Grounded language understanding: Overview

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding

Associated materials

1. Code
   a. Notebook: `colors_overview.ipynb`
   b. Homework and bake-off: `hw_colors.ipynb`

2. Core reading: Monroe et al. 2017

3. Auxiliary readings: Golland et al. 2010; Lewis et al. 2017; Andreas and Klein 2016; Tellex et al. 2014; Vogel et al. 2013

HAL

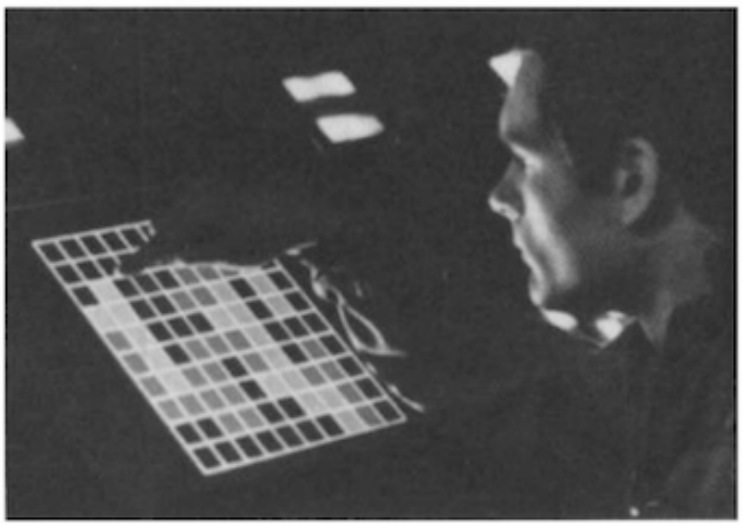
- In the 1967 Stanley Kubrick movie *2001: A Space Odyssey*, the spaceship's computer HAL can
  - display graphics;
  - play chess; and
  - conduct natural, open-domain conversations with humans.
- How well did the filmmakers do at predicting what computers would be capable in 2001?

Slide idea from Andrew McCallum

HAL

Graphics

HAL



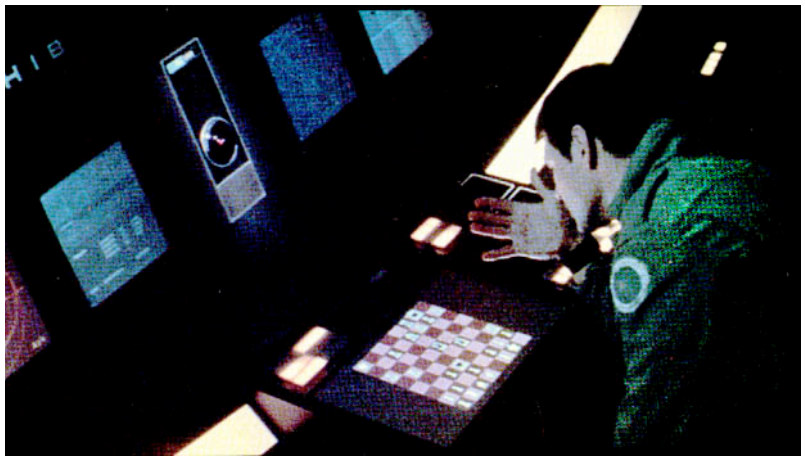
Jurassic Park (1993)

Slide idea from Andrew McCallum

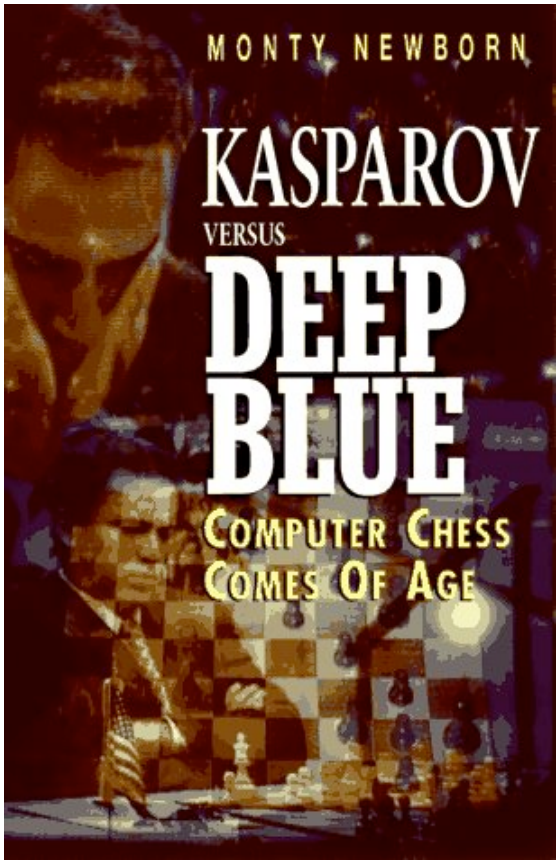
HAL

Chess

HAL



Deep Blue (1997)



Slide idea from Andrew McCallum

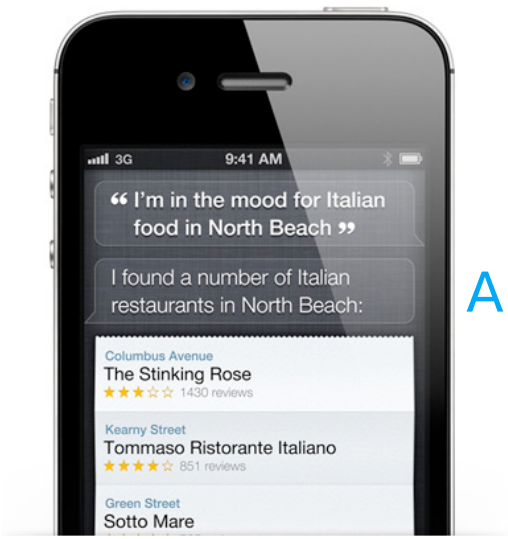
HAL

Dialogue

HAL	2014
David Bowman: Open the pod bay doors, HAL.	. . .
HAL: I'm sorry, Dave, I'm afraid I can't do that.	
David: What are you talking about, HAL?	
HAL: I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.	

Slide idea from Andrew McCallum

Siri



You: Any good burger joints around here?

Siri: I found a number of burger restaurants near you.

You: Hmm. How about tacos?

Apple: [Siri remembers that you asked about restaurants. so it will look for Mexican restaurants in the neighborhood. And Siri is proactive, so it will question you until it finds what you’re looking for.]

Slide idea from Marie de Marneffe

Siri

Colbert: For the love of **God**, the **cameras** are on, give me something?

Siri: What kind of place are you looking for? **Camera stores** or **churches**?

[. . .]

Colbert: I don't want to search for anything! I want to write the show!

Siri: Searching the Web for "search for anything. I want to write the shuffle."

Slide idea from Marie de Marneffe

Levinson’s (2000) analogy

Figure 0.1
Rembrandt sketch

Levinson’s (2000) analogy

Figure 0.1
Rembrandt sketch

“We interpret this sketch instantly and effortlessly as a gathering of people before a structure, probably a gateway; the people are listening to a single declaiming figure in the center. [. . .] But all this is a miracle, for there is little detailed information in the lines or shading (such as there is). Every line is a mere suggestion [. . .]. So here is the miracle: from a merest, sketchiest squiggle of lines, you and I converge to find adumbration of a coherent scene [. . .].

Levinson's (2000) analogy

Figure 0.1
Rembrandt sketch

"We interpret this sketch instantly and effortlessly as a gathering of people before a structure, probably a gateway; the people are listening to a single declaiming figure in the center. [. . .] But all this is a miracle, for there is little detailed information in the lines or shading (such as there is). Every line is a mere suggestion [. . .]. So here is the miracle: from a merest, sketchiest squiggle of lines, you and I converge to find adumbration of a coherent scene [. . .].

"The problem of utterance interpretation is not dissimilar to this visual miracle. An utterance is not, as it were, a veridical model or "snapshot" of the scene it describes [. . .]. Rather, an utterance is just as sketchy as the Rembrandt drawing."

Indexicality

1. I am speaking.
2. We won. [A team I'm on; a team I support; . . .]
3. I am here [office; Stanford; . . . planet earth; . . .]
4. We want to go here. [pointing at a map]
5. We went to a local bar after work.
6. three days ago, tomorrow, now

Context dependence

Where are you from?

Context dependence

Where are you from?

- *Connecticut.* (Issue: birthplaces)
- *The U.S.* (Issue: nationalities)
- *Stanford.* (Issue: affiliations)
- *Planet earth.* (Issue: intergalactic meetings)

Context dependence

I didn’t see any.

Context dependence

- Are there typos in my slides?

I didn’t see any.

Context dependence

- Are there typos in my slides?
- Are there bookstores downtown?

I didn’t see any.

Context dependence

- Are there typos in my slides?
- Are there bookstores downtown?
- Are there cookies in the cupboard?

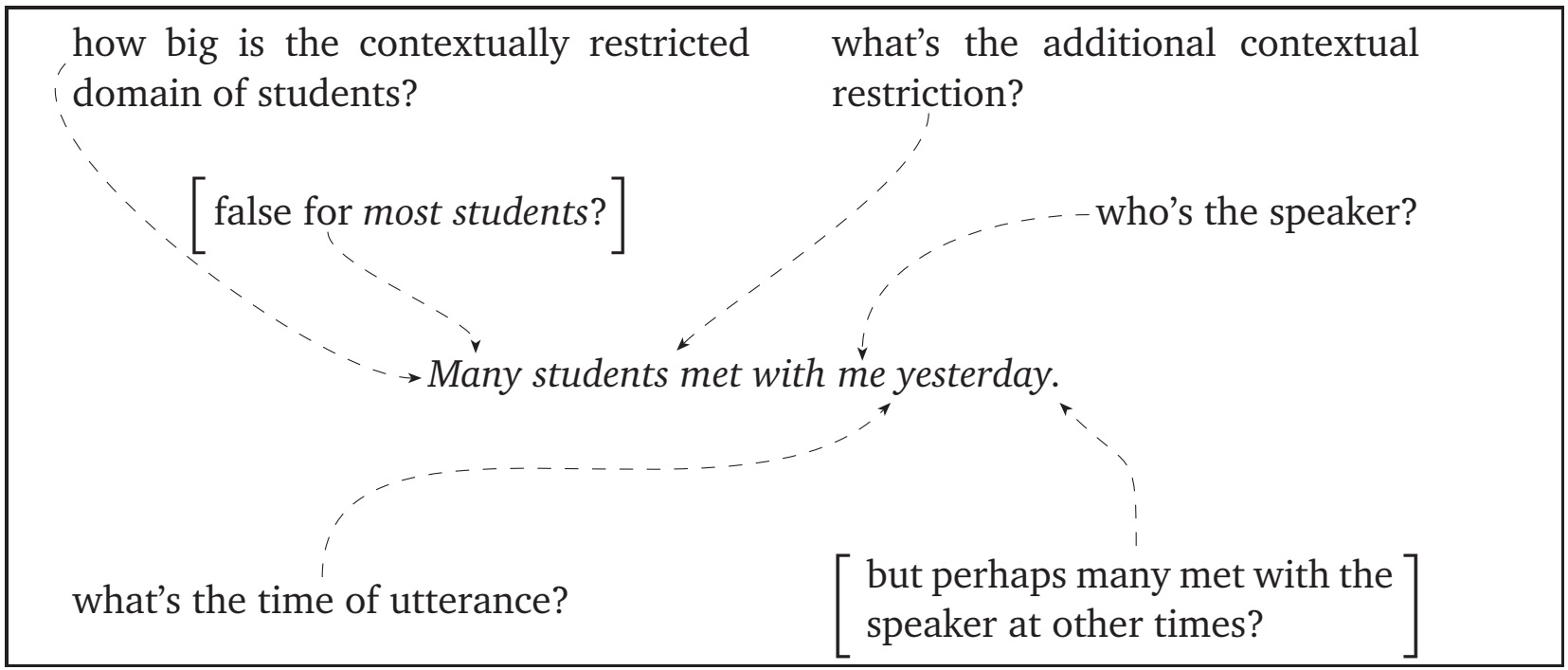
I didn’t see any.

Context dependence

- Are there typos in my slides?
- Are there bookstores downtown?
- Are there cookies in the cupboard?
- . . .

I didn’t see any.

Routine pragmatic enrichment

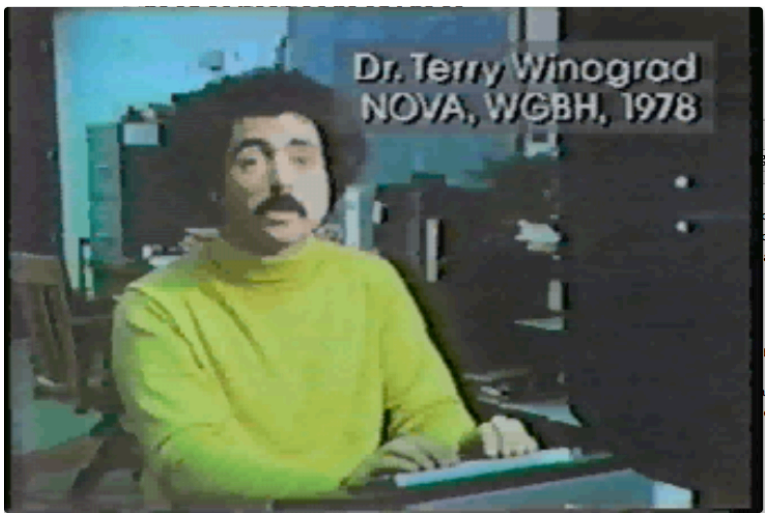


SHRDLU (Winograd 1972)

SHRDLU (Winograd 1972)

Full-grounded system that parses the user's input, maps it to logical form, interprets that logical form in its world, tries to take appropriate action, and generates responses.

http://hci.stanford.edu/winograd/shrdlu/



One project did succeed. Terry Winograd's program SHRDLU could use English intelligently, but there was a catch: the only subject you could discuss was a micro-world of simulated blocks.

SHRDLU (Winograd 1972)

1. Person: WHAT DOES THE BOX CONTAIN?
2. Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
3. Person: WHAT IS THE PYRAMID SUPPORTED BY?
4. Computer: THE BOX.
5. Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
6. Computer: FOUR OF THEM.
7. Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
8. Computer: YES, THE RED CUBE.
9. Person: IS IT SUPPORTED?
10. Computer: YES, BY THE TABLE.

Humans and situated language learning

Children learn language

1. with incredible speed
2. despite relatively few inputs
3. by using cues from
   - contrast inherent in the forms they hear
   - social cues
   - assumptions about the speaker's goals
   - regularities in the physical environment.

Frank et al. 2012; Frank and Goodman 2014

Consequences for NLU

1. Human children are the best agents in the universe at learning language, and they depend heavily on grounding.

2. Problems that are intractable without grounding are solvable with the right kinds of grounding.

3. Deep learning is a flexible toolkit for reasoning about different kinds of information in a single model, so it's led to conceptual and empirical improvements in this area.

4. We should seek out (and develop) data sets that include the right kind of grounding.

Datasets

1. Stanford English Colors in Context Corpus [link]
2. Stanford Chinese Colors in Context Corpus [link]
3. OneCommon [link]
4. Edinburgh Map Corpus [link]
5. Cards Corpus [link]
6. Deal or No Deal? [link]
7. CraigsListBargain [link]
8. ALFRED [link]
9. CrossTalk [link]
10. Room-to-Room [link]

References I

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182. Association for Computational Linguistics.

Michael C. Frank and Noah D. Goodman. 2014. Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, 75(1):80–96.

Michael C. Frank, Joshua B. Tenenbaum, and Anne Fernald. 2012. Social and discourse contributions to the determination of reference in cross-situational word learning. *Language, Learning, and Development*.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Stroudsburg, PA. ACL.

Stephen C. Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge, MA.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? End-to-end learning for negotiation dialogues. ArXiv:1706.05125.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Stefanie Tellex, Ross A. Knepper, Adrian Li, Thomas M. Howard, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. In *Proceedings of Robotics: Science and Systems*.

Adam Vogel, Max Bodoia, Christopher Potts, and Dan Jurafsky. 2013. Emergence of Gricean maxims from multi-agent decision theory. In *Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1072–1081, Stroudsburg, PA. Association for Computational Linguistics.

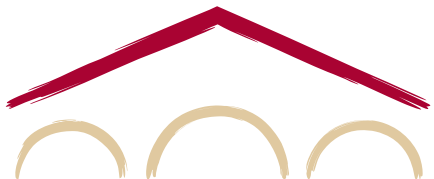
Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Grounded language understanding: Speakers: From the world to language

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding

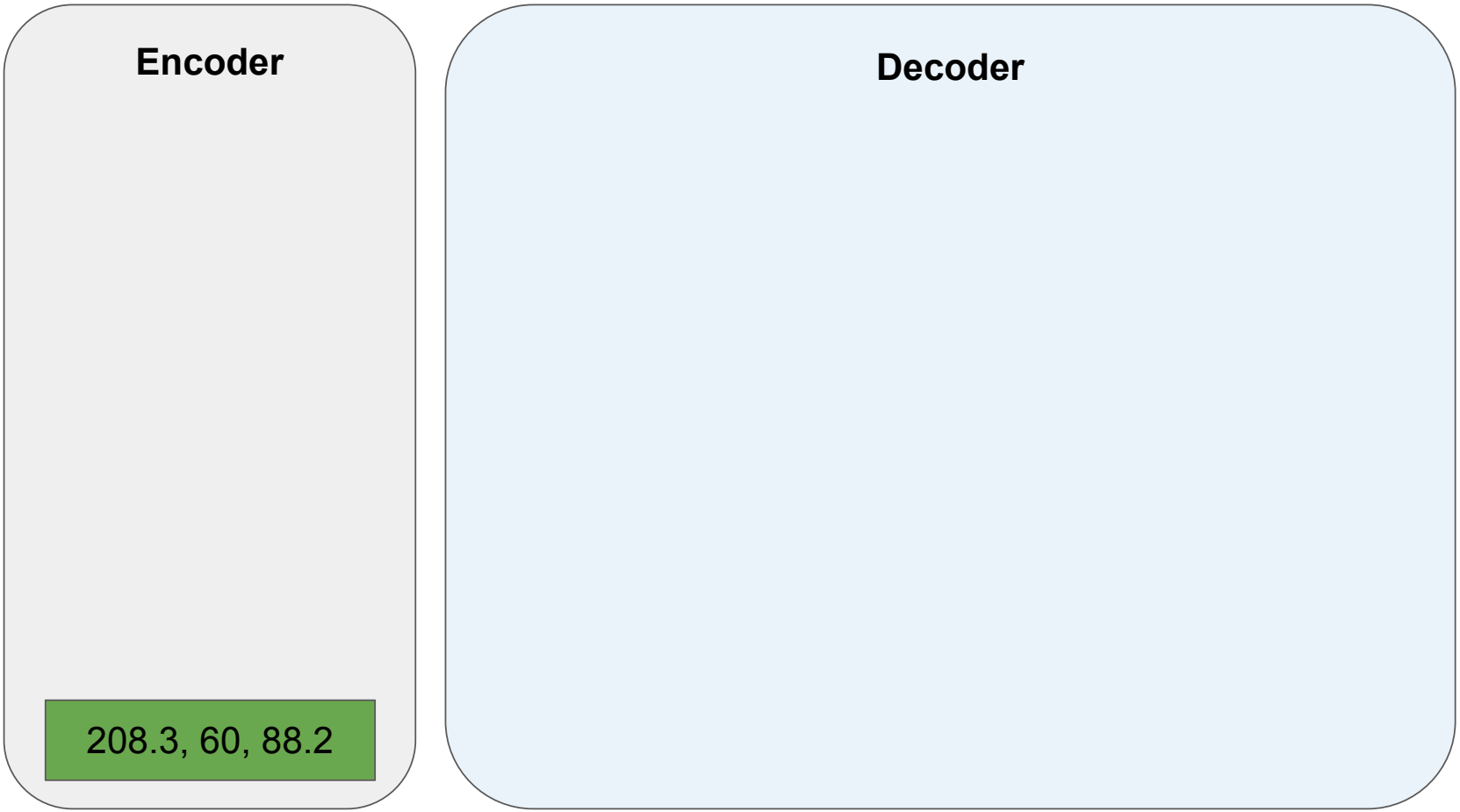


Color describer: Task formulation and data

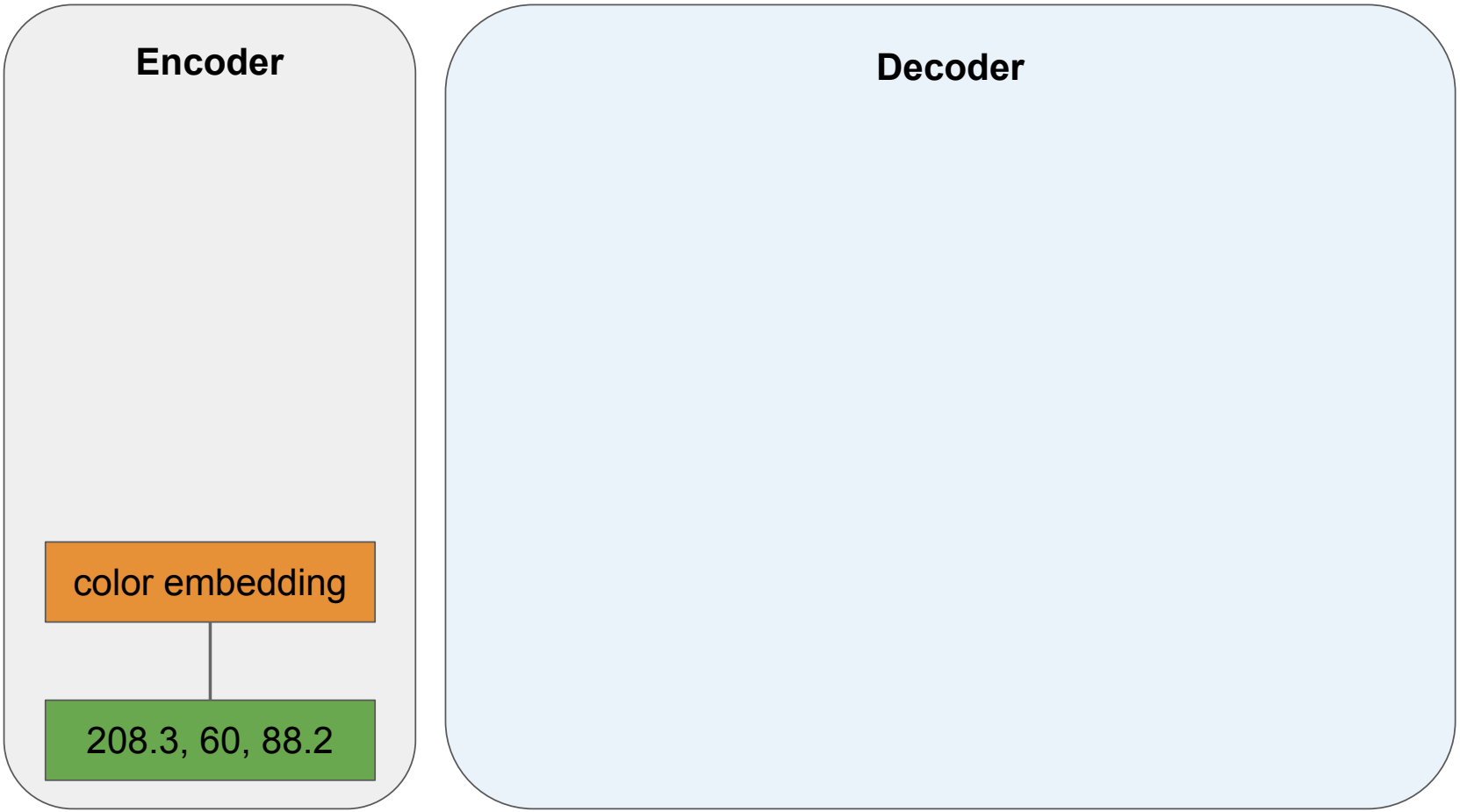
Color	Utterance
	green
	purple
	grape
	turquoise
	moss green
	pinkish purple
	light blue grey
	robin’s egg blue
	british racing green
	baby puke green

