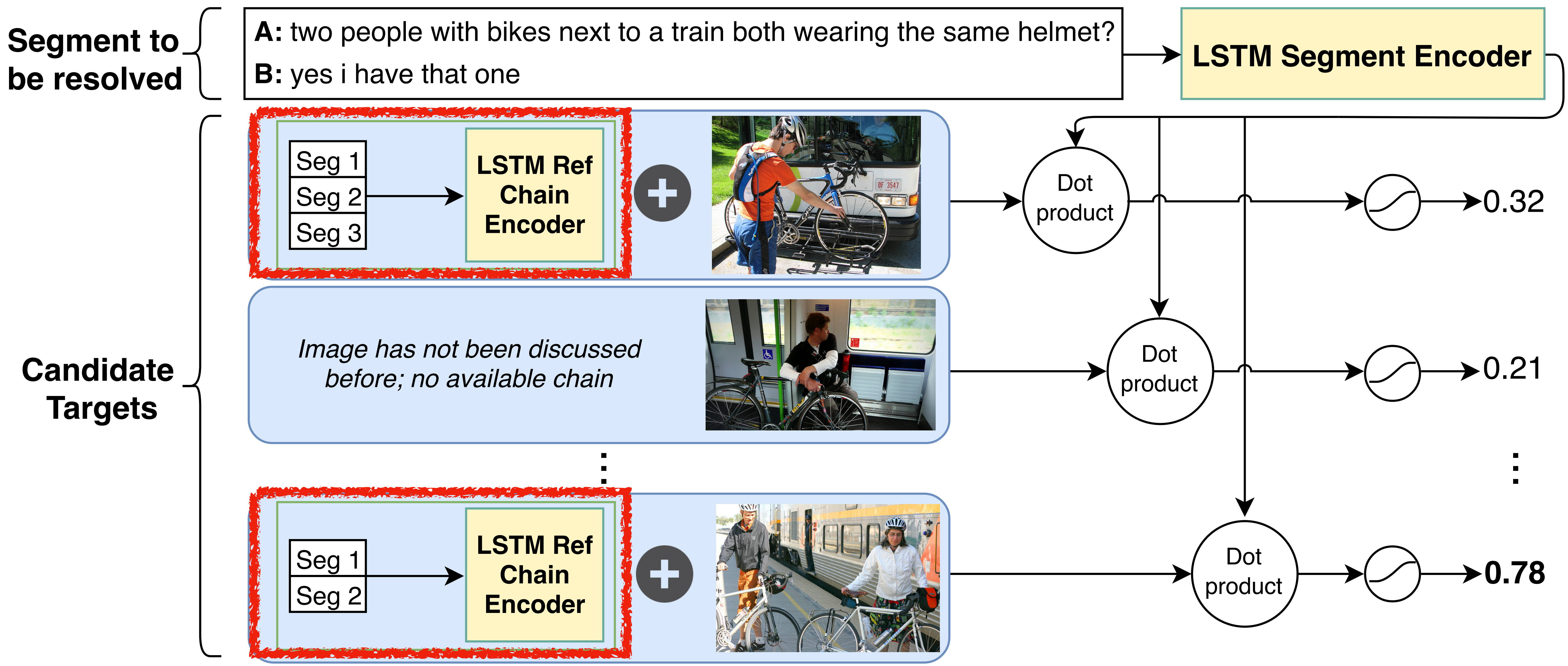
History condition



Besides visual information, each candidate target is represented with **conversational history**: how the image has been referred to before.

