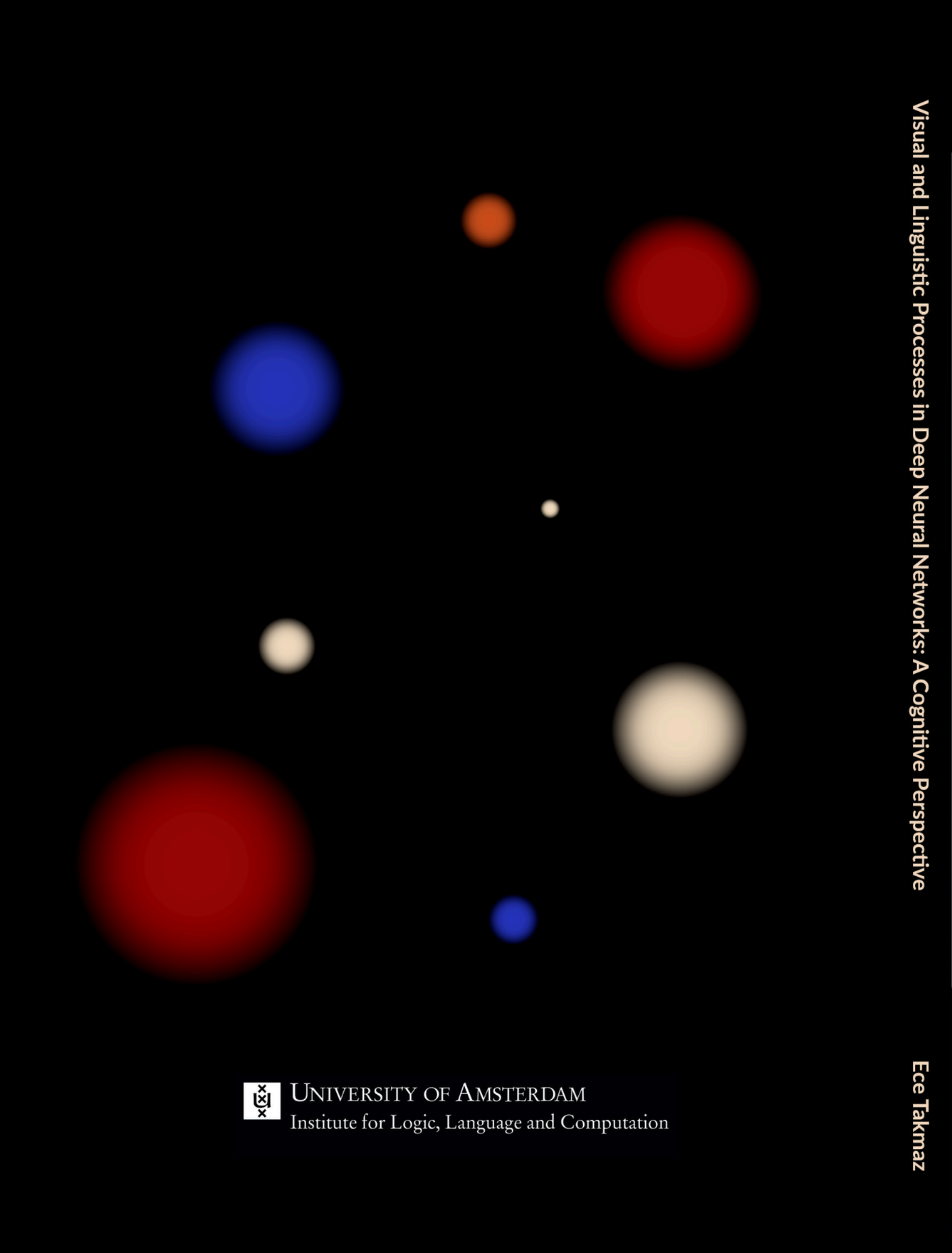


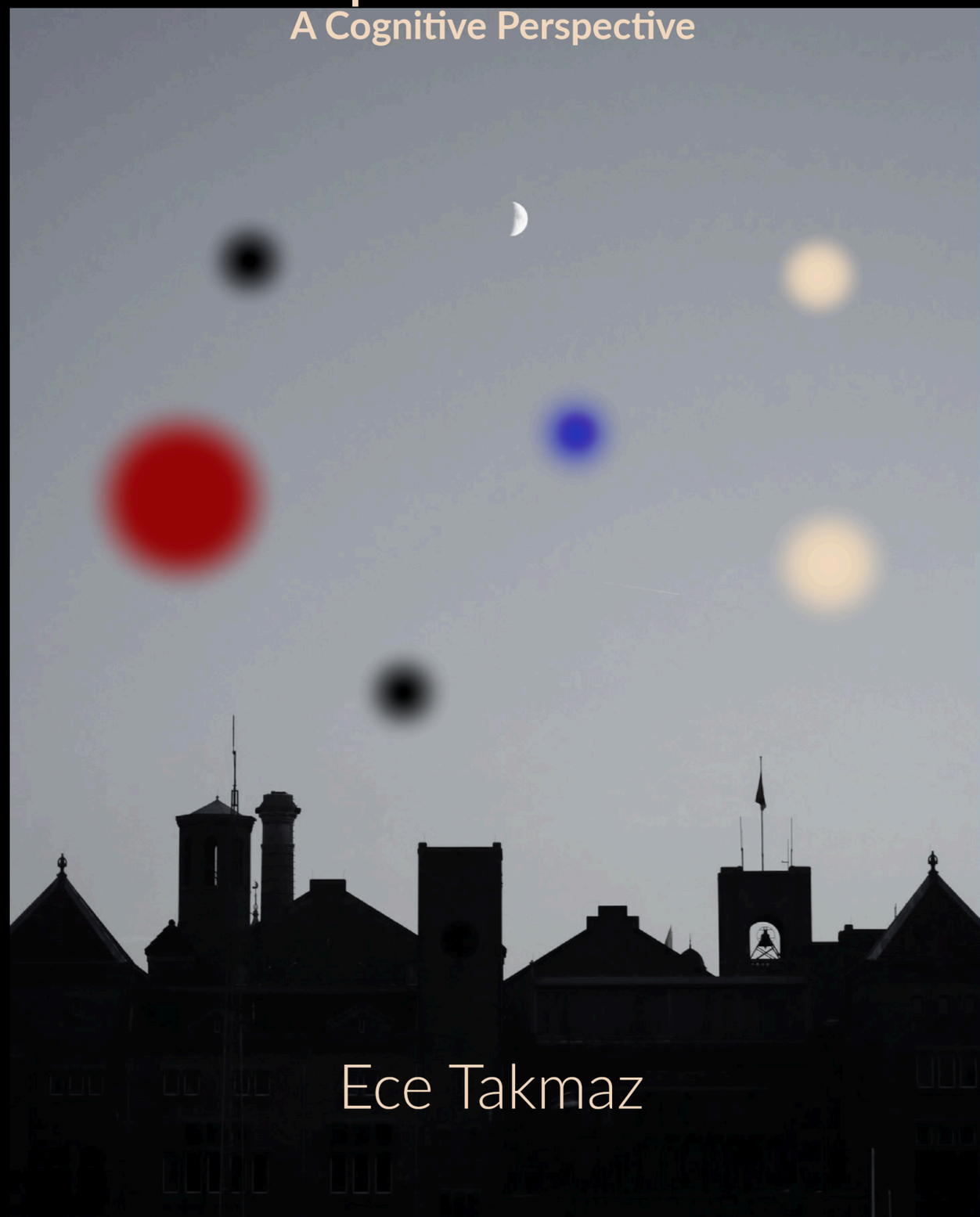
UNIVERSITY OF AMSTERDAM
Institute for Logic, Language and Computation

Visual and Linguistic Processes in Deep Neural Networks: A Cognitive Perspective

Ece Takmaz

Visual and Linguistic Processes in Deep Neural Networks:

A Cognitive Perspective



Ece Takmaz

Visual and Linguistic Processes in Deep Neural Networks:

A Cognitive Perspective

Ece Takmaz

Visual and Linguistic Processes in Deep Neural Networks:

A Cognitive Perspective

ILLC Dissertation Series DS-2024-04

Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

The research for this doctoral thesis has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 819455).

Cover design by Ece Takmaz.
Printed and bound by Ipskamp Printing.

ISBN: 978-94-6473-474-4

Visual and Linguistic Processes in Deep Neural Networks: A Cognitive Perspective

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 17 mei 2024, te 11.00 uur

door Ece Kamer Takmaz
geboren te Altındağ

Promotiecommissie

Promotor:	prof. dr. R. Fernández Rovira	Universiteit van Amsterdam
Copromotor:	dr. S. Pezzelle	Universiteit van Amsterdam
Overige leden:	prof. dr. A. Gatt	Universiteit Utrecht
	prof. dr. S. Zarrieß	University of Bielefeld
	dr. W.H. Zuidema	Universiteit van Amsterdam
	prof. dr. K. Sima'an	Universiteit van Amsterdam
	dr. E.V. Shutova	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Contents

Acknowledgments

I am extremely grateful to Raquel Fernández for being my supervisor and a great role model. I consider myself lucky that my path has crossed with hers. Thank you, Raquel, for guiding me and grounding my research ideas back down to earth, helping me bring structure to my work that culminated in this thesis.

I am equally indebted to my co-supervisor, Sandro Pezzelle; his out-of-the-box thinking and suggestions were always helpful in getting me out of a conundrum. It must also be known that he is the best acronym-finder and the title-conceiver I have ever had the chance to observe in the act.

Thanks to the understanding and encouragement provided by Raquel and Sandro, I was able to see the light at the end of the tunnel in my PhD; for this, I am immensely grateful.

I would like to thank Lisa Beinborn for always supporting me and providing tips on surviving academia. I appreciate your strong support that continued long after we worked together.

Arzu Çöltekin, thank you very much for taking the time to have semi-regular online meetings with me to talk about research and life from a perspective that resonated a lot with me. I sincerely thank you for your mentorship.

I give my warmest thanks to Ekaterina Shutova, Jelle Zuidema, Wilker Aziz, Khalil Sima'an, and Jelke Bloem, from whom I have learned a lot over my years at ILLC. I would also like to thank the members of the ILLC office for their help and friendliness.

Many thanks to Nora Hollenstein for the very nice month I spent at the University of Copenhagen, and also thanks to the people at the Centre for Language Technology for their warmth, which immediately made me feel welcome.

Past and current members and visitors of the Dialogue Modelling Group, Mario, Janie, Anna, Vera, Esam, Alberto, Michael, Adi, Joris, Jonas, Xinyi, and Nico, thank you for the collegiality and excellent discussions during meetings, as well as the fun chats during breaks and lunchtime, which often ventured far from academia. I am sure you all know by now that I can go on and write paragraphs

about each one of you, but I don't intend to add another chapter to this thesis. So, here, I will just say, it was wonderful being colleagues with you all smart and friendly people!

Many thanks to my paranymphs Jaap and Oskar. Jaap, I will never forget the time we suited up and had a fancy dinner sitting next to some big names in NLP & ML (I will not name-drop here), also the time we finally managed to play the guitar together, in Spain, of all places. Oskar, thank you for your quips on my Instagram stories, insights into (and humorous takes on) Dutch culture and for encouraging me to converse in Dutch.

Marianne, my initial paranymph, it was great having another fellow cognitive science enthusiast around, whom I vaguely recall meeting at a summer school as early as 8 years ago. Thank you also for accompanying me during my citizenship ceremony.

Thanks to Evgenia for being an amazing person; I would have loved to continue sharing an office with you, as you are the friendliest officemate one can imagine!

Thanks to current and former PhD candidates Rochelle, Charlotte, David, Bryan, Pedro, Alina, Simon, Puyu, and Julian for making working at ILLC fun!

UPF COLT team and all the attendees of the PhD retreats we organized in Spain; those were great times!

Ahmet, initially hailing from the amazing Groningen NLP team, thanks for taking the time to chat about the state of academia. Although you have moved to the industry, I am happy we continue discussing the state of NLP research.

Sam, Gulfaraz, and Akash, thank you for being friends since the first days of the AI master's at UvA. Back then, I would have never guessed that we would form a band with 3 guitars and a cello, play songs on stage, play Dungeons & Dragons online during a pandemic, and become Dutch together. Spending time with you has always been a pleasant diversion for me. Looking forward to 'putting the band back together' once I am finally done with this thesis.

I also thank my friends Masoumeh, Cecilia, and Christina, who kept in touch with me long after we graduated from the master's at UvA, even though we all have been busy in so many ways. Likewise, many thanks to my street photography buddy, Aranka, and my former neighbor, Ellen, for our lovely walks in Amsterdam and Utrecht.

Fuat abi of the famous Science Döner at Science Park, his son, and brother, I express my sincere thanks to you for never failing to ask me how I am doing when I pass by your truck on my way to the train station. Samimi sohbetleriniz ve destekleriniz için çok teşekkür ederim!

Finally, I am deeply grateful to my parents and my sister for their unwavering support and for being constant sources of motivation.

Amsterdam Ece Takmaz
January, 2024.

Chapter 1

Introduction

Humans are inundated with information from multiple sources that appeal to various sensory organs operating on distinct modalities. Unless we are distracted, under heavy cognitive load, or affected by a disorder, we manage to glean the relevant bits of information from these sources, which we integrate to make sense of the world. To do so, various perceptual and cognitive functions must be performed flawlessly and meticulously. Our attention is often directed towards novel, interesting, and informative regions in our visual field. We should keep what we looked at in memory long enough or redirect our attention. Human communication is also multimodal, spanning language production and comprehension involving speech, gaze, gestures, facial expressions, and body movements.

If we are talking about what we are seeing, complex visual and linguistic processes need to take place. These include understanding what we are looking at, retrieving words to name them, planning sentences, performing motor movements for speech, and monitoring if what we said at the end is what we initially set out to utter. Furthermore, if we are having a conversation with someone, we need to consult and review what we know about their beliefs and knowledge from our past interactions. Then, we should keep track of what has already been uttered in the current conversation, minding the nuances during the course of our whole interaction, and adapting our own utterances if necessary.

This thesis aims to provide support for the inclusion of information from diverse modalities in deep neural networks in the field of Natural Language Processing (NLP), while paying attention to the rich repertoire of human signals and strategies unraveled in cognitive science. One of the aspects I investigate is human gaze and its role in such models. Human gaze reflects visual attention, and it is considered a proxy for underlying cognitive processes (Rayner, 1977). As such, human gaze can be an informative signal for deep neural networks performing multimodal tasks involving vision and language, such as describing images or reading. Furthermore, having models that predict human gaze could provide crucial insights into human cognition itself, along with an exploration of the potential

to build well-performing models of visuo-linguistic processes.

In cognitive science, many studies reveal the intricacies of the cross-modal relation between language and vision. Both modalities have complex influences on each other as concurrent linguistic production and visual processes unfold (Griffin, 2004; Gleitman et al., 2007; Ferreira and Rehrig, 2019). Although the alignment between the two modalities is complex, there is a tight connection between them (Coco and Keller, 2012), e.g., when speakers describe an image, they tend to look at objects right before mentioning them (Griffin and Bock, 2000; Vaidyanathan et al., 2015).

The visuo-linguistic processes transpiring while we describe what we see in images are intricately related to the properties of the image being described (Jas and Parikh, 2015; Berger et al., 2023). These processes show ample variation, as manifested in human signals such as eye movements and *when* humans start to describe a given image. Having an understanding of the sources and the extent of such variation is valuable. For instance, the variation regarding when speakers start their descriptions could be informative about the relative cognitive complexity of the image (Coco and Keller, 2015a; Gatt et al., 2017). Despite the value of such signals of visuo-linguistic variation, they are virtually disregarded in current models, which motivates further investigation, particularly for Referring Expression Generation, a task where we model how speakers refer to images or entities in diverse visual contexts (Krahmer and van Deemter, 2012).

Another aspect I investigate in this thesis is how humans converse with each other, and how a 'common ground' is formed between the participants of a conversation (*interlocutors*) by collaboratively building on mutual knowledge and beliefs (Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Clark and Brennan, 1991; Clark, 1996; Brennan and Clark, 1996).

Various strategies and phenomena have been observed in dialogues. Consider, for instance, referring to an entity multiple times within different contexts with the same conversation partner. When speakers refer to the same objects or situations more than once, their later (*subsequent*) references (McDonald, 1978) depend on the shared knowledge that speakers accumulate during dialogue. Speakers establish 'conceptual pacts', i.e., particular ways of conceptualizing referents (Garrod and Anderson, 1987; Brennan and Clark, 1996), and continue to utilize established expressions to maintain cohesion and avoid communication problems in dialogue (Metzing and Brennan, 2003). Subsequent references exploit this established common ground accumulated by the interlocutors, and hence, have several interesting properties. Namely, they tend to be shorter and show lexical entrainment with previous mentions where effective phrasings are reutilized (Krauss and Weinheimer, 1967; Brennan and Clark, 1996). This trend has been confirmed in recent datasets made available in NLP (Shore and Skantze, 2018; Haber et al., 2019; Hawkins et al., 2020), where referring utterances become more compact, and yet dialogue participants are able to identify the intended referents.

When producing referring utterances grounded in visual and conversational contexts, we also adapt our language use to the perceived knowledge, information, and linguistic abilities of our interlocutors (Isaacs and Clark, 1987; Clark, 1996; Pickering and Garrod, 2004). When adults speak with children, for example, they use simplified expressions to ensure children can understand (Saxton, 2009); when computational linguists give a talk at a cognitive science conference, hopefully, they avoid making extensive use of NLP jargon, as that would prevent their audience from following through the presentation. Successful adaptation to the conceptual knowledge of conversational partners requires the ability to represent and reason about others' mental states (Tomasello, 2005), a socio-cognitive ability typically referred to as Theory of Mind (ToM; Premack and Woodruff, 1978). Yet, speakers do not always resort to explicitly modeling the knowledge of their dialogue partner: due to different cognitive costs and pressures, they sometimes plan their utterances egocentrically, i.e., only taking into account their own knowledge and abilities (Keysar, 2007).

My aim in this thesis is to contribute to the body of work on visually grounded language use, which requires considering its multimodal nature. I develop computational models of a range of tasks involving the interplay between vision and language, drawing inspiration from theories and findings from cognitive science and psycholinguistics. I aim to capture the intricate relation between non-linguistic modalities and language within deep neural networks, contributing to the line of research on multimodal NLP and cognitively inspired NLP (Mishra and Bhattacharyya, 2018; Beinborn and Hollenstein, 2024). I claim that human signals and strategies can serve as a beneficial source in two main ways. First, human data can be fed into deep learning models to help inform the models about how humans react to various stimuli to improve performance. Second, patterns of behavior and theories of human cognition can be a source of inspiration for how the models learn, and how the inputs are represented and processed.

The findings in this thesis advance our understanding of human visuo-linguistic processes by revealing that intricate strategies are at play in such processes and point to the importance of accounting for them when developing and utilizing multimodal models. In this way, advancements in artificial intelligence (AI) can lead to a better understanding of cross-modal processes, which can inform the development of better AI models. Next, I give an overview of the contents of the thesis.

1.1 Thesis Overview

This thesis consists of two parts with the overarching goal of advancing multimodal models by better understanding human visuo-linguistic processes. These two parts are preceded by a chapter introducing the background common to both parts.

Background (Chapter 2) In this chapter, I first provide background information for human gaze and visuo-linguistic processes in humans, with findings stemming from cognitive science, perception studies, and psycholinguistics. Then, I review visuo-linguistic processes in AI models, starting with task-based multimodal models followed by pretrained multimodal models. I exemplify the main datasets and metrics used in this line of research. Then, I review the use of gaze in NLP and cognitively inspired NLP research. Each chapter in the two parts of the thesis contains its own background section, providing more relevant and up-to-date information.

Part One: Modeling Human Gaze in Language Use In this part of the thesis, I investigate whether incorporating gaze into image description generation models enhances descriptions, whether pretrained *multimodal* models can capture the variation in human visuo-linguistic signals while describing images, and finally, whether pretrained *multilingual* models can capture the patterns in eye movements that accompany reading comprehension across languages. **Chapter 3** gives an overview of the research questions explored in this part.

In **Chapter 4**, I investigate the sequential cross-modal alignment between vision and language by modeling the image description generation process computationally using a corpus of Dutch image descriptions with concurrently collected eye-tracking data (van Miltenburg et al., 2018b). I take as the starting point a powerful image captioning system, which was state-of-the-art at the time of the project (Anderson et al., 2018). I then develop several model variants that exploit information from human gaze patterns recorded during language production. In particular, I propose the first approach to image description generation where visual processing is modeled *sequentially*. The experiments and analyses in this chapter confirm that better descriptions can be obtained by exploiting gaze-driven attention and shed light on human cognitive processes by comparing different ways of aligning the gaze modality with language production. The results reveal that processing gaze data sequentially leads to descriptions that are better aligned to those produced by speakers, more diverse, and more natural—particularly when gaze is encoded with a dedicated recurrent component.

In **Chapter 5**, using the same corpus of Dutch image descriptions as in Chapter 4, I explore the nature of the variation in visuo-linguistic signals and find for the first time significant correlations between different types of signals. Given this result, I hypothesize that variation stems partly from the properties of the images, and explore whether image representations encoded by pretrained vision encoders can capture such variation. The results indicate that pretrained models do so to a weak-to-moderate degree, suggesting that the models lack biases about what makes a stimulus complex for humans and what leads to variations in human outputs.

Finally, in the last chapter of this part, **Chapter 6**, I present the details of my

approaches that attained second place in the shared task of the ACL 2022 Cognitive Modeling and Computational Linguistics Workshop (Hollenstein et al., 2022). The shared task focuses on multi- and cross-lingual prediction of eye movement features in human reading behavior, which could provide valuable information regarding universal aspects of language processing as well as its language-specific properties (Liversedge et al., 2016; Hollenstein et al., 2021b). This task could help us gain insight into language-related eye movements and the predictive capabilities of models of human reading behavior. To this end, I train 'adapters' (Houlsby et al., 2019) inserted into the layers of frozen transformer-based pretrained language models (Conneau and Lample, 2019). The results reveal that multilingual models equipped with adapters perform well in predicting eye-tracking features. The outcomes suggest that utilizing language- and task-specific adapters is beneficial, and translating test sets into similar languages that exist in the training set could help with zero-shot transferability in the prediction of human reading behavior.

Part Two: Communication Strategies in Referential Tasks - Vision and Language in Dialogue In Part One, I show that human gaze is a significant factor when modeling language comprehension and production. In this part, I move on to conversational settings where interlocutors play a visually grounded reference game. In particular, I delve into quantifying and modeling referring utterances in visual and conversational contexts, mainly exploiting the PhotoBook dataset (Haber et al., 2019) introduced in **Chapter 7**.

In **Chapter 8**, I aim to shed light on the mechanisms employed by human speakers when referring to visual entities through the means of pretrained models. I quantify the degree of *descriptiveness* (how well an utterance describes an image in isolation) and *discriminativeness* (to what extent an utterance is effective in picking out a single image among similar images) of human referring utterances within multimodal dialogues. To this end, I use a transformer-based pretrained multimodal model, CLIP (Radford et al., 2021), to encode the images and referring utterances. Overall, the results show that utterances become less descriptive over time while their discriminativeness remains unchanged. The analyses in this chapter indicate that this trend could be due to participants relying on the previous mentions in the dialogue history, as well as being able to distill the most discriminative information from the visual context. In general, this study opens up the possibility of using pretrained models to quantify patterns in human data and to shed light on the underlying cognitive mechanisms and strategies utilized by interlocutors.

In **Chapter 9**, I tackle the generation of first and subsequent references in visually grounded dialogue in the PhotoBook dataset. I propose a generation model that produces referring utterances grounded in visual and conversational contexts. I also implement a reference resolution system to assess the effectiveness of the

referring utterances produced by the generation model. The experiments and analyses show that the model produces better, more effective referring utterances than a model not grounded in the dialogue context, and generates subsequent references that exhibit linguistic patterns akin to humans.

It is an open question how adaptation in dialogue can be modeled in computational agents. In the final chapter of this part, **Chapter 10**, I model a visually grounded referential game, based on the PhotoBook dataset, between a knowledgeable speaker and a listener with more limited visual and linguistic experience. Inspired by psycholinguistic theories, I endow the speaker with the ability to adapt its referring expressions via a simulation module that monitors the effectiveness of planned utterances from the listener's perspective. I propose an adaptation mechanism building on plug-and-play approaches to controlled language generation, where the simulator steers utterance generation on the fly without fine-tuning the speaker's underlying language model. The results and analyses show that the proposed approach is effective: the speaker's utterances become closer to the listener's domain of expertise, which leads to higher communicative success.

Conclusion (Chapter 11) I summarize the findings and contributions of the thesis, detail the limitations and potential future work, and touch upon ethical considerations. Overall, the research explained in this thesis has implications both for future multimodal models in AI and research into cognitive science, showing the importance of accounting for human cognitive processes when developing neural networks.

1.2 List of Publications

The contents of this thesis originate from the following papers, in the order they are presented in the chapters:

- Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4664–4677, Online. Association for Computational Linguistics.

- Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2024. Describing Images *Fast and Slow*: Quantifying and Predicting the Variation in Human Signals during Visuo-Linguistic Processes. To appear in *Proceedings of the 2024 Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

- Ece Takmaz. 2022. Team DMG at CMCL 2022 shared task: Transformer adapters for the multi- and cross-lingual prediction of human reading be-

havior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 136–144, Dublin, Ireland. Association for Computational Linguistics.

- Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2022. Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–42, Dublin, Ireland. Association for Computational Linguistics.

- Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

- Ece Takmaz*, Nicolo' Brandizzi*, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernández. 2023. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4198–4217, Toronto, Canada. Association for Computational Linguistics. *Shared first authorship.

I list below the other works that I contributed to during my PhD:

- Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2022. Time Alignment between Gaze and Speech in Image Descriptions: Exploring Theories of Linearization. Abstract presented at the 44th Annual Conference of the Cognitive Science Society.

- Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

- Gökhan Gönül, Ece Takmaz, and Annette Hohenberger. 2021. Preschool children's use of perceptual-motor knowledge and hierarchical representational skills for tool making. *Acta Psychologica*, 220:103415.

Chapter 2

Background

Human cognitive processes are grounded in multiple modalities, receiving information from various types of stimuli. Action, perception, and cognition are all grounded in information from various modalities we experience as we interact with the world in an embodied manner (Zwaan and Madden, 2005). Naturally, human communication is also multimodal. Face-to-face dialogue involves speech, gestures, and eye movements, conveying important and sometimes complementary information that we need to keep track of and interpret (Rasenberg, 2023). Clark and Krych (2004) list various non-verbal communication channels, including pointing, placing, eye and head movements that speakers use to monitor the addressees during conversation. These also modulate repairs to what the interlocutors uttered and modifications to what they are planning to say. These communicative channels can have critical effects in achieving joint attention and, thus, in how a conversation unfolds (Clark and Krych, 2004; Clark, 2003; Bangerter, 2004; Clark, 1996; Tomasello, 1999). The way in which these modalities interact with each other and how they are affected by our interlocutor's behavior leading to multimodal alignment during a conversation contribute to a complex system (Kendon, 2004; Rasenberg et al., 2020; Feyaerts et al., 2017; Brône and Oben, 2018).

In this thesis, when developing and evaluating AI models, I pursue a line of research inspired by the multimodality of the stimuli we are surrounded by. I argue that AI models should be made aware of the information that can be gathered through multiple modalities and be able to integrate such information meaningfully. Although current AI models can be made increasingly more multimodal, my exploration in this thesis covers visual and linguistic modalities, with the first half of the thesis extending the investigation into the inclusion of eye movements.