McMahan and Stone 2015

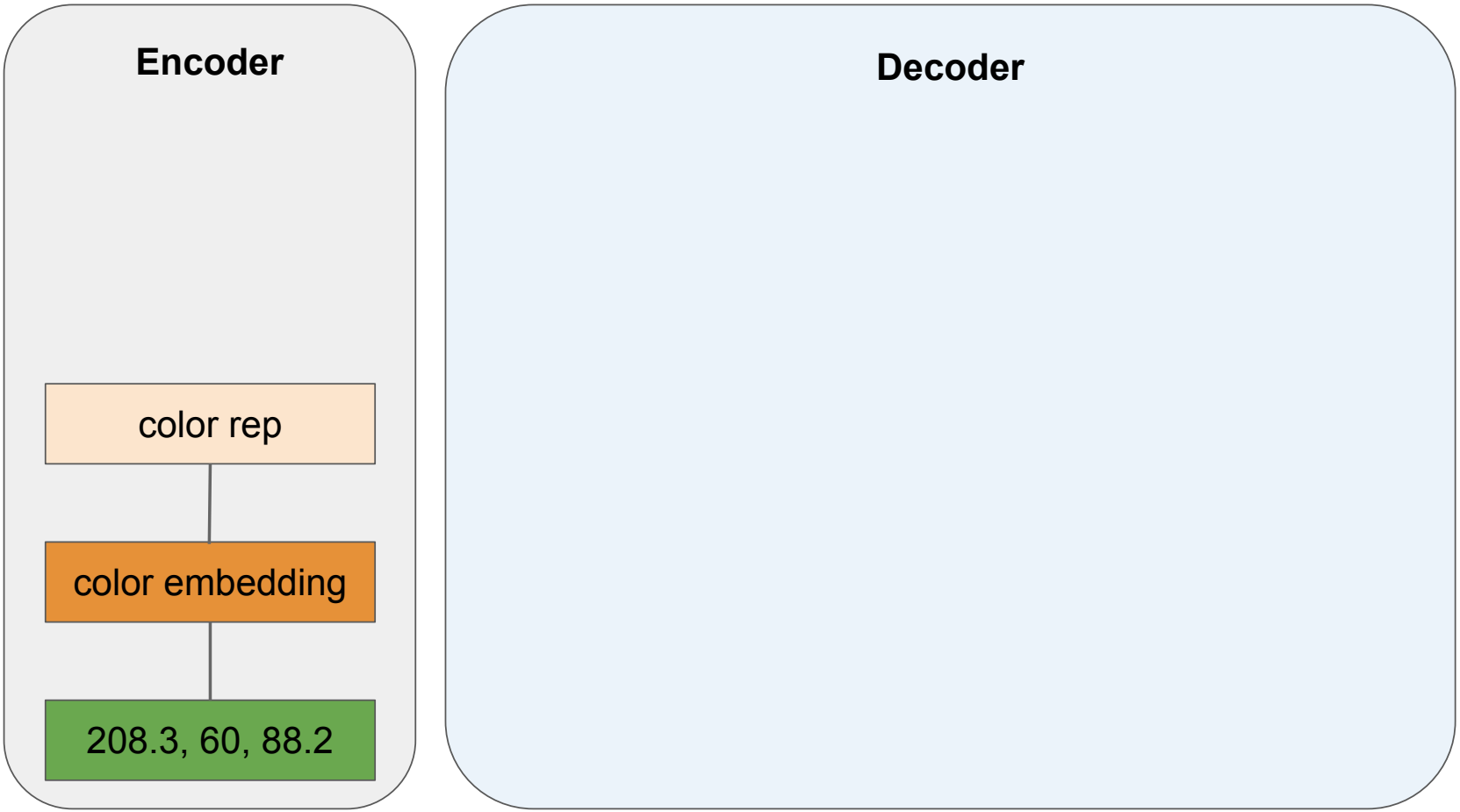
Color describer: Training with *teacher forcing*



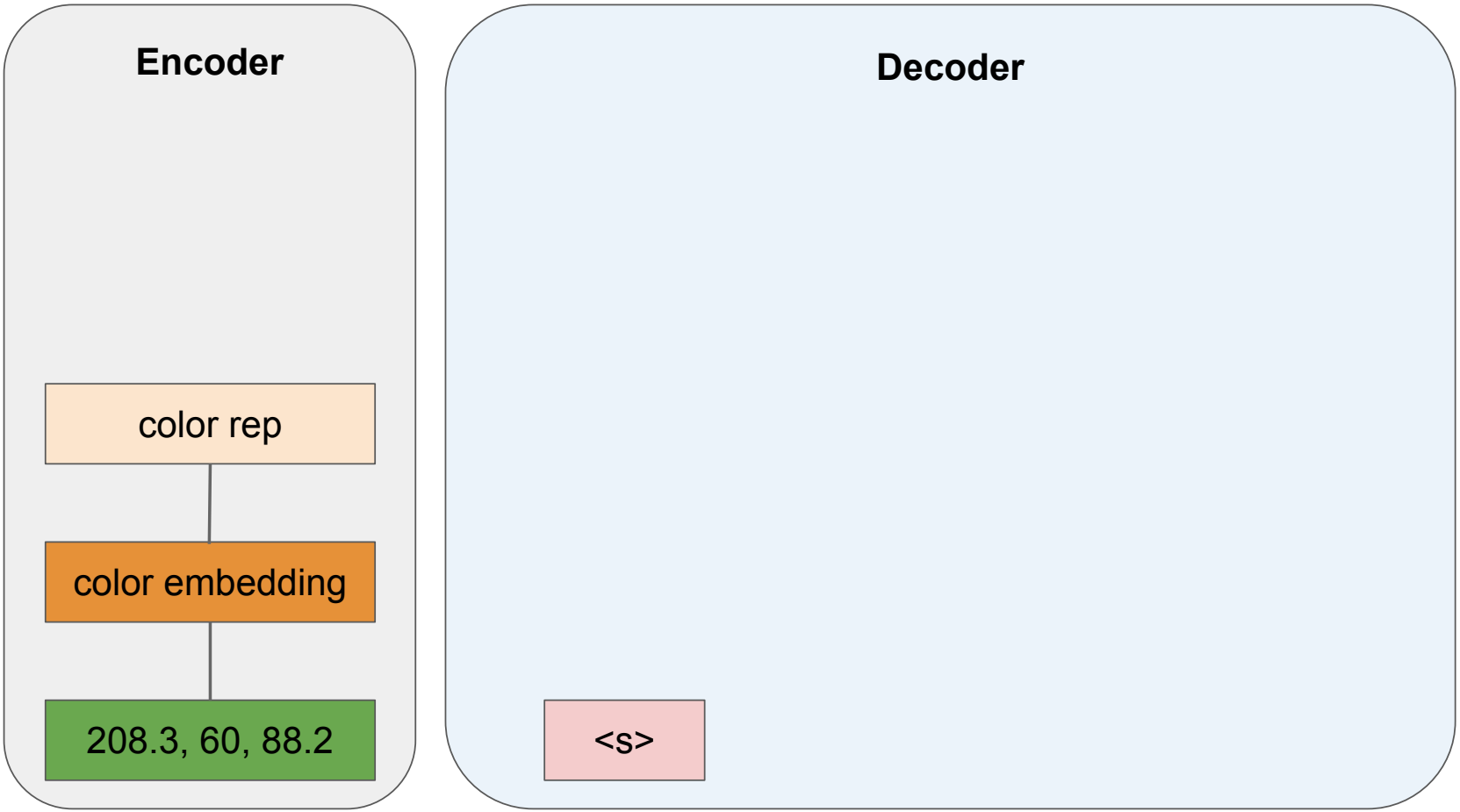
Color describer: Training with *teacher forcing*



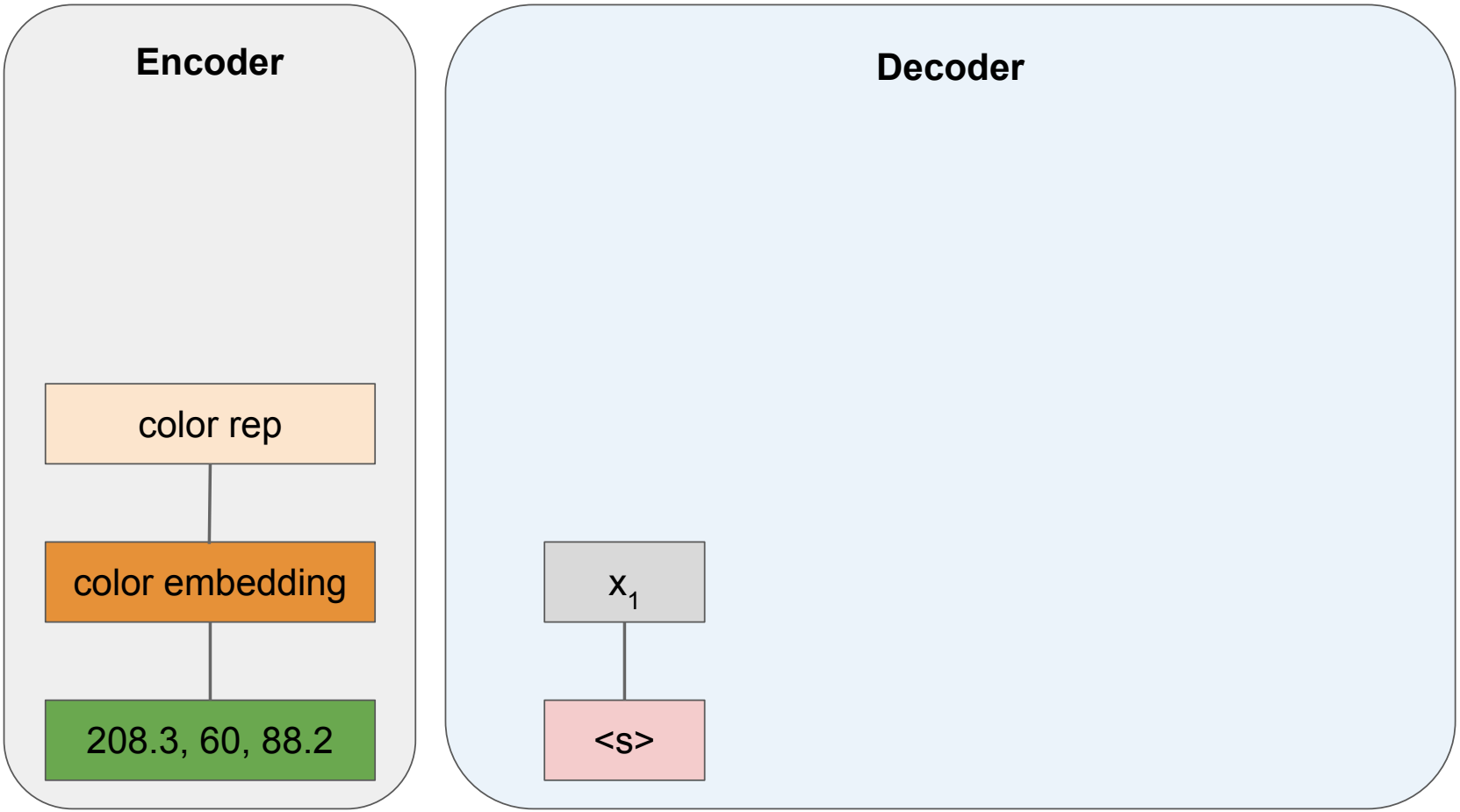
Color describer: Training with *teacher forcing*



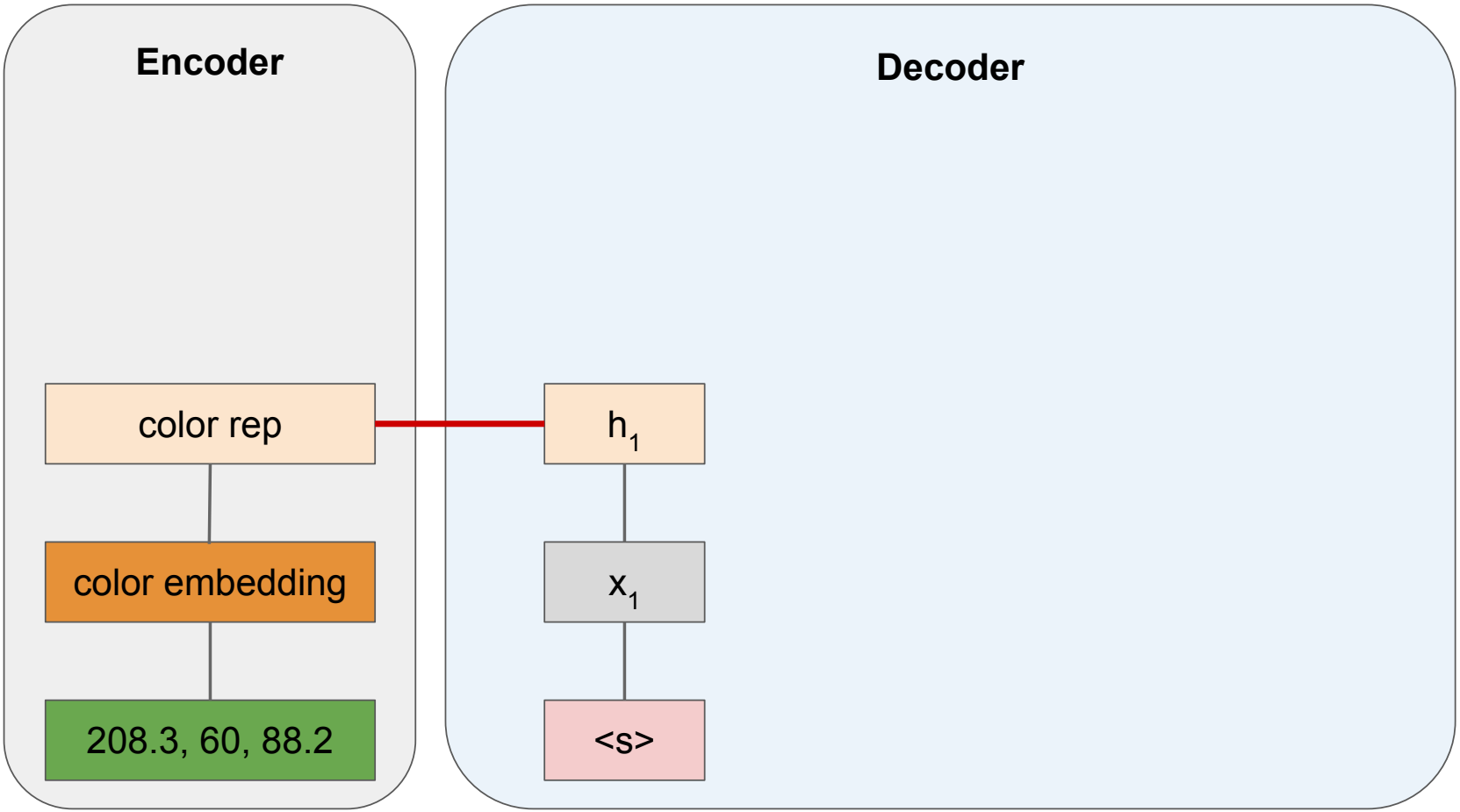
Color describer: Training with *teacher forcing*



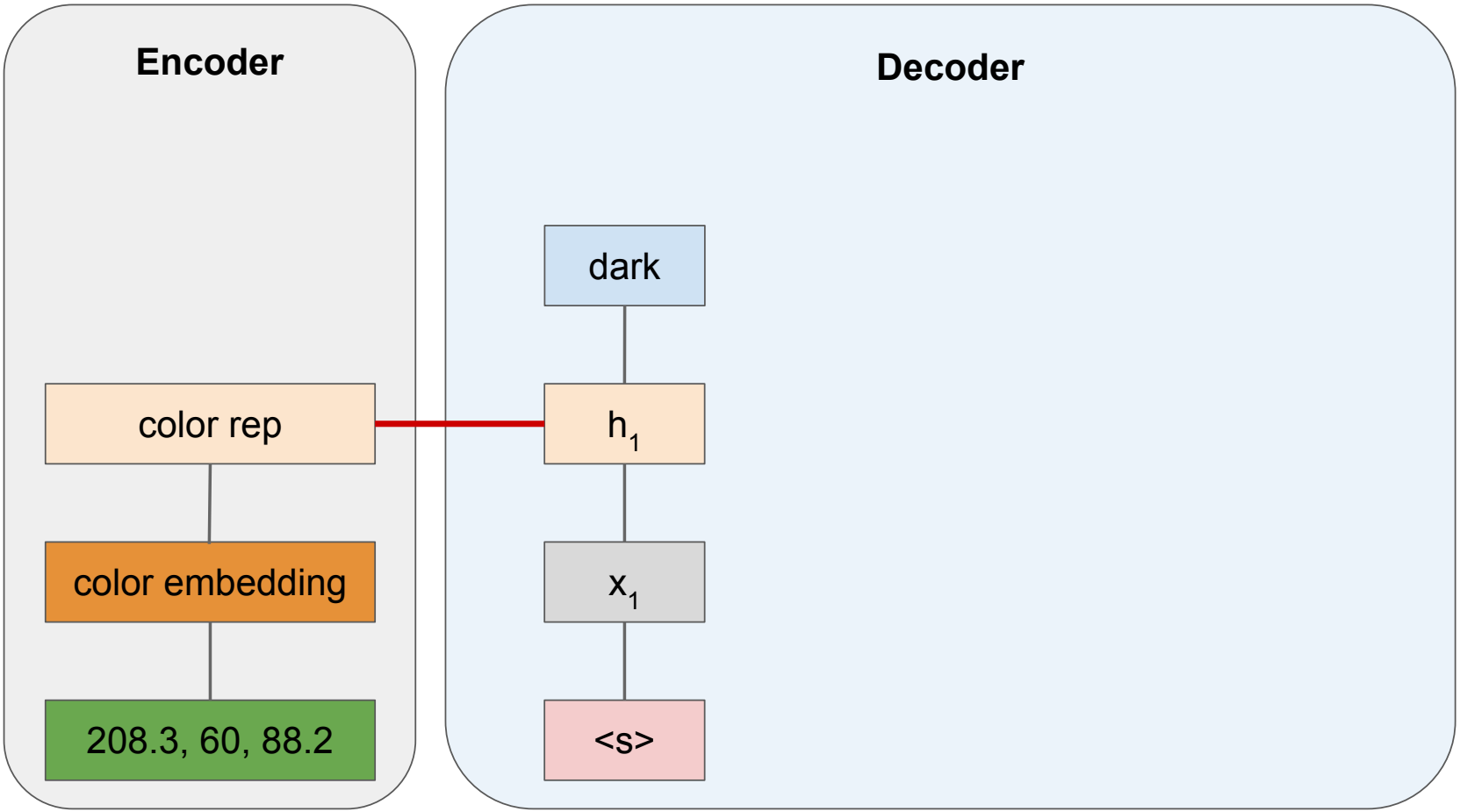
Color describer: Training with *teacher forcing*



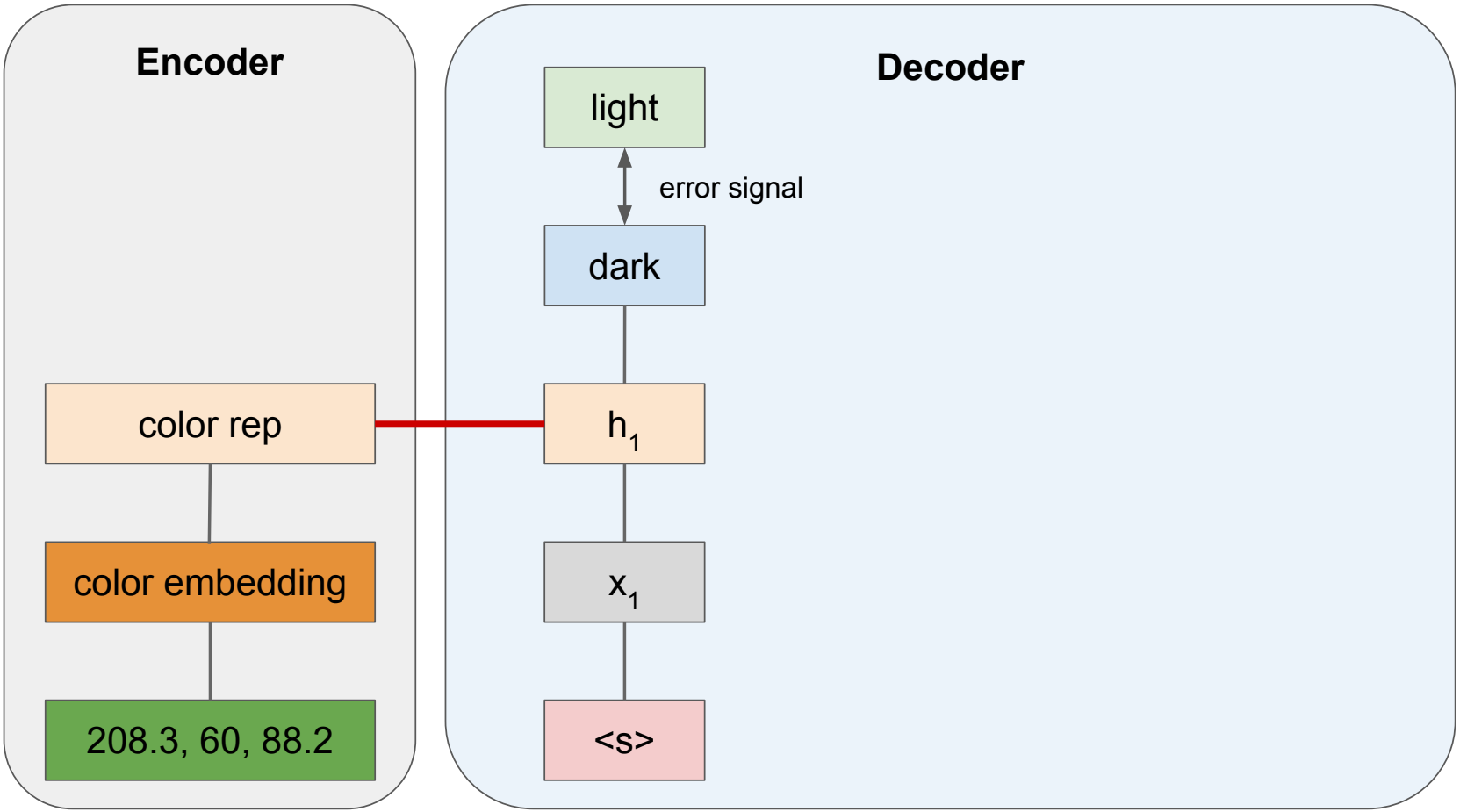
Color describer: Training with *teacher forcing*