Baseline models

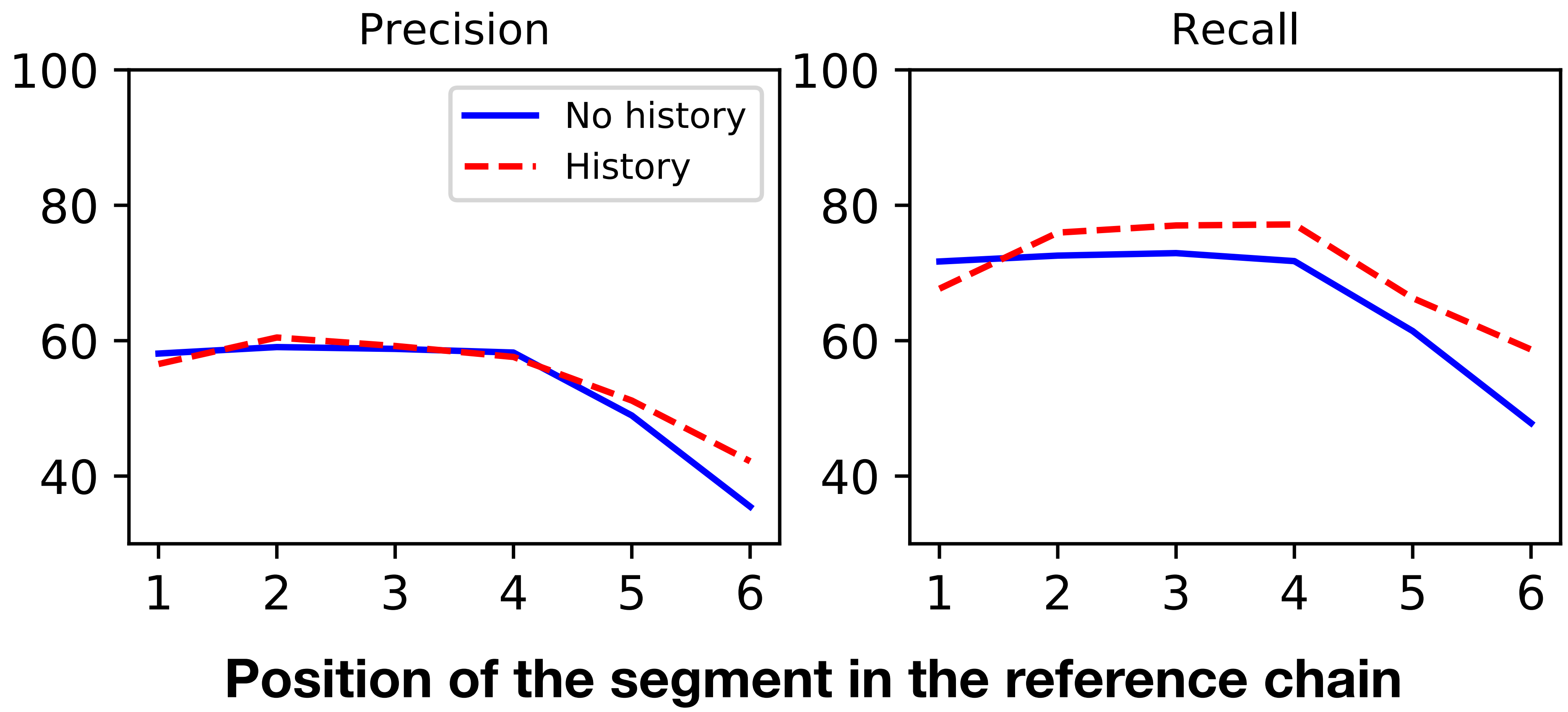
History condition



Besides visual information, each candidate target is represented with **conversational history**: how the image has been referred to before.