In this chapter, I provide background for why it is crucial to consider visuolinguistic processes and eye movements when developing AI models. In **Section 2.1**, I delve into the importance of what eye movements convey and the connection between visual and linguistic processes, providing a theoretical frame-

work behind the motivations for this thesis. Then, in **Section 2.2**, I detail the technical background underlying NLP models combining vision and language, giving an overview of the tasks, models, datasets, and metrics. Finally, in **Section 2.3**, I review assorted approaches to bridging human and machine processing features related to language and vision.

2.1 Visuo-Linguistic Processes in Humans

2.1.1 Eye Movements

Eye movements have been an intriguing source of information in research on visual processing in humans. Figuring out what draws humans' attention in the course of performing a task is a crucial research topic that has implications both for cognitive science and AI. In the literature on human visual cognition, methodologies such as eye-tracking are used to obtain insights into human gaze. The main focus of interest is the fixations: slow, stable eye movements maintained within a small spatial region (Holmqvist et al., 2011). Fixations indicate where attention falls, and they are considered to be representing the contents of cognitive processes (Salvucci and Goldberg, 2000; Nyström and Holmqvist, 2010). Therefore, tracking where people focus on, as well as what they skip, can reveal how certain aspects of human cognition work.

In the early years of eye gaze studies, crude and intrusive setups were used, where contact lenses and suction cups were employed to track where the eyes go (Buswell, 1935; Yarbus, 1967). Nowadays, more advanced and less intrusive eye-tracking devices capture raw gaze data based on infrared reflections from the eyes as well as face and eye position determination from video streams. During data collection, participants are usually shown visual stimuli such as images and text. The participants can be asked to inspect the stimuli freely, or they can be asked to perform a task such as describing images or searching for objects in images.

Before starting to analyze any collected gaze data, a crucial issue is deciding how to represent gaze meaningfully. The raw data is usually further processed to classify gaze samples into fixations (longer gazes over a region) and saccades (faster jumps to distant regions) based on features such as dispersion or the velocity of gaze (Salvucci and Goldberg, 2000; Nyström and Holmqvist, 2010). Two main representation types are then utilized: heatmaps and scanpaths. Heatmaps (also called saliency maps or attention maps) display the distribution of attention over the image, indicating the number or duration of fixations received by each region (Nyström and Holmqvist, 2010). Scanpaths can indicate the order of attention on regions, or if annotations are available, object bounding boxes (Nyström and Holmqvist, 2010). Such regions could also be defined for text, where each word can be considered an entity. In this way, we can obtain the sequences of

focus on areas of interest, as compared to static heatmaps.

When inspecting an image or reading a paragraph, humans do not pay attention to everything simultaneously and may utilize certain strategies. The unfolding of eye movements would depend on the properties of the stimulus as well as possible tasks and goals (Yarbus, 1967; Buswell, 1935; Henderson and Ferreira, 2013). If humans are instructed to *describe* an image by speaking or writing rather than simply *looking* at an image, they would selectively attend certain parts of the image, and there would be differences in attention (van Miltenburg et al., 2018b,c). Saliency is one aspect that attracts attention, which is due to low-level features of images such as colors and their intensities, the contrast in colors, brightness, and orientation (Itti and Koch, 2001, 2000; Itti et al., 1998). In addition to saliency, informative and particularly meaningful parts of an image also receive substantial attention (Henderson et al., 2018). See Castelhano and Williams (2021) for a survey on scene perception.

The difference between saliency and task-related attention is delineated in work proposing the existence of 'bottom-up' vs. 'top-down' attention in humans (Castelhano and Williams, 2021; Torralba et al., 2006; Paneri and Gregoriou, 2017). 'Bottom-up' attention corresponds to attention allocated with regard to the low-level features of images, as certain regions stand out as opposed to others. 'Top-down' attention, on the other hand, is informed by the goal of the task and the contextual information to allocate attention over regions that would be informative and meaningful with respect to the task. Gaze has been shown to be useful in inferring what task was performed by the participants, as the viewing task affects gaze (Castelhano et al., 2009; Castelhano and Williams, 2021; Coco and Keller, 2014).

An important claim in visual cognition is that subjects make a first pass over the whole image to obtain the gist of the scene very quickly (Buswell, 1935; Oliva, 2005; Oliva and Torralba, 2006). Fixations made in the first pass may serve the function of filling in a mental-spatial map. This would constitute a proxy for visual memory, as gaze can provide clues into how memories are formed, retrieved, and reconstructed (Ryan and Shen, 2020; Theeuwes et al., 2009), whose capacity, however, could be limited (Cohen et al., 2016). In relation to visual working memory, this capacity would depend on how the visual objects are represented and maintained even when they are not being attended (Ozimič and Repovš, 2020). When someone sees an image again, their attention over it would likely be different from the first time they had seen it. Repetitively seeing an image would increase its familiarity and how the subject reacts to it. Given a series of images, novelty and familiarity could condition where people prefer to look at, also regulated by curiosity and intrinsic motivation (Jaegle et al., 2019; Schmidhuber, 2010; Boden, 2003).

The cognitive state related to top-down cognitive processes can be reflected in gaze along with responses based on stimuli and oculomotor factors (Henderson et al., 2013). Such cognitive processes involve recalling from long-term memory,

reasoning, planning, and production. As a result, gaze could be an informative source when we aim to model the representational and algorithmic levels (as proposed by Marr, 1982) of cognitive processes. Interestingly, gaze can also affect problem-solving, having an overall effect on how the visual information is attended to and processed (Grant and Spivey, 2003; Pomplun et al., 1996). In addition, gaze could reflect confusion, inattention, and concentration (cognitive load) on the participant's side. Pupil dilation and blink rate can also reveal underlying processes related to task difficulty and mental effort or learning and goal-directed behavior spanning a range of cognitive structures and processes responsible for language (Eckstein et al., 2017).

Next, I review research into the effects of visual processing occurring in parallel with linguistic processes.

2.1.2 Eye Movements During Language Use

Eye movements are intricately connected to various facets of our environment and internal processes, as Richardson and Dale (2005) put it (p. 1143):

> *Eye movements are driven both by properties of the visual world and processes in a person's mind. Your gaze might also be influenced by what your friend is saying, what you say in reply, what is thought but not said, and where you agree and disagree.*

In the first half of this thesis, I focus on the relationship between eye movements and language use as they unfold over time. For instance, when speaking, we produce words one by one. Therefore, we need to map the contents of our thoughts into words in such a way that we can utter them sequentially. This is called the speaker's 'linearization' problem (Levelt, 1981; Ferreira and Rehrig, 2019). Similarly, when describing images, we need to map visual information onto a sequence of words by scanning the image, retrieving lemmas, encoding them phonologically, formulating the utterance, and finally, articulating it. In this case, our eye movements also need to be linearized due to attentional constraints – even though images are of multidimensional, non-sequential nature (Griffin, 2004; Griffin and Bock, 2000; Ferreira and Rehrig, 2019). In this way, certain properties of speech production and comprehension can be reflected in gaze as these processes unfold sequentially in parallel.

The existence and nature of cross-modal correlations between the stages of word-by-word sentence production and eye movement patterns (i.e., scanpaths) are actively being explored in the literature (Griffin and Bock, 2000; Gleitman et al., 2007; Coco and Keller, 2012; Ferreira and Rehrig, 2019; Henderson, 2017). The proposed theories regarding linearization suggest distinct explanations for what happens in sentence formulation and speech execution phases in relation to visual processes (incremental, holistic, or predictive processing; as reported in

Ferreira and Rehrig, 2019). Previous studies mainly utilize small sets of black-and-white line drawings or artificially-created sparse scenes containing few objects (Griffin and Bock, 2000; Gleitman et al., 2007), while later studies exploit naturalistic images (Ferreira and Rehrig, 2019).

In addition to visually conditioned language production, in the literature, there has also been substantial focus on language comprehension in visual contexts. Such studies are carried out utilizing the 'visual world paradigm', where participants look at an image while listening to verbal input pertaining to the image. These works investigate how vision and language comprehension work together, particularly, revealing predictive processes in human cognition by inspecting anticipatory eye movements (Kamide et al., 2003; Coco and Keller, 2015b).

There could also be behavioral differences in visuo-linguistic processes. For instance, some participants might start talking about an image immediately, whereas others prefer to observe the image for a longer duration and then start uttering. Such a phenomenon could even be observed in the trials of a single participant. There could be various explanations for this behavior, where more complex or ambiguous images require more deliberate thinking and time to access lexical items (leading to silent intervals). In contrast, certain images are described very easily, feeling almost instantaneously, albeit with potential mistakes and later corrections. These differences in behavior could be likened to the System-1 and System-2 models of thinking (dual process theory; Wason and Evans, 1974), which have been brought into a wider audience by Kahneman (2012). In the dual process theory, System-1 is claimed to be the system that reacts quickly and performs automatized actions based on intuitions or familiarity, whereas System-2 is the pathway taken when the task needs deliberation and reasoning – thinking about something in detail before producing an outcome.

It is also worthwhile to consider the extent to which verbal processes and visual attention are intertwined. For instance, Vogels et al. (2013) find that subjects choose *referents* depending on visual salience. On the other hand, their results indicate that during language production, *referring expressions* are not affected directly by visual salience. Bavelas and Chovil (2000) consider audible and visible cues as a whole and not as disjunctive signals; consequently, multimodal models would benefit from operationalizing this theory of 'visible acts'. Imagine the referring utterance 'the blue triangle', whose referent may be ambiguous in certain contexts that contain multiple blue triangles. However, this utterance coupled with gaze can be unambiguous, influenced by pragmatic effects, constraints, and task-based goals (Hanna and Tanenhaus, 2004; Hanna and Brennan, 2007).

Gaze also has social, collaborative, and referential roles in dialogue, and it can help determine whether the interlocutors are attending the same components of the visual context (Argyle et al., 1994; Somashekarappa et al., 2020). Research has indicated that a closer alignment between the focuses of attention of interlocutors can be a sign of better comprehension (Richardson and Dale, 2005; Richardson et al., 2007). Such an alignment can also alter how objects in the environment

are referred to when the interlocutors are aware of each other's focus of attention. In this way, dialogue participants may be able to predict the contents of others' mental states, maintaining shared mental representations that can be influenced by meta-linguistic processes (Brennan, 2005). This, in turn, would affect the time and space course of gaze itself by taking the conversation history (also what is *not* said) and visual context into account (Richardson and Dale, 2005). Finally, from a developmental perspective, attentive behavior in social contexts can help children 'perceive referential intent' from their parents and narrow down the set of possible mappings between words and meanings (Trueswell et al., 2016), also driven by infants' processes of physical reasoning and certain principles of object perception (Spelke, 1990).

In light of the literature reviewed above, I emphasize the importance of accounting for diverse modalities and processes to make sense of human cognition and the potential of modeling the inner workings of visuo-linguistic processes when developing AI systems that can communicate and produce human-like outputs.

2.2 Visuo-Linguistic Processes in AI Models

As explained in the previous section, human processing is multimodal, and AI models are increasingly developed to have multimodal input and output channels. Looking at the history of AI, we see that multiple modalities, which were initially modeled separately for a specific task, have come to be combined in contemporary models with overarching task-independent capabilities. In this thesis, I primarily focus on AI models that can process data in visual and linguistic modalities, often called Vision-and-Language (VL) models. Such models work at the intersection of computer vision (CV) and NLP, two fields that are nowadays converging with the use of similar architectures and training schemes.

A specific line of research pertinent to this thesis is Natural Language Generation (NLG), which aims to develop models that generate linguistic output. In the context of this thesis, I explore NLG models conditioned on visual contexts as well as conversational histories. In Section 2.2.1, I delve into detail about multimodal models trained to perform specific tasks such as automatic image description generation and conversing in visually grounded dialogue scenarios. Then, in Section 2.2.2, I review task-agnostic VL models developed to be powerful pretrained foundation models that transfer with good performance to downstream tasks. In Section 2.2.3, I describe a selection of cases in which multimodal models have been shown to come short.

2.2.1 Task-Specific Multimodal Models

Early multimodal models were trained to perform a specific task, such as image captioning or visual question answering (VQA). For example, in image captioning, the aim is to produce text that is relevant to the input image. This task requires a proper representation of the image, capturing scene semantics, objects, attributes, and relations, and finally, producing relevant information in a verbal form that is correct and sufficient. In humans, describing images involves understanding the image contents, mapping image features and language to each other, selecting the contents to be described, planning and articulating the final surface form of the description (Reiter and Dale, 1997; Levelt, 1981), which makes it a complex task to model and evaluate (Bernardi et al., 2016; Stefanini et al., 2023; Erdem et al., 2022).

Early methods for image description generation utilized explicit representations of image contents, exploiting a set of captions using retrieval, employing template-based generation, and handcrafted rules, which led to relatively inflexible setups (Bernardi et al., 2016; Stefanini et al., 2023). In contrast, this thesis commenced around a time when deep neural networks were gaining traction both in NLP and CV. The use of distributed representations in NLP provided fruitful outcomes by representing words with embedding vectors (Rumelhart et al., 1986; Mikolov et al., 2013a,b,c). In addition, the benefits of utilizing Long-Short Term Memory Recurrent Neural Networks (LSTM RNN) and their derivatives in processing sequences were well known to capture the sequential nature of various phenomena, including language (Sutskever et al., 2014; Hochreiter and Schmidhuber, 1997). In CV, when representing images, handcrafted, rule-based approaches were surpassed by end-to-end models based on deep neural networks, skipping intermediate processing stages, such as object detection and image segmentation (Ren et al., 2015a). As a result, visual features have started to be extracted from the later layers of models based on Convolutional Neural Networks (CNN) such as AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2015), and ResNet (He et al., 2016), pretrained to recognize or classify images.

The research in this thesis builds on such deep neural network-based efforts in multimodal models. In the case of automatic image description generation, the models show diversity in terms of how they represent and exploit images and linguistic data, although the underlying end-to-end structure tends to be similar. The models mainly follow the encoder-decoder architecture (Sutskever et al., 2014), where the crucial information in the image is captured by the *encoder*, and its output, usually a single fixed vector, is provided to the *decoder* to generate the description autoregressively, conditioned on the visual input (Vinyals et al., 2015). In such a model, the encoder could be a CNN-based pretrained image model, and the decoder could be an LSTM model. Then, the LSTM is trained to predict the next word conditioned on the encoded image representation and the previously generated words. At each step of generation, the hidden state of the

LSTM is projected to the dimensions of the vocabulary to sample a word based on softmax probabilities to form the predicted image description.

In the literature of deep learning, utilizing an attention mechanism, particularly in sequence-to-sequence models, has proven to be advantageous (Sutskever et al., 2014; Bahdanau et al., 2015). Attention mechanisms, originally proposed for machine translation (Bahdanau et al., 2015), help keep track of essential words in the source, prevent repetition, and provide coverage of all crucial information. This idea has also been transferred to image description generation, in a way representing the *eye movements* of the model over the input image. The influential work by Xu et al. (2015) applies attention over grid features representing the image extracted from a lower layer of a CNN as compared to the other works using a single feature vector from the final layer. Then, the model learns to dynamically apply attention over these features, selectively focusing on specific regions as it generates the description. Another influential model, inspired by how human attention works, is the Bottom-up Top-down image captioning model by Anderson et al. (2018). This model combines bottom-up visual saliency to obtain object regions with top-down task-related processes to apply attention over the regions in line with the goal of image caption generation. I base the models I develop in Chapter 4 on this model. For more details about image captioning models, I refer the reader to surveys reviewing the literature of image description generation over the years (Bernardi et al., 2016; Hossain et al., 2019; Erdem et al., 2022; Stefanini et al., 2023).

Going beyond image description generation models operating on a single image input and producing a single description, developing models in increasingly multimodal and social contexts is also an important research direction in AI (Bisk et al., 2020). To this end, task-specific models using the encoder-decoder approach with an attention mechanism have also been developed for visually grounded dialogue, where two participants have a conversation related to visual content (Das et al., 2017; de Vries et al., 2017). The exact architectures of these models diverge from those of image description generation models to accommodate the nature of visually grounded dialogue. For instance, the encoder is not only for the visual modality but also encompasses the linguistic modality, representing the dialogue history together with the visual context. In the second part of the thesis, I propose my own models for visually grounded dialogue and utilize task-agnostic models (reviewed in Section 2.2.2) to quantify and represent the input data in this task.

Task-specific models are trained in a supervised manner, with the input being images depicting mainly real-life scenes. The text corresponding to the image is used as the expected output. Therefore, task-specific models require task-specific datasets, whose creation and availability have been instrumental in the development of such models. A typical example of commonly used datasets in this field is MS COCO (Lin et al., 2014), which provides on average 5 captions for 120K images along with various types of annotations for object categories and locations. In contrast to MS COCO, which involves captions mentioning only

the entities that visibly exist in the images, other multimodal datasets including meta-information and commonsense knowledge in their descriptions have also been introduced (Alikhani et al., 2020; Kiela et al., 2020; Kruk et al., 2019). The datasets for visually grounded dialogue, on the other hand, particularly focus on task-oriented conversations about multimodal contexts eliciting realistic dialogue-related phenomena (Shore and Skantze, 2018; Haber et al., 2019; Suhr et al., 2019; Hawkins et al., 2020).

Task-specific models involving language generation are mainly evaluated on whether the generated text is sufficiently close to the reference sentences produced by humans. However, it is difficult to determine what contributes to the 'closeness' between the sentences. Two sentences could be deemed close to each other due to multiple factors such as semantics and syntax, as well as surface form information such as sentence length. Due to these potential dimensions in which 'closeness' can be measured and the fact that some dimensions are less straightforward to assess, evaluating the goodness of machine-generated text is still an open problem (Reiter and Dale, 1997; van Miltenburg, 2023). In many cases, human evaluations would be preferred; however, along with being costly, they are subjective—usually, there are multiple correct responses—and not perfect.

With the aim of evaluating generated text automatically, various metrics have been proposed. Following the approaches utilized in evaluating machine translation and text summarization, image captioning models have also been assessed with respect to the automatic measures based on the matches between generated text and ground-truth reference texts. Commonly-used metrics include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005). BLEU utilizes n-gram precision; METEOR performs unigram matching also using synonyms; ROUGE looks at recall and the longest common subsequence between the generated output and the ground truth references. CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016b) were specifically proposed for image captioning, with CIDEr using tf-idf weighing to assess the consensus among references, and SPICE using semantic scene graphs to measure the match between the text and the images.

These metrics have been shown to correlate with human evaluations to varying extents, mostly weakly, in tasks ranging from machine translation to image captioning (Callison-Burch et al., 2006; Liu et al., 2016; Novikova et al., 2017; Saphra et al., 2023; Hessel et al., 2021). Although the correlations are imperfect, these metrics are still used to compare models to past work and are reported frequently. Particularly, the metrics based on surface forms of the words have shortcomings when evaluating image captions, showing the importance of semantic metrics based on embeddings (Kilickaya et al., 2017). Therefore, there is a shift towards utilizing distributed representations and task-agnostic pretrained models in evaluating generated text, such as BERTScore (Zhang et al., 2020b) and CLIPScore (Hessel et al., 2021). In the next subsection, I provide a review of task-agnostic pretrained models relevant to multimodal NLP.

2.2.2 Task-Agnostic Pretrained Multimodal Models

The introduction of the transformer architecture has led to unprecedented progress in NLP and CV (Vaswani et al., 2017). Transformers allow for efficient computations based on non-recurrent operations, facilitating the management of long-range dependencies using self-attention over the input sequence, and connecting the encoder and a potential decoder with the cross-attention mechanism. The transformer architecture forms the backbone of influential pretrained large language models such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2019; OpenAI, 2023). The BERT model, for instance, was trained using large amounts of data in English in a self-supervised manner (Devlin et al., 2019). Its representations have been shown to transfer well to many downstream NLP tasks, whose objectives differ from the self-supervision objective of the BERT model (Devlin et al., 2019). Given these positive results, the overarching framework nowadays is to construct large-scale models that are task-agnostic, general-purpose foundation models trained to have good representations for a large range of potential applications (Bommasani et al., 2021).

The idea of training transformer-based models via self-supervision on large datasets has also transferred to the multimodal realm. The aim is to learn task-agnostic vision-and-language representations in a self-supervised manner, which would transfer well to potentially many multimodal downstream tasks. The training objectives are masked language modeling (where a subset of the tokens in the input is randomly masked and then predicted based on the available context as proposed by Devlin et al. (2019)), masked vision/image modeling (regressing to the image region vector or a masked bounding box or predicting its object category), image-sentence matching (binary label or an alignment score). In VL models, masked prediction objectives can be applied to either or both of the modalities: masking the visual input and predicting it with the help of surrounding visual regions as well as the language, or conversely, masking a word and predicting it from the sentential context along with the image. Once these models are pretrained, fine-tuning for specific downstream tasks is performed, in general, on much smaller datasets. Conversely, there have also been proposals to conduct the pretraining on multiple multimodal tasks, such as VQA, caption-based image retrieval, and reference resolution, to produce universal multimodal models (Lu et al., 2020).

A plethora of multimodal models have been built on the transformer architecture and trained using massive datasets via self-supervision. One type of such models fuses the vision and language modalities from the beginning, relaying the information through a single stream, as in models such as VL-BERT, VisualBERT, VideoBERT, UNITER, Unified VLP, OSCAR, and VinVL (Su et al., 2020; Li et al., 2019; Sun et al., 2019; Chen et al., 2020; Zhou et al., 2020; Li et al., 2020b; Zhang et al., 2021). On the other hand, dual-stream architectures have multiple encoders that handle modalities separately at first, as in ViLBERT and

LXMERT (Lu et al., 2019; Tan and Bansal, 2019). Then, the outputs of modality-specific streams are combined through fusion mechanisms, which show diverse structures in the literature: cross-modal attention, dot product, and concatenation. Bugliarello et al. (2021) unify both stream types in a single framework, revealing that they perform similarly under the same setup, while Hendricks et al. (2021) show the importance of multimodal attention in improving performance as compared to having deeper models applying modality-specific attention.

The language encoders of these models are initialized using BERT weights, following the preprocessing steps as in BERT tokenization to prepare the linguistic input. The visual input can be obtained from pretrained vision models such as object regions detected by Faster R-CNN (Ren et al., 2015b) or grid representations by ResNet (He et al., 2016). Li et al. (2020b) also incorporate object labels in the input as an anchor between image regions and descriptions. Zhang et al. (2021) improve on this model by proposing a more advanced object detector, enhancing the object labels to be used as cross-modal anchors.

A special case is proposed by Tan and Bansal (2020), where a language model is supervised by the visual task of predicting images for each token, which are later discarded during inference. Since visual context provides cues and contributes to linking meaning to words, object naming and grounding referential expressions are beneficial for downstream language-only use cases as well.

Contrastive learning of visual and textual representations as a training objective has also proven advantageous in models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). Via this objective, a model is trained to predict higher scores for a matching image and text pair. The CLIP model, for instance, uses separate transformers to encode vision and language and was trained on millions of image-text pairs collected from the internet (Vaswani et al., 2017; Dosovitskiy et al., 2021; Radford et al., 2021). Such automatically collected data crawled from the internet at a large scale could be noisy compared to the curated datasets used by the task-specific models, yet also advantageous for pretraining and learning general-purpose multimodal information (Radford et al., 2021).

CLIP's vision encoder has been shown to be powerful and robust, benefitting many VL tasks. For instance, when replacing previous image encoders or when it was used as a conditioning input for image caption generation, it improves or performs comparably with the state-of-the-art models in VQA and image captioning (Shen et al., 2022; Mokady et al., 2021). It can also help generate contrastive captions, acting in the role of a listener that picks the referred image given the caption (Ou et al., 2023). Given its representational power, the score output by the CLIP model when comparing an image and a text has been proposed as a metric for evaluating models of image caption generation (CLIPScore; Hessel et al., 2021). In this way, the captions can be assessed in direct comparison with the image, bypassing the need to have human-written references, resulting in scores highly correlated with human judgments.

The current state-of-the-art models exploit foundation models of different

modalities in an efficient manner, as in models such as LENS, Flamingo, MAPL, Frozen, BLIP-2, ClipCAP, and VisualGPT (Berrios et al., 2023; Alayrac et al., 2022; Mañas et al., 2023; Tsimpoukelli et al., 2021; Li et al., 2023b; Mokady et al., 2021; Chen et al., 2022). These recent approaches keep both the encoder and the decoder frozen, only training a mapping network with a small number of trainable weights, which projects the visual representation so that the frozen decoder can generate a correct image description. For instance, CLIP's visual encoder can be used to represent the image, and a GPT decoder can generate the description. Such models have recently achieved impressive results; one such model (PaLM-e; Driess et al., 2023) has also been integrated into embodied setups to enable interaction with the world.

Pretrained models have also been developed for visually grounded dialogue. For instance, Murahari et al. (2020) adapt ViLBERT for multi-turn visually grounded dialogue. Nowadays, instruction tuning is applied to improve the conversational capabilities of multimodal models based on foundation models, which is a line of work that gained substantial traction recently by fine-tuning models to follow human instructions, as in models such as LLaVA and KOSMOS-2 (Liu et al., 2023b,c; Peng et al., 2023). See Li et al. (2023a) for more details about multimodal foundation models.

Considering the advantages of these models, I employ them for multiple purposes in this thesis. Firstly, I exploit their encoders to represent inputs in different modalities. For instance, I use the CLIP model's vision encoder to represent images; CLIP's textual encoder, BERT and BERT-based models to represent linguistic input. Additionally, I utilize BERTScore to evaluate model-generated texts, whereas I use CLIPScore to quantify the strategies adopted by humans during visually grounded dialogue. I also investigate task-agnostic models' understanding of the variation in human visuo-linguistic data.

2.2.3 Shortcomings of Multimodal Models

Although large pretrained models have been shown to perform well on various benchmarks, they are sometimes prone to making mistakes that humans would usually not make. For instance, captioning models sometimes 'hallucinate' entities and mention objects that are not in the image (Rohrbach et al., 2018). There could be several reasons for this type of error. It could be due to the lack of visual perception or linguistic capabilities. Additionally, problems in bridging the two modalities could also cause such issues. To diagnose the root of the problems, various benchmarks have been proposed to evaluate multimodal models on phenomena related to bridging the modalities, with the construction of 'foil' captions that diverge from more suitable captions for a given image, such as the FOIL and VALSE datasets (Shekhar et al., 2017; Parcalabescu et al., 2022). Such benchmarks explore how models fare in tasks that require combining vision and language, as well as other phenomena such as counting objects in images and

mapping them to language properly (Parcalabescu et al., 2021).

In contrast, Li et al. (2020a) investigate the inner workings of a model, VisualBERT (Li et al., 2019), finding evidence for entity grounding connecting vision and language by examining its attention heads. Such zero-shot probing of the models and inspecting their attention patterns reveal what the models capture and what visuo-linguistic tasks they cannot perform, e.g., falling short in tasks including basic language abilities, counting, verb understanding, and generalizing to novel contexts (Hendricks and Nematzadeh, 2021; Parcalabescu et al., 2021; Cao et al., 2020; Chen et al., 2023). In addition, evaluating the learned representations of the models against human judgments also reveals the extent of the alignment between human and model representations. For instance, multimodal representations yield estimations closer to human judgments regarding the similarity between concrete words as compared to language-only representations, but not for abstract words (Pezzelle et al., 2021).