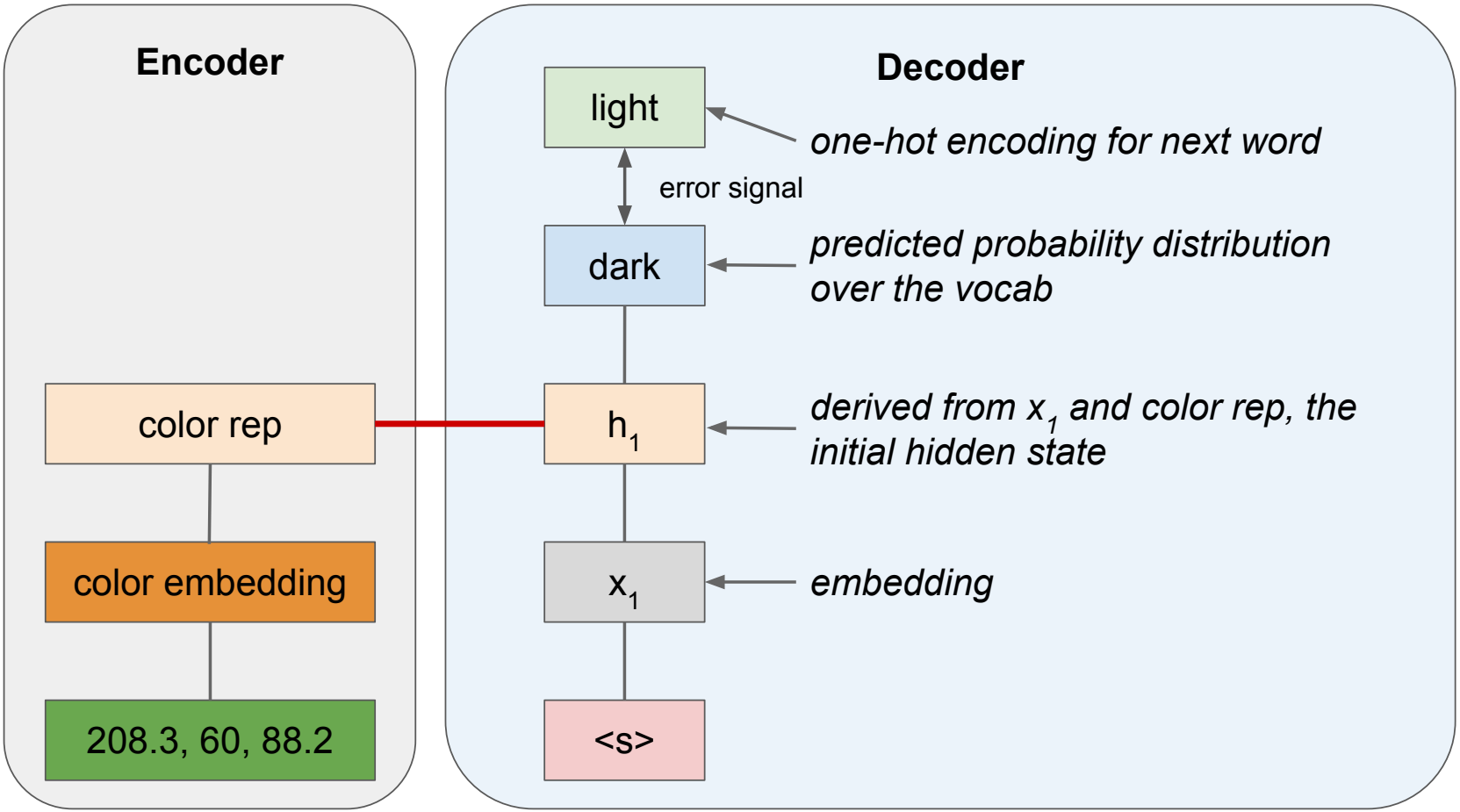
Color describer: Training with *teacher forcing*



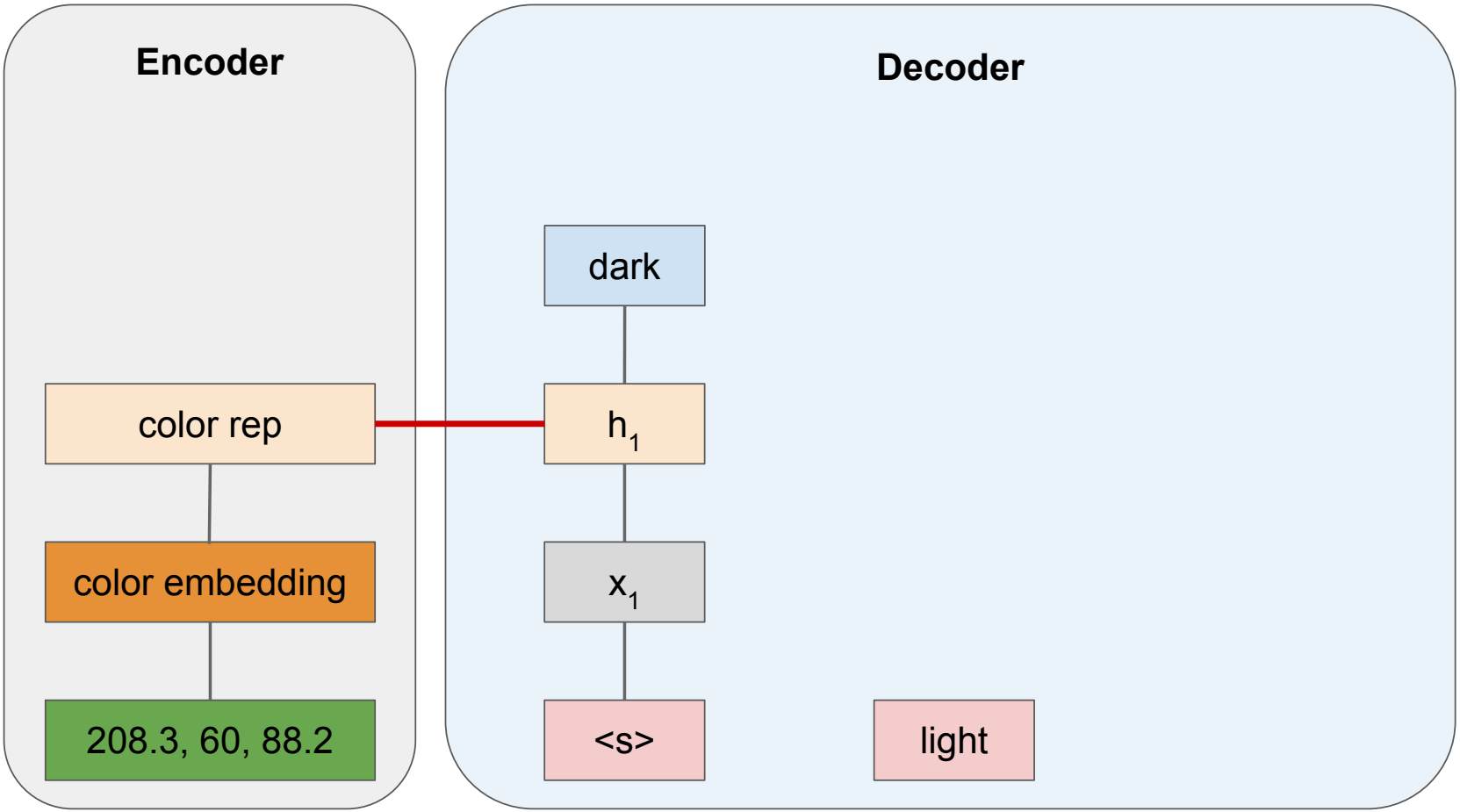
Color describer: Training with *teacher forcing*



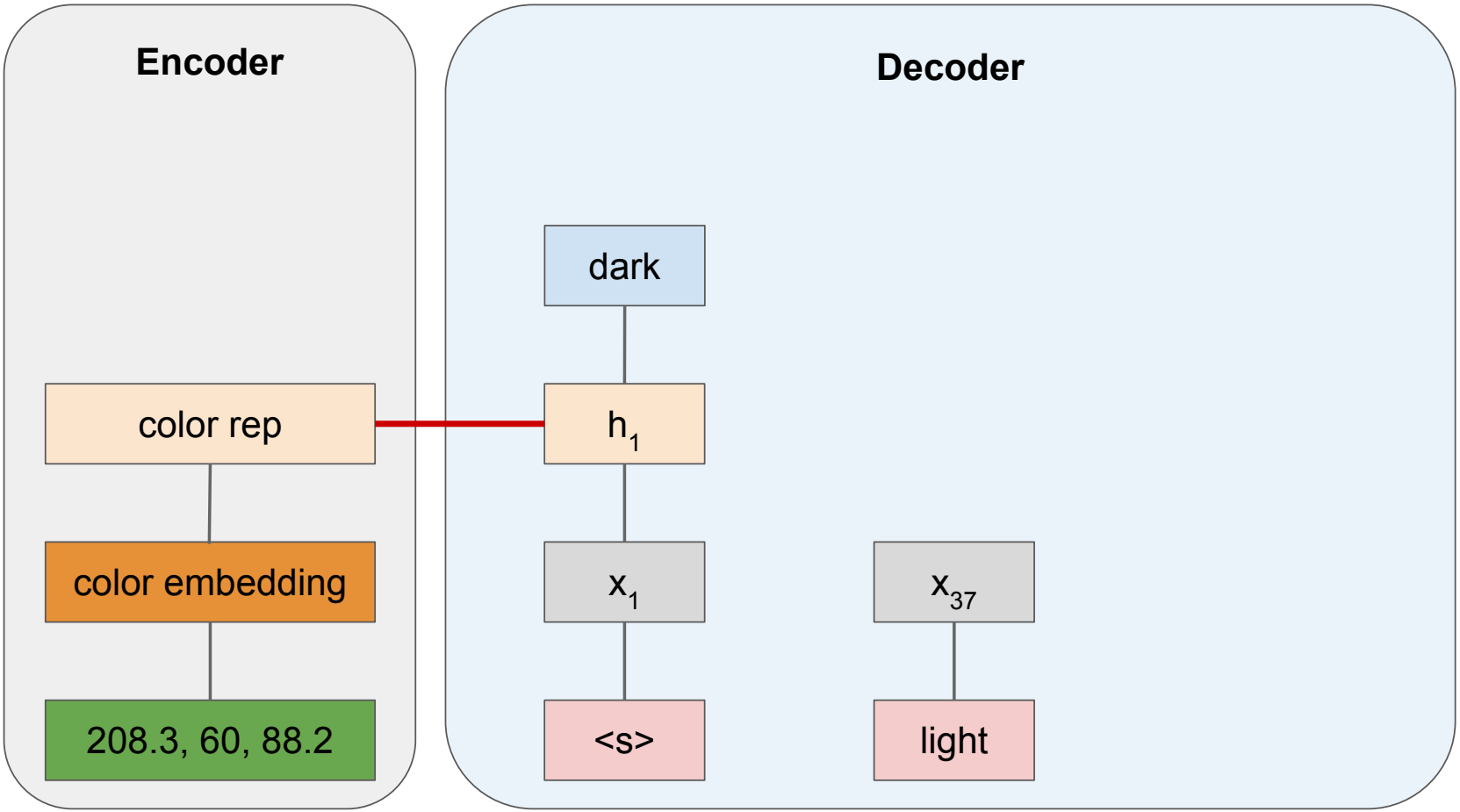
Color describer: Training with *teacher forcing*



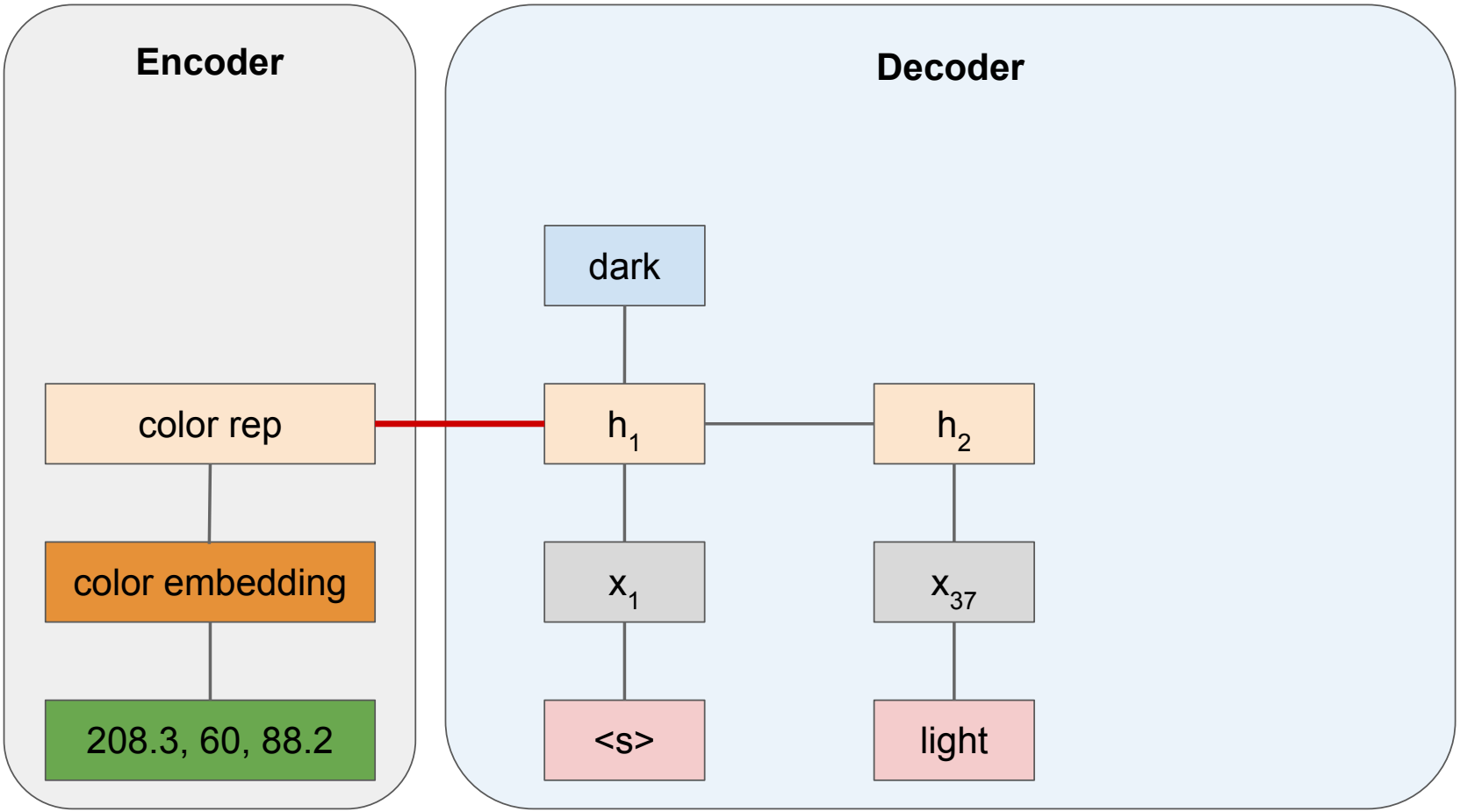
Color describer: Training with *teacher forcing*



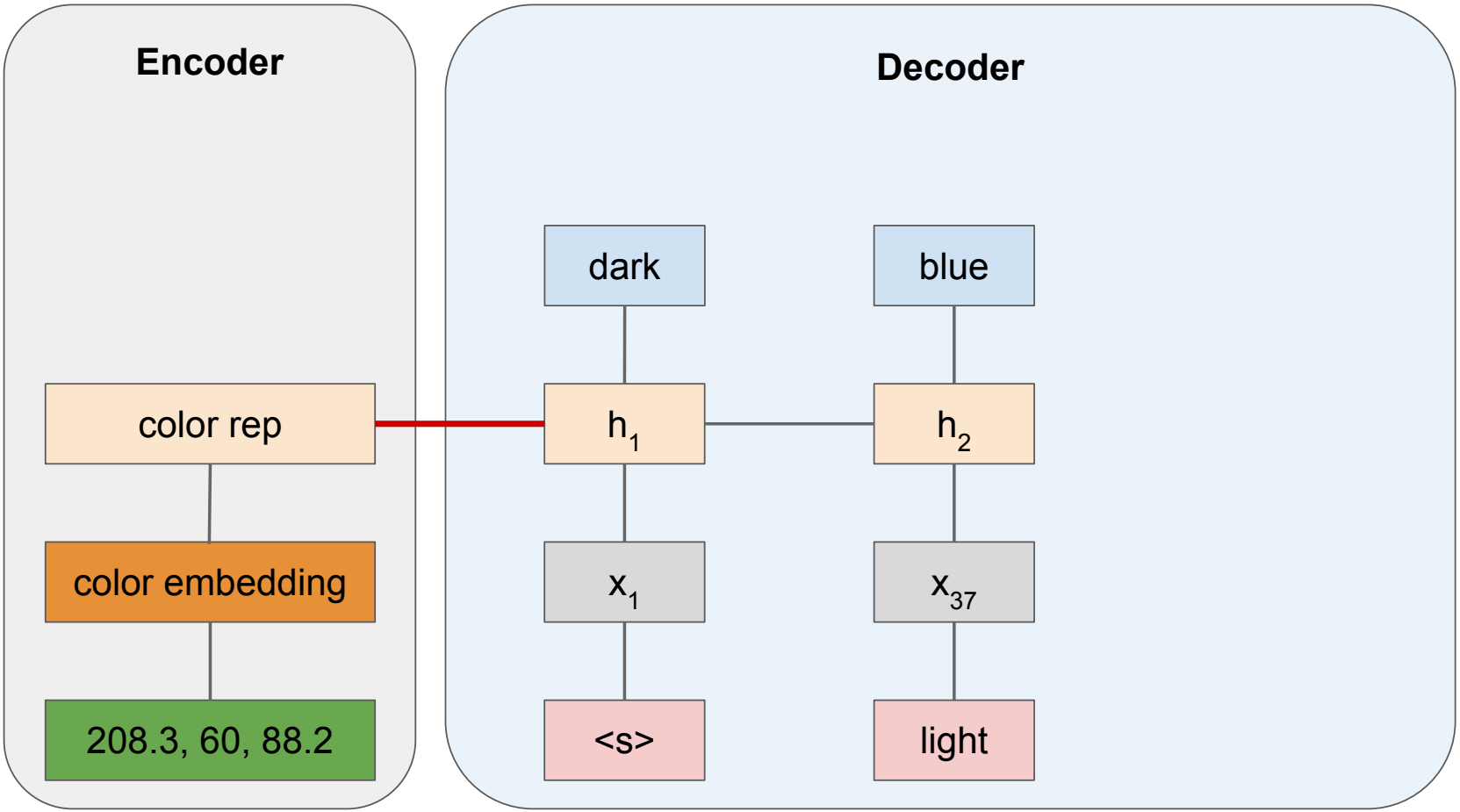
Color describer: Training with *teacher forcing*



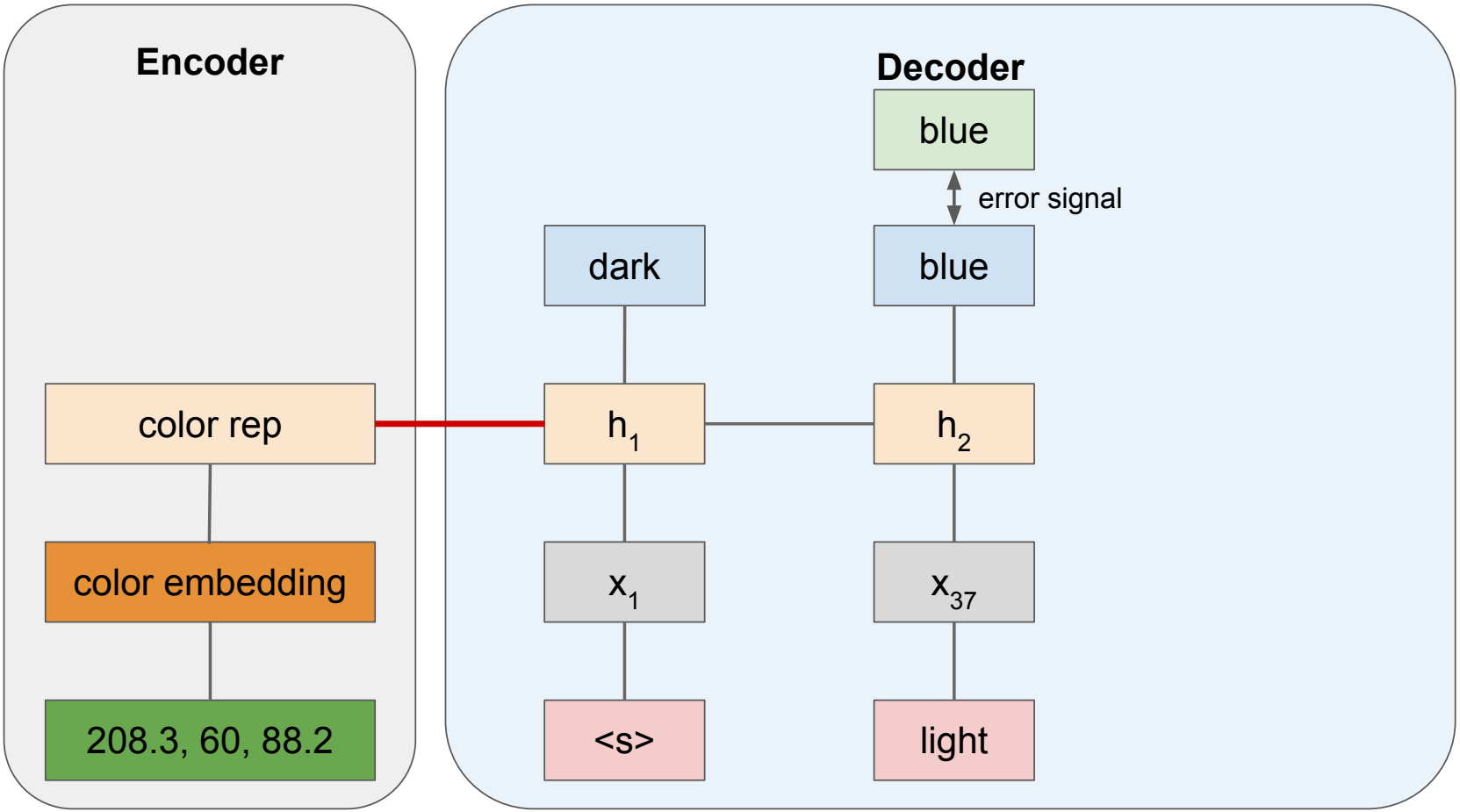
Color describer: Training with *teacher forcing*



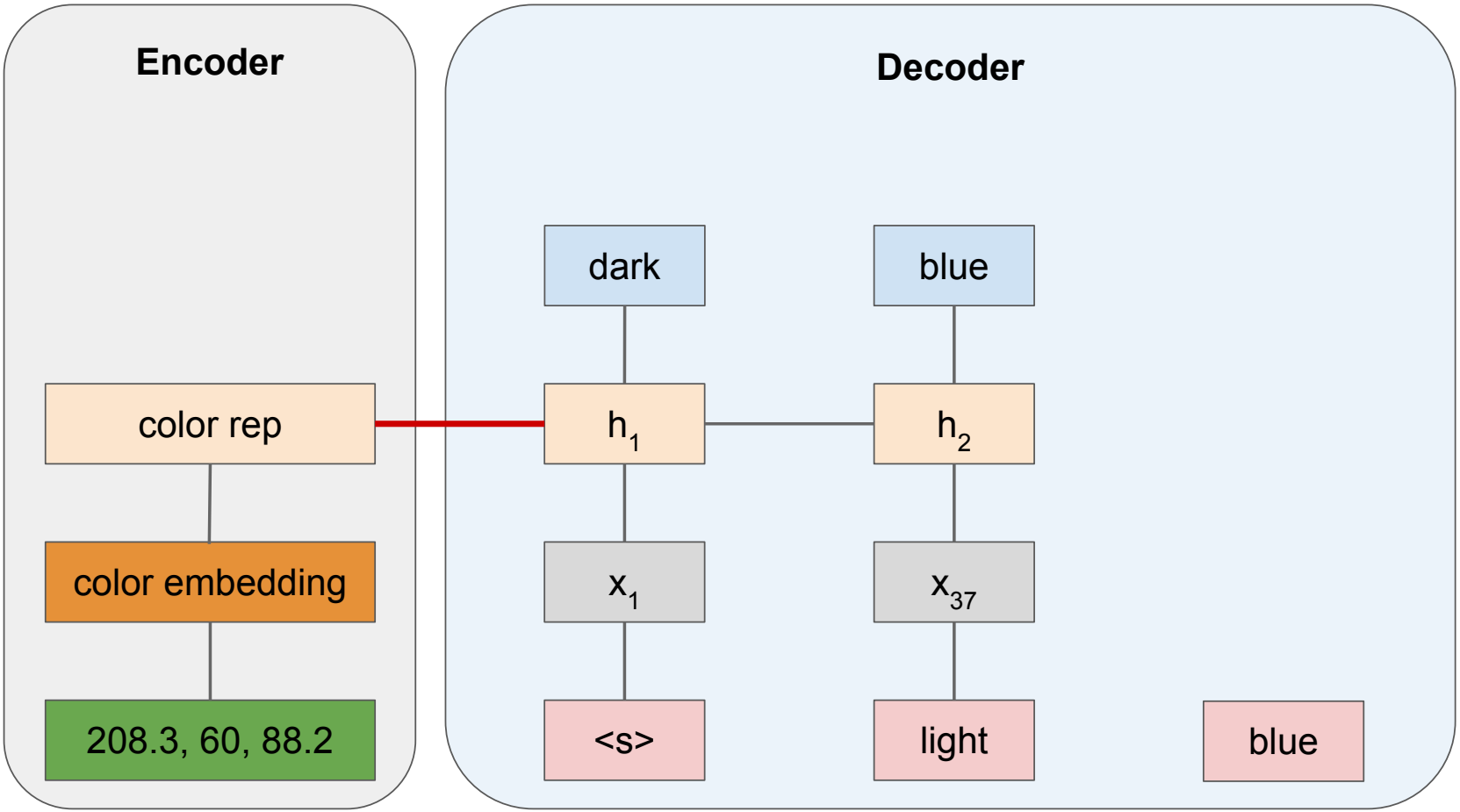
Color describer: Training with *teacher forcing*