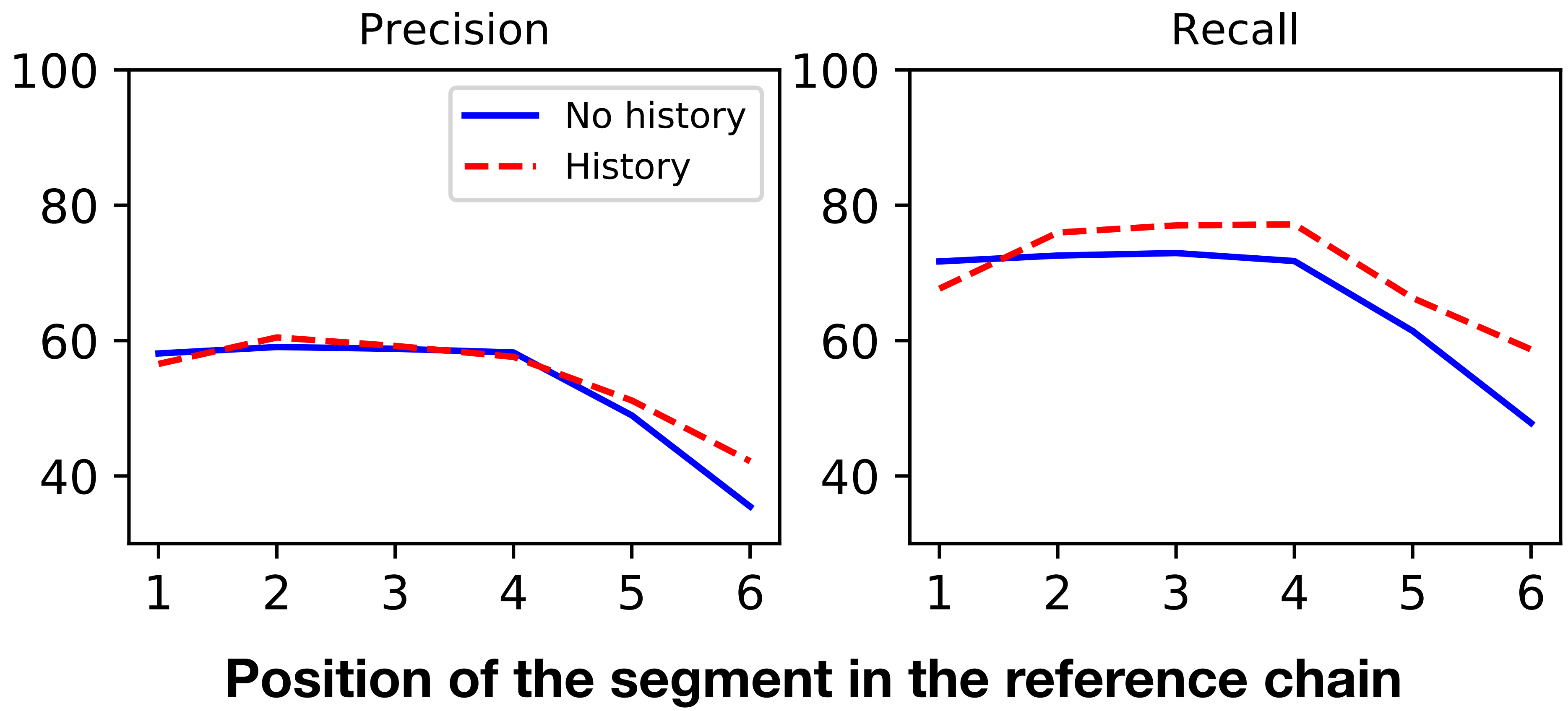
Results

Results for target images in the test set: F1 ~65% (random: 23.5%).



Results

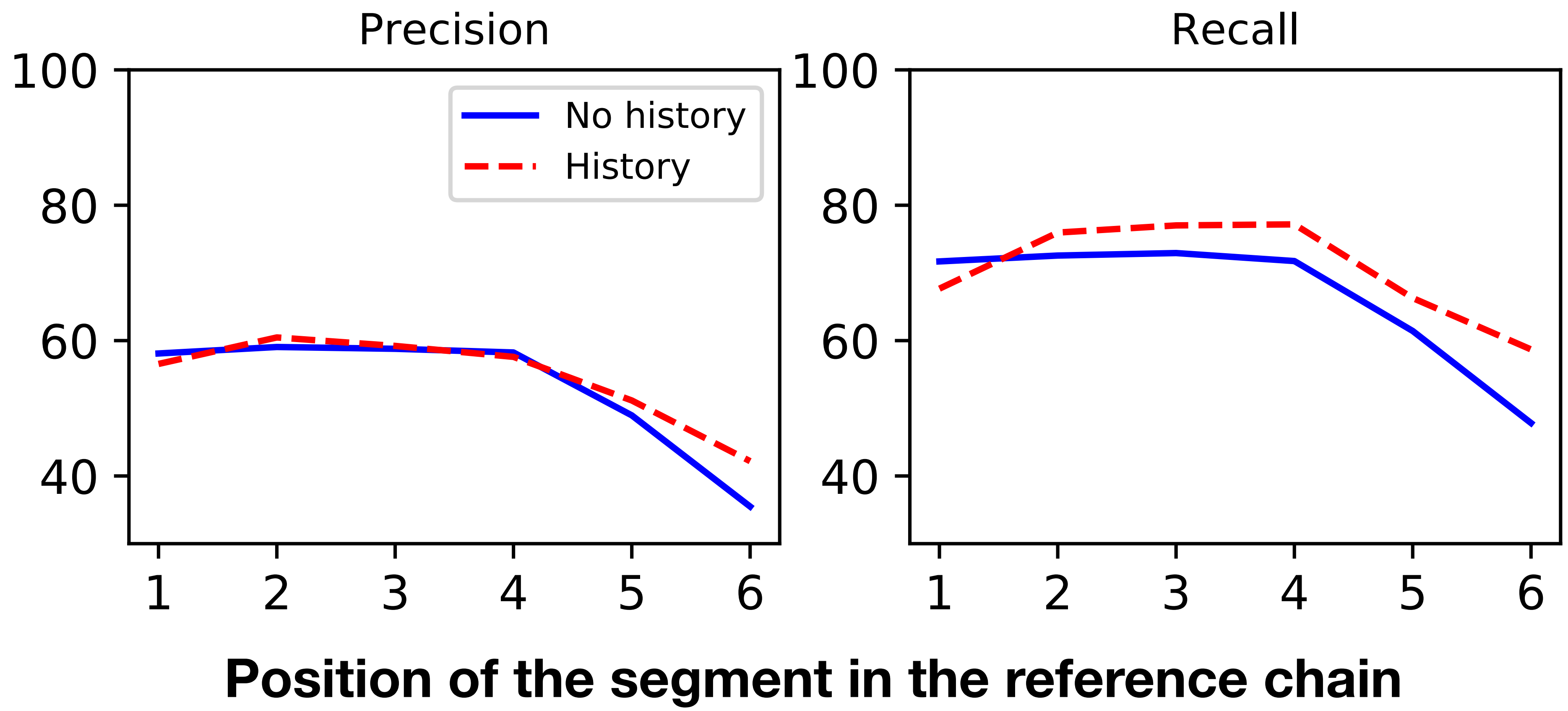
Results for target images in the test set: F1 ~65% (random: 23.5%).