A potential cause for shortcomings in multimodal models is attributed to modality dominance. In other words, the models might not lean on each modality with equal weights. Parcalabescu and Frank (2023) and Hessel and Lee (2020) have proposed metrics to investigate the extent to which a modality is utilized and contributes to the outcomes. The contributions of each modality have been represented by information flowing through the network and also by manipulating the input to observe the biases in the dataset. Via ablating parts of images and texts, Frank et al. (2021) show that multimodal models show asymmetric behavior across modalities.

A crucial ability that the models have been tested on is compositionality. Two captions composed of the same set of words but in different orders would likely exhibit differences in terms of how much they match a given image. As shown with the introduction of the Winoground dataset (Thrush et al., 2022), many powerful VL models fail to differentiate between such sentences and pick the one correctly describing the image. Such a task requires a strong understanding and combination of visual and textual elements, paying attention to grounding even small objects in the image, as well as commonsense reasoning (Thrush et al., 2022; Diwan et al., 2022).

These studies indicate that current models tend to be insensitive to—or unaware of—information that would be crucial for humans. For instance, in the framework of VQA, Agrawal et al. (2016) reveal that models pick up undesired correlations in the data, information that would be ignored or go unnoticed by humans, leading to biases in the answers generated by the models. One reason for this is that the training datasets, albeit of large scale, are not comprehensive enough to represent the complexity of multimodal processes as we observe them in humans (Erdem et al., 2022). In the following section, I focus on research that works on incorporating information about human cognitive processes in model development with the aim of alleviating models' non-human-like shortcomings.

2.3 Bridging Human and Machine Processing

Human signals and insights from cognitive science regarding human processing have been used to inform, improve, and analyze language-only, vision-only, and multimodal models. In this section, I first review a subset of efforts aiming to combine cognitive science and language-only research in NLP, then focus on vision-only work in CV in relation to human signals, and finally delve into multimodal models that are informed by human signals and processing.

Starting with language-only models, various efforts have been made to bridge human and machine processing, leading to a line of work on cognitively inspired NLP (Mishra and Bhattacharyya, 2018; Beinborn and Hollenstein, 2024; Hollenstein, 2021; Hollenstein et al., 2020). The reasoning behind this research is that by leveraging human processing data, NLP models can be improved to capture the intricacies of human cognition. Moreover, this line of research also enables the comparison between human and machine processing in order to reveal where they align and where they differ.

Research in this domain is inspired by cognitive processes and takes cognitive plausibility into consideration at various stages of model development (Beinborn and Hollenstein, 2024). Similarly, some other works incorporate human signals to inform models (Mishra and Bhattacharyya, 2018). For instance, aligning models with brain representations as well as predicting brain signals help with linguistic tasks (Schwartz et al., 2019; Toneva and Wehbe, 2019). Gaze is also one such signal reflecting underlying cognitive processes; as a result, it has been increasingly used in diverse facets of AI: CV, NLP, decision-making, and robotics; see Zhang et al. (2020a) for a survey. In the first part of this thesis, I focus on gaze in NLP, given the close interaction between eye movements and linguistic processes in humans, as reviewed in Section 2.1.

Gaze data has been utilized in language-only NLP tasks such as sentiment analysis, part-of-speech tagging, readability, named entity recognition, sarcasm detection, and grammatical error detection (Barrett et al., 2016, 2018; Ding et al., 2022; Ren and Xiong, 2021; Dong et al., 2022; Khurana et al., 2023; Sood et al., 2020a,b); see Mathias et al. (2020) for a survey. Gaze proves to be a good inductive bias for human-inspired attention mechanisms in NLP (Barrett et al., 2018), although similarity in attention does not necessarily lead to better performance in all cases (Sood et al., 2020a),

Predicting eye movements during reading is also a prolific line of work (Deng et al., 2023; Bolliger et al., 2023; Khurana et al., 2023), with gaze prediction being utilized as an auxiliary task in linguistic tasks such as sentence compression (Klerke et al., 2016). Multilingual models have also been fine-tuned to predict human reading behavior in the form of eye movements, and in this way, they capture patterns of reading and cognitive complexity across languages (Hollenstein et al., 2021a, 2022, 2021b; Pouw et al., 2023).

Regarding vision-only uses of gaze, in early CV works, the attention was on

predicting salient parts of an image based on bottom-up processes (Itti and Koch, 2001, 2000; Itti et al., 1998). Given the progress in deep learning, models have been used to predict gaze during free-viewing (Cornia et al., 2018b), performing VQA (Chen et al., 2021), and searching for entities in images (Yang et al., 2020). In addition to bottom-up and task-related factors, intrinsic human features can also help predict eye movements, with measures of curiosity and novelty contributing to the prediction of gaze when used as reward signals in reinforcement learning (Jaegle et al., 2019).

The use cases above provide evidence for the benefits of using human signals and processing features in single modalities. Moving towards multimodal setups, early efforts to utilize gaze in VL tasks made use of separately obtained gaze and descriptions (Yun et al., 2013), leading to the incorporation of static saliency maps from free-viewing of images in image captioning (Tavakoli et al., 2017; Cornia et al., 2018a; Chen and Zhao, 2018). More recent datasets involve concurrently collected gaze and image description data, enabling the analysis and modeling of how visuo-linguistic processes take place together sequentially over time in different languages (van Miltenburg et al., 2018b; He et al., 2019; Vaidyanathan et al., 2018). Given the advantages in language and vision modalities separately, as expected, VL tasks such as image captioning and VQA have also benefited from the use of human gaze (Sugano and Bulling, 2016; He et al., 2019; Takmaz et al., 2020b; Sood et al., 2021, 2023).

The fact that certain deep learning models learn to pay attention by themselves raises the question of whether the attention learned by such models is aligned with human attention and whether potential divergences in attention are partly responsible for the shortcomings of the models. There exist mixed findings in the literature regarding this subject: Gella and Keller (2018) find that there is a significant correlation between both types of attention in disambiguating verbs coupled with images, whereas Das et al. (2016) do not find such a significant correlation in the task of VQA. In an effort to inform model attention about human attention, models that use gaze as a more direct regularizer of attention have been developed (Barrett et al., 2018). It is crucial to note that there have also been findings that a better correlation with human attention does not consistently translate into better task performance by a model (Sood et al., 2021), whereas human-like attention generated from models trained on human attention can improve performance in multimodal tasks such as visual reasoning or VQA (Sharan et al., 2019; Qiao et al., 2018; Sood et al., 2023).

The work in this thesis moves in the direction of including human processing features and signals in model development. For instance, in Chapter 4, I show that including human gaze information helps an image captioning model generate more human-like captions containing specific words referring to even small objects in the images. In Chapters 9 and 10, I propose models that make use of common ground and ToM-based adaptation in visually grounded dialogue, helping make model outputs closer to human expectations, showing the benefits of

taking human processing into account when developing, training and evaluating multimodal models.

Part One

Modelling Human Gaze in Language Use

Chapter 3

Overview

This part investigates the modeling of human gaze in language use, covering both language production and comprehension under different settings. Chapters 4 and 5 are centered around visually conditioned language production: verbally describing images. Chapter 6 focuses on the process of reading instead. These studies contribute to our understanding of human visuo-linguistic processes, and help improve models in NLP by helping us comprehend their current capability to represent such processes in the form of deep neural networks. I investigate the following research questions:

- **Can human gaze help improve models generating textual descriptions of images? In what ways can it help such multimodal models in combining vision and language?** In **Chapter 4**, I explore the cross-modal interaction between vision and language, and inform a powerful image description generation model about where humans look while they utter descriptions for the images. I find that utilizing human gaze in a sequential, speaker-specific manner enhances such models and enables them to generate descriptions more similar to those of humans.

- **What is the extent of the variation in human signals while describing images? Can representations extracted from pretrained multimodal models capture the variation in human gaze and linguistic output?** In **Chapter 5**, I appraise the variation in human signals during image description generation to give a wider picture of visuo-lingustic processes. Furthermore, I check if we can predict human signals and the variation thereof using pretrained models. I find that there is substantial variation in the outcomes of visuo-linguistic processes even for a single image, and that pretrained vision encoders can only moderately capture features that relate to linguistic and visual variation.

- **Do pretrained multilingual models have the capability to predict human reading behavior as reflected in eye movements across lan-**

guages? In **Chapter 6**, I predict human reading behavior in the form of eye-tracking data in multi- and cross-lingual settings with the help of lightweight adapter layers inserted into pretrained multilingual models. My approach achieved second place in the leaderboards of the shared task of the ACL 2022 Cognitive Modeling and Computational Linguistics Workshop (Hollenstein et al., 2022).

Chapter 4

Generating Image Descriptions Guided by Sequential Human Gaze

The material in this chapter is based on: Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4664–4677, Online. Association for Computational Linguistics.

Contributions: Ece Takmaz: Implementing and running the experiments, writing and revising the paper. Sandro Pezzelle and Raquel Fernández: Supervising the research, writing, and revising the paper. Lisa Beinborn: Contributions to the conceptualization of the research in the early stages and revising the paper.

4.1 Introduction

Describing an image requires the coordination of different modalities. There is a long tradition of cognitive studies showing that the interplay between language and vision is complex. On the one hand, eye movements are influenced by the task at hand, such as locating objects or verbally describing an image (Buswell, 1935; Yarbus, 1967). On the other hand, visual information processing plays a role in guiding linguistic production (e.g., Griffin, 2004; Gleitman et al., 2007). Such cross-modal coordination unfolds sequentially in the specific task of image description (Coco and Keller, 2012)—i.e., objects tend to be looked at before being mentioned. Yet, the temporal alignment between the two modalities is not straightforward (Griffin and Bock, 2000; Vaidyanathan et al., 2015).

In this chapter, we follow up on these findings and investigate cross-modal alignment in image description by modeling the description generation process computationally. We take a model, which was the state-of-the-art system at the time of the project, for automatic image captioning (Anderson et al., 2018) and develop several model variants that exploit information derived from eye-tracking data. To train these models, we use a relatively small dataset of image descriptions in Dutch (DIDEC; van Miltenburg et al., 2018b) that includes information on gaze patterns collected during language production. We hypothesize that a system that encodes gaze data as a proxy for human visual attention will lead to better, more human-like descriptions. In particular, we propose that training such a system with eye movements sequentially aligned with utterances (see Figure 4.1) will produce descriptions that reflect the complex coordination across modalities observed in cognitive studies.[1]

To measure the level of semantic and sequential alignment between descriptions, we develop a novel metric and use it in two ways. First, we analyze cross-modal coordination in the DIDEC data, finding that the product of content and sequentiality better captures cross-modal correlations than content alone. Second, we test whether our models generate captions that capture sequential alignment along with semantic alignment. Our experiments show that exploiting gaze-driven attention helps enhance image caption generation, and that processing gaze patterns sequentially results in descriptions that are better aligned with those produced by speakers. The descriptions generated by gaze-driven models are also more diverse—both in terms of variability per image and overall vocabulary—particularly when gaze is encoded with a dedicated recurrent component that can better capture the complexity of the temporal alignment across modalities.

Overall, this work presents the first computational model of image description generation where both visual and linguistic processing are modeled sequentially, and lends further support to cognitive theories of sequential cross-modal coordi-

[1] Our preprocessed data and code are publicly available at `https://github.com/dmg-illc/didec-seq-gen`.

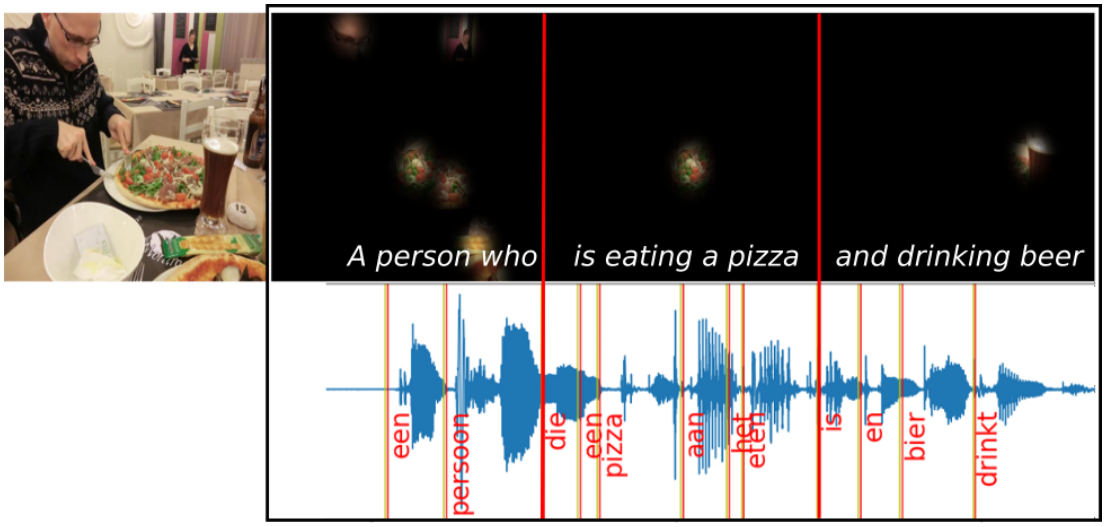


Figure 4.1: In our approach, an image captioning model is fed with a *sequence* of masked images encoding the gaze fixations of a *single* human speaker during language production. This diagram is a toy illustration.

nation.

4.2 Related Work

4.2.1 Image Captioning

Various models have been proposed to tackle the challenging task of generating a caption for a visual scene (Bernardi et al., 2016; Hossain et al., 2019; Stefanini et al., 2023). Contemporary approaches make use of deep neural networks and encoder-decoder architectures (Sutskever et al., 2014). In the influential model by Vinyals et al. (2015), a Convolutional Neural Network (CNN) is used to encode the input image into a feature representation, which is then decoded by a Long Short-Term Memory network (LSTM; Hochreiter and Schmidhuber, 1997) that acts as a generative language model. In recent years, there have been many proposals to enhance this basic architecture. For instance, via extracting features from a lower layer of a CNN, Xu et al. (2015) obtain representations for multiple regions of an image over which attention can be applied by the LSTM decoder. The 'Bottom-up and Top-down Attention' model by Anderson et al. (2018) further refines this idea by extracting multiple image features with the help of Faster R-CNN (Ren et al., 2015b), which results in the ability to focus on regions of different sizes better aligned with the objects in the image. Other models based on unsupervised methods (e.g., Feng et al., 2019) and Generative Adversarial Networks (Chen et al., 2019) have also been proposed.

We take as our starting point the model by Anderson et al. (2018) for two

main reasons: first, it was among the best-performing architectures on standard image captioning benchmarks at the time of the research explained in this chapter; second, its underlying idea (i.e., bottom-up and top-down attention) is explicitly inspired by human visual attention mechanisms (Buschman and Miller, 2007), which makes it suitable for investigating the impact of adding human gaze information.

Please note that at the time of writing the thesis, the models for image caption generation are mostly transformer-based multimodal models (Li et al., 2020b; Zhang et al., 2021), predominantly building on frozen pretrained language and vision models (Berrios et al., 2023; Alayrac et al., 2022; Mañas et al., 2023; Tsimpoukelli et al., 2021; Li et al., 2023b; Mokady et al., 2021; Chen et al., 2022). See Section 5.2.2 in the next chapter for a more up-to-date review of multimodal models and the use of eye-tracking in NLP.

4.2.2 Eye Tracking

In computer vision, human eye movements collected with eye-tracking methods have been exploited to model what is salient in an image or video for object detection (Papadopoulos et al., 2014), image classification (Karessli et al., 2017), image segmentation (Staudte et al., 2014), region labeling (Vaidyanathan et al., 2015, 2018), and action detection (Vasudevan et al., 2018). More relevant for the present study, gaze has also been used in automatic description generation tasks, such as video frame captioning (Yu et al., 2017b) and image captioning (Sugano and Bulling, 2016; Chen and Zhao, 2018; He et al., 2019). In all these approaches, gaze data from different participants is *aggregated* into a *static* saliency map to represent an abstract notion of saliency. This aggregated gaze data is used as supervision to train models that predict generic visual saliency.

In contrast, in our approach, we model the production process of a *single* speaker by directly inputting information about where that speaker looks at during description production, and compare this to the aggregation approach. In addition, we exploit the *sequential* nature of gaze patterns, i.e., the so-called scanpath, and contrast this with the use of static saliency maps. Gaze scanpaths have been used in NLP for diverse purposes: For example, to aid part-of-speech tagging (Barrett et al., 2016) and chunking (Klerke and Plank, 2019); to act as a regularizer in sequence classification tasks (Barrett et al., 2018); as well as for automatic word acquisition (Qu and Chai, 2008) and reference resolution (Kennington et al., 2015). To our knowledge, the present study is the first attempt to investigate sequential gaze information for the specific task of image description generation.

4.3 Data

We utilize the Dutch Image Description and Eye-Tracking Corpus (DIDEC; van Miltenburg et al., 2018b). In particular, we use the data collected as part of the description-viewing task in DIDEC, where participants produce a spoken description in Dutch for each image they look at, with no time limits.[2] The gaze of the participants is recorded with an SMI RED 250 eye-tracking device while they describe images. DIDEC contains spoken descriptions for 307 real-life images originating from the MS COCO dataset Lin et al. (2014), with high-quality eye-tracking data. Each of the 45 participants describe $\approx$ 102 images, resulting in 4604 descriptions in total. On average, each image has 15 descriptions. For each description, the audio, textual transcription, and the corresponding eye-tracking data are provided.

4.3.1 Preprocessing

We tokenize and lowercase the raw captions, exclude punctuation marks and information tokens indicating, e.g., repetitions (<rep>). We then use CMUSphinx[3] to obtain the time intervals of each word given an audio file for a description and its transcription. See Appendix A.1 for more details.

Gaze data in DIDEC is already classified into gaze events such as fixations, saccades or blinks. We discard saccades and blinks (since there is no visual input during these events), and use only the fixation gaze samples that fall within the borders of the actual image. We treat consecutive occurrences of such fixations as belonging to the same fixation window.

4.3.2 Saliency Maps

Using the extracted fixation windows, we create two types of saliency maps, *aggregated* and *sequential*, which indicate the prominence of certain image regions as signaled by human gaze.

Aggregated saliency maps *(per image)* The aggregated saliency map of an image is computed as the combination of all participants' gazes and represents what is generally prominent given the image description task. To create it, we first compute the saliency map of each participant who looked at a given image. Following Coco and Keller (2015a), for each fixation window of the participant, we create a Gaussian mask centered at the window's centroid with a standard deviation of 1° of visual angle. Given the data collection setup of DIDEC, this standard deviation corresponds to 44 pixels. We sum up the masks weighted

[2]The other task is 'free viewing', where the participants simply look at the images for 3 seconds.

[3]`https://cmusphinx.github.io/`

by relative fixation durations, and normalize the resulting mask to have values in the range $[0, 1]$. Finally, we sum up and normalize the maps of all relevant participants to obtain the aggregated saliency map per image.

Sequential saliency maps *(per image-participant pair)* This type of map consists of a sequence of saliency maps aligned with the words in a description, and represents the scan pattern of a given participant over the course of description production. Using the temporal intervals extracted from the audio files, we align each word with the image regions fixated by the participant right before the word was uttered. For each word w_t—using the same method described above for aggregated maps—we combine all the fixation windows that took place between w_{t-1} and the onset of w_t, and normalize them to obtain a word-level saliency map.[4] In this way, we obtain a sequence of saliency maps per description.

4.3.3 Masked Images and Image Features

The saliency maps are used to keep visible only the image regions that were highly attended by participants, and to mask the image areas that were never or rarely looked at (see Figure 4.1). We create each masked image by calculating the element-wise multiplication between the corresponding 2D saliency map and each RGB channel in the original image. We then extract the image features of the masked images using ResNet-101 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009). We take the output of the average pooling layer as the image features with 2048 dimensions to give as input to our models.

4.4 Evaluation Measures

We propose a novel metric to quantify the degree of both *semantic* and *sequential* alignment between two sentences. In our study, this metric will be leveraged in two ways: (1) to analyze cross-modal coordination in the DIDEC data (Section 4.5) and (2) to evaluate our generation models (Section 4.7). For context, we first briefly review several existing metrics for automatic image captioning.

4.4.1 Image Captioning Metrics

Image caption generation is evaluated by assessing some kind of similarity between the generated caption and one or more reference captions (i.e., those written by

[4]For the first word, we combine all the fixation windows that took place before its utterance. Some participants may scan larger portions of the image to obtain its gist before uttering the first word (Oliva and Torralba, 2006). However, we do not encode these differences in behavior explicitly. See Chapter 5 for analyses of the variation in the visuo-lingustic behavior in the DIDEC dataset.

human annotators). One of the most commonly used metrics for this purpose is CIDEr (Vedantam et al., 2015), which (a) computes the overlapping n-grams between the generated caption and the entire set of reference sentences for a given image, and (b) downweighs n-grams that are frequent in the entire corpus via *tf-idf* scores. Thus—regarding semantics and sequentiality—CIDEr scores can be affected by word order permutations, but not by the relative position of words in the entire caption nor by the presence of different but semantically similar words. Other metrics such as BLEU (which looks at n-gram precision; Papineni et al., 2002) and ROUGE-L (which considers n-gram recall; Lin, 2004) suffer from comparable limitations.

METEOR (Banerjee and Lavie, 2005) and SPICE (Anderson et al., 2016b) also make use of n-grams (or tuples in a scene's graph, in the case of SPICE) and take into account semantic similarity by matching synonyms using WordNet (Pedersen et al., 2004). This allows for some flexibility, but can be too restrictive to grasp overall semantic similarity. To address this, Kilickaya et al. (2017) proposed using Word Mover's Distance for image caption evaluation (Kusner et al., 2015), which builds on `word2vec` embeddings (Mikolov et al., 2013b). Several metrics capitalizing on contextual embeddings (Devlin et al., 2019) were proposed, such as BERTScore (Zhang et al., 2020b) and MoverScore (Zhao et al., 2019). More recently, a metric called CLIPScore has been proposed to evaluate captions independently of the existence of reference captions (Hessel et al., 2021).[5] This metric builds on the image and text representations of the pretrained CLIP model (Radford et al., 2021). However, these metrics neglect the sequential alignment of sentences.[6]

4.4.2 Semantic and Sequential Distance Metric

We propose *Semantic and Sequential Distance* (SSD), a metric which takes into account both semantic similarity and the overall relative order of words. Regarding the latter, SSD is related to *Ordering-based Sequence Similarity* (OSS; Gómez-Alonso and Valls, 2008), a measure used by Coco and Keller (2010) to compare sequences of categories representing gaze patterns.[7] Given two sequences of words, i.e., a generated sentence G and a reference sentence R, SSD provides a single positive value representing the overall *dissimilarity* between G and R:

[5] At the time of the research described in this chapter, neither this metric nor the capable pretrained model it is based on were published. It would be informative to consider such metrics and models within the context of this project in the future. See Chapter 5 for an initial exploration of the power of the pretrained vision encoders in capturing human visuo-linguistic signals, and Chapter 8, for a use case where we utilize the CLIPScore metric to quantify the multimodal properties of referring utterances.

[6] Moreover, metrics based on contextual embeddings have been shown to suffer with languages other than English.

[7] Despite its name, OSS is a *distance* measure. Note that it accounts for relative position, but not for semantic similarity.

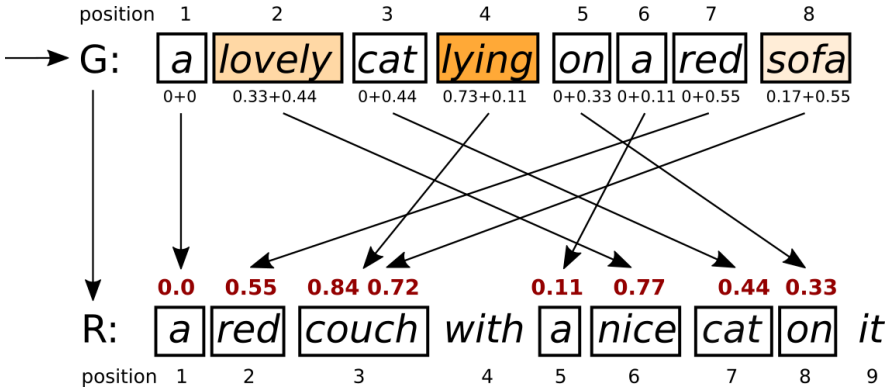


Figure 4.2: SSD. Computation of *gr* (Eq. 4.1). Sums below each word in G stand for $cos + pos$, darker shades of orange for higher *cos* distance. Value of *gr* is the sum of numbers in red (here 3.76). Best viewed in color.

the closer the value to 0, the higher the similarity between the two sentences (note that the value is unbounded). This single value is the average of two terms, *gr* and *rg*, which quantify the overall distance between G and R—the sum of their cosine (*cos*) and positional (*pos*) distance—from G to R and from R to G, respectively. The equation for *gr* is given below:

$$gr = \sum_{i=1}^{N} cos(G_{\mathrm{i}}, R_s(i)) + pos(G_{\mathrm{i}}, R_s(i)) \qquad [4.1]$$

where $R_s(i)$ is the semantically closest element to G_i in R, and *cos* in our experiments is computed over `word2vec` embeddings trained on the 4B-token corpus in Dutch, COW (Tulkens et al., 2016).

Figure 4.2 illustrates how the metric works in practice. Full details are in Appendix A.2. For simplicity, the diagram only shows the computation in the *gr* direction. For example, consider the second element in G, 'lovely'. Its closest embedding in R is 'nice' ($cos = 0.33$). For each of these elements, we retrieve their position index (i.e., 2 for 'lovely' in G and 6 for 'nice' in R), compute their positional distance, and normalize it by the length of the longest sentence in the pair (here R), obtaining $|2 - 6|/9 \approx 0.44$. We then sum up the cosine distance and the positional distance to obtain a score for 'lovely': $0.33 + 0.44 = 0.77$. To obtain the overall *gr* value, we add up the scores for all words in G. We compute *rg* in a similar manner and obtain SSD as follows: $SSD = (gr + rg)/2$.

4.5 Cross-Modal Coordination Analysis

To empirically motivate our generation models, as a preliminary experiment we investigate the level of coordination between visual attention and linguistic production in the DIDEC dataset. In particular, we test whether scanpath similarity

and sentence similarity are correlated, and whether taking into account the sequential nature of the two modalities results in higher cross-modal alignment.

We transform gaze data into time-ordered sequences of object labels, i.e., scanpaths, (e.g., S = 'cat', 'person', 'cat', 'table') using the annotations of object bounding boxes in the MS COCO image dataset. On average, scanpaths have a length of 23.4 object labels. As for captions, we simply take the full sentences and treat them as sequences of words (e.g., C = 'a cute cat cuddled by a boy'). Descriptions contain an average of 12.8 tokens.

Order-sensitive analysis (*sequential*) For each image, we take the set of produced descriptions and compute all pairwise similarities by using SSD (see Section 4.4). Similarly, we take the corresponding scanpaths and compute all pairwise similarities by using OSS (Gómez-Alonso and Valls, 2008). We then calculate Spearman's rank correlation (one-tailed) between the two similarity lists. In this way, we obtain a correlation coefficient and p-value for each of the 307 images in the dataset.

Bag of Words analysis (*BoW*) We compare the correlation observed in the order-sensitive analysis with a BoW approach. Here, we represent a sentence as the average of the `word2vec` embeddings of the words it contains and a scanpath as a term-frequency vector. We then perform the same correlation analysis described above.