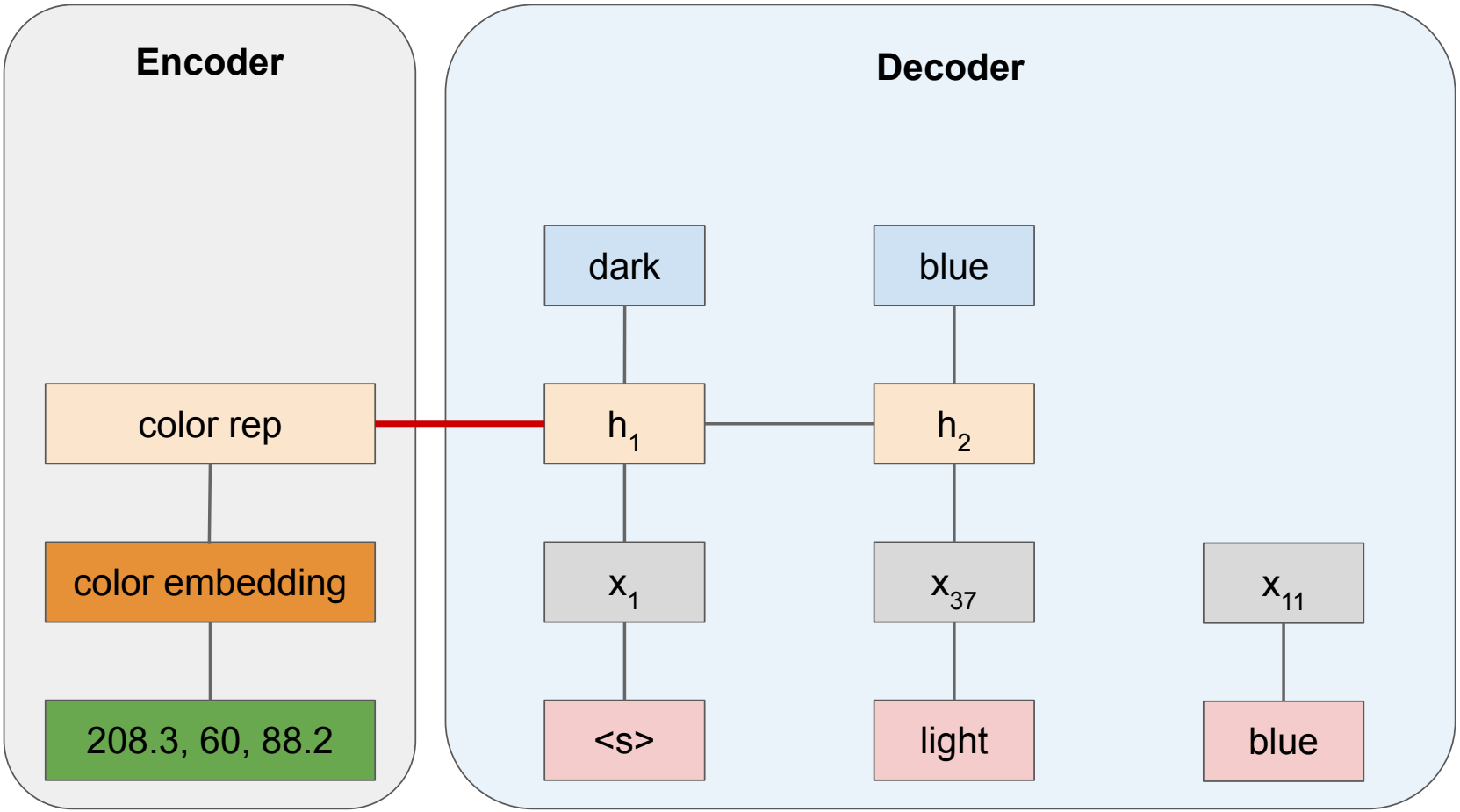
Color describer: Training with *teacher forcing*



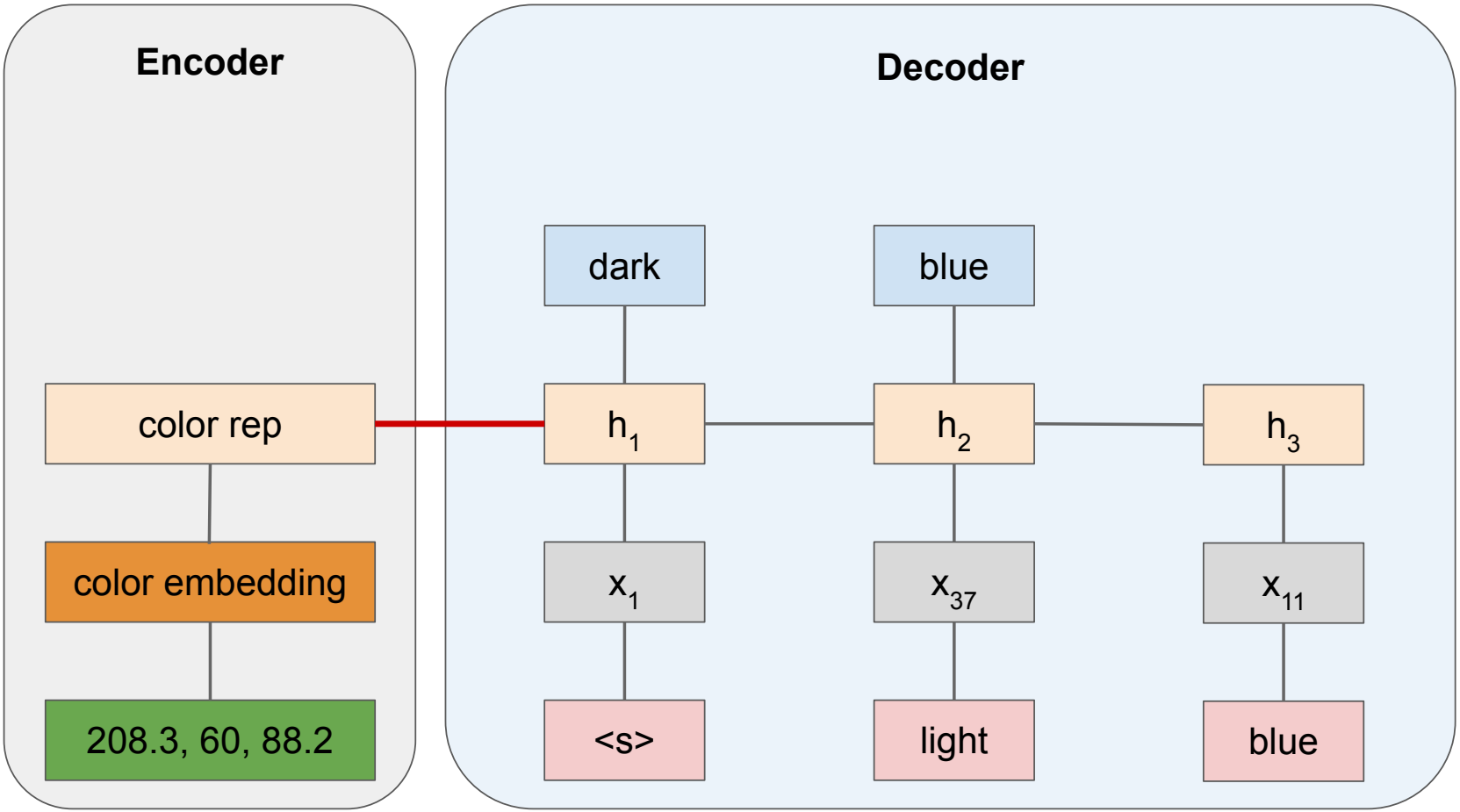
Color describer: Training with *teacher forcing*



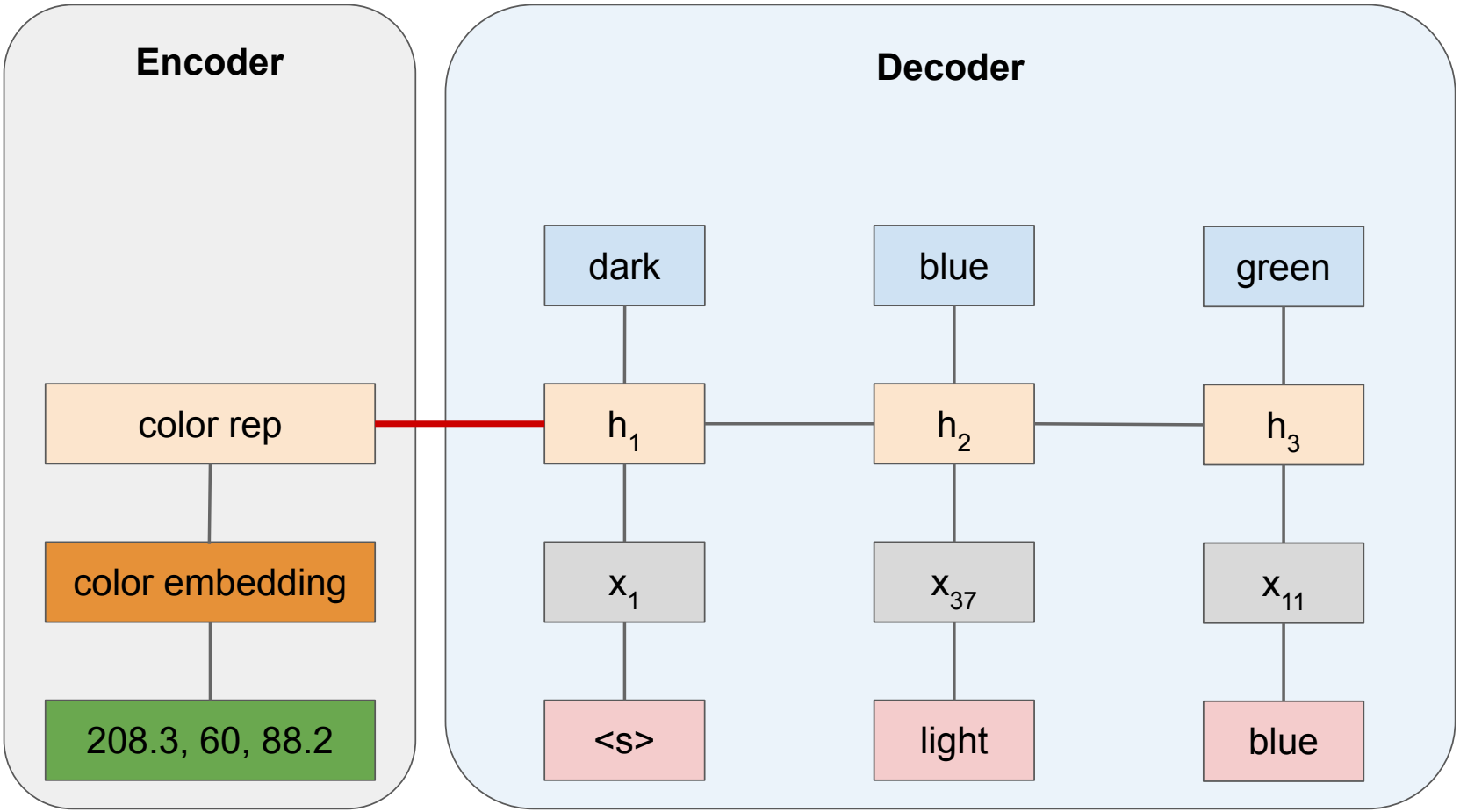
Color describer: Training with *teacher forcing*



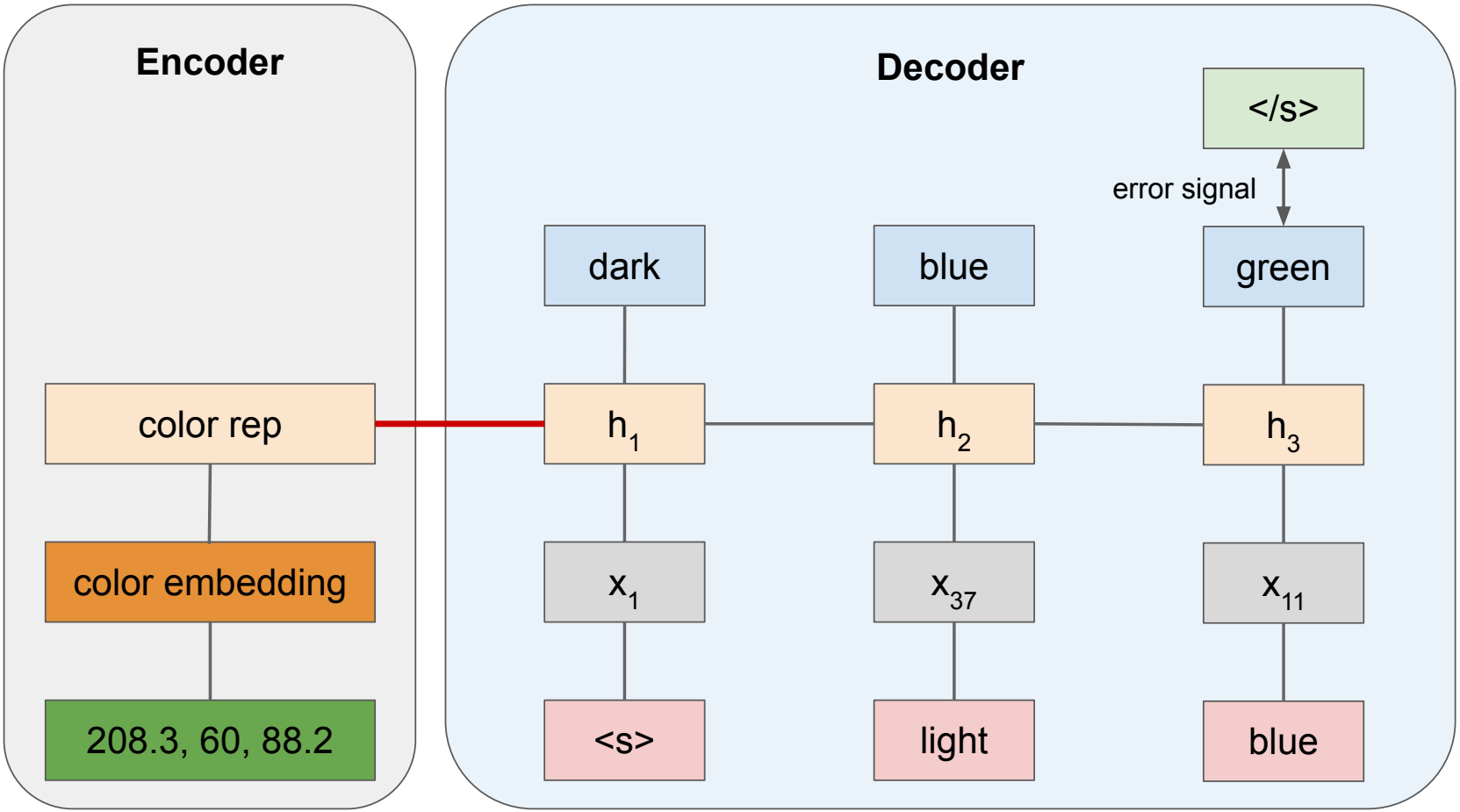
Color describer: Training with *teacher forcing*



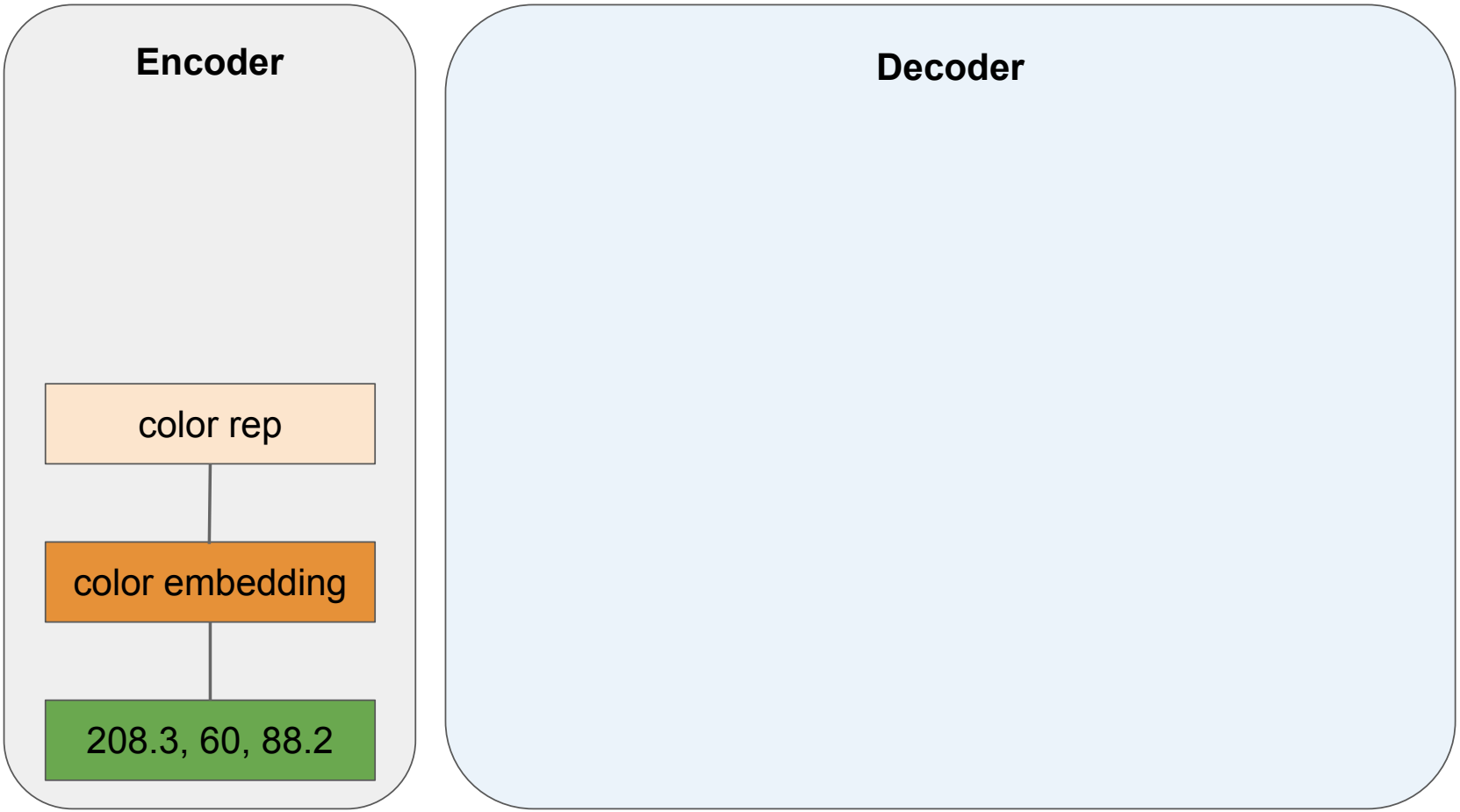
Color describer: Training with *teacher forcing*



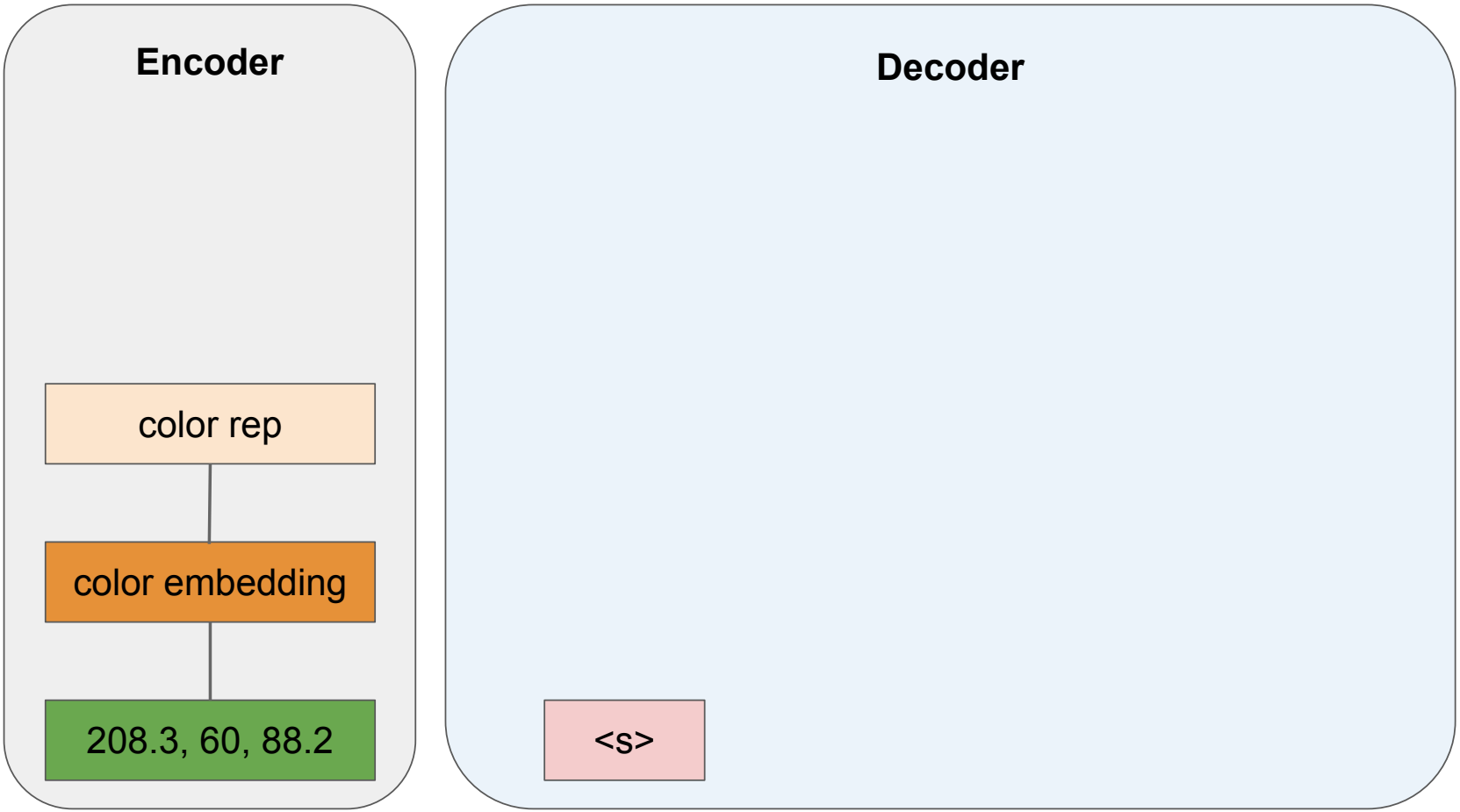
Color describer: Training with *teacher forcing*