- Later segments are more difficult to resolve for both models.

Results

Results for target images in the test set: F1 ~65% (random: 23.5%).



- Later segments are more difficult to resolve for both models.
- The **History** model achieves higher recall for positions > 1.

Qualitative analysis

When is conversational grounding critical?

Qualitative analysis

When is conversational grounding critical?

- When descriptions are not standard but are strongly visually grounded: both **History** and **No-History** models are effective.

Qualitative analysis

When is conversational grounding critical?

- When descriptions are not standard but are strongly visually grounded: both **History** and **No-History** models are effective.

“I see the carrot lady again”

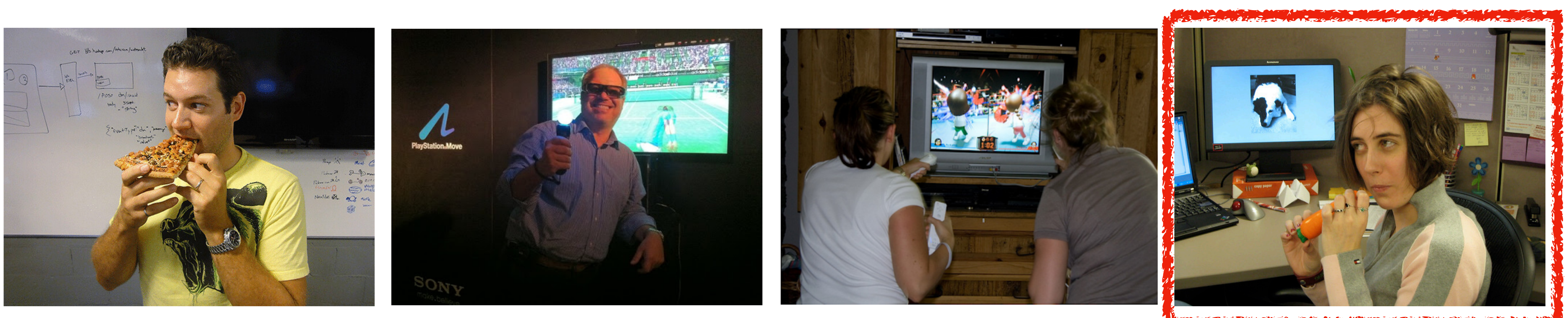
Set of candidate images (person + TV domain)

Qualitative analysis

When is conversational grounding critical?

- When descriptions are not standard but are strongly visually grounded: both **History** and **No-History** models are effective.

“I see the carrot lady again”



Set of candidate images (person + TV domain)

Qualitative analysis

When is conversational grounding critical?

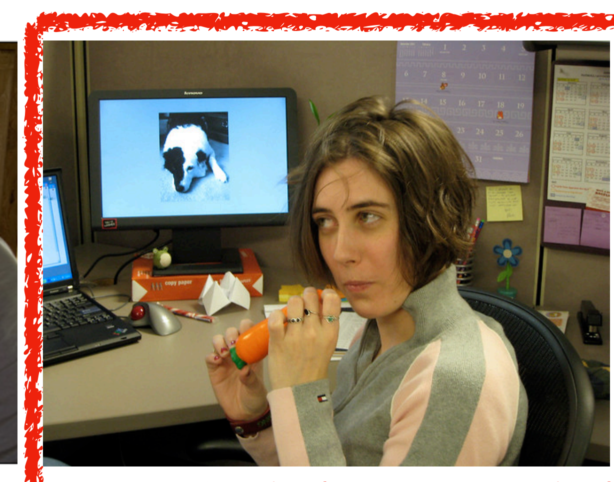
- When descriptions are not standard but are strongly visually grounded: both **History** and **No-History** models are effective.

“I see the carrot lady again”

First description

“A woman seating in front of a monitor with a dog wall paper while holding a plastic carrot”





Set of candidate images (person + TV domain)

Qualitative analysis

When is conversational grounding critical?

Qualitative analysis

When is conversational grounding critical?

- Descriptions relying on more abstract ‘conceptual pacts’ need to be grounded conversationally: **No-History** fails, **History** succeeds.

Qualitative analysis

When is conversational grounding critical?

- Descriptions relying on more abstract ‘conceptual pacts’ need to be grounded conversationally: **No-History** fails, **History** succeeds.

“strange one”

Set of candidate images (person + motorcycle domain)

Qualitative analysis

When is conversational grounding critical?

- Descriptions relying on more abstract ‘conceptual pacts’ need to be grounded conversationally: **No-History** fails, **History** succeeds.