Random baseline (*random*) As a sanity check, using the stricter order-sensitive measures, for each image, we re-compute the correlation between the two lists of similarities after randomly shuffling the sentences and corresponding scanpaths per image. We repeat this analysis 3 times.

4.5.1 Results

As shown in Table 4.1, the highest level of alignment is observed in the *sequential* condition, where a significant positive correlation between scanpath and sentence similarities is found for 81 images out of 307 (26%). In BoW, the level of alignment is weaker: a positive correlation is found for 73 images (24%), with lower maximum correlation coefficients (0.65 vs. 0.49). Substantially weaker results can be seen in the *random* condition. These outcomes are in line with those obtained by Coco and Keller (2012) in a small dataset of 576 English sentences describing 24 images.

Overall, the results of the analysis indicate that the product of content and sequentiality better captures the coordination across modalities compared to content alone. Yet, the fact that positive correlations are present for only 26% of the images suggests that coordination across modalities is (not surprisingly) more

	sequential	*BoW*	*random*
# positively corr.	**81**	73	52.3 ± 5.774
% positively corr.	**0.26**	0.24	0.17 ± 0.015
Spearman's ρ (min)	0.15	0.15	0.15 ± 0.002
Spearman's ρ (max)	**0.65**	0.49	0.50 ± 0.042

Table 4.1: Results of the correlation analysis: number and percentage of images with statistically significant ($p < 0.05$) positive correlations, and range of coefficients in the three conditions. For random, average over 3 runs.

complex than what can be captured by the present pairwise similarity computation, confirming the intricacy of the cross-modal temporal alignment (Griffin and Bock, 2000; Vaidyanathan et al., 2015). We take this aspect into account in our proposed generation models.

4.6 Models

The starting point for our models is the one by Anderson et al. (2018).[8] The main aspect that distinguishes this model from other image captioning systems is the use of Faster R-CNN (Ren et al., 2015b) as image encoder, which identifies regions of the image that correspond to objects and are therefore more salient—the authors refer to this type of saliency detection as "bottom-up attention". Each object region i is transformed into an image feature vector v_i. The set of region vectors $\{v_1, \ldots, v_k\}$ is utilized in two ways by two LSTM modules: The first LSTM takes as input the mean-pooled image feature $\overline{\boldsymbol{v}}$ (i.e., the mean of all salient regions) at each time step, concatenated with the two standard elements of a language model, i.e., the previous hidden state and an embedding of the latest generated word. The hidden state of this first LSTM is then used by an attention mechanism to assign weights to the vectors in $\{v_1, \ldots, v_k\}$—the authors refer to this kind of attention as "top-down". Finally, the resulting weighted average feature vector $\hat{v}_t$ is given as input to the second LSTM module, which generates the caption, one word at a time. Note that the set of region vectors $\{v_1, \ldots, v_k\}$ and the mean-pooled vector $\overline{\boldsymbol{v}}$ are constant over the generation of a caption, while the weights over $\{v_1, \ldots, v_k\}$ and hence the weighted average feature vector $\hat{v}_t$ do change dynamically at each time step since they are influenced by the words generated so far.

We take the original model as our baseline and modify it to integrate visual attention defined by gaze behavior. In particular, we replace the mean-pooled

[8] The original implementation of this model can be found at: `https://github.com/peteanderson80/bottom-up-attention`. We developed our models building on the PyTorch re-implementation of the model available at: `https://github.com/poojahira/image-captioning-bottom-up-top-down`.

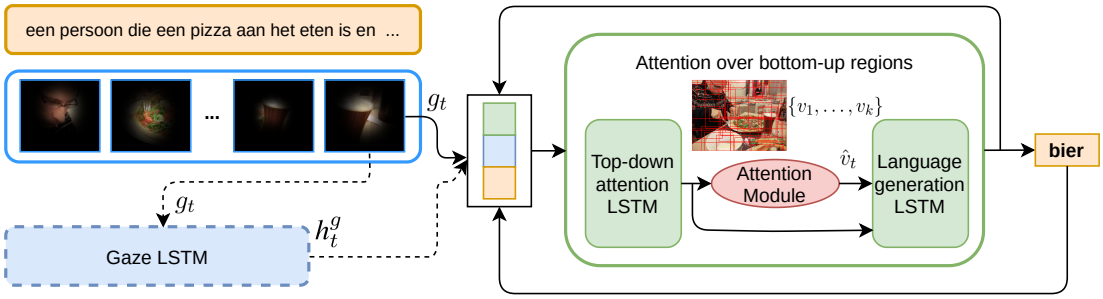


Figure 4.3: Architecture of the GAZE-SEQ and GAZE-2SEQ models. Dashed lines indicate that the connections to and from the Gaze LSTM are only present in the GAZE-2SEQ model.

vector $\overline{\boldsymbol{v}}$ by a gaze vector g computed from masked images representing fixation patterns as explained in Section 4.3. We do not directly modify the set of object regions $\{v_1, \ldots, v_k\}$ present in the original model (i.e., bottom-up attention is still present in our proposed models). However, the top-down attention weights learned by the models are influenced by the gaze patterns given as input. Concretely, we test the following model conditions:

- NO-GAZE: The original model as described above, with exactly the same image feature vectors used by Anderson et al. (2018).

- GAZE-AGG: The mean-pooled vector $\overline{\boldsymbol{v}}$ in the original model is replaced with a gaze image vector $\overline{g}$ computed on the image masked by the aggregated gaze saliency map. As explained in Section 4.3.2, this corresponds to the combination of all participants' fixations per image and hence remains constant over the course of generation.

- GAZE-SEQ: As depicted in Figure 4.3, we replace $\overline{\boldsymbol{v}}$ with g_t, which are features computed for the image that was masked by the participant-specific sequential gaze saliency map at time t. Hence, g_t differs at each time step t. Building on the results of the correlation analysis, this sequential condition thus offers a model of the production process of a speaker where visual processing and language production are time-aligned.

- GAZE-2SEQ: Cross-modal coordination processes seem to go beyond simplistic content and temporal alignment (Griffin and Bock, 2000; Vaidyanathan et al., 2015). To allow for more flexibility, we add an extra gaze-dedicated LSTM component (labeled 'Gaze LSTM' in Figure 4.3), which processes the sequential gaze vector g_t and produces a hidden representation h_t^g. This dynamic hidden representation goes through a linear layer, and then replaces $\overline{\boldsymbol{v}}$ at each time step t.

For the three GAZE models, we also considered a version where $\overline{\boldsymbol{v}}$ is concatenated with $\overline{g}$ or g_t as appropriate, rather than being replaced by the gaze vectors. Since they did not bring in better results, we do not discuss them further in this chapter.

4.7 Experiments

We experiment with the proposed models using the DIDEC dataset and report results per model type.

4.7.1 Setup

We randomly split the DIDEC dataset at the image level, using 80% of the 307 images for training, 10% for validation, and 10% for testing. Further details are available in Appendix A.3.

Pretraining Since DIDEC is a relatively small dataset, we pretrain all our models using a translated version of train/val annotations of MS COCO 2017 version. We machine-translated all the captions in the training and validation sets of MS COCO from English to Dutch using the Google Cloud Translation API.[9] We exclude all images present in our DIDEC validation and test sets from the training set of the translated MS COCO. We randomly split the original MS COCO validation set into validation and test. The final translated dataset in Dutch used for pretraining includes over 118k images for training, and 2.5k images for validation and testing, respectively, with an average of 5 captions per image.

Manual examination of a subset of translated captions showed that they are of good quality overall. Indeed, pretraining the NO-GAZE model with the translated corpus results in an improvement of about 21 CIDEr points (from 40.81 to 61.50) in the DIDEC validation set. Given that the MS COCO dataset is comprised of written captions compared to DIDEC, which includes spoken descriptions, these two datasets can have distinct characteristics. We expect the transfer learning approach to help mitigate this by allowing our models to learn the features of spontaneous spoken descriptions during the fine-tuning phase.

All results reported below were obtained with pretraining (i.e., by initializing all models with the weights learned by the NO-GAZE model on the translated dataset and then fine-tuning on DIDEC; also when applicable, training additional weights from scratch such as the Gaze LSTM in GAZE-2SEQ).

Vocabulary and hyperparameters We use a vocabulary of 21,634 tokens consisting of the union of the entire DIDEC vocabulary and the translated MS

[9] `https://cloud.google.com/translate/`

COCO training set vocabulary. For all model types, we perform parameter search focusing on the learning rate, batch size, word embedding dimensions and the type of optimizer. The reported results refer to models trained with a learning rate of 0.0001 optimizing the Cross-Entropy Loss with the Adam optimizer. The batch size is 64. The image features have 2048 dimensions and the hidden representations have 1024. The generations for the validation set were obtained through beam search with a beam width of 5. The best models were selected with respect to either SSD or CIDEr scores on the validation set, with an early-stopping patience of 50 epochs.[10]

More information regarding reproducibility can be found in Appendix A.4.

4.7.2 Results

The results obtained with different models are shown in Table 4.2. We report results on the test set, averaging over 5 runs with different random seeds, where we select the best models on the validation set based on either SSD or CIDEr. For reference, we also include scores for other metrics not used for model selection. This allows us to check whether scores for other metrics are reasonably good when the models are optimized for a certain metric; however, only scores in the shaded columns allow us to extract conclusions on the relative performance of different model types.

On average, the best GAZE models outperform the NO-GAZE model: 5.81 vs. 5.86 for SSD (lower is better) and 55.74 vs. 52.45 for CIDEr (higher is better). This indicates that eye-tracking data encodes patterns of attention that can contribute to the enhancement of image description generation. Zooming into the different gaze-injected conditions, we find that among the models selected with SSD, the sequential models perform better than GAZE-AGG (5.81 and 5.82 vs. 5.93). This shows that the proposed models succeed (to some extent) in capturing the *sequential* alignment across modalities, and that such alignment can be exploited for description generation. Interestingly, GAZE-2SEQ is the best-performing gaze model: it has the best average SSD across runs and the best absolute single run (5.70 vs. 5.79 and 5.80 by GAZE-SEQ and GAZE-AGG, respectively). This suggests that the higher flexibility and abstraction provided by the gaze-dedicated LSTM component offers a more adequate model of the intricate ways in which the two modalities are aligned.

As for the CIDEr-selected models, on average the gaze-injected models also perform better than NO-GAZE. The best results are obtained with GAZE-AGG (55.74). This is consistent with what CIDEr captures: it takes into account regularities across different descriptions of a given image; therefore, using a saliency map that combines the gaze patterns of several participants leads to higher scores

[10]We use the library at `https://github.com/Maluuba/nlg-eval` to obtain corpus-level BLEU and CIDEr scores.

Model	*selected with SSD*			*selected with CIDEr*		
	SSD	CIDEr	BLEU-4	**CIDEr**	SSD	BLEU-4
NO-GAZE	5.86 (0.25)	55.04 (4.31)	39.09 (2.16)	52.45 (3.43)	6.09 (0.15)	35.60 (2.56)
GAZE-AGG	5.93 (0.10)	53.39 (3.56)	38.84 (1.70)	**55.74** (3.74)	5.97 (0.12)	37.69 (1.71)
GAZE-SEQ	5.82 (0.03)	56.16 (1.62)	39.80 (1.24)	53.59 (2.03)	6.10 (0.14)	36.09 (3.01)
GAZE-2SEQ	**5.81** (0.15)	53.55 (1.69)	38.05 (1.88)	52.94 (2.27)	5.93 (0.14)	36.27 (3.04)

Table 4.2: Test set results (average over 5 runs, with standard deviations in brackets) for the models selected with SSD and with CIDEr. Scores for BLEU-4 and SSD/CIDEr when not used for model selection are shown for reference only. For SSD, lower is better; for CIDEr and BLEU-4, higher is better.

than inputting sequential saliency maps, which model the path of fixations of each speaker independently. This variability seems to have a negative effect on CIDEr scores of sequential models, which are lower than GAZE-AGG; yet higher than NO-GAZE (53.59 and 52.94 vs. 52.45).

It is worth noting that CIDEr and BLEU-4 scores obtained with the SSD-selected models are sensible, which indicates that the generated descriptions do not suffer with respect to distinct aspects evaluated by other metrics when the models are optimized with SSD. Indeed, the highest CIDEr score obtained among models selected via SSD (GAZE-SEQ: 56.16) is even higher than that obtained by the best CIDEr-selected one (GAZE-AGG: 55.74). However, this is likely due to CIDEr being sensitive to lexical differences between the test set and the validation set used for model selection, which could lead to slightly different patterns.

4.8 Analysis

This section presents an analysis of the descriptions generated by the models on the test set (446 descriptions). We focus on one single run per model.

Cross-modal sequential alignment Given what SSD captures, our results indicate that the captions generated by GAZE-2SEQ are better aligned—in terms of semantic content and the order of words—with the human captions than the ones generated by non-sequential models. Arguably, this enhanced alignment is driven by the specific information provided by the scanpath of each speaker. If this information is used effectively by the sequential models, then we should see more variation in their output. By definition, the non-sequential models generate only one single caption per image. Are the sequential models able to exploit the variation stemming from the speaker-specific scanpaths? Indeed, we find that GAZE-2SEQ generates an average of 4.4 different descriptions per image (i.e., 30% of the generated captions per image are unique).

Furthermore, we conjecture that tighter coordination between scanpaths and corresponding descriptions should give rise to more variation, since presumably

the scanpath has a stronger causal effect on the description in such cases. To test this, we take the 30 images in the test set and divide them into two groups: (A) images for which a significant positive correlation was found in the cross-modal coordination analysis of Section 4.5; (B) all the others. These groups include, respectively, 10 and 20 images. As hypothesized, we observe a higher percentage of unique captions per image in A (35%) compared to B (27%).

Quantitative analysis We explore whether there are any quantitative differences across models regarding two aspects, i.e., the average length in tokens of the captions, and the size of the vocabulary produced. No striking differences are observed regarding caption length: NO-GAZE produces slightly shorter captions (avg. 7.5) compared to both GAZE-2SEQ (avg. 7.7) and GAZE-AGG (avg. 8.1). The difference, however, is negligible. Indeed, it appears that equipping models with gaze data does not make sentence length substantially closer to the length of the reference captions (avg. 12.3 tokens).

In contrast, there are more pronounced differences regarding vocabulary. While GAZE-AGG has a similar vocabulary size (68 unique tokens produced) to NO-GAZE (63), GAZE-2SEQ is found to almost double it, with 109 unique tokens produced. Though this number is still far from the total size of the reference vocabulary (813), this trend suggests that a more diverse and perhaps 'targeted' language is encouraged when specific image regions are identified through gaze-based attention. The following qualitative analysis sheds some light on this hypothesis.

Qualitative analysis Manual inspection of all the captions generated by the models reveals interesting qualitative differences. First, captions generated by gaze-injected models are more likely to refer to objects—even when they are small and/or in the background—which are image-specific and thus very relevant for the caption. For example, when describing the top left image in Figure 4.4, NO-GAZE does not mention the word *donuts*, which is produced by both GAZE-AGG and GAZE-2SEQ. Second, gaze-injected models produce language that seems to reflect the uncertainty present in the visual input. For the top right image in Figure 4.4, e.g., both GAZE-AGG and GAZE-2SEQ generate disfluencies such as *uh* (interestingly, several participants' descriptions include similar disfluencies for this same image, which suggests some degree of uncertainty at the visual level); in contrast, in the entire test set no disfluencies are produced by NO-GAZE.

Finally, we find that GAZE-2SEQ is able to produce captions that somehow 'compress' a repetitive sequence (e.g., *a red bus and a bus*) into a shorter one, embedding a number (e.g., *two buses that are parked*; see the bottom left example in Figure 4.4). This phenomenon is never observed in the output of other models (crucially, not even in GAZE-SEQ). We thus conjecture that this ability is due to the presence of the gaze-dedicated LSTM, which allows for a more abstract processing of the visual input. However, the presence of gaze data does not fully

	specificity	**disfluency**
NO-G	een vrouw die in de keuken staat… *(a woman who is standing in the kitchen…)*	een foto van een straat met een aantal vogels *(a photo of a street with a number of birds)*
2SEQ	een vrouw in een keuken met **donuts** *(a woman in the kitchen with **donuts**)*	**uh uh uh uh** met een aantal vogels *(**uh uh uh uh** with some birds)*

	compression	**repetition**
NO-G	een rode bus en een bus *(a red bus and a bus)*	een straat met **auto's** en **auto's** *(a street with **cars** and **cars**)*
2SEQ	**twee** bussen die geparkeerd staan *(**two** buses that are parked)*	een straat in de stad met **auto's** en **auto's** *(a street in the city with **cars** and **cars**)*

Figure 4.4: Phenomena that are either particular to gaze models (specificity, disfluency, and compression) or common to all (repetition). Abbreviations NO-G and 2SEQ refer to NO-GAZE and GAZE-2SEQ, respectively.

solve the issue of words being repeated within the same caption, as illustrated by the bottom right example in Figure 4.4. Indeed, this weakness is common to all models, including the best performing GAZE-2SEQ.

4.9 Conclusion

We tackled the problem of automatically generating an image description from a novel perspective, by modeling the sequential visual processes of a speaker concurrently with language production. Our study shows that better descriptions—i.e., more aligned with speakers' productions in terms of content and order of words—can be obtained by equipping models with human gaze data. Moreover, this trend is more pronounced when gaze data is fed *sequentially*, in line with cognitive theories of sequential cross-modal alignment (e.g., Coco and Keller, 2012).

Our study was conducted using the Dutch language dataset DIDEC (van Miltenburg et al., 2018b), which posed the additional challenges of dealing with a small amount of data and a low resource language. We believe, however, that there is value in conducting research with languages other than English. In the future, our approach and new evaluation measure could be applied to larger eye-

tracking datasets, such as the English dataset by He et al. (2019). Since different eye-tracking datasets tend to make use of different gaze encodings and formats, the amount of preprocessing and analysis steps required to apply our method to other resources was beyond the scope of this chapter. We leave testing whether the reported pattern of results holds across different languages to future work.

Despite the challenges mentioned above, our experiments show that a state-of-art image captioning model can be effectively extended to encode cognitive information present in human gaze behavior. Comparing different ways of aligning the gaze modality with language production, as we have done in the present work, can shed light on how these processes unfold in human cognition. This type of computational modeling could help, for example, study the interaction between gaze and the production of filler words and repetitions, which we have not investigated in detail. Taken together, our results open the door to further work in this direction and support the case for computational approaches leveraging cognitive data.

In the next chapter, we conduct more analyses into the DIDEC data, focusing on quantifying the variation in language production and eye movements. We then investigate whether pretrained encoders frequently used in contemporary multimodal models capture the variation in human visuo-linguistic signals.

Chapter 5

Variation in Human Signals During Image Description Generation

The material in this chapter is based on: Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2023. Describing Images *Fast and Slow*: Quantifying and Predicting the Variation in Human Signals during Visuo-Linguistic Processes. To appear in *Proceedings of the 2024 Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Contributions: Ece Takmaz: Implementing and running the experiments, writing and revising the paper. Sandro Pezzelle and Raquel Fernández: Supervising the research, providing feedback on the paper.

5.1 Introduction

Humans can capture the gist of an image usually incredibly fast – 100 msec could be enough (Oliva, 2005; Oliva and Torralba, 2006); however, they would need more time to act on an image. For instance, human behavior while describing images illustrates the intricacies of visuo-linguistic processes. There may be repetitions, silent intervals and disfluencies, with considerable degrees of variation in what is uttered across speakers. The period prior to the utterance involves perceiving the image, conceptualizing the message, retrieving the labels of the entities to mention, formulating and preparing to articulate a grammatical and relevant utterance (Levelt, 1981; Slobin, 2003).

Min: 1.69 sec

Max: 7.07 sec

Figure 5.1: The images with the minimum and maximum mean speech onsets across speakers in the dataset. The image with the maximum onset also elicits the highest variation in the first nouns of the descriptions.

As a result, we observe variations in **speech onsets**, as in Figure 5.1, which could be indicative of the relative cognitive complexity induced by the images (Coco and Keller, 2015a; Gatt et al., 2017). In addition, different speakers might start their utterances with different words (**starting points**, see MacWhinney, 1977), continuing to produce a varied set of image descriptions (**linguistic variation**) with **variation in gaze**. These signify the intricate cross-modal relation between visual and linguistic processes in humans (Griffin and Bock, 2000; Ferreira and Rehrig, 2019).

Although human data can be rich in behavioral signals, current pretrained multimodal models virtually never receive information about such signals during training. The models generate descriptions without necessarily modeling how human processes unfold. For instance, deep neural networks can output words at the same rate even for images that would result in diverse speech behavior by humans due to complexity or ambiguity. Moreover, there is a gap between the manner in which humans perceive stimuli as compared to how large models process them. Model-predicted surprisal values for linguistic input can be lower

than human surprisal, possibly due to the massive size of the training data and the number of model parameters (van Schijndel and Linzen, 2021; Arehalli et al., 2022; Oh and Schuler, 2023a,b). Models also display different patterns of visual attention compared to humans (Das et al., 2016).

We argue that it is essential to consider human signals such as speech onsets and looking times, as they reflect the complexity and ambiguity of visuo-linguistic tasks (Coco and Keller, 2015a; Gatt et al., 2017; van der Meulen et al., 2001; Meyer and van der Meulen, 2000; van Miltenburg et al., 2018b). It is therefore desirable if models encode what leads to variations in such signals to help generate image descriptions in a way that is aligned with human processes and with types of variations observed in human data (van Miltenburg et al., 2018a). To this end, several applications have exploited human gaze to enhance image captioning, as in Chapter 4, and visual question answering models (Sugano and Bulling, 2016; He et al., 2019; Takmaz et al., 2020b; Sood et al., 2021, 2023). Still, the relation between gaze on images and language is not widely researched in NLP (Alacam et al., 2022).

We first explore the natural dynamics in visuo-linguistic processes using the same dataset we utilized in Chapter 4, the DIDEC dataset (van Miltenburg et al., 2018b). As this corpus provides gaze and speech data concurrently collected while participants describe images depicting real-life scenes, it is a rich resource to investigate our research questions in this chapter, as well. We preprocess the DIDEC dataset extensively, utilizing more recent methods as compared to the ones used in Chapter 4. We propose metrics to quantify the variation in visual and linguistic modalities, and reveal for the first time significant correlations between speech onsets, variation in starting points, descriptions and gaze.

We hypothesize that this variation is partly due to the properties of the images, and that similar images would elicit similar amounts of variation. Given the superior performance of pretrained encoders that are widely used in multimodal models, we investigate whether visual encoders such as CLIP Radford et al. (2021) and ViT Dosovitskiy et al. (2021) capture information regarding the variation in visuo-linguistic signals.[1] This is akin to probing pretrained models for meaningful syntactic and semantic information; see Conneau et al., 2018a. Using a similarity-based prediction method (Anderson et al., 2016a), we find that the pretrained encoders capture variation in signals to a limited extent. Our findings suggest that underlying factors leading to variation are encoded rather weakly by pretrained models. With our work, we aim to direct attention towards the importance of the information contained in such signals and the variation thereof when crowdsourcing data as well as during model development.

[1]Our preprocessed data and code are publicly available at `https://github.com/ecekt/visuolinguistic_signal_variation`.

5.2 Background

We first give an overview of visuo-linguistic processes in humans in Section 5.2.1, and then, in models in Section 5.2.2.

5.2.1 Visuo-Linguistic Processes in Humans

Cross-modal processes Describing images requires the linear unfolding of complex cross-modal processes between vision and language (Henderson and Ferreira, 2013; Griffin and Bock, 2000; Gleitman et al., 2007; Coco and Keller, 2012; Ferreira and Rehrig, 2019; Henderson, 2017). There exist several theories regarding how the 'linearization' (Levelt, 1981) takes place in sentence formulation in relation to visual processes (Griffin, 2004; Meyer, 2004; Ferreira and Rehrig, 2019). These theories consider the speaker's knowledge and expectation regarding the contents of the image, as factors affecting the allocation of gaze and the formulation of a description (Henderson, 2017; Ferreira and Rehrig, 2019). In addition, the way people look at an image changes based on the task at hand (Yarbus, 1967; Buswell, 1935; Castelhano et al., 2009), with similar sequences of fixations (*scanpaths*) leading to the production of similar sentences (Coco and Keller, 2012). Therefore, we hypothesize that the variation in language production and eye movements could be correlated.

Starting points A sentence must have a starting point, given that words need to be uttered in a linear order (Levelt, 1981). We focus on the first uttered noun as the starting point of image descriptions as they correspond to object categories, and gaze scanpaths are represented by the sequence of object or entity categories fixated by a participant, expressed as nouns. Additionally, the order of mention of these categories is the point of interest in linearization studies that investigate language production parallel to visual processes (Ferreira and Rehrig, 2019). Starting points can be selected based on a variety of factors (canonical word order of the language, perspective of the speaker, complexity of the planned sentence; see MacWhinney, 1977). When describing images, visual properties of an image influence how a sentence begins and unfolds (Bock et al., 2004). These findings signify how the selection of starting points can be influenced by a set of complex visuo-linguistic factors.

Variation in image descriptions People generally describe images with some variation. Jas and Parikh (2015) report that images with people and large objects tend to be described more specifically, whereas generic buildings, ambiguous scenes and images with less-important objects tend to elicit more varied descriptions. The degree to which the descriptions of an image vary is referred to as 'image specificity' by Jas and Parikh (2015), who propose an automatic metric to quantify it using the similarity scores between the WordNet paths of words in

descriptions (Miller, 1994). van Miltenburg et al. (2018b) explore image specificity in the corpus that we use in this study, utilizing word2vec vectors (Mikolov et al., 2013a) to compute the similarity scores. They find that the variation in descriptions is only to a limited extent due to the image's contents as there also seems to be an effect of language (English vs. Dutch). Additionally, their results indicate that attention maps extracted using gaze data do not help predict image specificity (van Miltenburg et al., 2018b). In this work, we also quantify and predict image specificity proposing different approaches.

Speech onsets Slower speech onsets indicate that a deliberate, effortful process is taking place, as compared to fast onsets; as claimed in the dual process theory (Wason and Evans, 1974; Kahneman, 2012). Various intertwined linguistic and visual processes modulate speech onsets and the latency of referring to an object (Meyer and van der Meulen, 2000; Coco and Keller, 2015a), such as the contents of an image and the locations of the objects (Gatt et al., 2017; Esaulova et al., 2019). This indicates that speech onsets are strongly linked to image features. Given the importance of speech onsets in relation to visuo-linguistic processes and the cognitive requirements of a task, the mean speech onset induced by an image across speakers is one of the signals we focus on.

5.2.2 Multimodal NLP

Pretrained models Many recent multimodal models employ frozen pretrained unimodal models and combine them with either no further training or via trained lightweight mapping networks (Berrios et al., 2023; Alayrac et al., 2022; Mañas et al., 2023; Tsimpoukelli et al., 2021; Li et al., 2023b; Mokady et al., 2021; Chen et al., 2022). Particularly, the visual encoder of the CLIP model (Radford et al., 2021) has been utilized in these models as a foundation model with strong zero-shot capabilities that improves multimodal models (Shen et al., 2022).

By training classifiers on top of visual encoders, Berger et al. (2023) predict the existence of linguistic features such as passive voice and the use of numeral expressions in image descriptions, and indicate that the selection of such linguistic features is constrained by visual features. These findings point to the underlying capabilities of pretrained models pertaining to human cognitive processes.