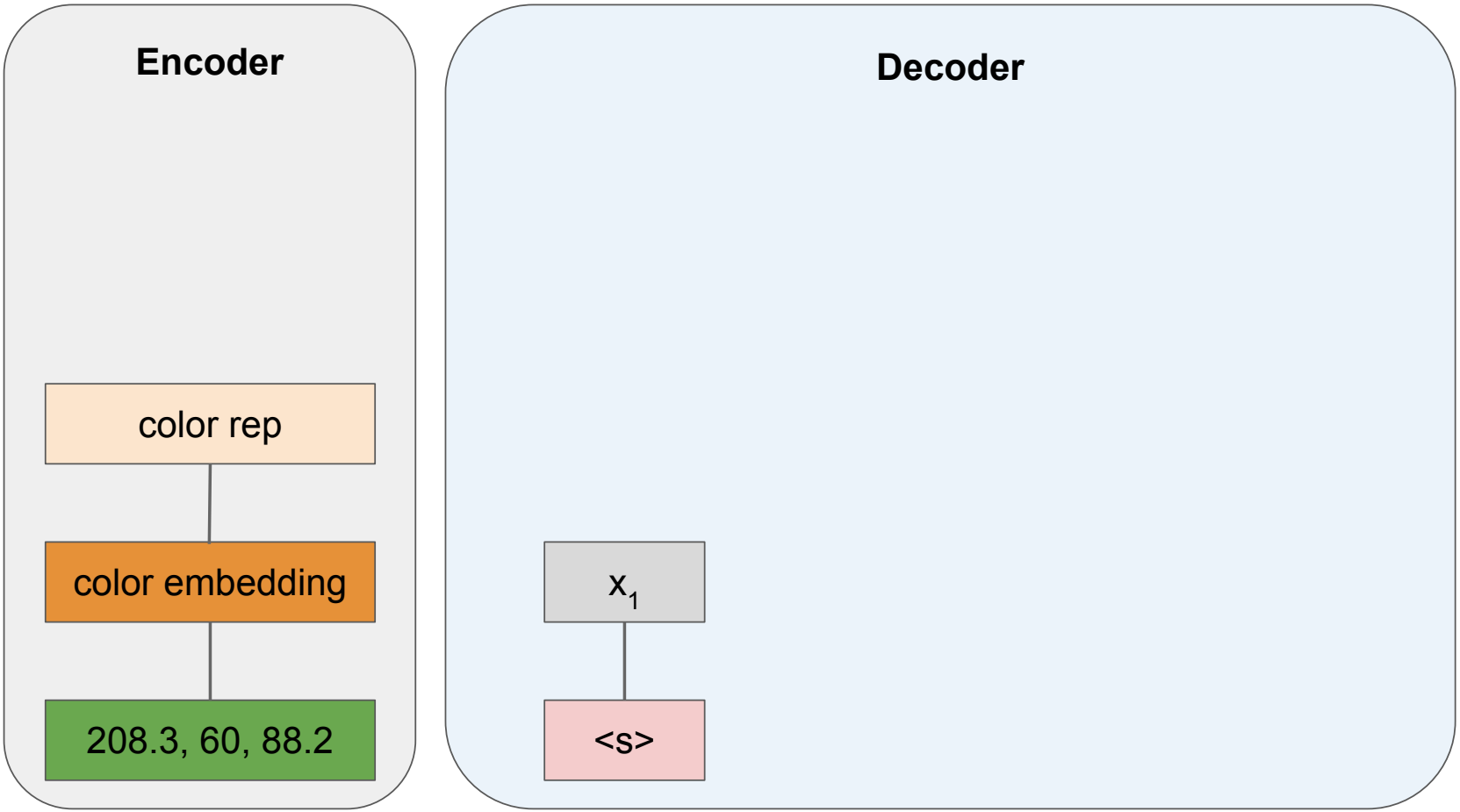
Color describer: Prediction



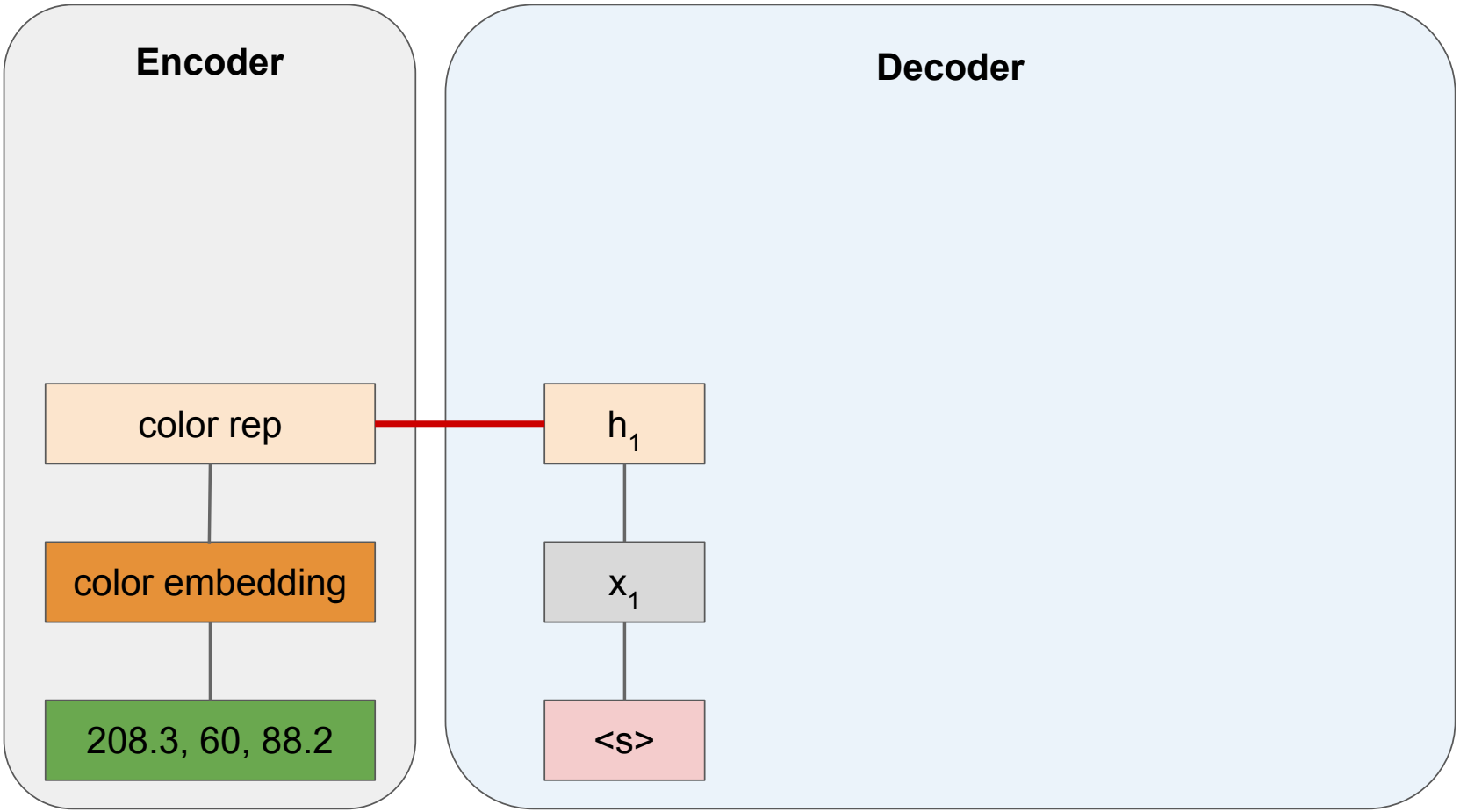
Color describer: Prediction



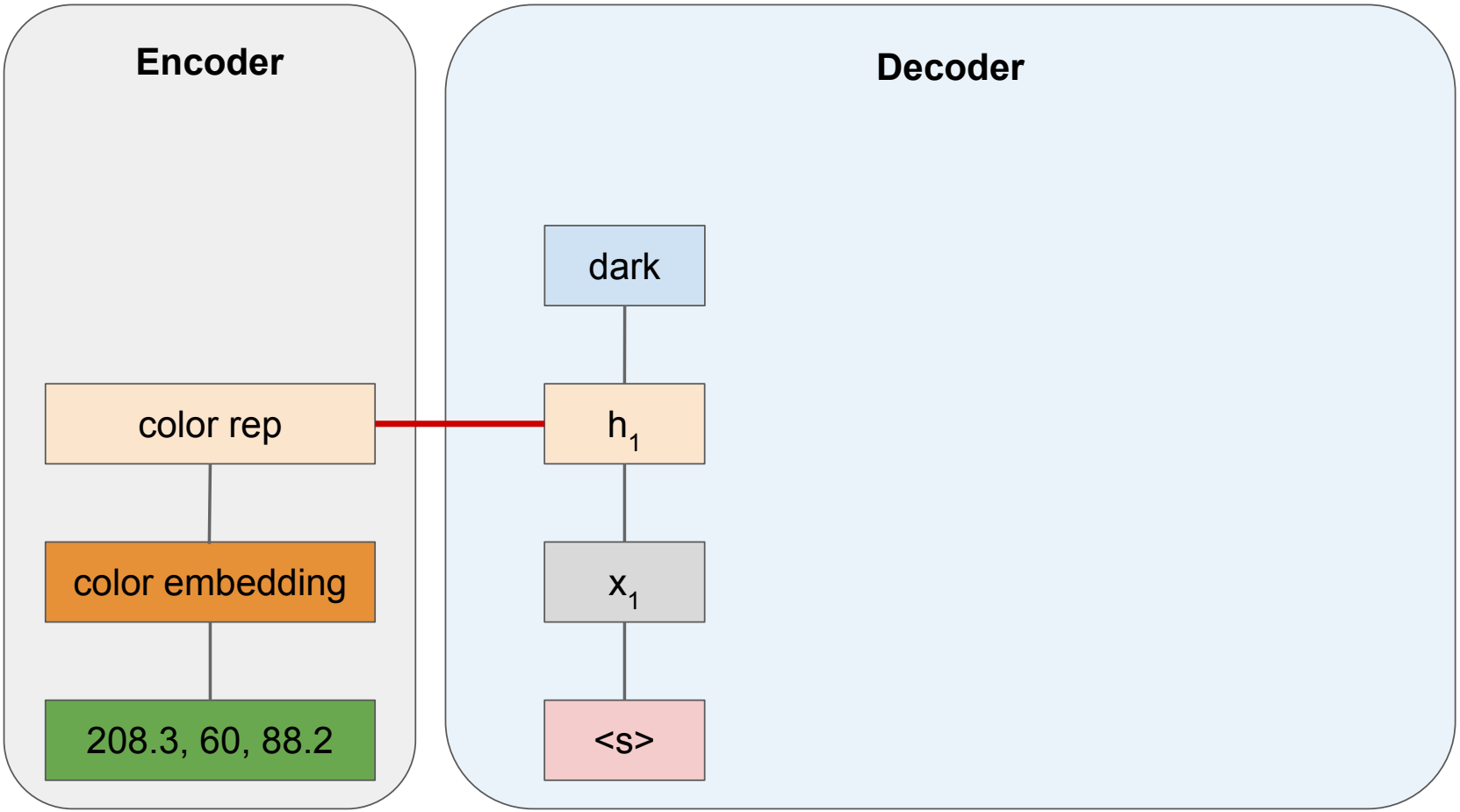
Color describer: Prediction



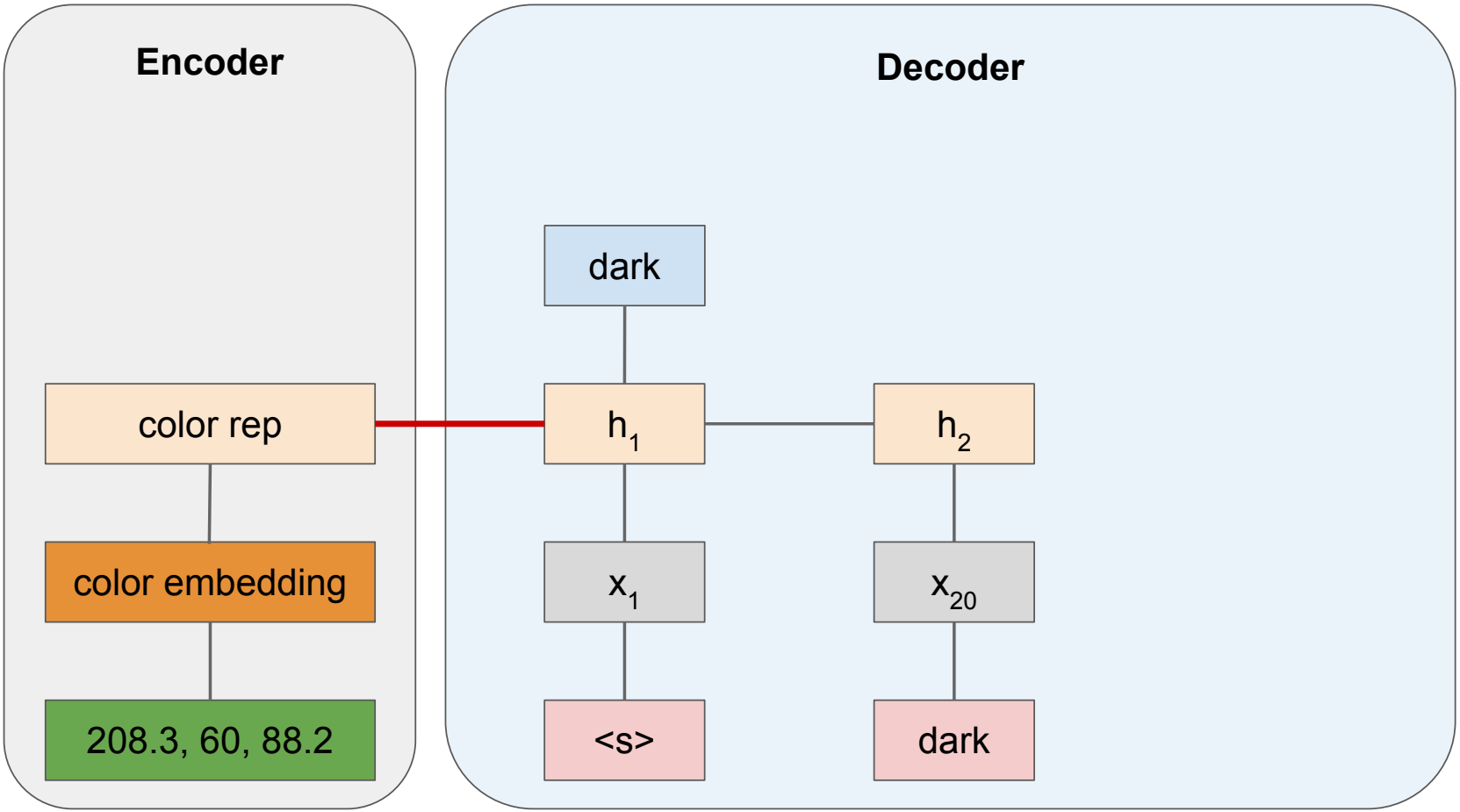
Color describer: Prediction



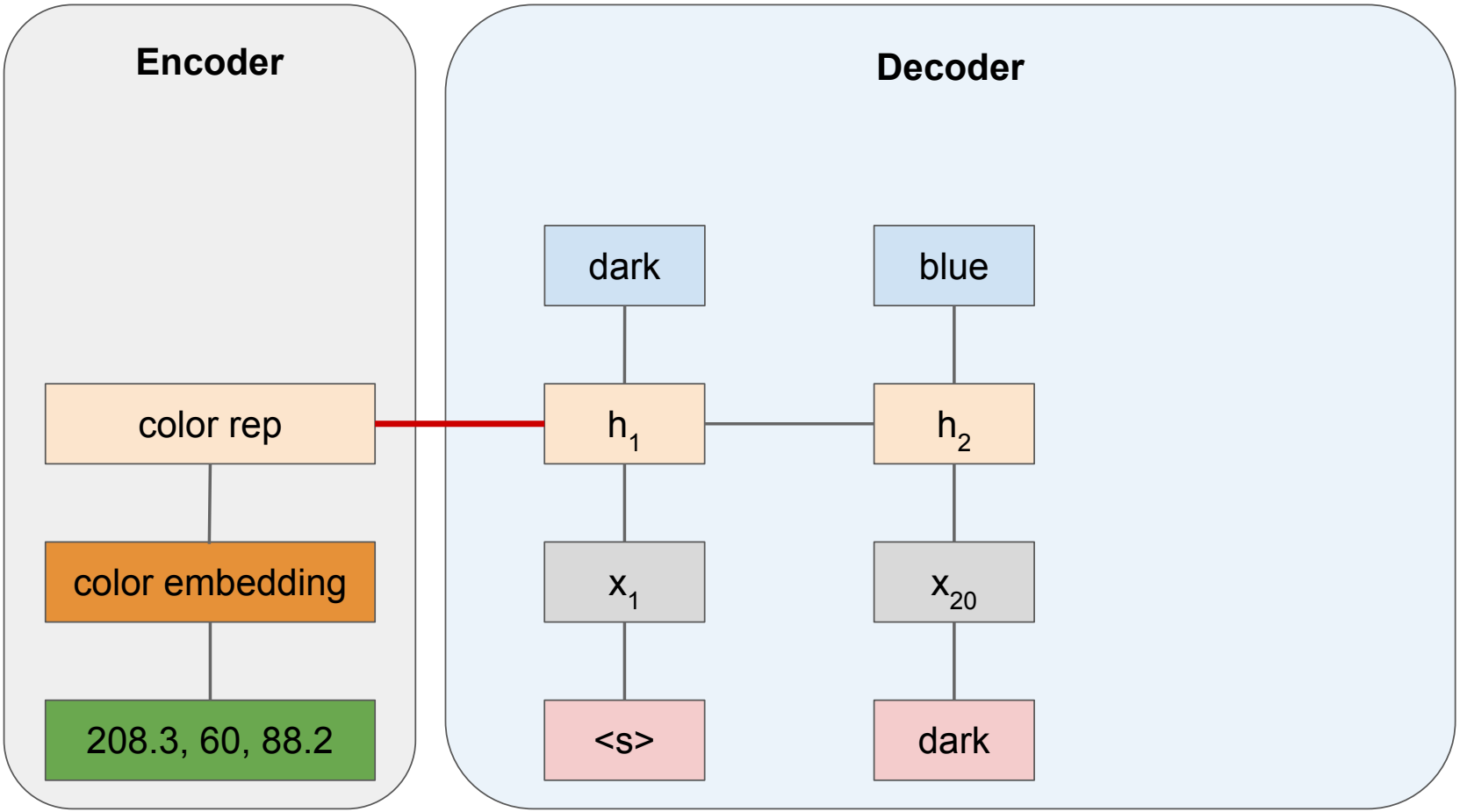
Color describer: Prediction



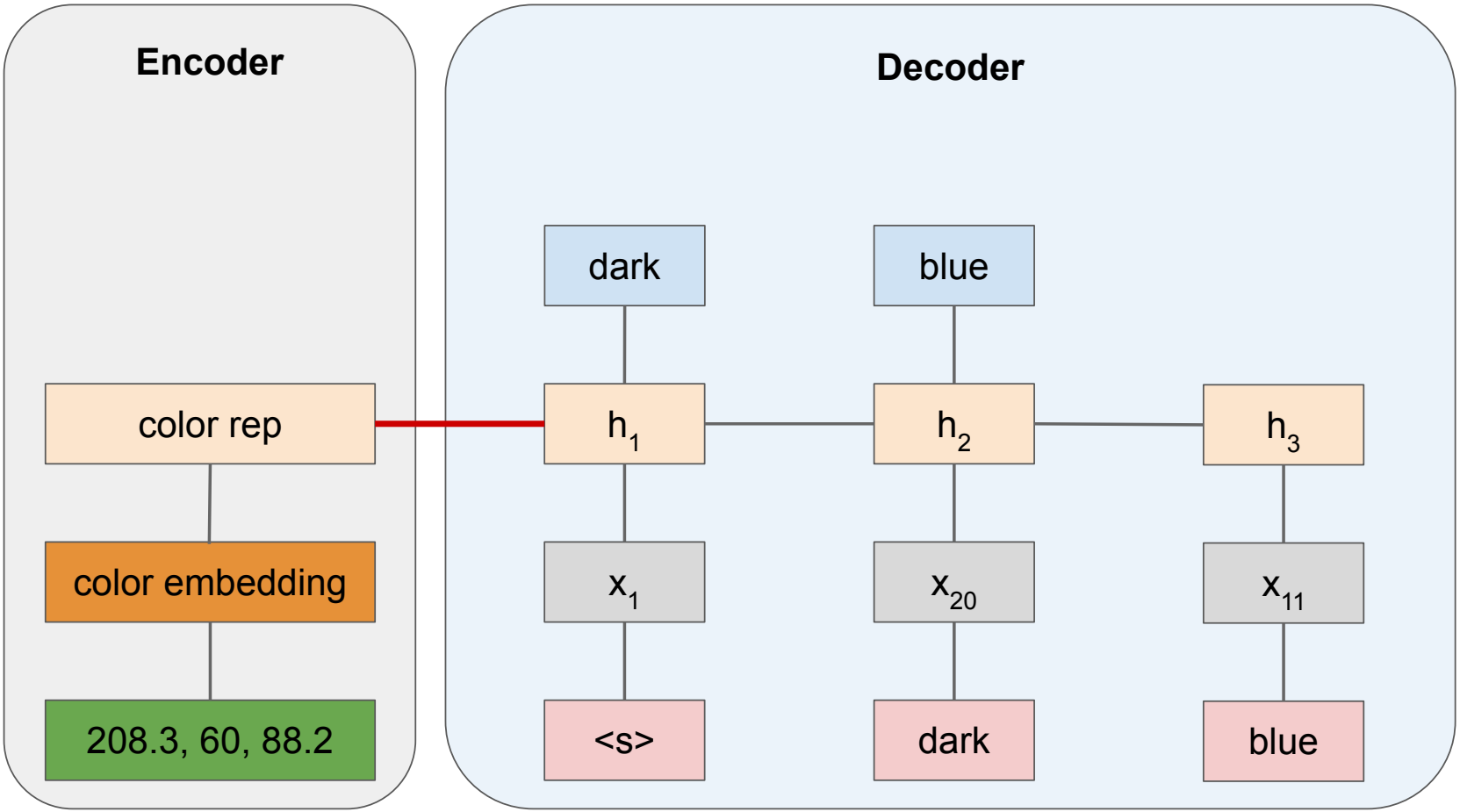
Color describer: Prediction



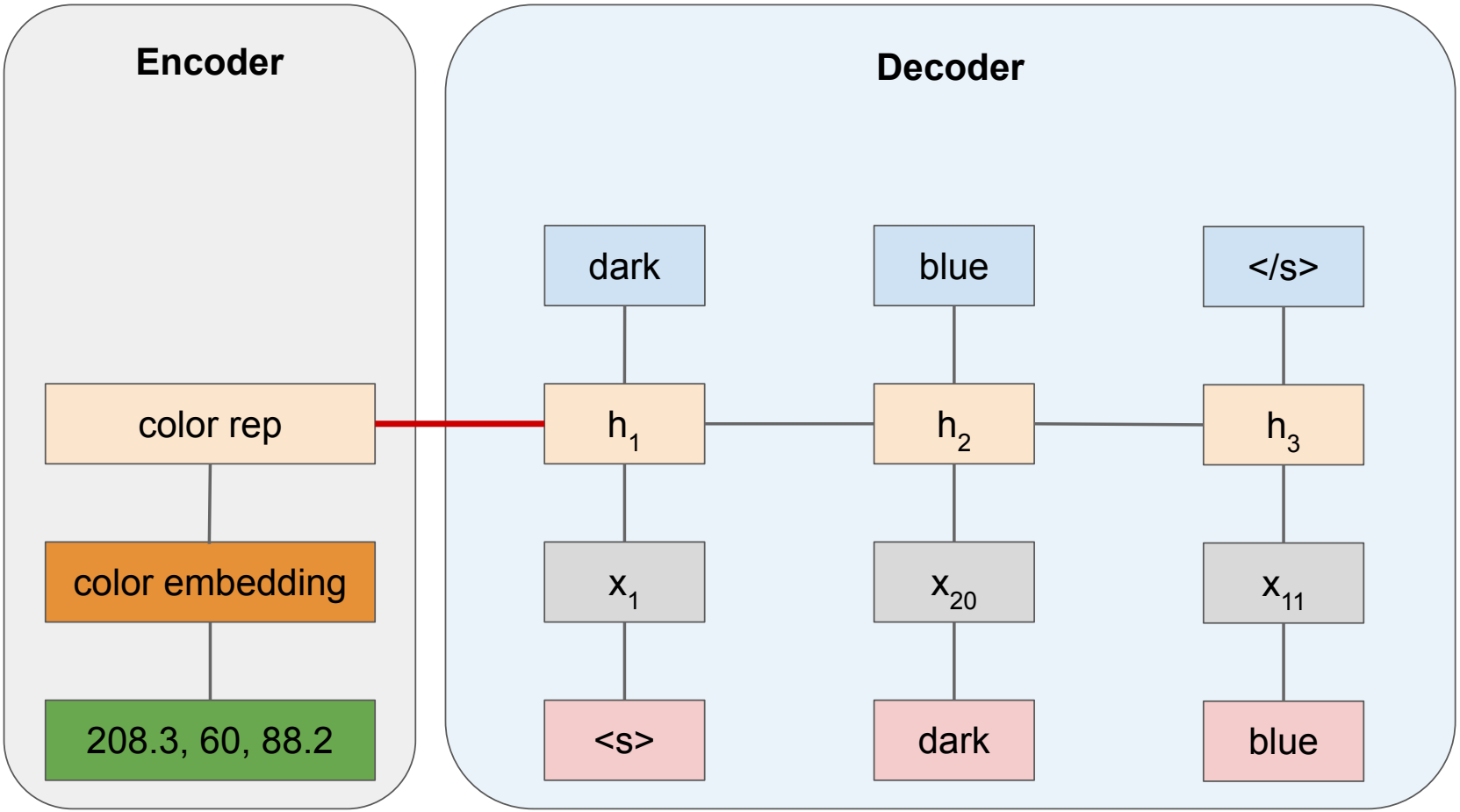
Color describer: Prediction



Color describer: Prediction



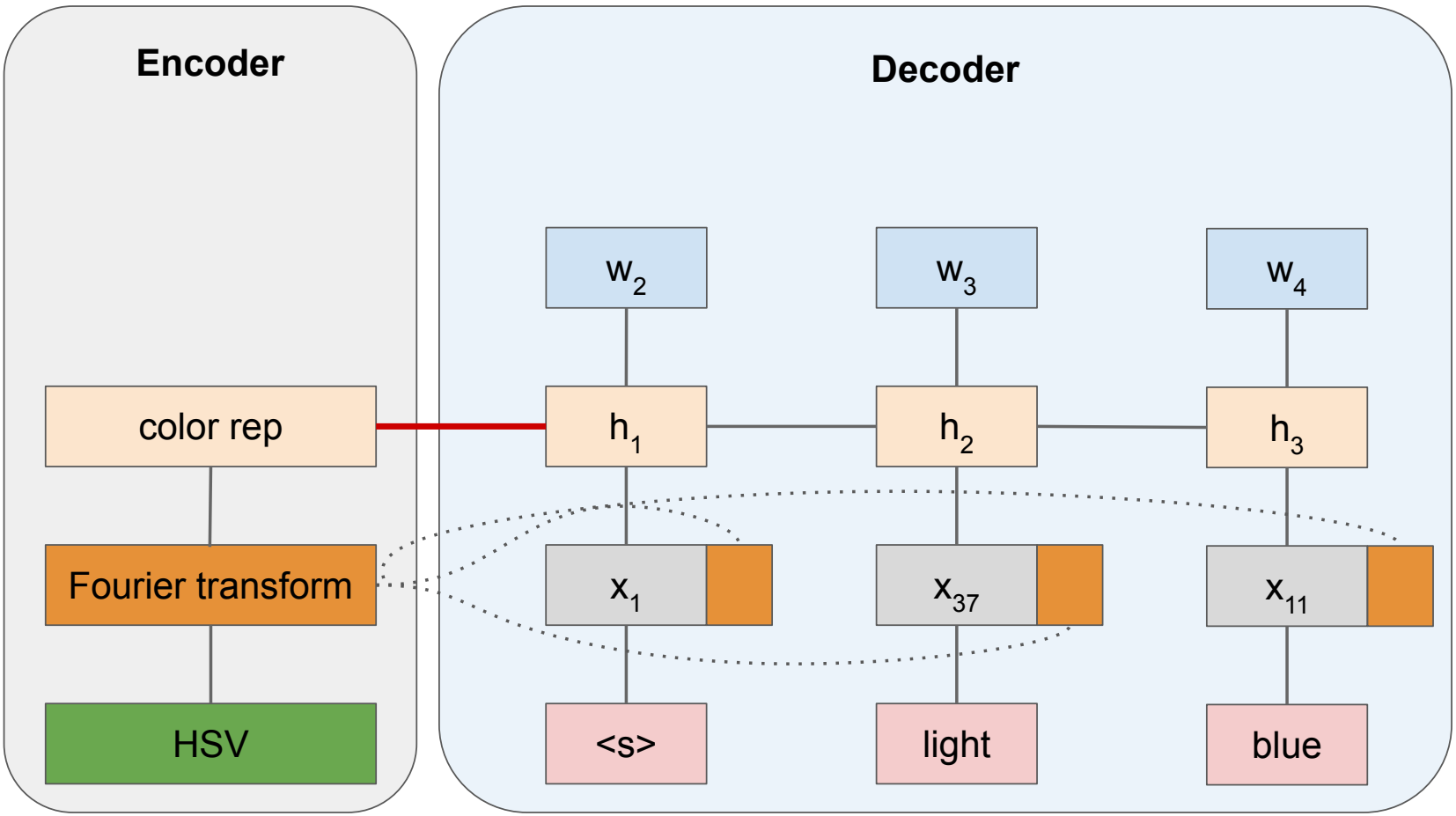
Color describer: Prediction



Miscellaneous design choices

- The Encoder and Decoder could have more hidden layers. We would expect the layer counts to match to facilitate the hand-off between Encoder and Decoder, though pooling or copying might work too.
- It seems very common at present for researchers to tie the embedding and classifier parameters (Press and Wolf 2017)
- During training, one might drop teacher forcing a small percentage of the time to encourage the model to explore.

Color describer of Monroe et al. (2016)



Related tasks

Non-linguistic representation ⇒ Language

- Image captioning
- Scene description
- Visual Question Answering
  (Image + Question-text ⇒ Answer-text)
- Instruction giving (State ⇒ Language)
- . . .

References I

Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Will Monroe, Noah D. Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248, Stroudsburg, PA. Association for Computational Linguistics.

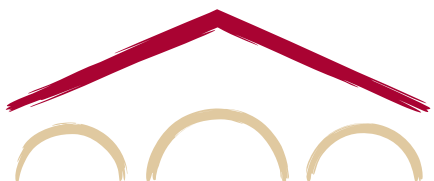
Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Grounded language understanding: Listeners: From language to the world

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding

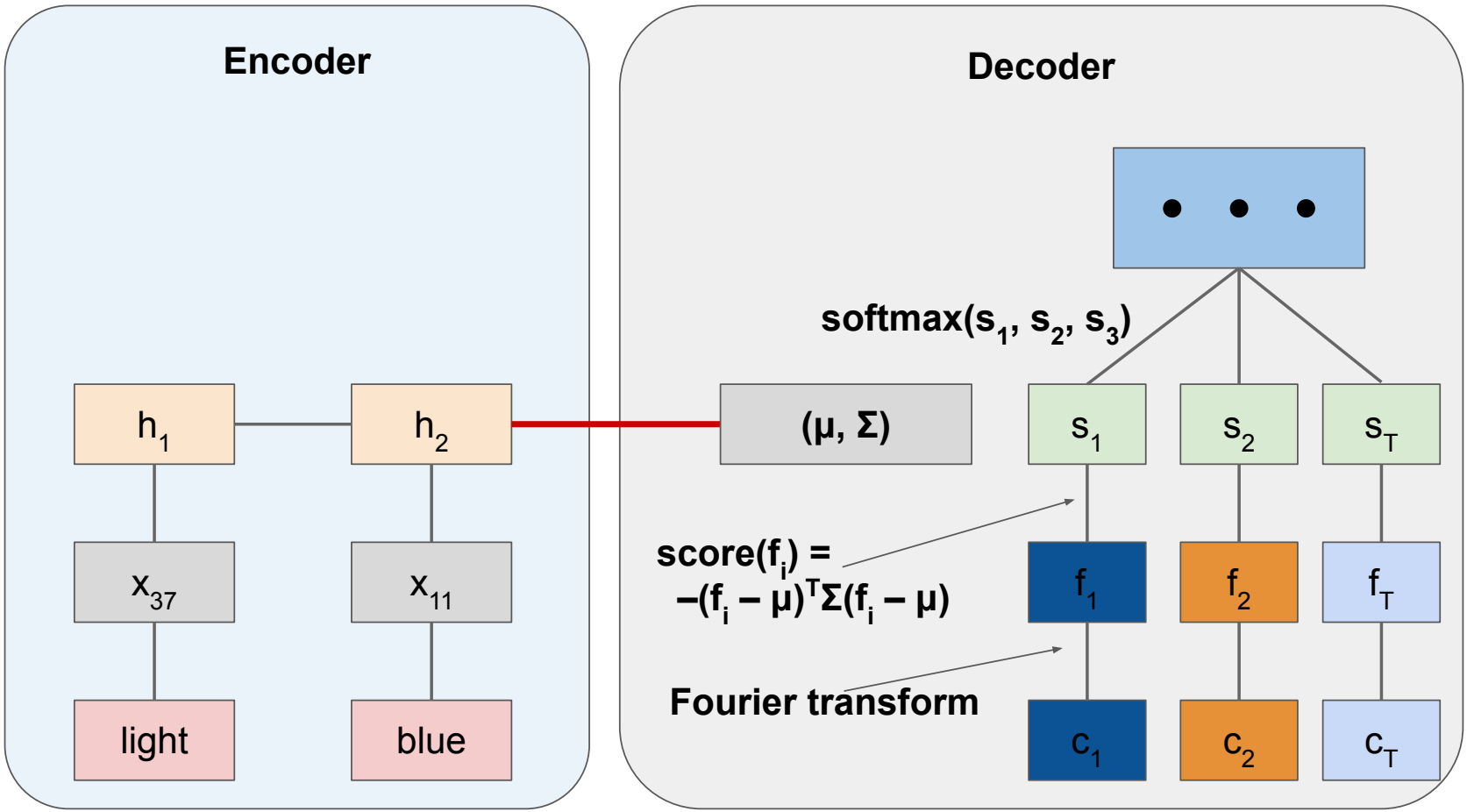


Color interpreter: Task formulation and data

Context	Utterance
	blue
	The darker blue one
	teal not the two that are more green
	dull pink not the super bright one
	not any of the regular greens
	Purple
	blue

Stanford Colors in Context corpus (Monroe et al. 2017)

A neural listener model



Other ideas and datasets

- NLU classifiers are very simple listeners: they consume language and make an inference in a structured space.
- Semantic parsers are very complex listeners: they consume language, construct rich latent representations, and predict into structured output spaces.
- Scene generation is the task of mapping language to structured representations of visual scenes (Seversky and Yin 2006; Chang et al. 2014, 2015).