Earlier descriptions

“strange one”

1. “*I have a strange bike with two visible wheels in the back*”
2. “*strange bike again yes*”

Set of candidate images (person + motorcycle domain)

Challenges of Dialogue

All levels of linguistic analysis (morphology, syntax, semantics, discourse…) are at play — plus more:

- Both ***understanding*** and ***generation***.
- Coordination among dialogue participants:
  - **When** to speak (turn taking)
  - **What** to say (content, function, coherence)
  - **How** to say it (style, adaptation)

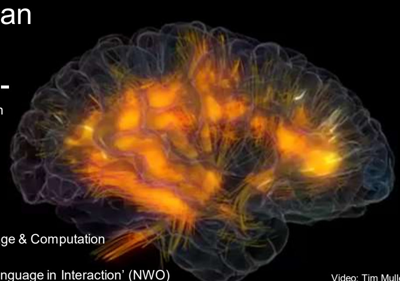
To know more

- Chapters on dialogue in Jurafsky and Martin, 3rd edition.
- Tutorials at recent *ACL conferences.
- Course on **Computational Dialogue Modelling** in block 5.

`http://www.illc.uva.nl/~raquel`

NLP & human language processing -

the virtuous cycle between cognitive science & AI



Willem Zuidema
Institute for Logic, Language & Computation
University of Amsterdam
Senior research fellow 'Language in Interaction' (NWO)

Video: Tim Mullen

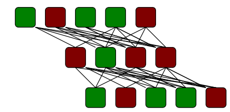
1

Mary knows John knows

Mary knows John lies

PN V PN V

S'



some further technical innovation is called for in neural network models, which will permit them to encode typed variables and the operation of instantiating them. (Jackendoff, 2002)

a few people in the connectionist paradigm have addressed these questions of combinatoriality [, but] none of these proposals have achieved currency in that community. (Jackendoff, 2007)

2

LSTMs

Karpathy, 2015: Character-based LSTM's generating bracket languages and Shakespeare