Human signals in NLP Most previous research into the use of human signals focuses on text-only cases (Klerke et al., 2016; Barrett et al., 2018, 2016; Mishra and Bhattacharyya, 2018; Hollenstein et al., 2021a, 2022, 2021b; Pouw et al., 2023; Ding et al., 2022; Ren and Xiong, 2021; Dong et al., 2022; Khurana et al., 2023; Mathias et al., 2020; Zhang et al., 2020a). However, the relationship between human gaze on images and language production, and its potential contribution to computer vision and NLP has been investigated even before the existence of

pretrained models (Yun et al., 2013). Research into whether the attention distributions in multimodal models correlate with human attention reveals contrasting findings (Das et al., 2016; Gella and Keller, 2018; He et al., 2019; Sood et al., 2021). Several works show that the use of human gaze enhances image captioning and visual question answering (Sugano and Bulling, 2016; He et al., 2019; Takmaz et al., 2020b; Sood et al., 2021, 2023). Yet, modeling gaze in conjunction with linguistic processes is still an under-explored area in NLP (Alacam et al., 2022).

In our work, we investigate the variation of a set of human signals in a corpus, as well as whether pretrained vision encoders can encode information related to these signals. Although such models are shown to be very effective in multimodal tasks, they are still under-explored from this point of view.

5.3 Data

We aim to explore the variation in human signals in visuo-linguistic processes and whether pretrained models can capture such variation in a realistic setup. A dataset consisting of simultaneous language production and eye movements over complex images would enable such an exploration. Therefore, we opt for using the DIDEC corpus (van Miltenburg et al., 2018b) instead of other existing image description datasets with eye-tracking, as this corpus allows us to delve into the dynamics of visual and linguistic processes in parallel. There exist few datasets containing such information, which we did not opt for utilizing, as they differ in their tasks (narratives (Vaidyanathan et al., 2018)), or the processing steps the authors have taken (e.g. only a small subset of the captions were checked manually (Vaidyanathan et al., 2018), the authors sample one gaze point every 4 points (He et al., 2019)). DIDEC dataset comprises manually checked descriptions of high quality, and the gaze data is provided in a raw format enabling custom processing.

As in Chapter 4, we use the 'production viewing' subset of DIDEC. Next, we explain how we extract features corresponding to human signals in visuo-linguistic processes from this dataset, to obtain 4586 descriptions with speech onsets, starting points, and fixated regions.

5.3.1 Visual Data

Using the raw gaze samples in DIDEC (van Miltenburg et al., 2018b) labeled as fixations, saccades, and blinks, we create fixation windows by treating saccades and blinks as boundaries (Salvucci and Goldberg, 2000). The gaze samples in the fixation window are then put into a list, skipping the ones that fall outside the boundaries of the images. To visually represent a fixation, we feed its gaze points as coordinate prompts to the Segment Anything Model (SAM; Kirillov et al., 2023). Using the prompts, this model predicts the objects the gaze corresponds

to, and outputs masks corresponding to fixated regions. We use the ViT-L version of the model building on vision transformers (Dosovitskiy et al., 2021), as it achieves good performance (Kirillov et al., 2023). We obtain a single mask per fixation window. The masks sometimes span non-contiguous regions; therefore, we utilize the bounding box based on the x-y limits of the predicted mask.

5.3.2 Linguistic Data

Speech onsets The dataset supplies audio files for spoken descriptions and their transcripts. To extract word-level timestamps, we use WhisperX (Bain et al., 2023) based on Whisper (Radford et al., 2023).[2] We relay the transcripts directly into the alignment function of WhisperX. The output contains the start and end timestamps of each word. This also allows us to extract information regarding when the participants start talking, i.e., speech onsets. The mean speech onset is 3.42 sec, and the median is 2.65 sec. We observe variation across participants and images, as the onsets can go up to 25.37 sec with a standard deviation of 2.45.

Starting points We use the spaCy library for tokenization, part-of-speech tagging, and lemmatization of the words in the descriptions.[3] For Dutch, the library provides 3 models (small, medium, and large). Upon manual inspection of 50 random samples from the data processed by each model, we opted for the large model, which yields the least number of errors. See Appendix B.1 for more details.

5.4 Variation in Human Signals

We first delve into the nature of the variation across humans per image in the DIDEC dataset. Our focus is on uncovering potential correlations between the variations in human signals in visuo-linguistic processes. We first explain how we quantify each signal and its variation, see Figure 5.2 for an example image with all of its variation scores. Then, we conduct pairwise correlation analyses between the 4 variables. If there exist correlations between variations across signals, one can speculate that at least part of the correlation stems from the image, with the rest being potentially due to factors such as viewing order, priming and cognitive load.

[2]The model for obtaining alignments for audio in Dutch: `jonatasgrosman/wav2vec2-large-xlsr-53-dutch`

[3]'nl_core_news_lg' pipeline from `https://spacy.io/`

Mean onset: 3.46 seconds
Variation in starting points: 11
Most common starting point: *pier*
Image specificity BLEU-2: 0.39
Variation in gaze: 38.47

*een **pier** waar het heel erg druk is uh rechts is een vis aquarium waar je vissen kan aanraken*
*(a **pier** where it is very busy uh on the right is a fish aquarium where you can touch fish)*

*een drukke **straat** met een aantal restaurants pier 39*
*(a busy **street** with a number of restaurants pier 39)*

***pier** waar veel mensen lopen*
*(**pier** where many people walk)*

*een drukbezette **pier***
*(a busy **pier**)*

*een toeristische **plaats** waar veel verschillende entertainment dingen te doen zijn*
*(a touristic **place** where there are many different entertainment things to do)*

*de **ingang** van een aquarium met veel mensen op een plein*
*(the **entrance** to an aquarium with many people in a square)*

Figure 5.2: An image with its variation scores, a subset of its descriptions (along with the English translations in parentheses), and the eye movements of a single participant. In the descriptions, the words in boldface indicate the starting points in Dutch and their equivalents in English.

5.4.1 Variation in Speech Onsets

We inspect the mean and standard deviation of speech onsets per image, see histograms in Appendix B.2. The mean onsets per image range between 1.69 and 7.07 seconds, constituting a non-normal distribution skewed towards shorter onsets ($p < .001$, 65.77% of the onsets shorter than the mean onset). For some images, some participants start talking immediately; whereas, in other cases, they wait for a considerable amount of time before speaking. This observation resonates with the fast and slow systems from the dual process theory (Wason and Evans, 1974; Kahneman, 2012), suggesting that more complex processes are recruited while describing certain images. However, even for a single image, the participants might start speaking at varying times (with SD per image ranging from 0.44 to 6.33). This suggests that various factors are at play while describing

images, such as contextual and speaker-specific effects.

To have a better picture of onset variation, we compare the onsets for an image against each other. Leaving one onset out of the set of onsets for an image, we calculate the average of the rest ($\approx$ 14 onsets). The difference between the average and the left-out onset corresponds to error. We perform this calculation for each sample. Then, we take the mean over all the samples, which yields an error of 1.625 seconds. This error is a proxy for the average variation over the participants, which suggests that there is a difference in response times across humans when prompted with the same image.

The DIDEC corpus comes with 3 mutually-exclusive image subsets called 'lists'. Each participant views only one list. We find that the mean onsets in List 2 are significantly shorter than the other two sets ($p < .001$, independent samples t-test). Since both the images and the participants are different across lists, it is not straightforward to separate their effects. See Appendix B.3 for a participant-based analysis of mean onsets.

5.4.2 Variation in Starting Points

Counting the first nouns of image descriptions reveals that there is an imbalance in the starting points in the data.[4] The participants utter words such as *man*, *people*, *woman*, *bus* and *street* most frequently as the first noun of a description (370, 238, 221, 174, 141, respectively, constituting in total 25% of the samples). This is potentially due to the salience of such entities and their frequency. We represent the variation in starting points by the number of unique starting points uttered per image, yielding $mean = 6.45$, $min = 1$, $max = 13$. These values indicate that some images elicit the same first nouns, whereas some others prompt the production of a range of starting points. See Figure 5.3 for the images with the minimum and maximum number of unique starting points.

5.4.3 Variation in Full Descriptions

Each image can be described in distinct ways, both in terms of the words uttered and their order. We quantify the *linguistic variation* in image descriptions, following a different approach compared to Jas and Parikh (2015) and van Miltenburg et al. (2018b). We adopt a widely used NLG metric, BLEU (Papineni et al., 2002). This metric computes n-gram-based precision scores between a generated sentence and a set of references. We opt for the bigram version (BLEU-2), since we are mostly interested in the surface form variation of words, and to a limited extent, the sequences of words. BLEU-2 allows us to measure the linguistic

[4]Although it would also be interesting to consider synonyms since they would be referring to the same object, lexical choices reflect categorization and conceptualization of objects that can be affected by the visual context in which the object is situated (Gualdoni et al., 2023). Therefore, this is the type of variation of interest for starting points.

Min: 1

Max: 13

Figure 5.3: Variation in the number of unique starting points. For the image with the minimum score, all the speakers start with *keuken*, meaning kitchen. The image with the maximum score has descriptions starting with a variety of words: bureau, fitness, huiskamer, springding, atletiek, balk, hoek, tafel, plek, turnattribuut, restaurant, bank, turnobject.

variation in descriptions independently of a pretrained model.[5] We calculate the BLEU-2 score between a description and the remaining descriptions for the image constituting the reference set. Then, we take the average over all descriptions of an image.[6] This method yields an extensive range of normally distributed scores ($\mu = 0.53, min = 0.25, max = 0.81$). Figure 5.4 depicts the images with the minimum and maximum variation in the descriptions.

5.4.4 Variation in Gaze

The variation in eye movements has been quantified in various ways in the literature: scanpath complexity, dispersion of the heatmap of gaze on an image, entropy of the gaze distribution (Coco and Keller, 2015a). We propose a distance metric based around the contents of fixated regions and their orders. We represent a scanpath in the form of a sequence of fixation bounding boxes represented as (x_1, y_1, x_2, y_2). Given two scanpaths S_1 and S_2, for each fixation box in S_1, we find the most similar box in S_2 that yields the highest ratio of intersection over union (IoU) between the bounding boxes. The IoU dissimilarity $(1 - IoU)$ as well

[5] See Appendices B.4 and B.5 for a semantic variation metric we propose using Dutch BERT-based representations (BERTje; de Vries et al., 2019), another combining BERTje and BLEU-2-based variation, as well as a comparison to human annotations provided by Jas and Parikh (2015).

[6] This metric is similar to Self-BLEU (Zhu et al., 2018), which was proposed to calculate the diversity of the sentences generated by a model. In Self-BLEU, each generated sentence is compared to the rest of the generated sentences, and an average of the whole set is computed to indicate how varied a model's generations are.

Min: 0.248

Max: 0.811

Figure 5.4: BLEU-2-based linguistic variation scores. The image with the maximum BLEU-2 score elicited the most similar set of descriptions in the dataset.

as the normalized positional distance between these boxes are summed up. This step is performed for all fixation boxes in S_1. The total gives us a comparison score for two scanpaths. We compare S_1 to all the other scanpaths for the same image and then, take the average. Each scanpath for the image is compared to the rest of the related scanpaths in the same way. This yields 15 image-scanpath variation scores, whose mean corresponds to the gaze variation score of a single image. The higher this score is, the more variation exists in the gaze modality. We obtain a range of gaze variation scores for the whole set ($mean = 24.00$, $min = 11.22$, $max = 38.79$). Figure 5.5 illustrates the images with the minimum and maximum variation in gaze.

Min: 11.22

Max: 38.79

Figure 5.5: Variation in gaze. The image with the minimum score elicited more similar scanpaths across speakers than the one with the maximum score.

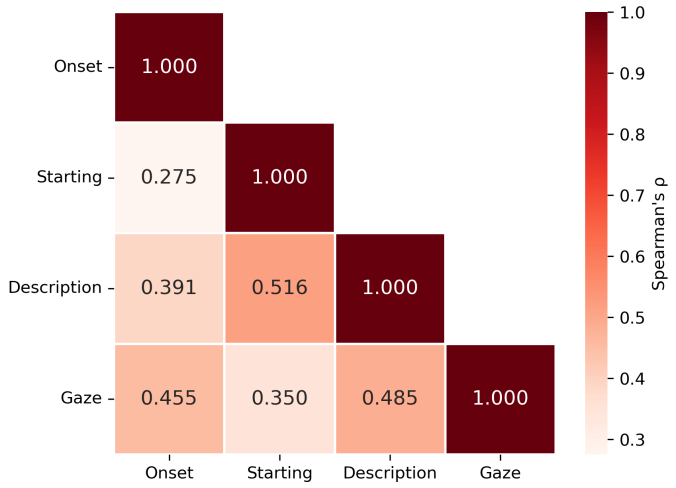


Figure 5.6: Spearman's correlation coefficients between the mean onsets per image (Onset), the variation in starting points (Starting), BLEU-2-based variation in full descriptions (Description), and the variation in gaze (Gaze) in the full dataset. Since higher BLEU scores mean less variation unlike the trends in the other measures, we utilize $1 - BLEU$ for better interpretability. All of the correlations are significant, $p < .001$.

5.4.5 Correlation between Variations

In the previous subsections, we have quantified the variation in speech onsets, starting points, descriptions and gaze per image. We now turn to the correlation between the variation types. Since the initial common point is the image itself, we hypothesize that image features contribute to varying levels of variation in different modalities. We run Spearman's correlation between each type of variation.[7] When interpreting the magnitudes of the correlation coefficients, we use the terminology suggested by Prion and Haerling (2014). See Figure 5.6 for all correlation results.

We find a significant negative correlation, approaching moderate effect, between BLEU-2-based linguistic variation and the mean onset of an image (Spearman's $\rho = -0.391, p < .001$, see Appendix B.6 for the regression line). This means that speakers start describing images that yield more similar descriptions earlier.[8] In addition, as starting points vary, image descriptions become less similar (moderate, Spearman's $\rho = -0.516, p < .001$), indicating that initial

[7] We conduct Spearman's rank correlation analysis to uncover monotonic relations in the data. This type of correlation does not assume a particular distribution of the data (non-parametric, as opposed to Pearson's normality assumption). Since some of the signals we have investigated are non-normally distributed (e.g., speech onsets), and the dataset is relatively small, we opted for Spearman.

[8] Unlike this correlation, we find that speech onsets are not correlated with how many words or nouns are uttered.

deviations continue until the end of language production.

We find that the variation in gaze significantly correlates with speech onsets (moderate, Spearman's $\rho = 0.455, p < .001$); the variation in starting points (weak, Spearman's $\rho = 0.350, p < .001$); and the variation in full descriptions (moderate, Spearman's $\rho = -0.485, p < .001$). These outcomes indicate that high variation in gaze tends to co-occur with longer onsets, high variation in starting points, and less similarity in descriptions.[9]

The correlations reveal a connection between the variation in visual and linguistic modalities. We hypothesize that the underlying reasons for such variation partly reside in the features of an image, echoing the claims by Jas and Parikh (2015) and Berger et al. (2023). In this sense, similar images are expected to elicit similar amounts of variation. Hence, the results motivate our research into whether image features as encoded by pretrained models can capture the variation in gaze and language.

5.5 Similarity-based Prediction

In light of the correlation findings in Section 5.4, we expect image features to be predictive of the variation in visuo-linguistic signals to some extent. We explore if the similarity scores between image features encoded by pretrained models would be meaningful when capturing variation in human signals. In particular, we hypothesize that the signals that are more internal to the pretrained models' training objectives would be captured better. For instance, CLIP was trained with respect to an image-to-text alignment objective (Radford et al., 2021); hence, it would be reasonable to expect that signals that are more inherent to the visual and language data could be encoded better compared to speech onsets, which are never seen by the model.

Approach We employ an approach that was proposed as an alternative to training regression models and representational similarity analysis, for predicting fMRI signals given linguistic input (Anderson et al., 2016a). Using the similarities between model-encoded stimuli (embeddings of concepts) and the corresponding fMRI responses, the authors predict the fMRI signals for novel stimuli for which embeddings exist. This approach has been utilized to assess the extent to which deep neural networks capture brain representations in language-only and visually grounded setups (Anderson et al., 2017; Bruera et al., 2023; Bruera and Poesio, 2023). We explain how we operationalize this extrapolation method for our purposes in Section 5.5.1. As this approach does not require training, it is suitable for shedding light on the predictive power of pretrained image representations,

[9]Investigating the correlation between these types of variation and the number of objects in an image is not straightforward, as current object detection algorithms annotate images exhaustively, yielding a high number for many images.

given the small size of the dataset we use. We determine the splits based on the images. Hence, to mitigate imbalance issues, we create 50 random split setups with 90% training (277 images) and 10% test sets (30 images), and report results on the average of these 50 setups. Across setups, the training sets have similar representative powers in terms of their CLIP vector similarities to the images in the corresponding test sets.

Visual encoders To encode the images, we exploit three visual encoders: CLIP, ViT, and a randomly initialized CLIP model (without training at all). We use the ViT-B/32 version of CLIP's visual encoder (Radford et al., 2021), and extract the final 512-dimensional output for each image. Since this encoder has been trained in coordination with CLIP's textual encoder (Radford et al., 2021), we expect it to capture not only vision-related features, but also properties that are aligned with language. In addition, we test the representations of a purely visual encoder trained on object recognition, ViT (Dosovitskiy et al., 2021). We extract the last hidden states from ViT, and use the vector corresponding to the `[CLS]` token as the image representation. Finally, we also experiment with a randomly-initialized version of CLIP (RndCLIP), along the lines of what Berger et al. (2023) did to avoid the information learned during pretraining.

5.5.1 Predicting the Variation in Descriptions

From the training set, we retrieve k images that are closest to the target image—the image for which we predict a signal variation score—based on their representational similarities, echoing the k-nearest neighbors algorithm. The final score is the weighted average of the variation found in the neighboring images. The weights correspond to the similarity scores between the retrieved images and the target image.

As depicted in Table 5.1, we find significant, yet weak, positive correlations for almost half of the 50 split configurations both for CLIP and ViT, with no meaningful correlations for RndCLIP. CLIP slightly outperforms ViT, suggesting that language alignment in the visual modality yields a potential benefit in estimating the variation in descriptions.

The loss corresponds to the average difference between the predicted and target scores across the dataset. The losses are similar across encoder types despite the differences in correlations. Since this method makes predictions based on the ground truth outputs of the retrieved set, it is likely that the predictions remain in a similar range.

5.5.2 Predicting Onset

We perform the similarity-based prediction approach outlined in Section 5.5.1 to predict mean speech onsets per image. Since longer onsets can be associated with

Model	Coefficient	Sig.	Loss
CLIP	**0.3380**	**27**	0.0738
ViT	0.3135	23	**0.0723**
RndCLIP	0.0472	3	0.0744

Table 5.1: Predicting variation in descriptions with the similarity-based approach, $k = 277$. Averages over 50 random splits. 'Coefficient' and 'Sig.' correspond to Spearman's ρ correlation coefficient and how many runs out of 50 yield significant correlations with $p < 0.05$.

more cognitively demanding images, we are interested in the average onset elicited by each image. The results (see Table 5.2) indicate that, by using a larger sample of CLIP-encoded images, we can obtain predictions weakly correlating with the target onsets. The differences in the results when using different k values suggest that the choice of the retrieval set limits the boundaries of the predictions, even though the median image similarity score for $k = 1$ is 0.77 in the dataset.

Model	Coefficient	Sig.	Loss	Range
CLIP-277	**0.2981**	**18**	0.8216	3.37 - 3.50
CLIP-10	0.2500	10	**0.7989**	2.60 - 4.37
CLIP-5	0.2265	14	0.8149	2.26 - 4.81
CLIP-1	0.0640	4	1.0746	1.69 - 6.39
ViT	0.2428	17	0.8072	3.11 - 3.67
RndCLIP	0.0350	3	0.8249	3.38 - 3.47

Table 5.2: Predicting mean speech onsets with the similarity-based approach. The numbers in the model names correspond to k when retrieving closest images from the training set. RndCLIP and ViT with $k = 277$. 'Range' is the range of the predictions for the test set.

When we use 277 images encoded with ViT to obtain the image similarities, the correlation is weaker than the same setup with CLIP. When we encode the images with RndCLIP, although the loss is quite similar to the other setups, there is no meaningful correlation. The predictions in general center around the mean onset, as they are based on the outputs from the retrieval set.

5.5.3 Predicting Starting Points

We utilize the similarity-based prediction algorithm to predict the first uttered nouns of the descriptions. Since this is a subtask of generating descriptions, we consider this an interesting use case. For each image, we represent the most common first noun as a one-hot vector (with the dimensions being 739, corresponding to the size of the first-noun vocabulary of the whole dataset). We report the accuracy of predicting the correct starting point.

Model	$k = 277$	$k = 10$
CLIP	13.00%	**31.73**%
ViT	**26.47**%	30.53%
RndCLIP	11.27%	10.40%
Baseline - Random	4.00%	4.00%
Baseline - Most common	11.27%	11.27%

Table 5.3: Predicting starting points with the similarity-based approach and the baselines, percentage of correctly identified starting points for different k values.

As illustrated in Table 5.3, all setups attain scores that outperform the baseline where we predict random starting points (theoretically, for a uniform distribution of starting points, $1/739 = 0.14\%$). We also predict the most common starting point ('man'), which performs similarly to RndCLIP. With pretrained encoders, it is better to utilize lower k to attain better accuracy, since very similar images likely contain similar objects that are mentioned earlier in the utterances. Both CLIP and ViT show similar performances when $k = 10$, hinting at the relation between their training objectives and starting points, which often correspond to the most salient entity in the image.

5.5.4 Predicting the Variation in Gaze

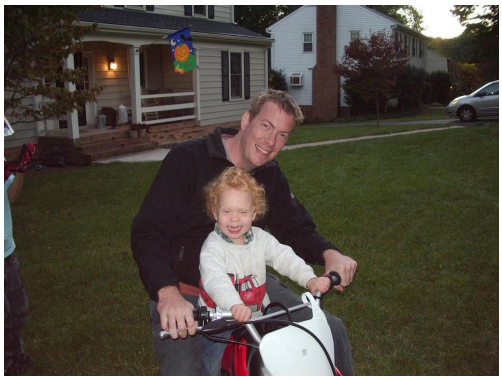
We apply the similarity-based approach to predict the variation in gaze. The results (Table 5.4) reveal that the gaze variation can be approximated to a moderate extent with CLIP. Using a smaller retrieval set is beneficial, suggesting a strong link between image properties and the variation in gaze. Since CLIP has a powerful visual encoder (Shen et al., 2022), it is reasonable that the similarities between image features encoded by CLIP seem to be more meaningful when approximating the variation in gaze.

Model	Coefficient	Sig.	Loss	Range
CLIP-277	0.4035	30	4.0200	23.55 - 24.45
CLIP-10	0.4253	35	3.5774	17.05 - 29.63
CLIP-5	0.4435	33	**3.5707**	15.43 - 32.92
CLIP-1	**0.4687**	**39**	3.8889	11.22 - 38.79
ViT	0.3801	28	3.8847	22.62 - 25.67
RndCLIP	0.0109	2	4.0571	23.76 - 24.26

Table 5.4: Predicting gaze variation using the similarity-based approach. Targets range between 11.22 and 38.79.

The outcomes are in line with our hypothesis that signals that could be considered more internal to the models' training objectives would be captured better,

whereas external signals can be captured weakly. For instance, speech onsets and surface form variation in descriptions can be deemed external to CLIP's space. Therefore, we claim that there could be room for incorporating such external signals when training or fine-tuning pretrained multimodal models, and the models would benefit from such signals. It should be noted, though, since human processes are complex, there could be extraneous factors beyond image features that influence variation, which makes it difficult for models to capture these signals perfectly.



Min: 3.381

Max: 3.488

Figure 5.7: Images with the minimum and maximum predicted mean onsets. The image with the minimum was also predicted to elicit the lowest variation in gaze.

5.5.5 Examples

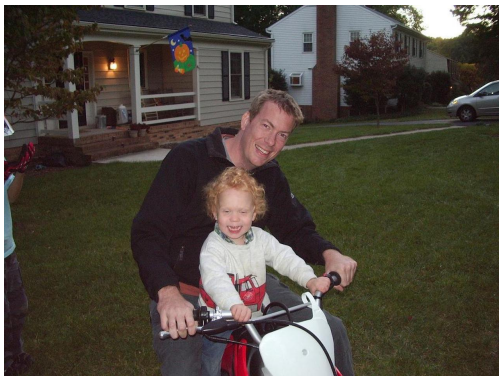
We illustrate the images with the minimum and maximum mean onsets as predicted by the similarity-based approach in Figure 5.7. Figure 5.8 depicts predicted variation in descriptions, and Figure 5.9 the predicted variation in gaze. We see a tendency to predict shorter speech onsets, more similar descriptions and gaze patterns in images containing a couple of people compared to scenes of streets with no visible or salient humans, a finding resonating with the conclusions drawn by Jas and Parikh (2015). This is potentially due to the salient and non-ambiguous nature of humans in images, as opposed to general street scenes with cars, buses and non-salient humans.

5.6 Conclusion

We quantified the variation in speech onsets, starting points, descriptions and gaze using a Dutch dataset of image descriptions with eye-tracking data. Our findings revealed the extent of variation in the process of describing images, and that variations in different signals correlate with each other. Furthermore, using

Min: 0.529 Max: 0.541

Figure 5.8: BLEU-2-based linguistic variation scores as predicted by the similarity-based approach. Lower BLEU-2 scores mean more diversity.



Min: 23.666 Max: 24.308

Figure 5.9: Variation in gaze as predicted by the similarity-based approach. Higher scores indicate more diverse gaze patterns.

a similarity-based prediction approach, we showed that image representations encoded by pretrained vision encoders capture variation in visuo-linguistic behavior to a weak-to-moderate extent. This pattern can be interpreted in light of models' pretraining objectives, as the predictions correlated more strongly for signals more internal to the objectives. Our study has implications for how human processes unfold as well as pretrained models' capabilities to represent such processes.

Human and machine processing have differences, and we are motivated by the potential benefits of making the models increasingly knowledgeable of the multimodal landscape of human data. Although the impact of fine-tuning an already powerful pretrained model on a small-scale dataset with human signals could be quite modest, we hope that our work motivates the collection of more signals during crowdsourcing. For instance, it would be beneficial to take into

account how long it took participants to complete a task given a certain stimulus, indicating the relative complexity and the uncertainty induced by the task as well as the stimulus. By inducing biases based on human signals, models can further take advantage of the information contained within such signals. Although it would be difficult to capture the full extent of the intricacies of human processing, this could help, for instance, a model interacting with human users to generate responses more aligned with human expectations.

In the next chapter, we investigate the representational power of multilingual models when predicting eye-tracking features during reading.

Chapter 6

Multi- and Cross-Lingual Prediction of Human Reading Behavior

The material in this chapter is based on a paper that received the **'Best Shared Task Paper Award'** at CMCL 2022: Ece Takmaz. 2022. Team DMG at CMCL 2022 shared task: Transformer adapters for the multi- and cross-lingual prediction of human reading behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 136–144, Dublin, Ireland. Association for Computational Linguistics.

Contributions: Ece Takmaz: Implementing and running the experiments, writing the paper.

6.1 Introduction

In the previous chapters, we delved into the relation between language production and eye movements over images, and how findings regarding this topic would help inform multimodal models in NLP. In this chapter, we turn to predicting gaze during reading, which reflects cognitive processes and attention during language comprehension (Rayner, 1977). Modeling gaze in relation to reading could provide insights into language-related eye movements, as well as revealing the potential of using computational means to model human reading behavior. This was also the aim of the shared task of CMCL 2022, focusing on the prediction of multi- and cross-lingual eye movements during reading (Hollenstein et al., 2022). Such a task would necessitate capturing universal as well as language-specific aspects of human reading behavior (Liversedge et al., 2016; Hollenstein et al., 2021b). In this chapter, we describe our approaches that attained second place in the shared task of CMCL 2022 (Hollenstein et al., 2022).