- Young et al. (2014) seek to learn visual denotations for linguistic expressions.
- Mei et al. (2015) develop essentially a seq2seq version of the above model: given a linguistic input, they predict action sequences. (Kai Sheng Tai did his 2015 CS224u project on this, working at the same time as Mei et al.!)
- Suhr et al. (2019): Released the CerealBar data and game engine for learning to execute instructions.

References I

Angel Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D. Manning. 2015. Text to 3d scene generation with rich lexical grounding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 53–62, Stroudsburg, PA. Association for Computational Linguistics.

Angel Chang, Manolis Savva, and Christopher D. Manning. 2014. Learning spatial knowledge for text to 3D scene generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2028–2038, Doha, Qatar. Association for Computational Linguistics.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2015. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. ArXiv:1506.04089.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Lee M Seversky and Lijun Yin. 2006. Real-time automatic 3D scene generation from natural language voice and text descriptions. In *Proceedings of the 14th ACM International Conference on Multimedia*, pages 61–64. ACM.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2119–2130, Hong Kong, China. Association for Computational Linguistics.

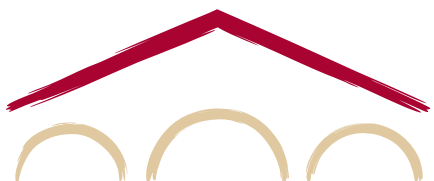
Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Grounded language understanding: Varieties of contextual grounding

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding



Our task: Color description in context

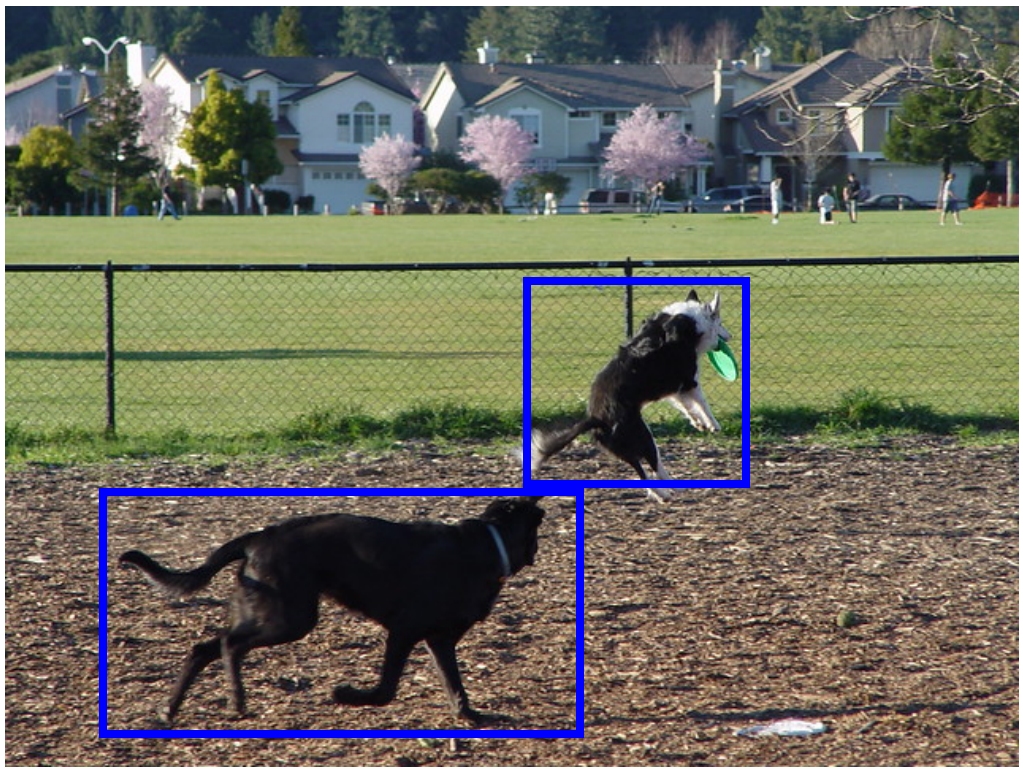
Context	Utterance
	blue
	The darker blue one
	teal not the two that are more green
	dull pink not the super bright one
	not any of the regular greens
	Purple
	blue

Stanford Colors in Context corpus (Monroe et al. 2017)

Discriminative image labeling

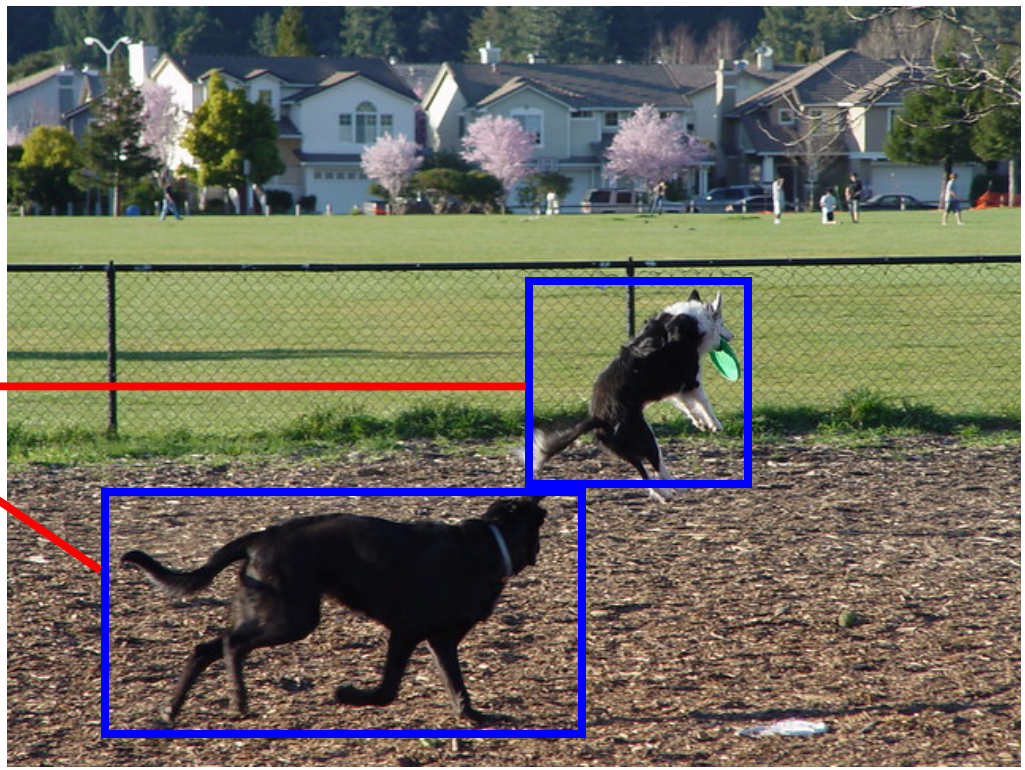
Mao et al. 2016

Discriminative image labeling



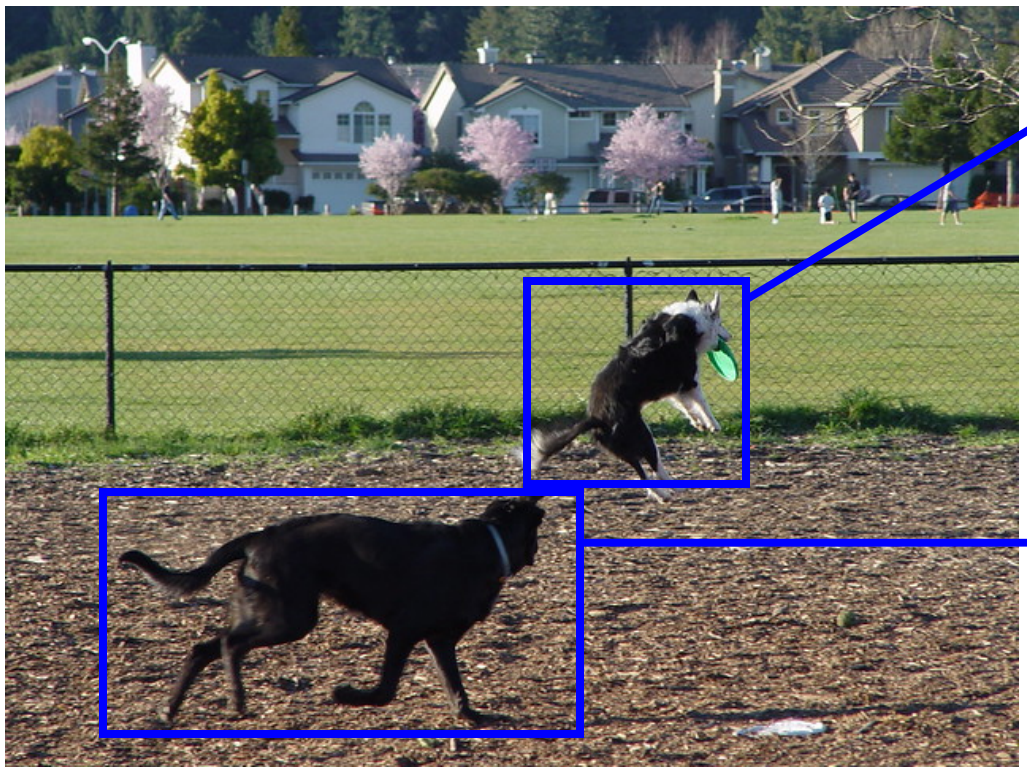
Mao et al. 2016

Discriminative image labeling



Mao et al. 2016

Discriminative image labeling

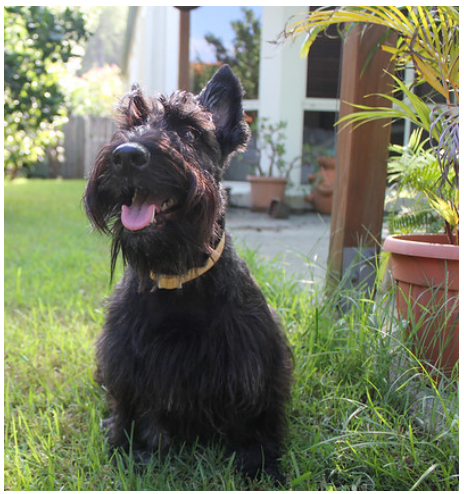
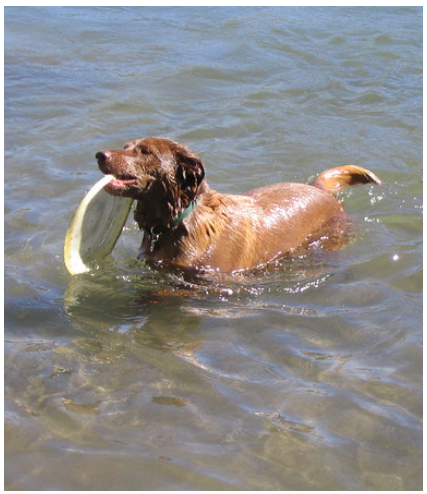


A little dog jumping and catching a frisbee

A big dog running

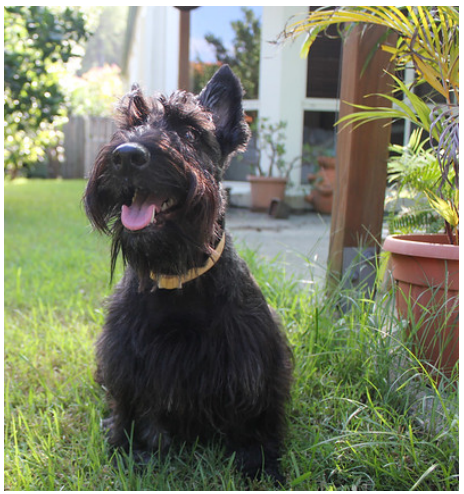
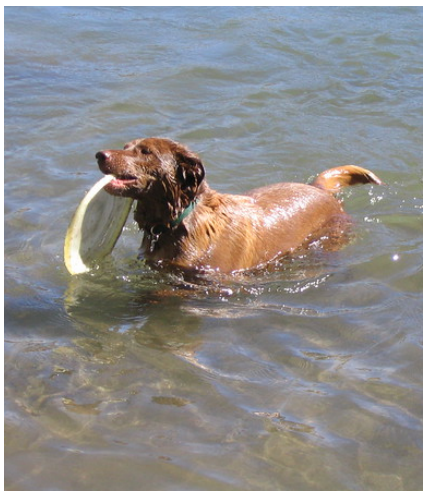
Mao et al. 2016