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.
```

3

Neural language models: predicting next word

Model	Test Perplexity
Sigmoid-RNN-2048 (Ji et al., 2015a)	68.3
Interpolated KN 5-gram, 1.1B n-grams (Chelba et al., 2013)	67.6
Sparse Non-Negative Matrix LM (Shazeer et al., 2015)	52.9
RNN-1024 + MaxEnt 9-gram features (Chelba et al., 2013)	51.3
LSTM-512-512	54.1
LSTM-1024-512	48.2
LSTM-2048-512	43.7
LSTM-8192-2048 (No Dropout)	37.9
LSTM-8192-2048 (50% Dropout)	32.2
2-Layer LSTM-8192-1024 (BIG LSTM)	30.6
BIG LSTM+CNN Inputs	**30.0**

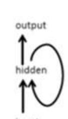
Jozefowicz et al'16, 1B words benchmark

4

Background: Gating in Recurrent Networks

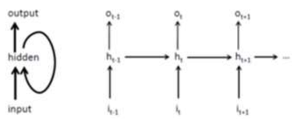
5

Recurrent network



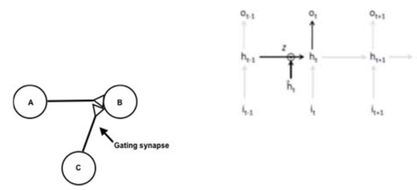
6

Unfold in time



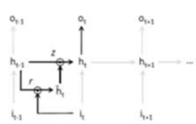
7

Add update gate



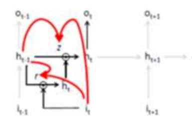
8

Add reset gate



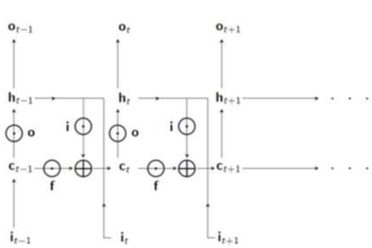
9

Make gates trainable: GRU (Cho et al., 2015)



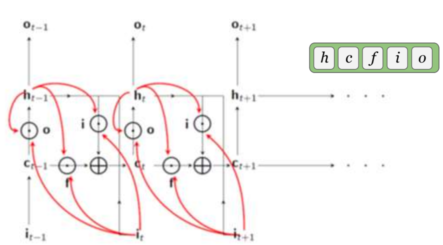
10

Add memory cell & forget gate

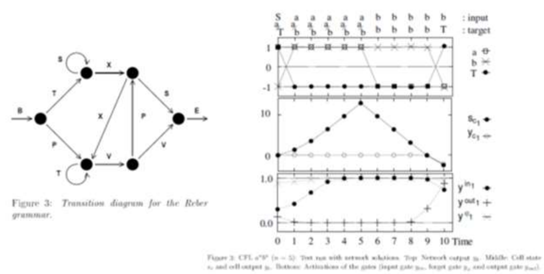


11

Long-short term memory (Hochreiter & Schmidhuber 1997)



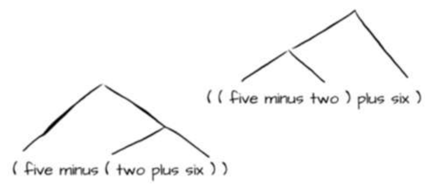
12



(Hochreiter & Schmidhuber 1997) (Gers & Schmidhuber 2001)

13

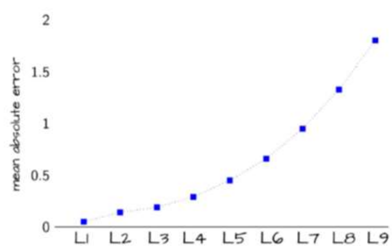
Arithmetic Language



(Veldhoen et al. 2016; Hupkes, Veldhoen & Zuidema, 2018, JAIR)

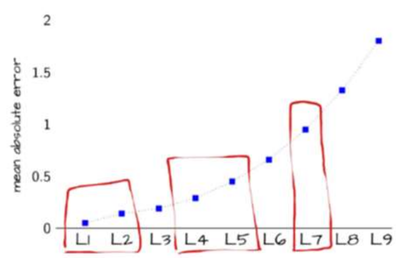
14

Results



15

Results



16

a.k.a. "probing classifiers"

Case study 1: Diagnostic Classification

(Veldhoen et al. 2016; Hupkes, Veldhoen & Zuidema, 2018, JAIR)

How do neural language models represent grammar, and how do we find out?

Giulianelli, Harding, Mohnert, Hupkes & Zuidema, 2018
Best Paper award BlackboxNLP @EMNLP

17

Subject-verb agreement

Subject Attractor Verb

Any **bias** in the **articles** almost certainly **relates** to ...

Linzen et al. 2016
Gulordava et al. 2018

18

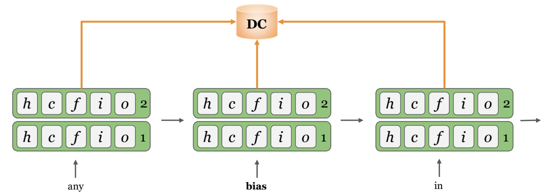
Experimental Setup

- Pretrained Neural Language Model from (Gulordava et al. 2018) with 2 LSTM-layers, with 650 hidden units each
- Wikipedia dependency dataset (Linzen et al. 2016)
- Extract activations for components $\mathbf{h_t}, \mathbf{c_t}, \mathbf{f_t}, \mathbf{i_t}, \mathbf{o_t}$ during forward pass of the LSTM

(Giulianelli, Harding, Mohnert, Hupkes & Zuidema, 2018)

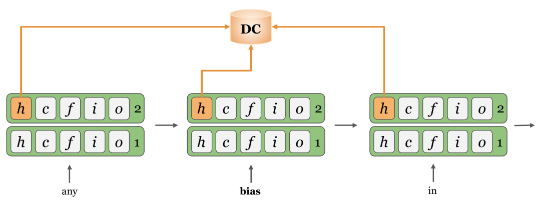
19

Diagnostic Classification



20

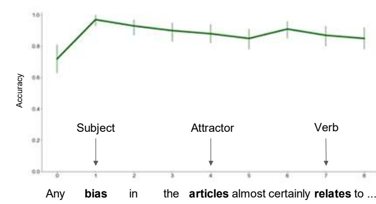
Diagnostic Classification



21

Diagnostic Classification to Predict Number

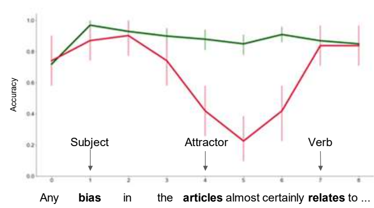
Train: 1000 sentences, context size 5, at least 1 word before subject and at least 1 words after verb.
Test: Two sets of circa 100 sentences with 1 agreement attractor, according to correct/wrong number prediction.



22

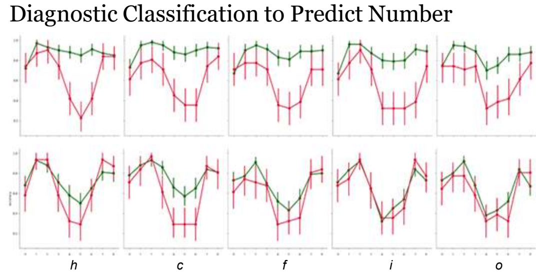
Diagnostic Classification to Predict Number

Train: 1000 sentences, context size 5, at least 1 word before subject and at least 1 words after verb.