Various approaches have been proposed for the modeling of human reading behavior (Rayner, 1998; Reichle et al., 1998; Hahn and Keller, 2016). The CMCL 2021 shared task focused on the prediction of 'monolingual' reading behavior and the participants applied various methodologies to predict eye-tracking features, e.g. gradient boosting, ensembling, using handcrafted features, deep learning (Hollenstein et al., 2021a; Bestgen, 2021; Li and Rudzicz, 2021; Oh, 2021; Vickers et al., 2021).

With regard to deep learning-based approaches, there exist findings suggesting that, as compared to transformer-based models (Vaswani et al., 2017), recurrent neural networks exhibit attention patterns closer to human attention (Sood et al., 2020a). However, more recently, transformer-based models have been shown to better account for human reading behavior than recurrent neural networks (Merkx and Frank, 2021). Moreover, pretrained language models (PLM) such as BERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) can predict multilingual human reading behavior well (Hollenstein et al., 2021b), in addition to having advanced the state-of-the-art in many downstream NLP tasks.

The CMCL 2022 shared task (Hollenstein et al., 2022) consists of predicting four eye-tracking features for data containing sentences in 6 different languages as well as transferring to a new language. For this purpose, we train 'adapters' inserted into transformer layers of frozen PLMs (Houlsby et al., 2019). We find that training adapters for each language separately within multilingual transformers leads to good performance, attaining second place in the leaderboard. In addition, we show that such models can transfer to new languages via simply translating the new test sets into closely related languages (e.g. lexically or grammatically) that the model was exposed to during training.[1]

[1]The repository: `https://github.com/ecekt/cmcl2022_dmg`

6.2 Background

6.2.1 Data and Subtasks

The CMCL 2022 shared task consists of 2 subtasks. The data for Subtask 1 includes publicly available eye-tracking corpora for 6 languages (English, Chinese, Russian, Hindi, German, Dutch). These corpora differ in size as well as the nature of the sentences they contain (i.e. news articles, scientific texts, Wikipedia entries). The data is already partitioned into train, validation and test splits. For Subtask 2, we are only supplied with a test set comprised of Danish sentences. We only use the data provided in the shared task and preprocess the textual input utilizing the tokenizers of PLMs. For more details, see Appendix C.1.

The eye-tracking features provided in the data correspond to 'first fixation duration' (FFD, duration of the first fixation on the current word) and 'total reading time' (TRT, total duration of all fixations on the current word including regressions). The values of these features were provided per token entry, averaged across all the readers: **FFDAvg** and **TRTAvg**. In addition, to account for the individual differences between readers, the data also includes the standard deviations of these features across readers: **FFDStd** and **TRTStd**.

The aim of the subtasks is to predict these 4 features for each token. The submissions to the shared task system are ranked with respect to test-set Mean Absolute Error (MAE): the average of the absolute differences between the ground-truth values and the values output by the model. We implement MAE as below:

$$\frac{\sum_{i=1}^{N} |o_i - t_i|}{N} \quad [6.1]$$

where N is the number of tokens in the data, o_i is the value output by the model for a given token, and t_i is the ground-truth value for this token. We calculate MAE for all 4 eye-tracking features and take their average to obtain the final MAE. The shared task system also reports coefficients of determination (R^2), which we provide in Appendix C.4.

6.2.2 Adapters

The common method for using PLMs in downstream tasks is to fine-tune them for each task. If there are multiple tasks the model should handle at the same time, this could lead to some issues (Pfeiffer et al., 2021). For instance, learning tasks in parallel could cause interference, and the model might learn a certain task better than the others. In the case of sequential training, we might observe catastrophic forgetting, where the model forgets the previously learned tasks. In addition, usually the whole model is fine-tuned; hence, we might need to save a new model per task, which increases compute and memory requirements.

To overcome these issues, 'adapters' have been proposed (Houlsby et al., 2019; Bapna and Firat, 2019). Adapters are bottleneck layers consisting of new weights integrated into each layer of a transformer model. They first project down ($W_D \in \mathbb{R}^{h \times d}$) the dimensions of the transformer hidden state h_l at layer l, apply a non-linearity, and then project the activations back up ($W_U \in \mathbb{R}^{d \times h}$) to the original dimensions. The outcome is then summed up with the residual r_l via a skip-connection to obtain the output of the adapter A_l:

$$A_l = W_U(ReLU(W_D h_l)) + r_l \quad [6.2]$$

Keeping the pretrained model frozen and only training adapters have been shown to yield performances close to those of fully-fine-tuned models while also maintaining efficiency (Houlsby et al., 2019; Bapna and Firat, 2019; Rücklé et al., 2021). Various types of adapters, insertion and training schemes have been proposed for machine translation, multi-task settings and cross-lingual transfer (Ansell et al., 2021; Pfeiffer et al., 2020b, 2021; Philip et al., 2020; Üstün et al., 2020, 2021; Poth et al., 2021).

Given their relevant advantages, we use Adapters from the AdapterHub framework (Pfeiffer et al., 2020a)[2] built on HuggingFace Transformers (Wolf et al., 2020), to insert trainable adapters into frozen PLMs for the prediction of eye-tracking features. Then, we train language- and task-specific adapters and store their trained weights along with a single model. The details of the models and adapters used in Subtasks 1 and 2 are provided in Sections 6.3 and 6.4, respectively. For reproducibility, the hyperparameters for the best models selected with respect to their MAE scores on the validation set, and the details of the development environment are provided in Appendix C.2.

6.3 Subtask 1: Multi-lingual

In this subtask, the aim is to predict eye-tracking features for data from 6 languages, for which we have training, validation and test sets. We focus on comparing a single setup for all languages vs. separate setups for different languages.

6.3.1 Methodology

Single adapter for all languages We first train a single task-specific adapter integrated into a frozen PLM on all the languages per eye-tracking feature. We utilize the XLM-RoBERTa-base (XLM-R) model (Conneau et al., 2020), which is a multilingual version of RoBERTa (Liu et al., 2019), trained with the masked language modeling objective on 100 languages covering all of the shared task languages.[3]

[2] `https://adapterhub.ml`

[3] `https://huggingface.co/xlm-roberta-base`

We place a token-level regression head on top of XLM-R. We then train this head and the adapters to predict eye-tracking features for each contextualized token in a given sentence. Since we keep the underlying model frozen, this method only learns a small set of parameters for the eye-tracking features, which we expect would capture universal patterns in human reading behavior.

Language-specific adapters When a single model is trained on multiple languages, its capacity for certain languages might decrease, which is called 'the curse of multilinguality' (Conneau et al., 2020; Pfeiffer et al., 2020b). To avoid this issue, we increase the language-specific capacity by training adapters separately for each language.

In this approach, we train a single adapter that is specific to a language-task pair (yielding $6 * 4 = 24$ adapters) integrated into frozen XLM-R. In addition, we also implement another setup where we *stack* language- and task-specific adapters on top of each other (Pfeiffer et al., 2020b). In the latter setup, per language, we utilize a frozen language-specific adapter that was trained on Wikipedia articles with the masked language modeling objective, as provided on AdapterHub (Pfeiffer et al., 2020b, 2021).[4] We train the new task-specific adapter and the token regression head to predict eye-tracking features specific to each language. For Dutch, AdapterHub did not have a language adapter trained on Wikipedia; therefore, we only use a single new adapter.[5]

PLM tokenizers produce multiple wordpieces for some tokens. For such tokens, the models output predictions for each wordpiece. During training and validation, we calculate the MAE loss taking into account every wordpiece, where each wordpiece of the same token is assigned the same target value. For the test set predictions, we calculate the average output of the wordpieces, and assign it as the prediction for the whole token entry. To explore whether the way the wordpieces are treated has an effect on accuracy, we also train and test the stacked setup only keeping the first wordpiece to represent the full token entry.

6.3.2 Results

In the top half of Table 6.1, we present the results for Subtask 1. Overall, our models outperform the mean baseline,[6] and seem to predict FFD features better

[4] `https://adapterhub.ml/explore/text_lang/` The names of the language-specific adapters are '{x}/wiki@ukp', where {x} is to be replaced by the abbreviation corresponding to the language, e.g. 'en/wiki@ukp' for English.

[5] We also experiment with training two new adapters stacked together for Dutch to make the setups more comparable. See Appendix C.3 for the outcomes of additional models including the use of RoBERTa and XLM-RoBERTa-large.

[6] For every word in the test set, the mean of each eye-tracking feature over the training set is predicted. Hollenstein et al. (2022) also report a stronger baseline, where the mean is calculated based on the language subsets of the data. Our models also outperform this baseline.

than TRT features. XLM-R with new adapters trained from scratch on all languages together performs the worst. XLM-R with new language-specific adapters further improves the results, in particular decreasing the MAE of features corresponding to averages.

The XLM-R setup that stacks adapters per language yields our best results for Subtask 1, achieving second place in the leaderboard of the shared task (MAE = 3.6533, our second submission to the system). The breakdown of results per language is provided in Table C.2 in Appendix C.3. It can be observed from this table that the model performs well for languages such as German and Dutch, yet struggles with languages such as Chinese and Russian, which could be due to the differences in their typologies, the nature of the corpora, vocabulary size and the issues that might have been caused by the multilinguality of the underlying PLM.

Finally, utilizing only the first wordpieces seems to degrade the performance across the features (MAE = 3.7261, our third submission). This finding indicates that retaining all wordpieces provides a better picture of the value to be predicted, as each wordpiece might contribute to the processing of the full token, affecting fixation duration times.

Model setup	**FFDAvg**	**FFDStd**	**TRTAvg**	**TRTStd**	**MAE**
All languages together	3.1449	1.9697	6.4339	4.6253	4.0434
Language-specific	2.8563	1.9741	5.5682	4.6956	3.7736
Language-specific-stack	2.6086	1.9219	5.6542	4.4284	**3.6533**
First wordpiece-only	2.6876	1.9609	5.7059	4.5501	3.7261
Zero-shot	3.4955	2.7370	7.1336	7.1502	5.1291
Translate train	14.6278	4.4001	19.8624	14.2824	13.2932
Translate test - EN	13.7903	5.1338	20.9214	13.5084	13.3385
Translate test - EN (without Provo)	4.5843	3.9382	9.3022	6.8426	6.1668
Translate test - DE	5.4512	1.7349	6.9036	5.7730	**4.9657**
Mean baseline	5.6858	2.5395	8.8200	5.8877	5.7332

Table 6.1: Test set results for Subtask 1 and Subtask 2. The best models per subtask are indicated in bold.

6.4 Subtask 2: Cross-lingual

For this subtask, we conduct various experiments to obtain results for the Danish test set in the absence of training and validation data in this language.

6.4.1 Methodology

Zero-shot We first feed the Danish test set directly into the XLM-R model with the trained all-languages adapters for each eye-tracking feature. Since the

adapters in this case are expected to have learned universal eye movement features, and XLM-R includes Danish in its training, we expect this model to transfer well to Danish without being exposed to eye-tracking data in this language.

Translate train In this approach, we translate the training and validation set from their source language into the target language to be used in the training of a new model (Conneau et al., 2018b). We have chosen English as the source language, as it constitutes almost half of the whole shared task data and XLM-R performs well in English (Conneau et al., 2020). We translate the English training and validation data word-by-word[7] into Danish using the MarianMT en-da model.[8] Since, at the time of this project, AdapterHub did not host a language-specific adapter for Danish, we did not implement stacking, and only trained task-specific adapters for Danish.

Translate test In this setup, we translate the test set into a language for which we have training and validation data (Conneau et al., 2018b) using MarianMT models. We first translate the Danish test set into English word-by-word. Using the best English model we obtained in Subtask 1, we generate predictions for the translated test set. In addition, we notice that the Provo corpus (Luke and Christianson, 2018) in the English subset has rather higher values for the features as compared to the other English corpora existing in the data. As a result, we retrain the best English setup using the same hyperparameters and skipping the Provo data.

In our final setup for Subtask 2, we translate Danish into German, and utilize the best German model from Subtask 1 to obtain predictions. The main reason for opting for German was to better account for the effects of word order, e.g. inversions in main and subordinate clauses, exploiting mainly the syntactic similarities between Danish and German.

6.4.2 Results

The bottom half of Table 6.1 provides the results for Subtask 2. First of all, the translate train approach does not seem to be a viable option, as its accuracy is much lower than the mean baseline (MAE = 13.2932, our first submission). Using the translate test approach in English yields very similar results. However, as we hypothesized, removing the Provo corpus from the training improves the translate test performance substantially (MAE = 6.1668, our second submission), albeit still underperforming. The zero-shot setup, on the other hand, yields a MAE

[7]Sentence-by-sentence translation could yield more reliable outcomes; however, it may cause issues in word order and count: source and translated text would need to be aligned to keep the eye-tracking data intact.

[8]`https://huggingface.co/docs/transformers/model_doc/marian`

score better than the mean baseline, suggesting that our adapters learn universal eye-tracking feature across languages combined with the multilingual pretraining of XLM-R.

Finally, the translate test setup in German yields our best results for this subtask achieving second place in the leaderboard (MAE = 4.9657, our third submission). These results indicate that the selection of source language and data has an effect on the results. Furthermore, it can be claimed that translate test is a viable option for adapters integrated into PLMs for achieving good transfer to a test set in a new language, without being exposed to actual eye-tracking data in this language.

6.5 Conclusion

We have trained language- and task-specific adapters for the prediction of eye-tracking features reflecting human reading behavior in multi- and cross-lingual settings. Our best models performed well, attaining second place in the CMCL 2022 leaderboard. This suggests that pretrained language models enhanced with small adapter layers possess the capability to predict eye-tracking features.

In addition to our setups, other methods such as dropping adapters or adapter fusion could be implemented (Rücklé et al., 2021; Pfeiffer et al., 2021). It would also be informative to consider autoregressive models, and the possibility of making use of various lexical and syntactic features and additional cognitive signals. The prediction of each eye-tracking feature could also be informed by other eye-tracking features, as each of them represents different aspects of human reading behavior. Similar approaches could also be of help in the modeling of other human cognitive signals, opening up novel ways of predicting and inspecting cognitive processes in humans.

Part Two

Communication Strategies in Referential Tasks - Vision and Language in Dialogue

Chapter 7

Overview

In Part One, I have shown the importance of accounting for human gaze when modelling language production and comprehension. Part Two explores how to model communication strategies in referential tasks in visually grounded dialogue. To this end, I build models quantifying, generating, resolving, and adapting utterances in referential tasks within visual and conversational contexts.

The dataset that I mainly use in this part of the thesis is PhotoBook (Haber et al., 2019). This dataset is a collection of task-oriented visually grounded English dialogues between pairs of participants who communicate via written chat. See Figure 7.1 for a sample dialogue from PhotoBook. The PhotoBook task is a game comprising 5 rounds where two participants see their own private sets ('photobooks') of 6 real-life images belonging to the same visual domain. The participants interact freely using a chat interface with the aim of picking the images they have in common without seeing each other's visual contexts. In the later rounds of the game, previously seen images can reappear in different visual contexts. This allows the players to refer to such images again, facilitating the production of subsequent references to the same images. This feature of the PhotoBook dataset makes it a valuable resource for modeling the incremental development of conversational common ground between interlocutors. Within this setup, I investigate the following research questions:

- **Can pretrained multimodal models help quantify referring utterances to reveal patterns resonating with human strategies? Chapter 8** quantifies referring utterances with pretrained multimodal models to reveal human strategies in visually grounded referring utterance generation. I find that CLIP (Radford et al., 2021), a pretrained multimodal model, can capture the strategies deployed by speakers when referring to images multiple times in isolation, and also in the context of similar images, allowing us to measure the *descriptiveness* and *discriminativeness* of the utterances.

- **How can we model referring utterance generation and resolution**

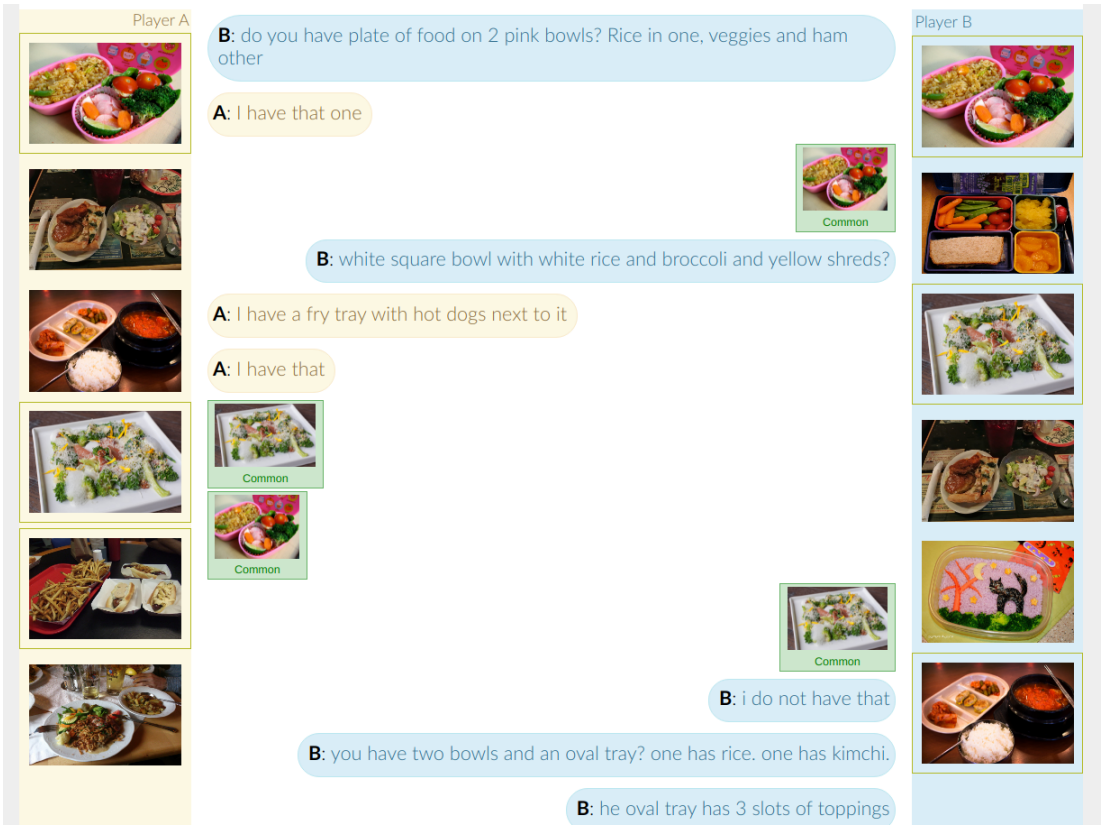


Figure 7.1: Sample dialogue from a PhotoBook game. Each player only sees their own set of 6 images in addition to the dialogue.

with deep neural networks? What would be the impact of incorporating previous utterances in such models? In **Chapter 9**, I propose models of referring utterance generation and resolution, and explore the influence of visual and conversational contexts in generation and resolution. The results show that when models exploit previous references, the outcomes exhibit human-like patterns and improve task performance.

- **How can we adapt pretrained referring utterance generation models to produce outputs addressing listeners with different knowledge backgrounds? Chapter 10** details how referring utterance generation models can be modulated to generate utterances better understood by various types of listeners in visually grounded referential games. I propose an adaptation mechanism for the generation model so that it can adapt its outputs on the fly according to its understanding of the listeners to increase communicative success.

Chapter 8

Quantifying the Properties of Multimodal Referring Utterances

The material in this chapter is based on: Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2022. Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–42, Dublin, Ireland. Association for Computational Linguistics.

Contributions: Ece Takmaz: Implementing and running the experiments, writing and revising the paper. Sandro Pezzelle: Supervising the research, writing and revising the paper. Raquel Fernández: Supervising the research, writing and revising the paper.

Figure 8.1: Referring utterance chain from PhotoBook (Haber et al., 2019). The chain has 4 ranks (4 references to the target image, in red outline). For simplicity, only the 5 distractor images from rank 1 are shown.

8.1 Introduction

Speakers can refer to an entity multiple times during a conversation (e.g., the girl in Figure 8.1). This leads to further expressions based on previous mentions that are more compact and less descriptive (Krauss and Weinheimer, 1967; Brennan and Clark, 1996), yet still remain pragmatically informative so that the participants are able to identify the intended referent (Shore and Skantze, 2018; Haber et al., 2019; Hawkins et al., 2020).

Several approaches (Mao et al., 2016; Cohn-Gordon et al., 2018; Schüz et al., 2021; Luo et al., 2018, i.a.) have tackled the generation of image captions from the perspective of pragmatic informativity; Coppock et al. (2020) have compared the informativity of image captions and of referring expressions; and Haber et al. (2019); Hawkins et al. (2020) have explored how dialogue history contributes to discriminativeness. However, no work to date has investigated how these two dimensions, *descriptiveness* and *discriminativeness* or pragmatic informativity, interact in referring expressions uttered in dialogue.

In this chapter, we use a transformer-based pretrained multimodal model to study the interplay between descriptiveness and discriminativeness in human referring utterances produced in dialogue. Due to their unprecedented success in numerous tasks, pretrained V&L models—such as LXMERT (Tan and Bansal, 2019), VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020) and ALIGN (Jia et al., 2021)—have attracted a lot of interest aimed at understanding the properties and potential of their learned representations as well as the effect their architectures and training setups have (Bugliarello et al., 2021; Hendricks et al., 2021). These include probing such models in a zero-shot manner, i.e., without any specific fine-tuning (Hendricks and Nematzadeh, 2021; Parcalabescu et al., 2021); quantifying the roles of each modality (Frank et al., 2021); inspecting

attention patterns (Cao et al., 2020); and evaluating their learned multimodal representations against human judgments (Pezzelle et al., 2021).

We focus on CLIP as our model of choice (Radford et al., 2021), which learns via contrasting images and texts that can be aligned or unaligned with each other. This contrastive objective makes CLIP particularly suitable for modeling referential tasks that inherently include such comparisons. Here, we use CLIP to gain insight into the strategies used by humans in sequential reference settings, finding that although the descriptiveness of referring utterances decreases significantly, the utterances remain discriminative over the course of multimodal dialogue.[1].

8.2 Data

The data we use in this chapter originates from PhotoBook (Haber et al., 2019), which consists of 2,500 games, 165K utterances, and 360 unique images from MS COCO (Lin et al., 2014).[2] We conduct the experiments in this chapter on a subset of 50 PhotoBook games with manually annotated referring utterances, which contains 364 referential chains about 205 unique target images. We refer to this subset as PB-GOLD.[3] Although a dataset of automatically-extracted chains using all PhotoBook data is also made available (Takmaz et al., 2020a), these chains may contain errors (see Chapter 9). We therefore opt for using the smaller but higher-quality PB-GOLD subset since we are interested in analyzing human strategies, and this option helps prevent any issues that might be caused by the automatic annotation of chains. Given that we use a pretrained model without fine-tuning, experimenting with large amounts of data is not a requisite.

PB-GOLD's chains contain 1,078 utterances, i.e., 2.96 utterances per chain on average (min 1, max 4). We henceforth use the term 'rank' to refer to the position of an utterance in a chain. The average length of utterances in terms of tokens is 13.34, 11.03, 9.23, and 7.82, respectively, for ranks 1, 2, 3, and 4.[4] This decreasing trend, which is statistically significant at $p < 0.01$ with respect to independent samples t-tests between the ranks, is in line with the trend observed in the whole dataset (Haber et al., 2019). PB-GOLD's vocabulary consists of 926 tokens.

8.3 Model

We use CLIP (Radford et al., 2021), a model pretrained on a dataset of 400 million image-text pairs collected from the internet using a contrastive objective to learn

[1]The code to reproduce our results is available at `https://github.com/ecekt/clip-desc-disc`

[2]The PhotoBook dataset is available at `https://dmg-photobook.github.io`.

[3]We use the gold set of the utterance-based chains version 2.

[4]To tokenize, we use TweetTokenizer: `https://www.nltk.org/api/nltk.tokenize.html`

strong transferable vision representations with natural language supervision.[5] In particular, we employ the ViT-B/32 version of CLIP, which utilizes separate transformers to encode vision and language (Vaswani et al., 2017; Dosovitskiy et al., 2021; Radford et al., 2019, 2021).

As the model learns to align images and texts, this enables zero-shot transfer to various V&L tasks such as image-text retrieval and image classification, and even certain non-traditional tasks in a simple and efficient manner (Radford et al., 2019; Agarwal et al., 2021; Shen et al., 2022; Cafagna et al., 2021; Hessel et al., 2021). This makes it an intriguing tool to investigate the properties of visually grounded referring utterances. CLIP has been used in frozen form to build recent multimodal models (Berrios et al., 2023; Alayrac et al., 2022; Mañas et al., 2023; Tsimpoukelli et al., 2021; Li et al., 2023b; Mokady et al., 2021; Chen et al., 2022). In this work, we also freeze CLIP's weights and do not fine-tune the model or perform prompt engineering, since we aim to exploit the model's pretrained knowledge for the analysis of human referring strategies.

8.4 Descriptiveness

In our first experiment, we investigate the degree of descriptiveness exhibited by referring utterances in the PhotoBook game, i.e., the amount of information they provide about the image out of context. We consider each target image and corresponding referential utterance at a given rank *in isolation*, i.e., without taking into account the other competing images nor the dialogue history. We quantify descriptiveness as the alignment between an utterance and its image referent using `CLIPScore` (Hessel et al., 2021), assuming that a more descriptive utterance will attain a higher score. For all the target image-utterance pairs in the chains of PB-GOLD, we use CLIP to obtain a vector t representing the utterance and a vector v representing the image. `CLIPScore` is then computed as the scaled cosine similarity between these two vectors, with range $[0, 2.5]$:[6]

$$\texttt{CLIPScore}(t, v) = 2.5 * max(cos(t, v), 0) \quad [8.1]$$

We compute the average `CLIPScore` per rank over the whole PB-GOLD dataset.

Results We find that earlier utterances are better aligned with the target image features and that there is a monotonically decreasing trend over the 4 ranks (Figure 8.4, blue bars). The differences between all pairs of ranks are statistically significant (according to independent samples t-tests, $p < 0.01$), except for the comparison between the last 2 ranks ($p > 0.05$). Since earlier referring utterances

[5] `https://github.com/openai/CLIP`

[6] The scaled factor was introduced by Hessel et al. (2021) to account for the relatively low observed cosine values.

1. *girl lying on a bed surfing the internet on a laptop computer*
2. *a girl sleeping on her belly on top of a bed looking at a laptop.*
3. *woman laying on her stomach on a bed in front of a laptop.*
4. *a girl with long brown hair with streaks of red lays on a bed and looks at an open laptop computer.*
5. *a young girl laying on a bed using her laptop.*

Figure 8.2: Set of captions from COCO (Lin et al., 2014), where the order of captions is arbitrary as they are provided separately by different annotators.

tend to be longer (see Section 8.2), we check to what extent length may be a confounding factor. We find that there is only a weak correlation between length in tokens and `CLIPScore` (Spearman's $\rho = 0.29, p < 0.001$).