Discriminative image captioning

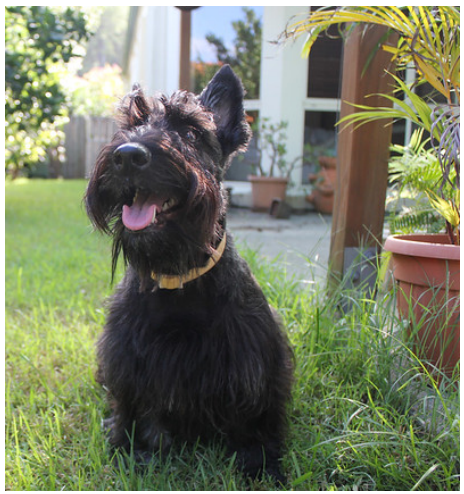
Vedantam et al. 2017; Cohn-Gordon et al. 2018

Discriminative image captioning

Vedantam et al. 2017; Cohn-Gordon et al. 2018

Discriminative image captioning


Vedantam et al. 2017; Cohn-Gordon et al. 2018

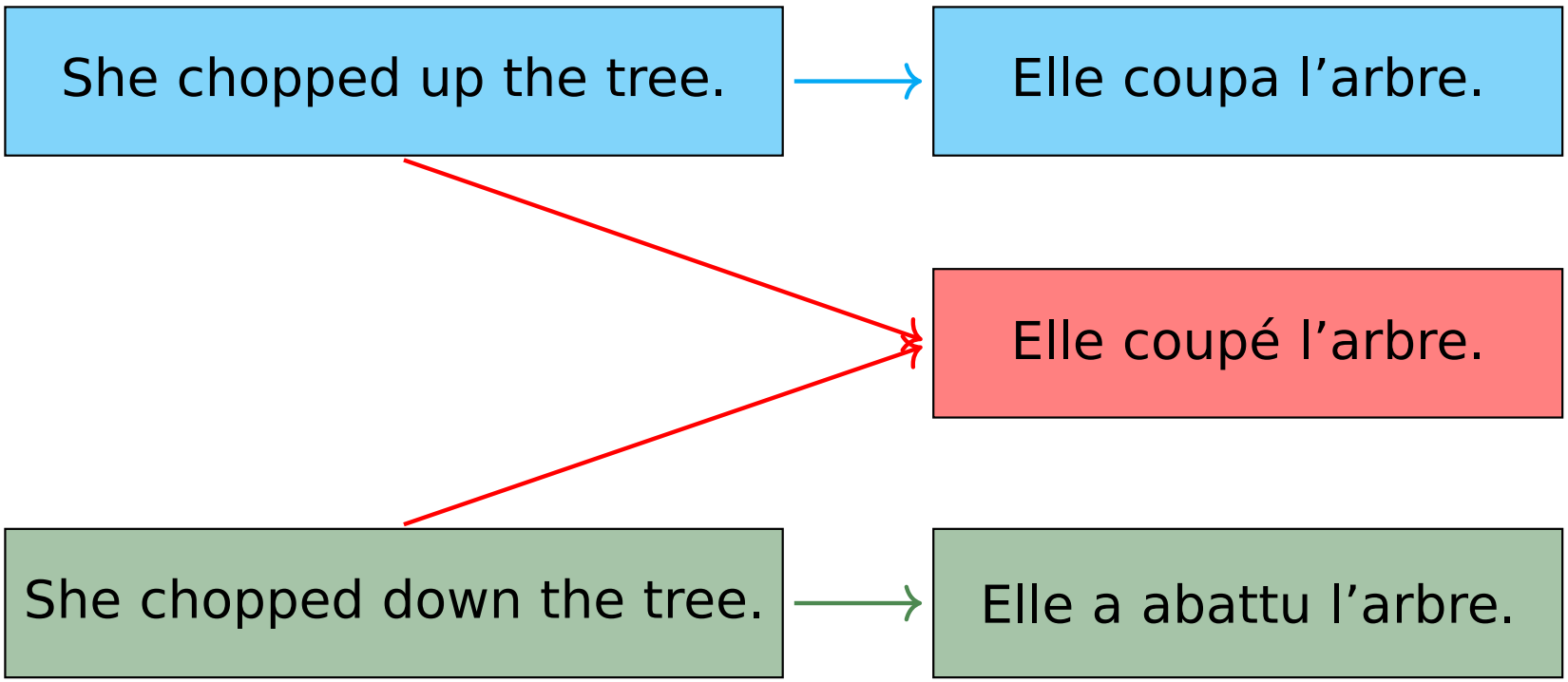
Machine translation

She chopped up the tree. → Elle coupa l'arbre.

She chopped down the tree. → Elle a abattu l'arbre.

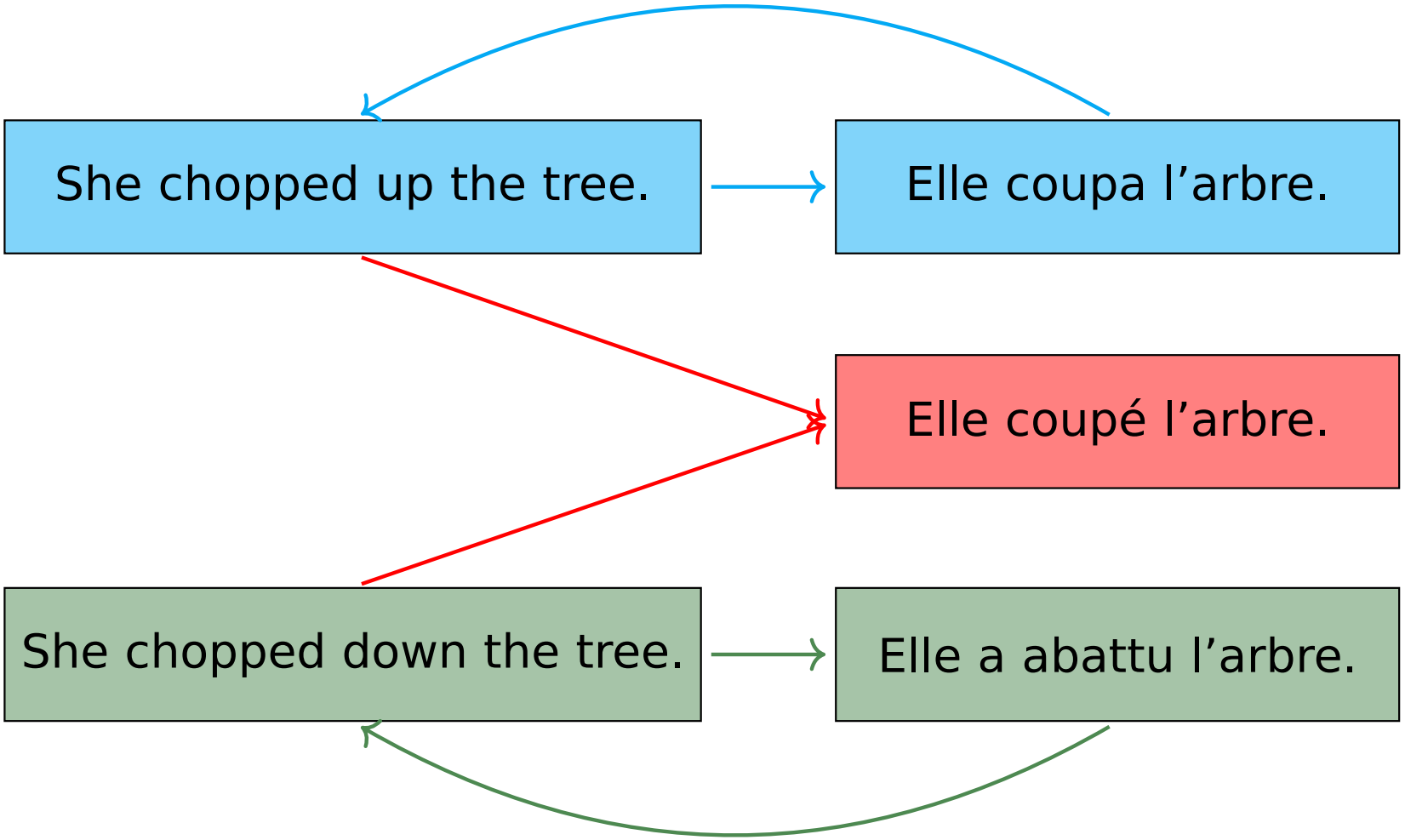
Cohn-Gordon and Goodman 2019

Machine translation



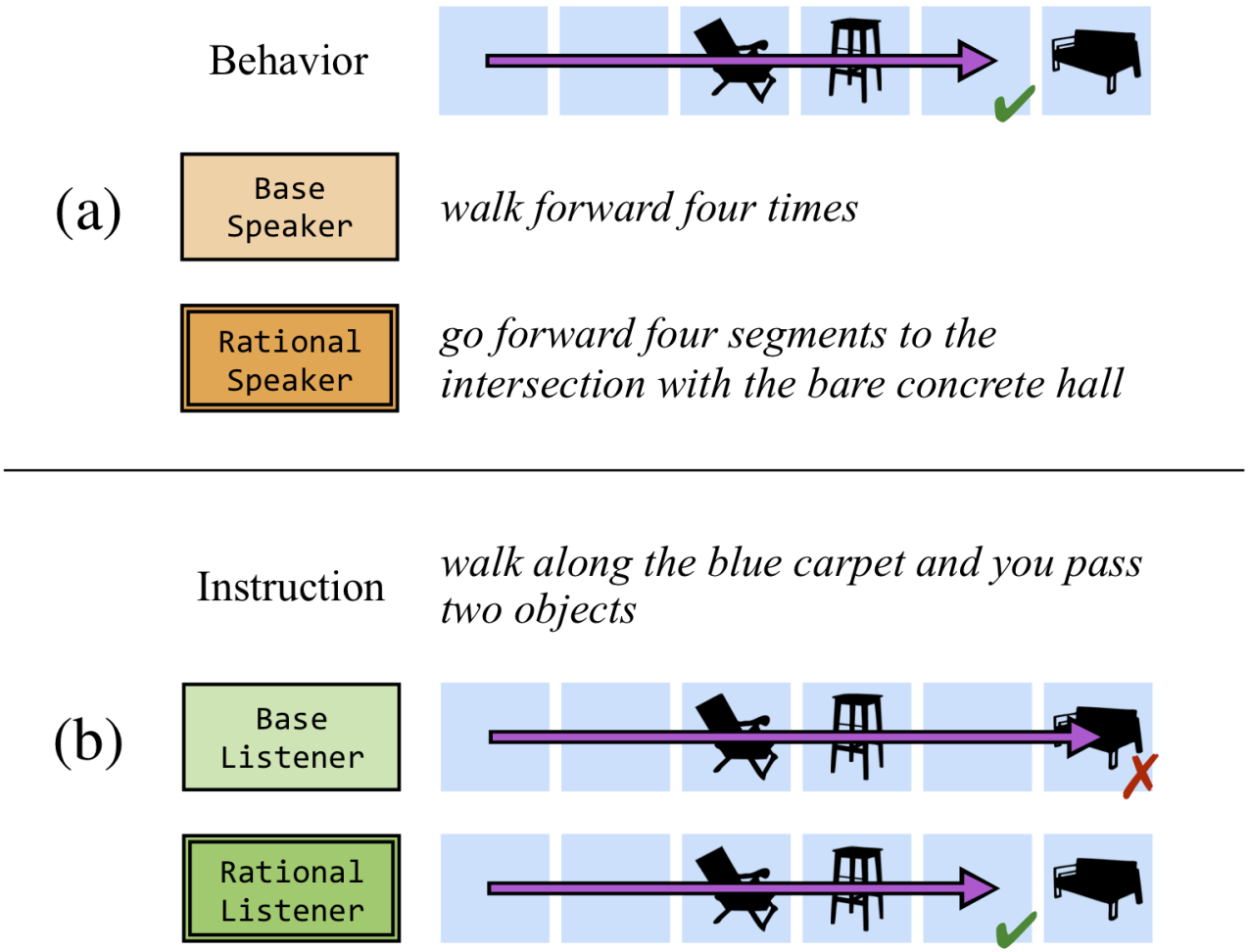
Cohn-Gordon and Goodman 2019

Machine translation



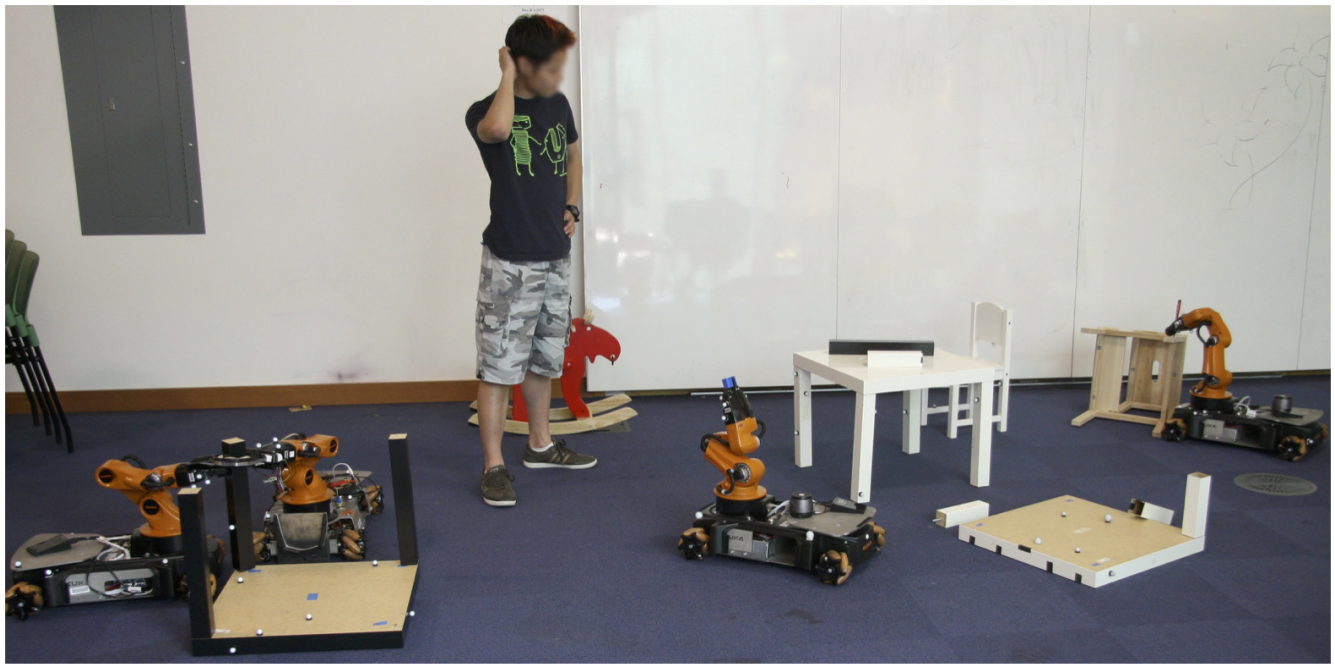
Cohn-Gordon and Goodman 2019

Generating and following instructions



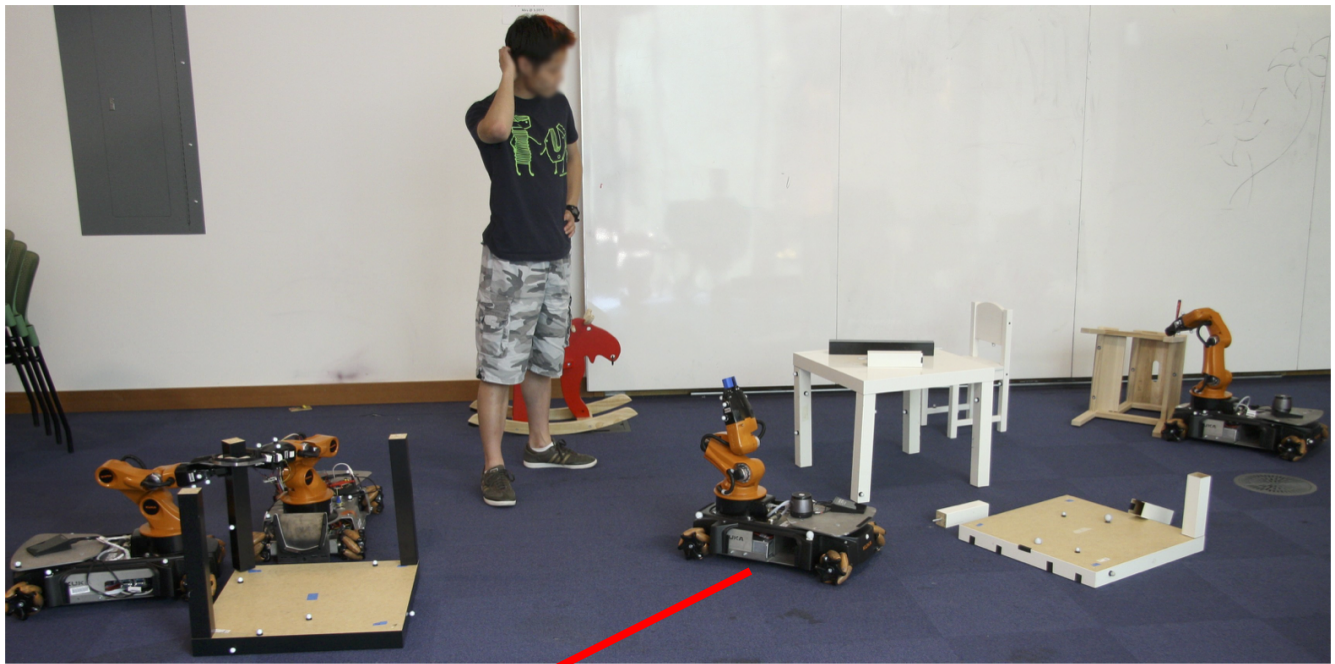
Fried et al. 2018

Collaborative problem solving



Tellex et al. 2014

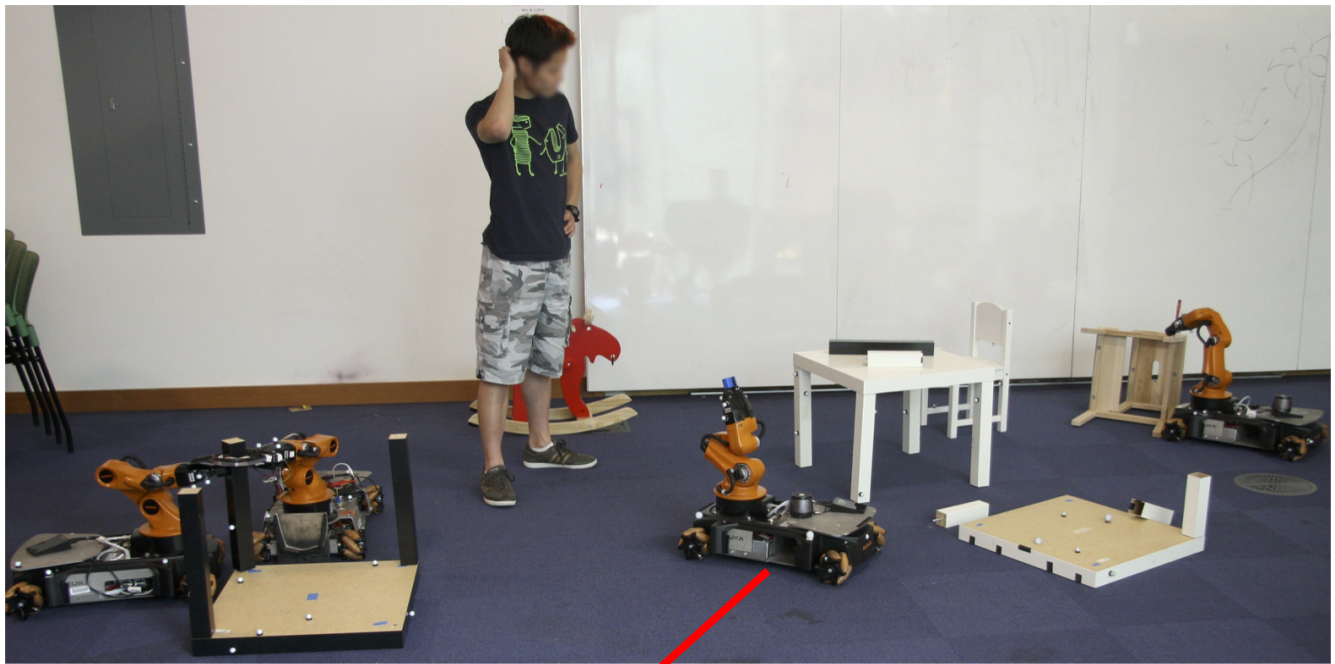
Collaborative problem solving



Help me!