Test: Two sets of circa 100 sentences with 1 agreement attractor, according to correct/wrong number prediction.



23

Diagnostic Classification to Predict Number



24

How is number agreement information processed across timesteps?

25


Characterizing the dynamics of mental representations: the temporal generalization method

J-R. King[1,2,3] and S. Dehaene[1,2,4,5]

[1] Cognitive Neuroimaging Unit, Institut National de la Santé et de la Recherche Médicale, U992, F-91191 Gif/Yvette, France
[2] NeuroSpin Center, Institute of BioImaging Commissariat à l'Energie Atomique, F-91191 Gif/Yvette, France
[3] Institut du Cerveau et de la Moelle Épinière Research Center, Institut National de la Santé et de la Recherche Médicale, U975 Paris, France
[4] Université Paris 11, Orsay, France
[5] Collège de France, F-75005 Paris, France

Parsing a cognitive task into a sequence of operations is a central problem in cognitive neuroscience. We argue that a major advance is now possible owing to the

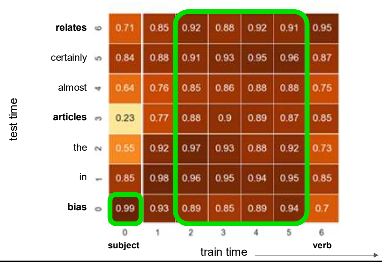
series of steps from those operating as a continuous flow or 'cascade' of overlapping stages [3].

More recently, the advent of brain-imaging techniques

Trends in Cognitive Science, 2014

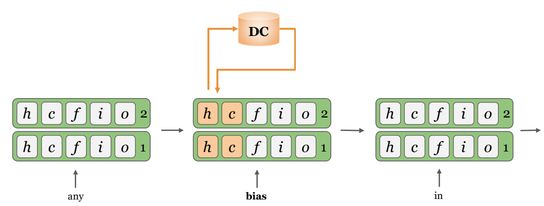
26

Diagnostic Classification to Predict Number



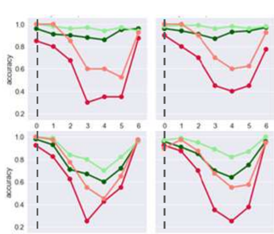
27

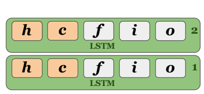
Influencing language models with diagnostic classifiers



28

Diagnostic Classification to Predict Number

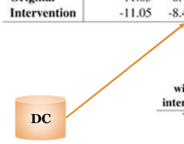




29

Influencing language models with diagnostic classifiers

	An	official	estimate	issued	in	2003	suggests	suggest
Original		-11.05	-8.426	-8.472	-1.243	-3.951	-5.753	**-5.6979**
Intervention		-11.05	-8.426	-8.472	-1.268	-3.97	**-5.691**	-6.4361



without intervention	with intervention
78.0	85.4

30

Take-home points

- Gated Recurrent Neural Networks are capable of learning hierarchical structure, and are an attractive model for how the human brain does it: distributed & using run-of-the-mill circuitry!
- Diagnostic Classifiers allow us to track the dynamics of subject-verb agreement in an LSTM-based language model
- Temporal Generalization Method shows the LSTM represents number information in at least two different ways
- An intervention study allows us to go beyond correlation, but shows a causal role for the representations we identified

31

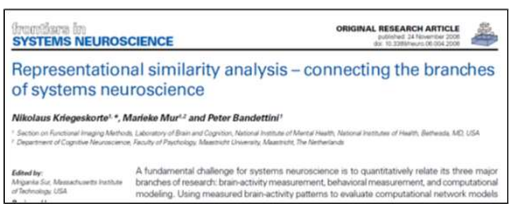
Case study 2: Representational Similarity & Stability Analysis

How similar are representations learned by different models, and how similar are they to representations in the brain?

(Abnar, Beinborn, Choenni & Zuidema, 2019)

BlackboxNLP @ACL2019

32


frontiers in SYSTEMS NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 24 November 2008
doi: 10.3389/neuro.06.004.2008

Representational similarity analysis – connecting the branches of systems neuroscience

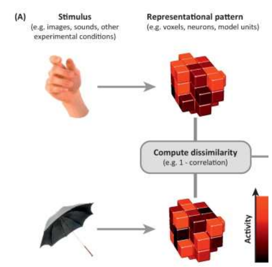
Nikolaus Kriegeskorte[1,*], Marieke Mur[1,2] and Peter Bandettini[1]