We compare these results on PhotoBook with text-to-image alignment computed with the same method on two other datasets: (1) COCO (Lin et al., 2014),[7] which includes 5 captions per image provided independently by different annotators as shown in Figure 8.2; here we do not expect to find significant differences in the level of descriptiveness across the captions, and (2) Image Description Sequences (IDS, Ilinykh et al., 2019)[8] where one participant describes an image incrementally as shown in Figure 8.3, by progressively adding sentences with further details; here, we do expect a pattern similar to the pattern found in PhotoBook, albeit for different reasons (because participants add less salient information in later sentences; Ilinykh et al., 2019).

Figure 8.4 shows that these expectations are confirmed. Based on CLIPScore values, COCO captions (green bars) are more descriptive than IDS descriptions and PhotoBook referring utterances, and are equally aligned with the image across 'ranks' (the order is arbitrary in COCO). In contrast, IDS incremental descriptions (yellow bars) are intrinsically ordered and show a significant decreasing trend similar to PhotoBook.

[7] We use the set of COCO images in PB-GOLD (N=205).
[8] The images are from ADE20k corpus (Zhou et al., 2017)

1. This is a kitchen with white cupboards and worktop.
2. There is a red wall painted behind the cooker section.
3. There is a wooden table to the right with pale floor tiles on the floor.
4. There is a sink to the left with a window near the sink.
5. There is a bunch of flowers beside the window.

Figure 8.3: Sequential description from Image Description Sequences (Ilinykh et al., 2019).

8.5 Discriminativeness

In order for a listener to select the target image among distractor images, a referring utterance should be discriminative in its visual context. Our results in the previous section show that descriptiveness decreases over time—what is the trend regarding discriminativeness? To address this question, in our second experiment we use CLIP from the perspective of reference resolution.

We focus on local text-to-image alignment, initially ignoring the previous dialogue history. To this end, we feed CLIP a single referring utterance together with the visual context of the speaker who produced that utterance. CLIP yields softmax probabilities for each image contrasted with the single text. As a metric, we use accuracy: 1 if the target image gets the highest probability; 0 otherwise.

Results The overall accuracy is 80.15%, which is well above the random baseline of 16.67%. In Figure 8.5, we break down the results per rank (blue bars). A 4×2 chi-square test (4 ranks vs. correct/incorrect) did not yield significant differences in accuracy between the ranks, $p > 0.05$. Thus, although descriptiveness decreases over time, discriminativeness is not significantly affected. An analysis of the entropy of the softmax distributions over the visual context reveals that entropy increases monotonically over the ranks (this difference is statistically significant according to an independent samples t-test between ranks 1 and 4; $H_1 = 0.62$, $H_4 = 0.79$, $p < 0.01$). That is, the model is more uncertain when trying to resolve less descriptive utterances. There is indeed a negative correla-

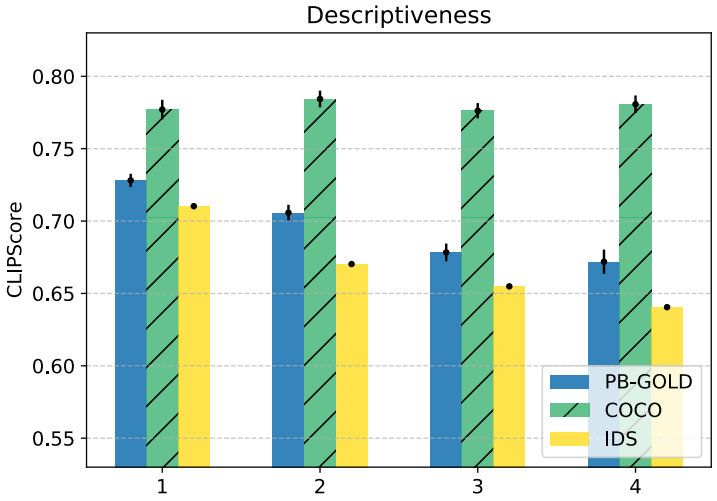


Figure 8.4: Descriptiveness (`CLIPScore`) for PB-GOLD, COCO and IDS. We only plot the first 4 'ranks' (x-axis) for COCO and IDS for comparability with PB-GOLD. The error bars illustrate the standard error.

tion between entropy and `CLIPScore` computed between the target image and the corresponding utterance (Spearman's $\rho = -0.5, p < 0.001$).

8.6 Analysis

The findings indicate that the players in PhotoBook manage to maintain discriminativeness while decreasing descriptiveness. We investigate potential factors contributing to this strategy. Do the players rely on previous mentions in the dialogue history? Do they refine their referring utterances by distilling the most discriminative information in a given context?

8.6.1 Dialogue History

The results of our experiments in the previous section show that utterances in isolation are effective at referring; yet, uncertainty increases when less descriptive utterances are considered out of context. To reduce such uncertainty, participants may rely on the dialogue history (Brennan and Clark, 1996; Shore and Skantze, 2018; Takmaz et al., 2020a). We consider a scenario where participants keep in memory the previous mention when processing the current referring utterance. We model this scenario by prepending the previous referring utterance in the chain to the current utterance and feeding this into the reference resolution model described in Section 8.5. As shown in Figure 8.5, the resulting discriminativeness is similar to the one obtained earlier (the differences are not significant; chi-square test, $p < 0.05$) and, as before, remains stable across ranks (chi-square test, $p >$

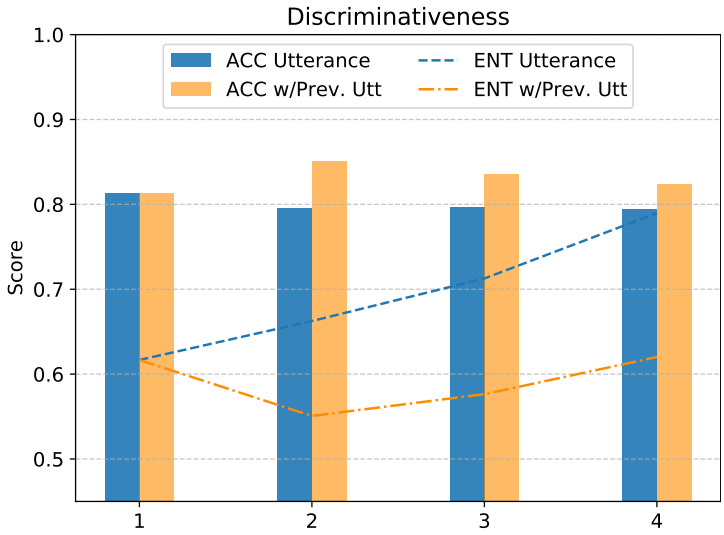


Figure 8.5: Discriminativeness (reference resolution accuracy, ACC) per rank with PB-GOLD utterances (Utterance) and utterances with history (w/Prev. Utt), along with their respective entropies (ENT).

0.05). However, taking into account the previous mentions leads to a significant reduction of the entropy in general: e.g., at the last rank $H_4 = 0.79$ vs. $H'_4 = 0.62$ (t-test, $p < 0.05$). This suggests that relying on the dialogue history allows speakers to use less descriptive utterances by reducing discriminative uncertainty.

8.6.2 Most Discriminative Information

Besides exploiting the dialogue history, participants may refine their referring strategy by distilling the most discriminative information in a given context. To gain insight into this hypothesis, we explore what is discriminative in the images: we compute the discriminative features v_d of a target image by taking the average of the visual representations of distractor images to obtain the mean context vector and then subtracting this vector from the visual representation of the target image. We encode all 926 words in the vocabulary of PB-GOLD using CLIP, and retrieve the top-10 words whose representations are the closest to v_d in terms of cosine similarity (amounting to 1% of the vocabulary). We take these words to convey the most discriminative properties of the target image in the provided visual context. We analyze whether at least one of these retrieved words is mentioned exactly in the corresponding referring utterance, finding that this is indeed the case for a remarkable 60% of utterances.[9] As an illustration, for the example in Figure 8.1, the words *walking* (mentioned at rank 1) and *blue* (used at ranks 1, 2, 3, 4) are among the top-10 most discriminative words, while

[9]Randomly sampling 10 words from the vocabulary for each utterance yields 11% (average of 5 random runs).

the word *water* (mentioned at ranks 1, 2, 3, 4) is close to the word *beach*, which is also retrieved as one of most discriminative words in this case.

The most discriminative words are likely to be reused in later utterances, even though the visual context changes from rank to rank. For instance, the most discriminative words mentioned at rank 1 constitute 60% of the discriminative words at rank 2, indicating that entrainment is likely for words that have high utility across contexts. We also find a significant increase in the proportion of discriminative content words to all the content words per utterance (only between ranks 1 and 4, 14% vs. 19%, $p < 0.01$).

8.7 Conclusion

We used a pretrained multimodal model claimed to be a reference-free caption evaluator, CLIP (Radford et al., 2021), to quantify descriptiveness and discriminativeness of human referring utterances within multimodal dialogues. We showed that (i) later utterances in a dialogue become less descriptive in isolation while (ii) remaining similarly discriminative against a visual context.

We found that the addition of dialogue history helps decrease and control the entropy of resolution accuracy even when the speakers produce less descriptive referring utterances. In addition, we found that the proportion of discriminative words increases over the ranks. These findings suggest that participants playing the PhotoBook game (Haber et al., 2019) show a tendency towards distilling discriminative words and utilizing the dialogue history to keep task performance stable over the dialogue. This outcome resonates with the findings by Giulianelli et al. (2021) who observe that PhotoBook dialogue participants tend to limit fluctuations in the amount of information transmitted within reference chains, in line with uniform information density principles (e.g., Genzel and Charniak, 2002; Jaeger and Levy, 2007).

In the next chapter, we develop models of referring utterance generation and resolution by taking into account the visual context and the recent dialogue history, following human strategies to obtain human-like outcomes.

Chapter 9

Generating Subsequent References in Visual and Conversational Contexts

The material in this chapter is based on: Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Contributions: Ece Takmaz: Implementing and running the experiments, writing and revising the paper. Mario Giulianelli: Creating the dataset of chains, conducting the analysis, writing and revising the paper. Arabella Sinclair: Conducting the analysis, writing and revising the paper. Sandro Pezzelle: Conducting the analysis, supervising the research, writing and revising the paper. Raquel Fernández: Supervising the research, writing and revising the paper.

Referring utterances extracted from dialogue 1

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

Referring utterances extracted from dialogue 2

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

Figure 9.1: Two chains of referring utterances for the same target image from two PhotoBook games with different participants, including the first description of the image in that dialogue and two subsequent references (⇝). In the game, each participant sees 5 additional images besides the target shown here. The distractor images change at every round of the game, i.e., each co-referring utterance within a dialogue is produced in a different visual context.

9.1 Introduction

In the previous chapter, we have shown that when speakers refer to the same entities more than once in a conversation, they may distill and reuse discriminative information, and depend on the dialogue history to maintain communicative success. The accumulated shared knowledge between the interlocutors affects the production of such subsequent references (McDonald, 1978). For example, dialogue participants may first mention *"a white fuzzy dog with a wine glass up to his face"* and later refer to it as *"the wine glass dog"*, as shown in Figure 9.1, dialogue 1. While *"the wine glass dog"* may be odd as a standalone description, it is an appropriate referring expression in the above conversational context, as the speakers established a 'conceptual pact' regarding that expression (Garrod and Anderson, 1987; Brennan and Clark, 1996). Yet, uttering it in a different context (such as dialogue 2 in Figure 9.1, after the participants had successfully referred to the image as *"the dog on the red chair"*) may lead to communication problems as it disrupts the cohesion of the dialogue (Metzing and Brennan, 2003).

In this chapter, we tackle the generation of referring utterances—i.e., utterances that contain referring descriptions, as in Figure 9.1—grounded both in the visual environment and the dialogue context. These utterances have several inter-

esting properties that make their automatic generation challenging. First, they are produced with the communicative goal of helping the addressee identify the intended referent. Second, because humans operate under cognitive and time-bound constraints, dialogue participants will aim to fulfill this communicative goal while optimizing the use of their limited cognitive resources. This results in two common features of subsequent mentions: (1) *Reduction:* Utterances tend to become shorter—a well-attested phenomenon since the work of Krauss and Weinheimer (1967)—as a result of interlocutors' reliance on their common ground (Stalnaker, 2002): As more shared information is accumulated, some information becomes predictable and can be left implicit (Grice, 1975; Clark and Wilkes-Gibbs, 1986; Clark and Brennan, 1991; Clark, 1996). Sentence compression also takes place in discourse, as predicted by the entropy rate principle (Genzel and Charniak, 2002; Keller, 2004). (2) *Lexical entrainment:* Speakers tend to reuse words that were effective in previous mentions (Garrod and Anderson, 1987; Brennan and Clark, 1996), possibly due to priming effects (Pickering and Garrod, 2004). Thus, besides being a challenging problem intriguing from a linguistic and psycholinguistic point of view, computationally modeling the generation of subsequent references can contribute to better user adaptation in dialogue systems and to more natural human-computer interaction.

Since the PhotoBook dataset was developed to elicit subsequent references to the same images within task-oriented dialogue, it allows us to address our questions regarding the generation of referring utterances within conversational and visual contexts (Haber et al., 2019). To isolate the issues we are interested in, we base our research on 'chains', which was also the case in Chapter 8. In this chapter, we automatically extract chains referring to a given image from each dialogue, resulting in a dataset of dialogue-specific chains of co-referring utterances. For example, Figure 9.1 shows two chains of co-referring utterances from two different dialogues, both referring to the same image. Figure 9.2 shows another example. We then formulate the problem as the generation of the next utterance in a chain given the current visual context and the common ground established in previous co-referring utterances (whenever these are available). To computationally model this problem, we propose three variants of a generation system based on the encoder-decoder architecture (Sutskever et al., 2014). We evaluate their outputs with metrics commonly used in the domain of NLG and with several linguistic measures. In addition, to assess the communicative effectiveness of the generated references, we implement a reference resolution agent playing the role of the addressee.[1]

We find that conditioning the generation of referring utterances on previous mentions leads to better, more effective descriptions than those generated by a model that does not exploit the conversational history. Furthermore, our quantitative and qualitative analyses show that the context-aware model generates

[1]Our data, code, and models are available at `https://dmg-photobook.github.io`.

subsequent references that exhibit linguistic patterns akin to those of humans, regarding markers of new vs. given information, reduction, and lexical entrainment, including novel noun-noun compounds.

9.2 Related Work

Generation of distinguishing expressions Our work is related to Referring Expression Generation (REG), a task with a long tradition in computational linguistics that consists of generating a description that distinguishes a target from a set of distractors—Krahmer and van Deemter (2012) provide an overview of early approaches. Follow-up approaches focused on more data-driven algorithms exploiting datasets of simple visual scenes annotated with symbolic attributes (e.g., Mitchell et al., 2013a,b, among others). More recently, the release of large-scale datasets with real images (Kazemzadeh et al., 2014) has made it possible to test deep multimodal models on REG, sometimes in combination with referring expression comprehension (Mao et al., 2016; Yu et al., 2017a). While REG typically focuses on describing objects within a scene, a few approaches at the intersection of REG and image captioning (Bernardi et al., 2016) have aimed to generate discriminative descriptions of full images, i.e., image captions that can distinguish the target image from a pool of related ones (Andreas and Klein, 2016; Vedantam et al., 2017; Cohn-Gordon et al., 2018). Similarly to these approaches, in this chapter, we generate utterances that refer to a full image with the aim of distinguishing it from other distractor images. In addition, our setup has several novel aspects: the referring utterances are the result of interactive dialogue between two participants and include subsequent references.

Generation of subsequent references Follow-up work within the REG tradition has extended the early algorithms to deal with subsequent references (Gupta and Stent, 2005; Jordan and Walker, 2005; Stoia et al., 2006; Viethen et al., 2011). These approaches focus on content selection (i.e., on generating a list of attribute types such as `color` or `kind` using an annotated corpus) or on choosing the type of reference (definite or indefinite noun phrase, pronoun, etc.) and do not directly exploit visual representations. In contrast, we generate the surface realization of first and subsequent referring utterances end-to-end, grounding them in continuous visual features of real images.

Our work is also related to a recent line of research on reference *resolution* in visually grounded dialogue, where previous mentions have been shown to be useful (Shore and Skantze, 2018; Haber et al., 2019; Roy et al., 2019). Here, we focus on *generation*. To our knowledge, this is the first attempt at generating visually grounded referring utterances taking into account earlier mentions in the dialogue. Some work on generation has exploited dialogue history in order to make lexical choice decisions that align with what was said before (Brockmann et al.,

2005; Buschmeier et al., 2009; Stoyanchev and Stent, 2009; Lopes et al., 2015; Hu et al., 2016; Dušek and Jurčíček, 2016). Indeed, incorporating entrainment in dialogue systems leads to an increase in the perceived naturalness of the system responses and to higher task success (Lopes et al., 2015; Hu et al., 2016). As we shall see, our generation model exhibits some lexical entrainment.

Dialogue history in visual dialogue Recent work in the domain of visually grounded dialogue has exploited dialogue history in encoder-decoder models trained on large datasets of question-answering dialogues (Das et al., 2017; de Vries et al., 2017; Chattopadhyay et al., 2017). Recently, Agarwal et al. (2020) showed that only 10% of the questions in the VisDial dataset (Das et al., 2017) genuinely require dialogue history in order to be answered correctly, which is in line with other shortcomings highlighted by Massiceti et al. (2018). More generally, visually grounded dialogue datasets made up of sequences of questions and answers lack many of the collaborative aspects that are found in natural dialogue. In the PhotoBook dataset, however, dialogues are less restricted, and the common ground accumulated over the dialogue history naturally plays an important role Haber et al. (2019).

9.3 Data

DIALOGUE FRAGMENT AND IMAGES VISIBLE TO PARTICIPANT A IN THE FIRST ROUND OF A GAME

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_4>`

RESULTING REFERRING UTTERANCE CHAIN WITH SUBSEQUENT REFERENCES EXTRACTED FROM THE FOLLOWING GAME ROUNDS

1. I see a guy taking a picture. What about you?
2. guy with camera
3. I have the guy with camera
4. The last one is the camera guy.

Figure 9.2: Example from our new dataset of referring utterance chains. Given a target image selected by a participant (here `<common img_4>`), the utterances in the dialogue prior to that selection action are scored by their likelihood of referring to the target. In this example, the utterance in bold is selected as the first description. To construct the reference chain, subsequent references are extracted in a similar manner from the dialogue in the following game rounds. The set of distractor images available to a participant changes across rounds.

Although the dialogues in PhotoBook include different types of dialogue acts that may provide useful information, we abstract away from this aspect and concentrate on referring utterances, as in the previous chapter.[2] To create the data for our generation task, we extract utterances that contain an image description and their corresponding image target from the dialogues as follows. Within a game round, we consider all the utterances up to the point where a given image i has been identified by the participants[3] as candidate referring utterances for i – see Figure 9.2. We then compare each candidate against a reference set of descriptions made up of the MS COCO (Lin et al., 2014) captions for i (since all images in PhotoBook originate from MS COCO), and the attributes and relationship tokens of i in the Visual Genome dataset (Krishna et al., 2017). We score each candidate utterance with the sum of its BERTScore[4] (Zhang et al., 2020b) for captions and its METEOR score (Banerjee and Lavie, 2005) for attributes and relationships. The top-scoring utterance in the game round is selected as a referring utterance for i and used as an additional caption for extracting subsequent references in the following game rounds. As a result of this procedure, for a given dialogue and an image i, we obtain a reference chain made up of the referring utterances—maximum one per round—that refer to i in the dialogue. Since images do not always reappear in each round, chains can have different lengths. Two examples of chains of length 3 are shown in Figure 9.1 and a chain of length 4 in Figure 9.2. Given that each utterance in a chain belongs to a different game round, each utterance was produced in a slightly different visual context with different distractor images. Figure 9.2 shows the visual context available to participant A in the first round of a game, when the participant produced the first description in the dialogue for target image number 4. The other three descriptions in the chain were produced while seeing different distractors.

We evaluate the referring utterance extraction procedure and the resulting chains using 20 dialogues hand-annotated by Haber et al. (2019) with labels linking utterances to the target image they describe, part of PB-GOLD from Chapter 8. Using our best setup, we obtain a precision of 0.86 and a recall of 0.61. The extracted chains are very similar to the human-annotated ones in terms of chain and utterance length.

Our new dataset is made up of 41,340 referring utterances and 16,525 chains (i.e., there are 16,525 first descriptions and 24,815 subsequent references). The median number of utterances in a chain is 3. We use the splits defined by Haber et al. (2019) to divide the dataset into Train, Validation, and Test, and all hand-annotated dialogues are excluded from these splits. Table 9.1 reports relevant

[2]In a similar vein, Haber et al. (2019) extracted coreference chains made up of multi-utterance dialogue excerpts. However, the chains in this thesis contain single utterances, which are more suitable for generation.

[3]Image identification actions are part of the metadata.

[4]BERTScore uses contextualized embeddings (Devlin et al., 2019) to assess similarity between a target sentence and one or more reference sentences.

Split	Games	*First*		*Later*	
		N	**Length**	**N**	**Length**
Train	1725	11540	10.52 (4.80)	17393	7.52 (4.15)
Val	373	2503	10.49 (4.81)	3749	7.70 (4.22)
Test	368	2482	10.52 (4.85)	3673	7.59 (4.17)

Table 9.1: Number of games and referring utterances in the splits of our dataset with their average length in tokens (standard deviation in brackets), broken down by first mentions vs. subsequent ('Later') references.

descriptive statistics of the dataset. More details about the extraction procedure and the dataset are available in Appendix D.1. Appendix D.2 describes how the dataset is further processed to be used in our models.

9.4 Models

With the new dataset of referring utterance chains in place, we operationalize the problem of generating a referring utterance, taking into account the visual and conversational context as follows. The model aims to generate a referring utterance given (a) the *visual context* in the current game round made up of 6 images from the perspective of the player who produced the utterance, (b) the *target* among those images, and (c) the *previous co-referring utterances* in the chain (if any). Besides being contextually appropriate, the generated utterance has to be informative and discriminative enough to allow an addressee to identify the target image. We thus also develop a reference resolution model that plays the role of the addressee. The two models are trained independently.

9.4.1 Generation Models

We propose three versions of the generation model, which all follow the encoder-decoder architecture (Sutskever et al., 2014). These versions differ from each other with respect to whether and how they exploit earlier referring utterances for the target image: (1) a baseline model that does not use the dialogue context at all (henceforth, **Ref**); (2) a model that conditions the generation on the previous referring utterance, if available, and operates attention over it (**ReRef**); (3) a model that builds on (2) by adding a 'copy' mechanism similar to the mechanism proposed by See et al. (2017) (**Copy**). We describe them below in detail.

Ref This model is provided only with the information about the visual context in the current game round–and not with the linguistic context in previous rounds, see Figure 9.3 for a depiction of its architecture. We encode each image in the context by means of visual features extracted from the penultimate layer of

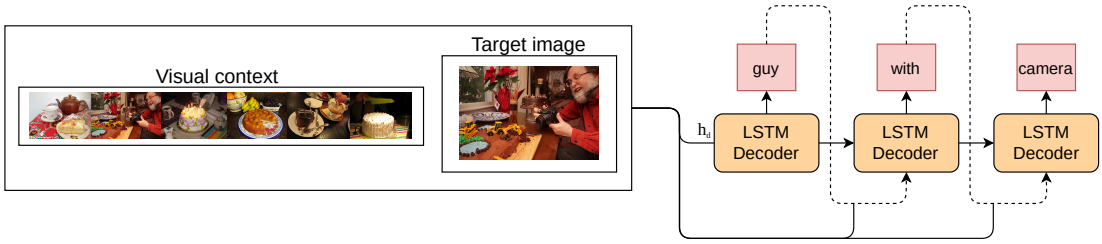


Figure 9.3: Architecture of the **Ref** referring utterance generation model.

ResNet-152 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009). First, the visual features of the 6 candidate images are concatenated. This concatenated vector goes through dropout, a linear layer and ReLU (Nair and Hinton, 2010). The same process is applied for the target image. We then concatenate the final visual context vector with the target image vector, apply a linear transformation, and use the resulting hidden representation h_d to initialize an LSTM decoder (Hochreiter and Schmidhuber, 1997), which generates the referring utterance one word at a time, t_t. At each timestep, the input to the decoder is a multimodal vector, i.e., the concatenation of h_d and the word embedding of token t_t. The weights of the embeddings are initialized uniformly in the range $(-0.1, 0.1)$ and learned from scratch for the task at hand.

ReRef With this model, we aim to simulate a speaker who is able to *re-refer* to a target image in accordance with what has been established in the conversational common ground (Clark, 1996; Brennan and Clark, 1996). This model enriches **Ref** by incorporating linguistic information into the encoder (in addition to visual information), and an attention mechanism applied over the hidden states of the encoder during decoding, see Figure 9.4. The model thus generates a new utterance conditioned on both the visual and the linguistic context.

The encoder is a one-layer bidirectional LSTM initialized with the same visual input fed to **Ref** (consisting of a representation of the visual context and the targe image). Instead of initializing the decoder as in the prior model, here, the visual data initializes the encoder. In addition, the encoder receives as input the previous referring utterance used in the dialogue to refer to the target image,[5] or else is fed the special <nohs> token, indicating that there is no conversational history for the target image yet. The embeddings of this input go through dropout.

We concatenate the last hidden states of the forward and backward directions of the BiLSTM encoder. This concatenated vector is then projected to hidden dimensions and used to initialize the decoder. The input to the decoder during training is an embedding of the ground-truth utterance.

During decoding, we utilize the attention mechanism proposed by Bahdanau

[5]The latest description seems to contain the most relevant information. Including all referring utterances in the chain up to that point in the dialogue did not lead to improvements.

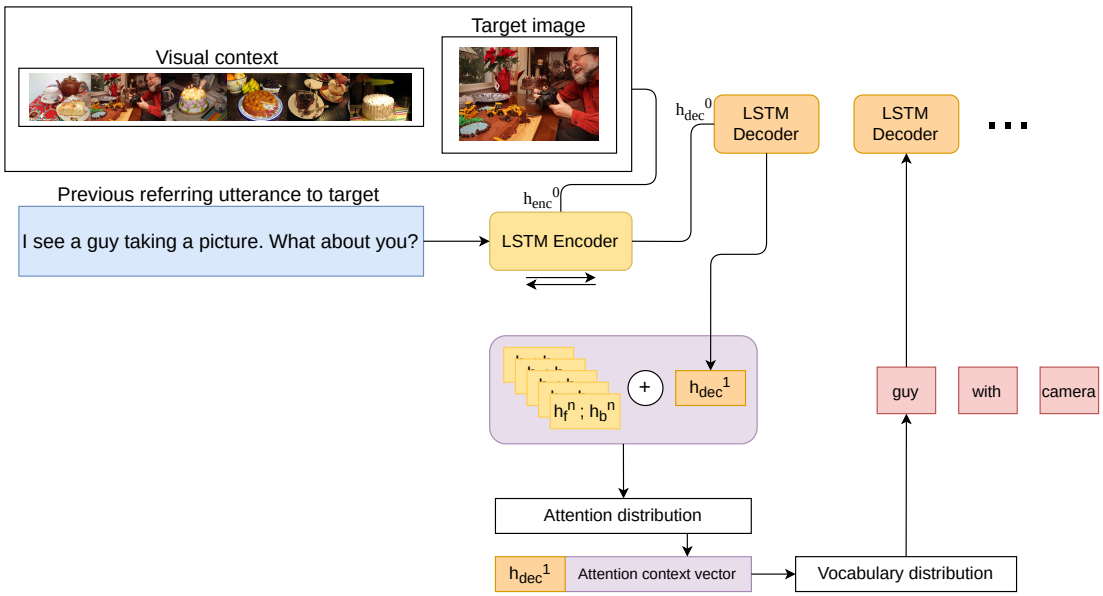


Figure 9.4: Architecture of the **ReRef** referring utterance generation model.

et al. (2018) and used by See et al. (2017). The attention contributes to determining which aspects of the multimodal context are most critical in generating the words of the next referring utterance. We expect this attention mechanism to be able to identify the words in a previous utterance that should be present in a subsequent reference, resulting in lexical entrainment.