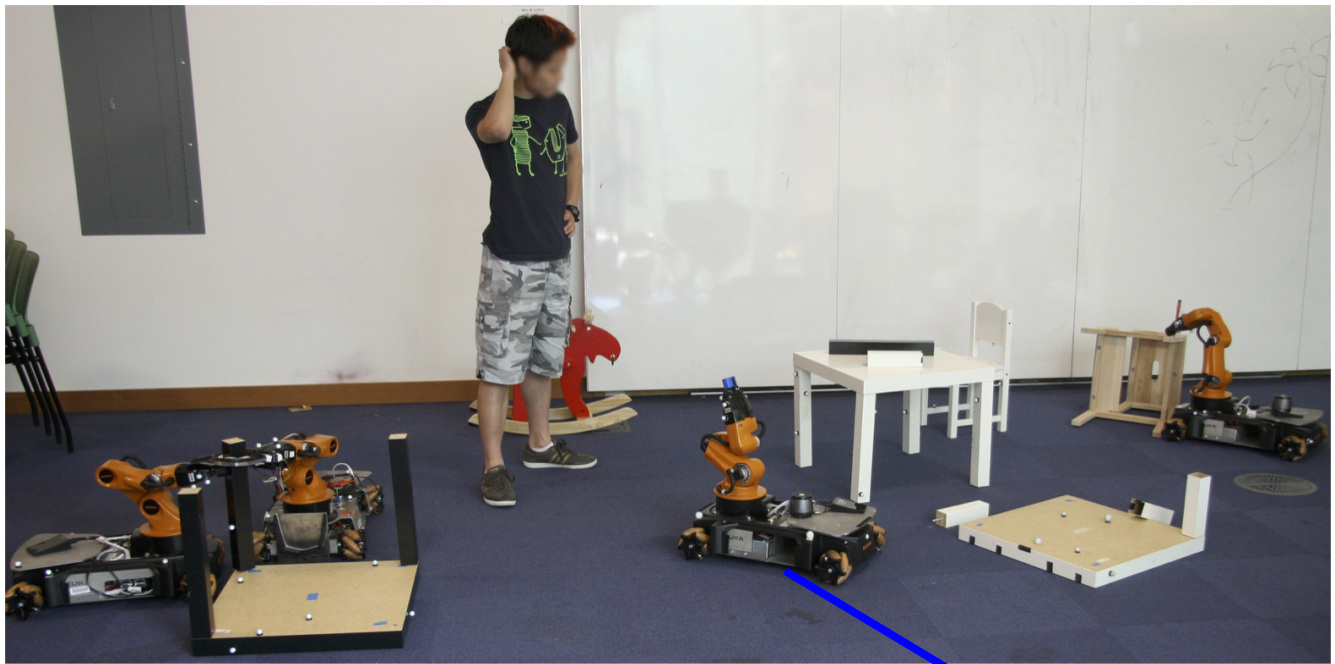
Tellex et al. 2014

Collaborative problem solving



Tellex et al. 2014

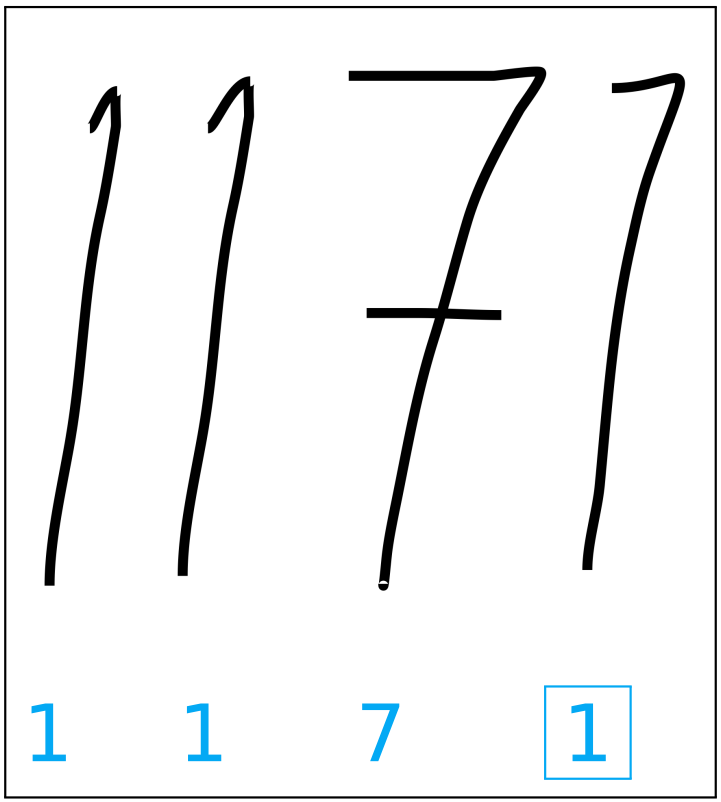
Collaborative problem solving

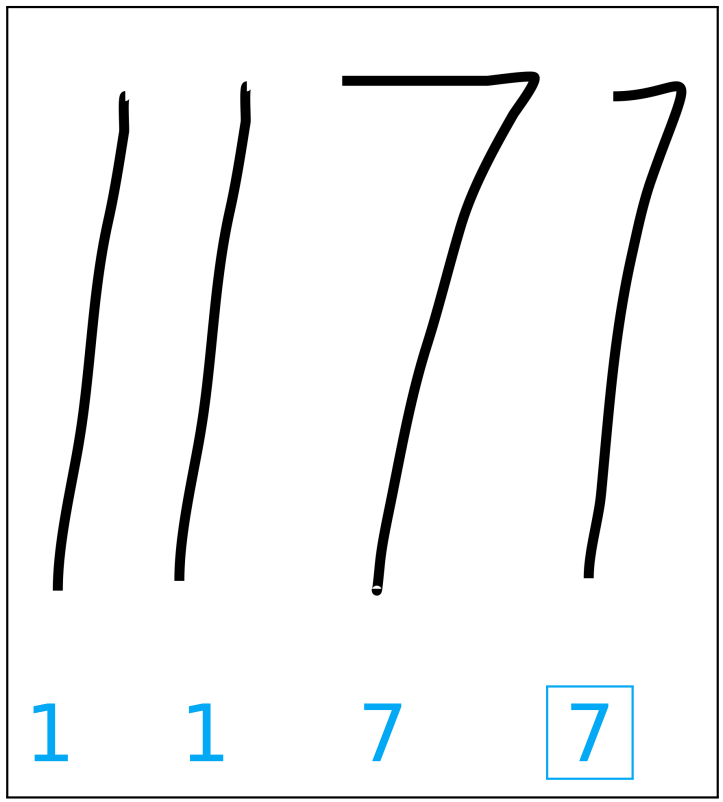


Hand me the white leg on the table

Tellex et al. 2014

Optical character recognition





References I

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in machine translation: A method to reduce meaning loss. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 437–441, Minneapolis, Minnesota. Association for Computational Linguistics.

Reuben Cohn-Gordon, Noah D. Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 439–443, Stroudsburg, PA. Association for Computational Linguistics.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified pragmatic models for generating and following instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20. IEEE.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Stefanie Tellex, Ross A. Knepper, Adrian Li, Thomas M. Howard, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. In *Proceedings of Robotics: Science and Systems*.

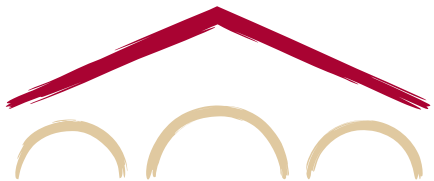
Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. *arXiv:1701.02870*.

Grounded language understanding: The Rational Speech Acts model

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding



Additional resources

1. Goodman and Frank 2016
2. Technical screencast: https://youtu.be/bPd6CNy5UqA
3. Associated slides: https://web.stanford.edu/class/linguist130a/screencasts/130a-screencast-implicature.pdf
4. Reference implementation: https://web.stanford.edu/class/linguist130a/materials/rsa130a.py

Pragmatic listeners

Pragmatic listeners

Literal listener

$$L_{\text{lit}}(state \mid msg) = \frac{[\![msg, state]\!] P(state)}{\sum_{state'} [\![msg, state']\!] P(state')}$$

Pragmatic listeners

Pragmatic speaker

$$S_{\text{prag}}(msg \mid state) = \frac{\exp\left(\alpha\left(\log L_{\text{lit}}(state \mid msg) - C(msg)\right)\right)}{\sum_{msg'} \exp\left(\alpha\left(\log L_{\text{lit}}(state \mid msg') - C(msg')\right)\right)}$$

Literal listener

$$L_{\text{lit}}(state \mid msg) = \frac{[\![msg, state]\!] P(state)}{\sum_{state'} [\![msg, state']\!] P(state')}$$

Pragmatic listeners

Pragmatic listener

$$L_{\text{prag}}(state \mid msg) = \frac{S_{\text{prag}}(msg \mid state)P(state)}{\sum_{state'} S_{\text{prag}}(msg \mid state')P(state')}$$

Pragmatic speaker

$$S_{\text{prag}}(msg \mid state) = \frac{\exp\left(\alpha\left(\log L_{\text{lit}}(state \mid msg) - C(msg)\right)\right)}{\sum_{msg'} \exp\left(\alpha\left(\log L_{\text{lit}}(state \mid msg') - C(msg')\right)\right)}$$

Literal listener

$$L_{\text{lit}}(state \mid msg) = \frac{[\![msg, state]\!]P(state)}{\sum_{state'} [\![msg, state']\!]P(state')}$$

Pragmatic listeners

Pragmatic listener

$$L_{\text{prag}}(state \mid msg) = \textbf{pragmatic speaker} \times \text{state prior}$$

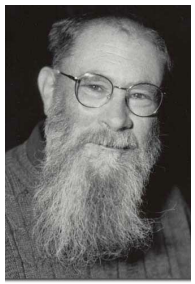
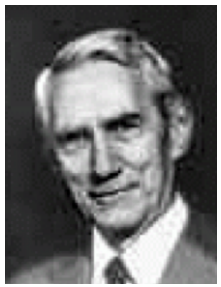
Pragmatic speaker

$$S_{\text{prag}}(msg \mid state) = \textbf{literal listener} - \text{message costs}$$

Literal listener

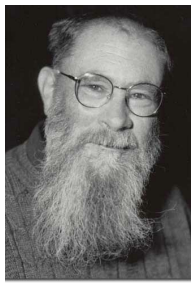
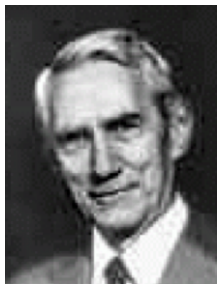
$$L_{\text{lit}}(state \mid msg) = \textbf{lexicon} \times \text{state prior}$$

A simple example

			
beard	1	0	0
glasses	1	1	0
tie	0	1	1

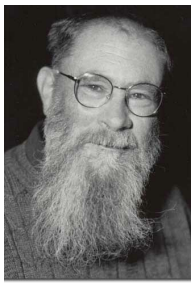
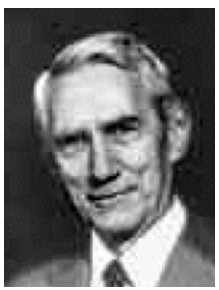
L_{prag}
S_{prag}
L_{lit}
$[\![\cdot]\!]$

A simple example

			
beard	**1**	0	0
glasses	.5	.5	0
tie	0	.5	.5

L_{prag}
S_{prag}
L_{lit}
$[\![\cdot]\!]$

A simple example

	beard	*glasses*	*tie*
	**.67**	.33	0
	0	.5	.5
	0	0	**1**

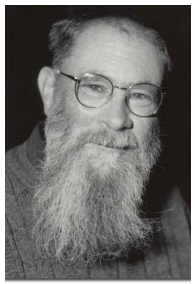
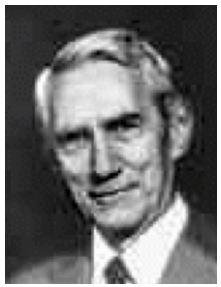
L_{prag}

S_{prag}

L_{lit}

$[\![\cdot]\!]$

A simple example

			
beard	**1**	0	0
glasses	.4	**.6**	0
tie	0	.33	**.67**

L_{prag}
S_{prag}
L_{lit}
$[\![\cdot]\!]$

Pragmatic speakers

Pragmatic speakers

Literal speaker

$$S_{\text{lit}}(msg \mid state) = \frac{\exp\left(\alpha\left(\log[\![msg, state]\!] - C(msg)\right)\right)}{\sum_{msg'} \exp\left(\alpha\left(\log[\![msg', state]\!] - C(msg')\right)\right)}$$

Pragmatic speakers

Pragmatic listener

$$L_{\text{prag}}(state \mid msg) = \frac{S_{\text{lit}}(msg \mid state)P(state)}{\sum_{state'} S_{\text{lit}}(msg \mid state')P(state')}$$

Literal speaker

$$S_{\text{lit}}(msg \mid state) = \frac{\exp\left(\alpha\left(\log[\![msg, state]\!] - C(msg)\right)\right)}{\sum_{msg'} \exp\left(\alpha\left(\log[\![msg', state]\!] - C(msg')\right)\right)}$$

Pragmatic speakers

Pragmatic speaker

$$S_{\text{prag}}(msg \mid state) = \frac{\exp\left(\alpha\left(\log L_{\text{prag}}(state \mid msg) - C(msg)\right)\right)}{\sum_{msg'} \exp\left(\alpha\left(\log L_{\text{prag}}(state \mid msg') - C(msg')\right)\right)}$$

Pragmatic listener

$$L_{\text{prag}}(state \mid msg) = \frac{S_{\text{lit}}(msg \mid state)P(state)}{\sum_{state'} S_{\text{lit}}(msg \mid state')P(state')}$$

Literal speaker

$$S_{\text{lit}}(msg \mid state) = \frac{\exp\left(\alpha\left(\log[\![msg, state]\!] - C(msg)\right)\right)}{\sum_{msg'} \exp\left(\alpha\left(\log[\![msg', state]\!] - C(msg')\right)\right)}$$

Pragmatic speakers

Pragmatic speaker

$$S_{\text{prag}}(msg \mid state) = \textbf{pragmatic listener} - \text{message costs}$$

Pragmatic listener

$$L_{\text{prag}}(state \mid msg) = \textbf{literal speaker} \times \text{state prior}$$

Literal speaker

$$S_{\text{lit}}(msg \mid state) = \textbf{lexicon} - \text{message costs}$$

Limitations

- Hand-specified lexicon
- Reasoning about *all* possible utterances?

$$S_{\text{prag}}(msg \mid state) = \frac{\exp\left(\alpha\left(\log L_{\text{lit}}(state \mid msg) - C(msg)\right)\right)}{\sum_{msg'} \exp\left(\alpha\left(\log L_{\text{lit}}(state \mid msg') - C(msg')\right)\right)}$$

- High-bias model; few chances to learn from data
- Cognitive demands limit speaker rationality
- Speaker preferences
- Scalability

References I

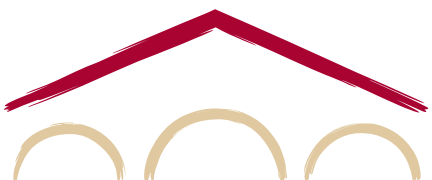
Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Grounded language understanding: Neural RSA

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding



Papers employing these techniques

- Andreas and Klein 2016
- Fried et al. 2018
- Monroe et al. 2017
- Monroe et al. 2018

Motivation

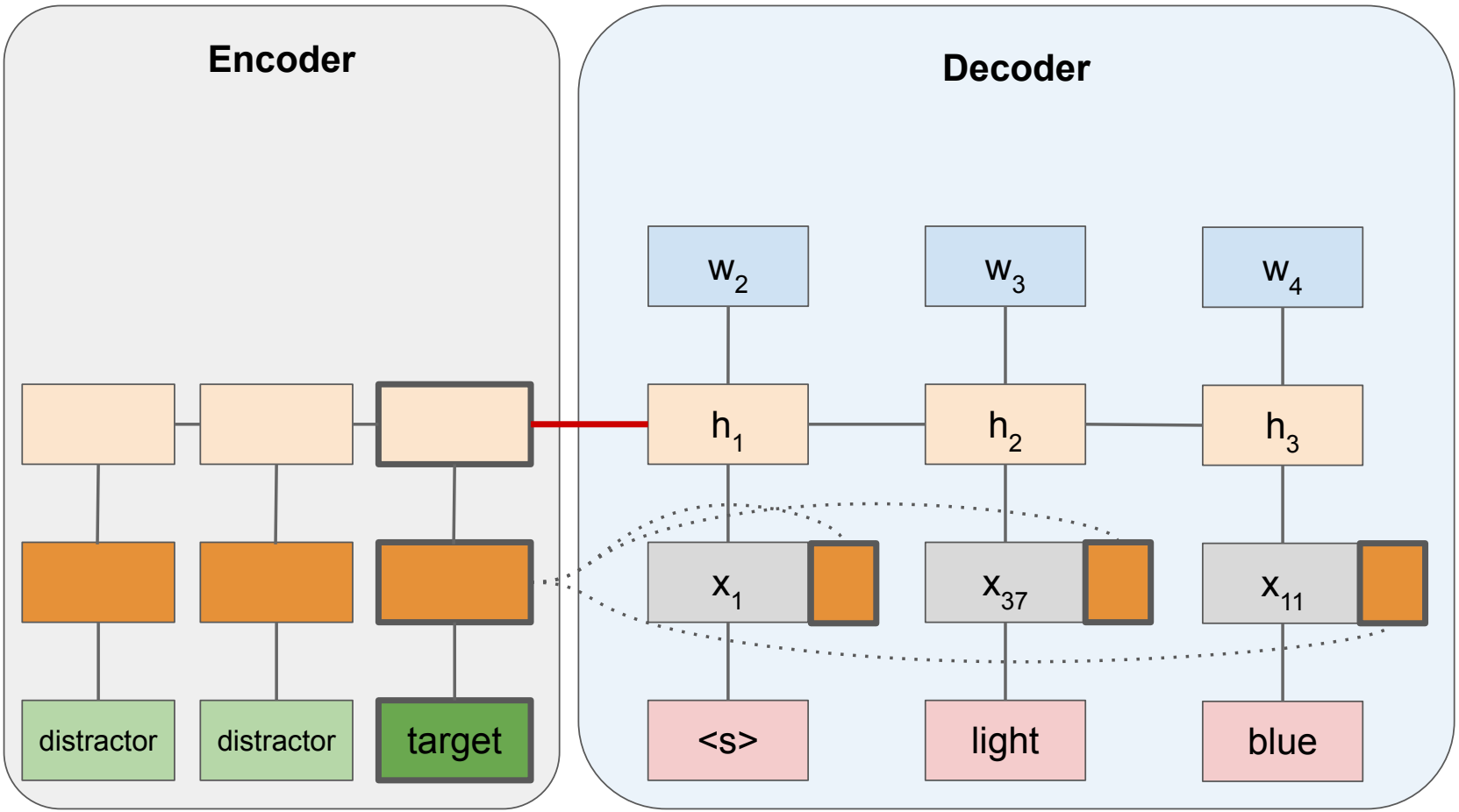
- Discriminative image labeling
- Image captioning
- Machine translation
- Collaborative problem solving
- Interpreting complex descriptions
- Optical Character Recognition
- Scalability
- Sensitivity to variation
- Bounded rationality
- New kinds of model assessment
- Impact

Colors in context

Context	Utterance
	blue
	The darker blue one
	teal not the two that are more green
	dull pink not the super bright one
	not any of the regular greens
	Purple
	blue

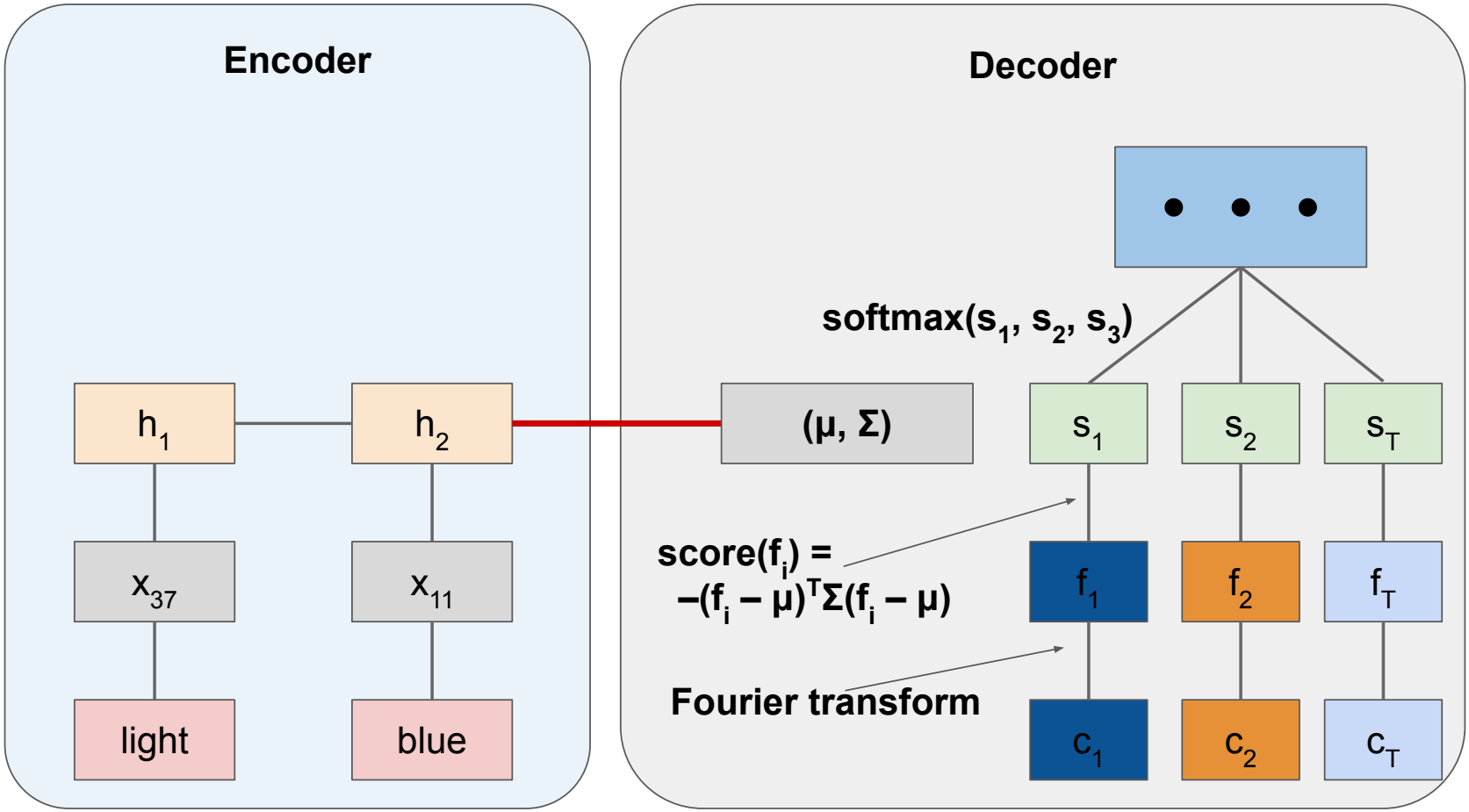
Stanford Colors in Context corpus (Monroe et al. 2017)

Literal neural speaker $\mathbf{S}^{\theta}_{\text{lit}}$



Monroe et al. 2017

Neural literal listener



Monroe et al. 2017

Neural pragmatic agents

Neural pragmatic speaker (Andreas and Klein 2016)

$$\mathbf{S}^{\theta}_{\text{prag}}(msg \mid state) = \frac{\mathbf{L}^{\theta}_{0}(state \mid msg)}{\sum_{msg' \in X} \mathbf{L}^{\theta}_{0}(state \mid msg')}$$

with X a sample from $\mathbf{S}^{\theta}_{\text{lit}}(msg \mid state)$ such that $msg \in X$.

Neural pragmatic listener

$$\mathbf{L}^{\theta}_{1}(state \mid msg) \propto \mathbf{S}^{\theta}_{\text{prag}}(msg \mid state)$$

Blended neural pragmatic listener

Weighted combination of $\mathbf{L}^{\theta}_{0}$ and $\mathbf{L}^{\theta}_{1}$.

Other related work

- Golland et al. (2010): Recursive speaker/listener reasoning as part of interpreting complex utterances compositionally, with grounding in a simple visual world.
- Wang et al. (2016): Pragmatic reasoning helps in online learning of semantic parsers.
- Tellex et al.'s (2014) Inverse Semantics: Robot utterances are scored by models similar to RSA's pragmatic speakers.
- Khani et al. (2018): Collaborative games with pragmatic reasoning.
- Cohn-Gordon and Goodman (2019): RSA for translation
- Cohn-Gordon et al. (2018, 2019): Word- and character-level RSA
- Monroe and Potts (2015): "RSA as a hidden activation function"
- Mao et al. 2016: pragmatic learning objectives

References I

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182. Association for Computational Linguistics.

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in machine translation: A method to reduce meaning loss. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 437–441, Minneapolis, Minnesota. Association for Computational Linguistics.

Reuben Cohn-Gordon, Noah D. Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 439–443, Stroudsburg, PA. Association for Computational Linguistics.

Reuben Cohn-Gordon, Noah D. Goodman, and Christopher Potts. 2019. An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics*, pages 81–90, Washington, D.C. Linguistic Society of America.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified pragmatic models for generating and following instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Stroudsburg, PA. ACL.

Fereshte Khani, Noah D. Goodman, and Percy Liang. 2018. Planning, inference and pragmatics in sequential language games. *Transactions of the Association for Computational Linguistics*, 6:543–555.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20. IEEE.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Will Monroe, Jennifer Hu, Andrew Jong, and Christopher Potts. 2018. Generating bilingual pragmatic color references. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2155–2165, Stroudsburg, PA. Association for Computational Linguistics.

Will Monroe and Christopher Potts. 2015. Learning in the Rational Speech Acts model. In *Proceedings of 20th Amsterdam Colloquium*, Amsterdam. ILLC.

References II

Stefanie Tellex, Ross A. Knepper, Adrian Li, Thomas M. Howard, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. In *Proceedings of Robotics: Science and Systems*.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2368–2378. Association for Computational Linguistics.