[1] Section on Functional Imaging Methods, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA
[2] Department of Cognitive Neuroscience, Faculty of Psychology, Maastricht University, Maastricht, The Netherlands

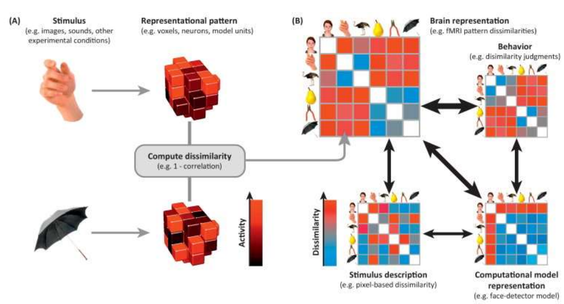
Edited by:
Mriganka Sur, Massachusetts Institute of Technology, USA

A fundamental challenge for systems neuroscience is to quantitatively relate its three major branches of research: brain-activity measurement, behavioral measurement, and computational modeling. Using measured brain-activity patterns to evaluate computational network models

33

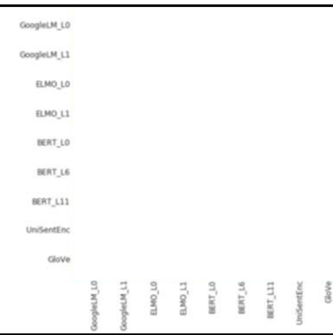


34



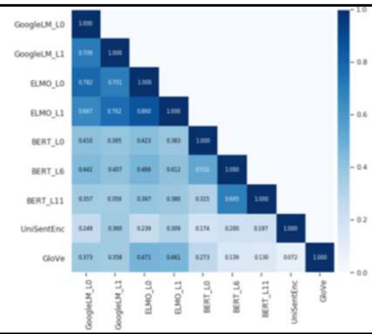
35

RSA across models



36

RSA
across
models



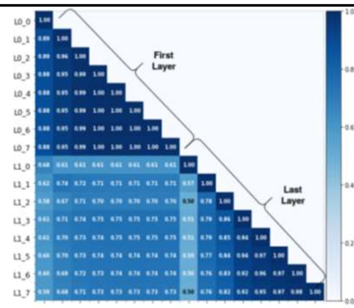
37

Representational
Stability Analysis

(a) GoogleLM

38

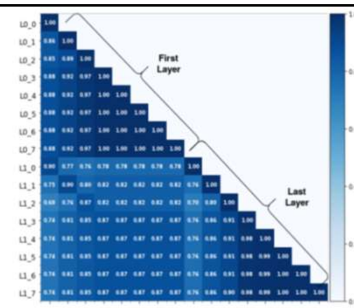
Representational
Stability Analysis



(a) GoogleLM

39

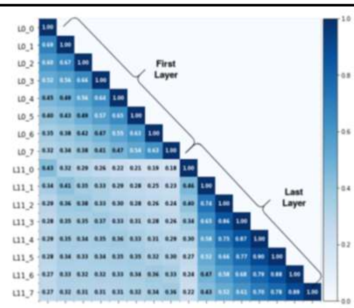
Representational
Stability Analysis



(a) ELMO

40

Representational
Stability Analysis

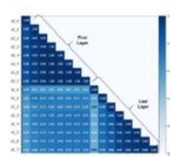


(a) BERT

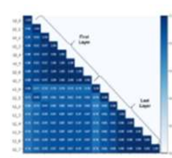
41

Representational Stability Analysis

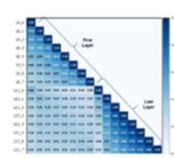
great differences between current deep language models in their dependence on context



(a) GoogleLM



(b) ELMO

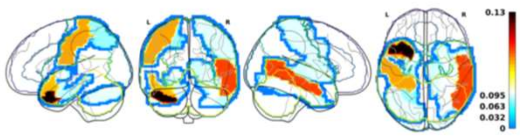


(c) BERT

42

Representational Similarity Analysis

great differences between brain areas in their similarity to the models

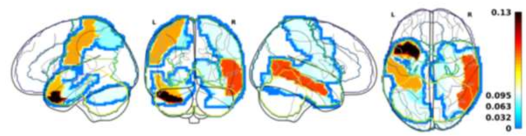


http://projects.illc.uva.nl/LaCo/clclab/

43

Representational Similarity Analysis

great differences between brain areas in their similarity to the models



http://projects.illc.uva.nl/LaCo/clclab/

44

Discussion – Interpretability - How and why?

- All state-of-the-art models in NLP are based on deep learning
- Presents us with the blackbox problem, making it difficult to:
  - Generate *explanations* to users and *justify* decisions based on the systems
  - Allow users to *interact* with the learned solutions and *adapt* them to their needs
  - Use prior knowledge to *augment* machine learned solutions
- Diagnostic classification is a way to test specific hypotheses on what information is represented; should be applied with as much rigor as model testing in (cognitive) neuroscience

45

Interpretability: How and why?

- Representational Similarity Analysis is a way to compare models across paradigms, and test the sensitivity of the learned representations to parameter choices
- There is no silver bullet: the excellent performance of current models is found away from the easily interpretable points in hypothesis space
- We need to systematically apply the ever increasing toolbox of interpretability tools and see how far we get!

46

http://projects.illc.uva.nl/LaCo/clclab/

Lisa Beinborn

Samira Abnar

Mario Giulianelli

Jack Harding

Florian Mohnert

Dieuwke Hupkes

Willem Zuidema
zuidema@uva.nl

47