The attention is applied as follows: Each hidden output of the encoder ${h_{enc}}^t$ (concatenation of forward and backward hidden states for timestep t) goes through a linear layer that projects it from double the size of hidden dimensions to the attention dimensions. In addition, the current hidden state of the decoder ${h_{dec}}^c$ is projected from the hidden dimensions to the attention dimensions.

$$enc^t = W_e {h_{enc}}^t \quad [9.1]$$

$$dec^c = W_d {h_{dec}}^c \quad [9.2]$$

$$e_t = v_a(tanh(enc^t + dec^c)) \quad [9.3]$$

Attention weights are calculated based on the sum of enc^t and dec^c, on which we apply *tanh* non-linearity and a linear layer. Padded tokens are masked, and softmax is applied over all remaining encoder timesteps i:

$$a_i = softmax(e_i) \quad [9.4]$$

$$h^* = \sum_i a_i h^i_{enc} \quad [9.5]$$

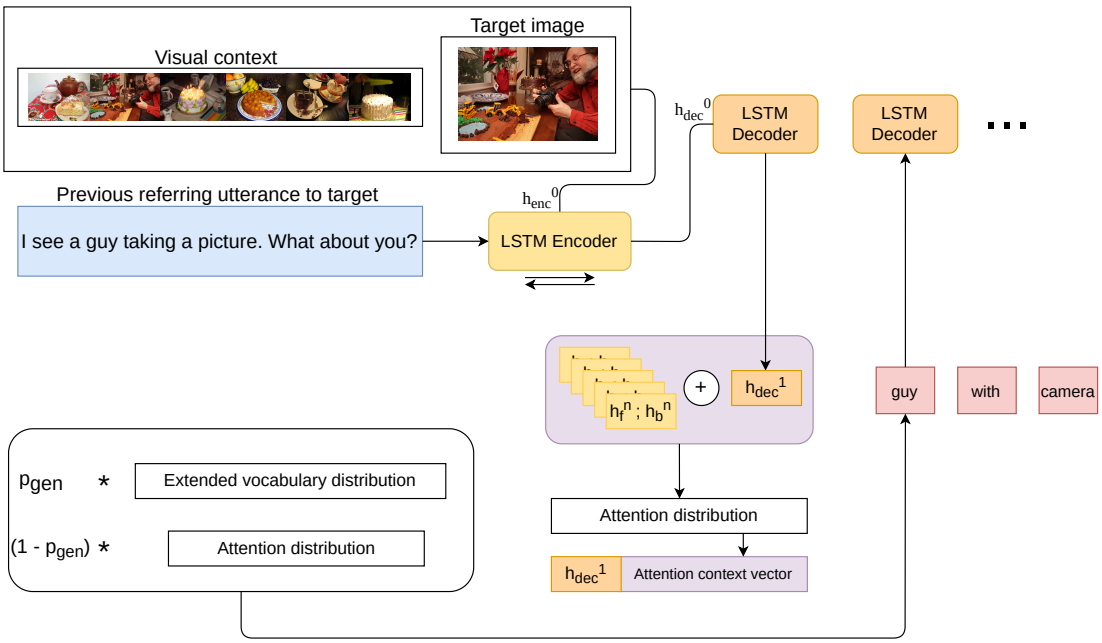


Figure 9.5: Architecture of the **Copy** referring utterance generation model.

To predict the word that the decoder will generate, we concatenate the decoder's current hidden state ${h_{dec}}^c$ with the weighted average from the encoder, i.e., encoder context vector h^*. This concatenation is projected to the size of the vocabulary minus 1, as we do not want the model to predict the <nohs> token.

Copy This model builds on **ReRef** and incorporates a means of simulating lexical entrainment more explicitly, by regulating when a word that was used in the previous mention should be used again in the current referring utterance (i.e., should be produced by the decoder). Given the shortening property of subsequent references, our task bears some similarity to text summarization. We thus draw inspiration from the summarization model proposed by See et al. (2017). As illustrated in Figure 9.5, we equip the model with the 'copy' mechanism utilized by See et al. (2017), which combines the probability of copying a word present in the encoded input with the probability of generating (p_{gen}) that word from the vocabulary. We expect this mechanism to contribute to generating rare words present in preceding referring utterances that are part of a 'conceptual pact' (Brennan and Clark, 1996) between the dialogue participants, but may have low generation probability overall.

The encoder part of this model is the same as that of the **ReRef** model. However, this model uses various versions of the input, and the decoder is altered to accommodate the copy mechanism.

First of all, we keep track of the unknown tokens in the input to provide the ability to predict them in the decoder phase. For this, we map the previous utter-

ance from the history to temporary indices in a new extended vocabulary. This extended vocabulary contains the unknown words existing in the input utterance in their original forms appended to the end of the original vocabulary. Since we do not want <nohs> to be predicted, we take additional precautions when it exists in the encoder input. The decoder input stays the same with unknown embeddings; nevertheless, the target utterance can include temporary indices assigned to unknown words encountered in the given input utterance so that we can calculate the loss according to them, as well.

The attention mechanism works in the same manner as in the **ReRef** model. However, we change what comes afterwards in line with the copy mechanism, where the attention for each word in the input utterance is added to their generation probabilities in the vocabulary. Here, we scatter the attention scores for the temporary indices of unknown words onto the distribution of the extended vocabulary, as well. For this reason, we maintain multiple versions of the input and output (mapped to the reduced vocabulary and mapped to the full vocabulary), as well as keeping track of the set of unknown words in the previous utterance and their temporary indices. Crucial here is the calculation of the generation probability p_{gen}, which requires the addition of several more linear layers that process the encoder context vector h_t^*, decoder input x_t, and the current decoder state s_t. As compared to the calculation of p_{gen} by See et al. (2017), we altered the formula for this value by adding *tanh* non-linearities:

$$p_{gen} = \sigma(tanh(w_{h^*}^T h_t^*) + tanh(w_s^T s_t) + tanh(w_x^T x_t)) \quad [9.6]$$

9.4.2 Reference Resolution Model

Given an utterance referring to a target image and a 6-image visual context, our reference resolution model predicts the target image among the candidates, see Figure 9.6. This model is similar to the resolution model proposed by Haber et al. (2019) for the PhotoBook dataset, but includes several extensions: (1) We use BERT embeddings from the uncased base BERT model (Devlin et al., 2019; Wolf et al., 2020) to represent the linguistic input rather than LSTMs;[6] (2) The input utterance is encoded taking into account the visual context: We compute a multimodal representation of the utterance by concatenating each BERT token representation with the visual context representation, obtained in the same way as for the generation models;[7] (3) We apply attention over the multimodal representations of the utterance in the encoder instead of using the output from a language-only LSTM encoder. The utterance's final representation is given

[6]In the generation models, we did not use BERT due to the difficulties of using contextualized embeddings in the decoder, and the desirability of using the same word embeddings in both the encoder and the decoder.

[7]We also tried using multimodal representations obtained via LXMERT (Tan and Bansal, 2019). No improvements were observed.

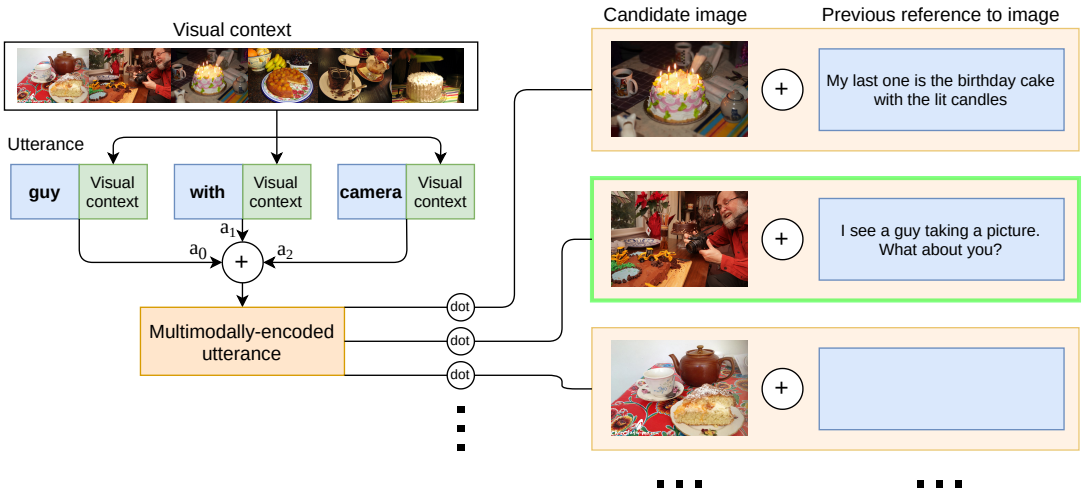


Figure 9.6: Architecture of the reference resolution model.

by the weighted average of these multimodal representations with respect to the attention weights.

In this model, BERT embeddings go through a dropout layer, and then a linear layer projecting the size to hidden dimensions. Finally, ReLU is applied (Nair and Hinton, 2010).

All 6 images in the context are concatenated, and the concatenation goes through dropout, a linear layer and ReLU to produce the final visual context vector. We then concatenate each of the BERT representations with the visual context vector to obtain multimodal token representations. This multimodal vector goes through a linear layer and ReLU, which finalizes the multimodal input vectors. The model then determines the attention to be paid to each of the multimodal vectors as indicated below:

$$e_i = v_a(tanh(W_e h_i)) \quad [9.7]$$

h_i is the multimodal output for each token, W_e is a linear layer projecting from hidden dimensions to attention dimensions, v_a is a linear layer that projects the output from the attention dimensions to a scalar. The model then masks the pad tokens before applying softmax over e_i scores to obtain the attention weights a_i:

$$a_i = softmax(e_i) \quad [9.8]$$

The final multimodally-encoded utterance representation is then the weighted average of all h_i, given their attention weights a_i:

$$h_L = \sum_i a_i h_i \quad [9.9]$$

Each candidate image is represented by its ResNet-152 features (He et al., 2016) or, if it has been previously referred to in the dialogue, by the sum of the visual

features and the representation of the previous utterance (obtained via averaging its BERT embeddings).[8] The images first separately go through dropout and a linear layer before the addition of possible linguistic history.

The history of each candidate image is determined by looking at their respective chains in the given game. Crucially, we only look at the chain items that were uttered before the current utterance we are trying to resolve. We take only the last utterance in the history, if such a history exists for a candidate image. In this case, we take the average of the BERT representations in the last utterance for that image. This average then goes through dropout, a linear layer and ReLU.

The final history representation for a candidate image is added to this image's final visual representation to obtain its final candidate representation. As the last step, we apply RELU and normalize the outcomes for each candidate separately with L2 normalization. Please note that not all images in the context necessarily have histories associated with them. Therefore, some candidate representations will be multimodal, whereas the others will remain in the visual domain, with no linguistic history being added.

To determine the target image, we take the dot product between the candidate representations and the multimodally encoded input utterance representation. The candidate with the highest dot-product value is then predicted to be the referent of the input utterance.[9]

Ablation As an ablation of the model described above, we train another type of model where the history is not added to the candidate images. Hence, the candidates are always represented only in the visual modality.

Baseline This model only uses one-hot vectors based on image IDs. These vectors go through the same operations as the image features go through in the models described above (dropout, linear layer, ReLU and normalization). At the end, instead of the dot product, the outputs for the candidates are projected to scalar values, and the model tries to predict the target by applying softmax directly over these scalars.

9.4.3 Model Configurations

For each model, we performed hyperparameter search for batch size, learning rate, and dropout; also, the search included different dimensions for the embedding, attention, and hidden layers. All models were trained for up to 100 epochs (with a patience of 50 epochs in the case of no improvement to the validation performance)

[8] Thus, some of the candidate images have multimodal representations (if they were already mentioned in the dialogue), while others do not.

[9] See Wu et al., 2023 for an implementation of a listener agent that utilizes the full dialogue history to directly play a game from PhotoBook, i.e., to predict if an image exists in the set of the interlocutor.

using the Adam optimizer (Kingma and Ba, 2015) to minimize the Cross-Entropy Loss with sum reduction. BERTScore F1 (Zhang et al., 2020b) in the validation set was used to select the best model for the generation task, while we used accuracy for the resolution task. In the next section, we report average scores and standard deviations over 5 runs with different random seeds. Further details on hyperparameter selection, model configurations, and reproducibility can be found in Appendix D.4.

9.5 Results

9.5.1 Evaluation Measures

We evaluate the performance of the reference resolution model by means of both accuracy and Mean Reciprocal Rank (MRR). As for the generation models, we compute several metrics that are commonly used in the domain of NLG. In particular, we consider three measures based on n-gram matching: BLEU-2 (Papineni et al., 2002),[10] ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015). We also compute BERTScore F1 (Zhang et al., 2020b) (used for model selection), which in our setup compares the contextual embeddings of the generated sentence to those of the set of referring utterances in the given *chain*. Further details of the metrics are in Appendix D.3.

All these measures capture the degree of similarity between generated referring utterances and their human counterparts. In addition, to assess the extent to which the generated utterances fulfill their communicative goal, we pass them to our reference resolution model to obtain accuracy and MRR. While this is not a substitute for human evaluation, we take it to be an informative proxy. In Section 9.6, we analyze the generated utterances with respect to linguistic properties related to phenomena that are not captured by any of these metrics.

9.5.2 Reference Resolution Results

Our reference resolution model achieves an accuracy of 85.32% and MRR of 91.20% on average over 5 runs. This is a substantial result. A model that predicts targets at random would yield an accuracy of roughly 16.67% (as the task is to pick one image out of 6 candidates), while the baseline that simply takes projected versions of one-hot representations of the image IDs in the context achieves 22.37% accuracy.[11]

[10] BLEU-2, which is based on bigrams, appears to be more informative than BLEU with longer n-grams in dialogue response generation (Liu et al., 2016).

[11] The fact that this is slightly higher than random accuracy seems due to the different frequencies of images being the target in the dataset.

Subset	ACC	MRR	Instances
First	80.27 (0.46)	87.78 (0.28)	2482
Later	88.74 (0.18)	93.51 (0.09)	3673
Overall	85.32 (0.19)	91.20 (0.10)	6155

Table 9.2: Test set scores of the reference resolution model: averages of 5 runs with the best configuration, with the standard deviations in parentheses.

In Table 9.2, the results are presented by breaking down the test set into two subsets: the *first* referring utterances in a chain, and *later* referring utterances, i.e., subsequent references where the target image has linguistic history associated with it. The model performs better on subsequent references. Exploiting dialogue history plays a role in this boost: the ablated version of the model that does not have access to the linguistic history of subsequent references yields an accuracy of 84.82% for the *Later* subset, which is significantly lower than the 88.74% obtained with our model ($p < 0.01$ independent samples t-test). This confirms the importance of accessing information about previous mentions in visually grounded reference resolution (Haber et al., 2019). We use the best model run to assess the communicative effectiveness of our generation models.

9.5.3 Generation Model Results

As we did for the reference resolution model, we break down the test set into first referring utterances in a chain and subsequent references, for which generation is conditioned on a previous utterance. The outcomes of this breakdown are provided in Table 9.3, where we report the test set performances of our three generation models. Overall results on the validation set are available in Appendix D.5.

ReRef obtains the highest scores across all measures, followed by **Copy**, while **Ref** achieves substantially lower results. Regarding the comparison between first and subsequent references, the context-aware models **ReRef** and **Copy** attain significantly higher results when generating later mentions vs. first descriptions ($p < 0.001$, independent samples t-test). As expected, no significant differences are observed in this respect for **Ref**.[12]

As for the communicative effectiveness of the generated utterances as measured by our resolution model, both accuracy and MRR are particularly high (over 90%) for **ReRef**. Across all model types, generated *subsequent* references are easier to resolve by the model, in line with the pattern observed in Table 9.2 for the human data.

All in all, the addition of the copy mechanism does not provide improvements over **ReRef**'s performance that can be detected with the current evaluation mea-

[12]While first descriptions do not require linguistic context, **ReRef** and **Copy** perform better on first description generation than **Ref**. This is likely due to their higher complexity.

Model	Subset	BLEU-2	ROUGE	CIDEr	BERT-F1	ACC	MRR
Ref	**First**	20.80 (1.02)	29.74 (1.59)	41.26 (3.14)	54.48 (1.38)	57.12 (4.85)	72.47 (3.19)
	Later	23.06 (1.20)	31.88 (1.66)	40.79 (2.83)	55.54 (1.40)	60.94 (2.67)	75.34 (1.59)
ReRef	**First**	33.09 (0.79)	42.32 (0.42)	94.63 (2.12)	62.55 (0.12)	90.36 (1.73)	94.49 (1.14)
	Later	52.15 (1.19)	56.74 (0.63)	143.59 (5.84)	71.25 (0.39)	92.21 (0.73)	95.62 (0.45)
	baseline	36.66 (0.92)	45.37 (0.57)	96.41 (2.69)	64.13 (0.24)	90.14 (2.28)	94.38 (1.41)
Copy	**First**	25.25 (0.40)	33.31 (0.50)	60.51 (1.21)	57.61 (0.36)	81.36 (0.53)	88.70 (0.49)
	Later	43.08 (0.36)	48.79 (0.41)	128.45 (1.98)	66.07 (0.17)	83.96 (0.53)	90.60 (0.32)

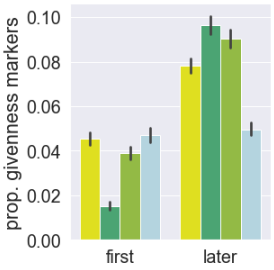
Table 9.3: Test set scores of the generation models (averaged over 5 runs) for first vs. subsequent references, including word-overlap metrics, BERTScore F1, and accuracy and MRR obtained by our resolution model on the generated utterances. ReRef *baseline* uses the first generated description verbatim in all later mentions. All differences across model types are statistically significant ($p < 0.001$, independent samples t-test).

sures. We do find, however, that the **Copy** model uses a substantially larger vocabulary than **ReRef**: 1,791 word types vs. 760 (the human vocabulary size on the test set is 2,332, while **Ref** only uses 366 word types). An inspection of the vocabularies shows that **Copy** does generate a good deal of low-frequency words, in line with what is expected from the dedicated copy mechanism (less desirably, this also includes words with spelling errors). Further analysis also shows that **Copy** generates utterances that include more repetitions: 18% of the utterances generated by **Copy** in the test set contain two identical content words e.g. *"do you have the runway runway woman?"*, while only 7% of those generated by **ReRef** do.[13] Adding a means to control for repetitions, such as the 'coverage' mechanism by See et al. (2017), could be worth exploring in the future.

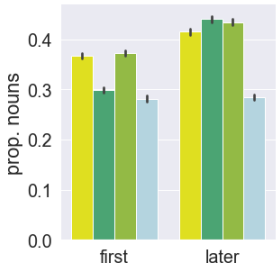
We compare our best performing model **ReRef** to a baseline consisting in reusing the first generated utterance verbatim in later mentions. In this case, the model does not learn how to reuse previous referring utterances taking into account the changing visual context, but simply keeps repeating the first description it has generated. We expect this baseline to be relatively strong given that experiments in psycholinguistics studies have shown that dialogue participants may stick to an agreed description even when some properties are not strictly needed to distinguish the referent in a new visual context (Brennan and Clark, 1996; Brown-Schmidt et al., 2015). The results (reported in Table 9.3 *baseline*) show that the model significantly outperforms this baseline when generating later mentions.

Overall, our results confirm that referring utterances do evolve during a dialogue and indicate that the models exploiting the conversational context are able to learn some of the subtle modifications involved in the re-referring process. In the next section, we look into the linguistic patterns characterizing this process.

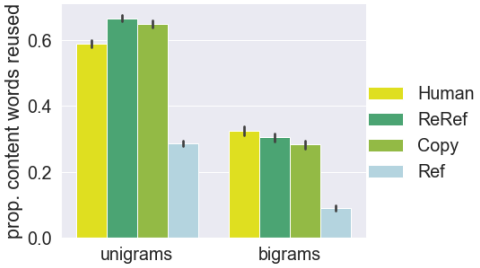
[13]The **Ref** model is even more repetitive: 21% of the generated utterances contain repeated content words.



(a) Givenness markers



(b) Proportion of nouns



(c) Lexical entrainment in later references

Figure 9.7: Linguistic patterns in human referring utterances and in referring utterances generated by our three models. Givenness markers and proportion of nouns per utterance are displayed for first and later references.

9.6 Linguistic Analysis

We analyze the linguistic properties of the utterances generated by the best-performing run of each of our models and compare them with patterns observed in the human data. Extensive descriptive statistics are available in Appendix D.6.

9.6.1 Main Trends

Givenness markers We first look into the use of markers of new vs. given information, in particular indefinite and definite articles as well as particles such as *again* or *before* (as in *"I have the X one again"* or *"the X from before"*), which are anaphoric and presuppose that an image has been discussed previously in the dialogue. Figure 9.7a shows the proportion of givenness markers (*the, one, same, again, also, before*) in first vs. subsequent references. Not surprisingly, this proportion increases in the human subsequent references. **ReRef** and **Copy** both display an amplified version of this trend, while **Ref**, which cannot capture any given information, shows no difference.

Reduction Regarding referring utterance length, we observe a significant shortening in subsequent mentions in human dialogues (11.3 vs. 8.3 tokens on average

in first and subsequent mentions, respectively). This shortening is also observed in the utterances generated by **ReRef** (11.3 vs. 7.2) and **Copy** (10.8 vs. 7.8). **Ref** tends to generate longer utterances across the board (13.7 vs. 13.6).

Shortening may be linked to compression, i.e., to an increase in information density (Shannon, 1948). To analyze this aspect, we consider the proportion of content words in the utterances, since such proportion can capture mechanisms such as syntactic reduction (e.g., the removal of the complementizer *that*), which has been shown to be a good predictor of information density increase (Jaeger and Levy, 2007). Haber et al. (2019) reported a rise in the proportion of content words for all utterance types in later rounds of the PhotoBook games. We also observe such an increase in our referring utterance chains, and a similar trend is exhibited as well by the output of the **ReRef** and **Copy** models: In particular, generated subsequent references contain a significantly higher proportion of nouns and adjectives compared to first descriptions. Figure 9.7b shows this pattern for nouns, which are the most prominent type of content word in our data.

Entrainment In order to analyze the presence of lexical entrainment, we compute the proportion of expressions in subsequent references that are reused from the previous mention. We compare reuse at the level of unigrams and bigrams. Figure 9.7c shows this information focusing on content words. Around 60% of content tokens are reused by humans. The proportion is even higher in the utterances generated by our context-aware models. Digging deeper into the types of content tokens being reused, we find that nouns are reused significantly more than other parts of speech by humans. This is also the case in the subsequent references generated by the **ReRef** and **Copy** models.

Humans also reuse a substantial proportion of content word bigrams—as do, to a smaller degree, the context-aware models. For example, given the gold description *"pink bowls rice and broccoli salad next to it"*, **ReRef** generates the subsequent reference *"pink bowls again"*. Noun-noun compounds are a particularly interesting case of such bigrams, which we qualitatively analyze below.

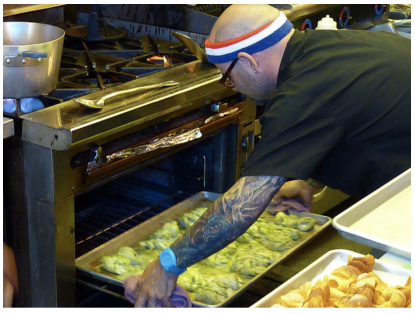
9.6.2 A Case Study: Noun-Noun Compounds

A partial manual inspection of the human utterances in our chains reveals that, as they proceed in the dialogue, participants tend to produce referring expressions consisting of a noun-noun compound.[14] For example, in Figure 9.2 we observe the compound *"camera guy"* being uttered after the previous mention *"guy with camera"* (reused nouns are underlined). Another example is *"wine glass dog"* in Figure 9.1. This is in line with Downing (1977), who argues that novel (i.e., not

[14]This is consistent with the fact that the proportion of noun-noun bigrams is significantly higher in subsequent references (0.05 vs. 0.08 on average in first and subsequent references, respectively; $p < 0.001$ independent sample t-test).

P: lady with basket?

⇝ **ReRef:** basket lady?



P: do you have headband guy?

⇝ **ReRef:** tattoo guy?

Figure 9.8: Two examples from the test set where **ReRef** generates a noun-noun compound based on the previous human mention (***P***). Left: a genuine *reuse* case; right: a *non-reuse* case. Reused words are underlined.

yet lexicalized) noun-noun compounds can be built by speakers on the fly based on a temporary, implicit relationship tying the two nouns, e.g., 'the guy *taking a picture with* a camera'. Such noun-noun compounds are thus prototypical examples of reuse and reduction: On the one hand, the novel interpretation (which needs to be pragmatically informative, diagnostic, and plausible; Costello and Keane, 2000) can only arise from the established common ground between speakers; on the other hand, compounds are naturally shorter than the 'source' expression since they leave implicit the relation between the nouns.

We check whether our best performing generation models produce compounds as humans do, i.e., by reusing nouns that were previously mentioned while compressing the sentence. We perform the analysis with a qualitative focus, by manually inspecting a subset of the generated utterances.[15] In Figure 9.8, we show two noun-noun compounds generated by **ReRef** (similar cases were observed for **Copy**). The example on the left is a noun-noun compound, *"basket lady"*, that is consistent with the dialogue context: both nouns are indeed reused from the previous mention. In contrast, the compound on the right does not build on the conversational history; the noun *"tattoo"* is not in the previous mention and never uttered within the reference chain (not reported), and thus may be perceived as breaking a conceptual pact (Metzing and Brennan, 2003). The compound is grounded in the image, but not in the conversational context.

[15]The subset is obtained by applying simple heuristics to the set of generated utterances, such as length and POS tags.

9.7 Conclusion

We have addressed the generation of descriptions that are (1) discriminative with respect to the visual context, and (2) grounded in the linguistic common ground established in previous mentions. To our knowledge, this is the first attempt at tackling this problem at the level of surface realization within a multimodal dialogue context.

We proposed an encoder-decoder model that is able to generate both first mentions and subsequent references by encoding the dialogue context in a multimodal fashion and dynamically attending over it. We showed that our best-performing model is able to produce better, more effective referring utterances than a variant that is solely grounded in the visual context. Our analysis revealed that the generated utterances exhibit linguistic properties that are similar to those observed in human utterances regarding the reuse of words and reduction. Generating subsequent references with such properties has the potential to enhance user adaptation and successful communication in dialogue systems.

Yet, in our approach, we abstracted away from important interactive aspects such as the collaborative nature of referring in dialogue (Clark and Wilkes-Gibbs, 1986), which was considered by Shore and Skantze (2018) for the task of reference resolution. In the present work, we simplified the interactive aspects of reference by extracting referring utterances from the PhotoBook dialogues and framing the problem as that of generating the next referring utterance given the previous mention. We believe that the resulting dataset of referring utterance chains can be a useful resource to analyze and model other dialogue phenomena, such as saliency or partner specificity, both on language alone or on the interaction of language and vision.

In the next chapter, we go beyond the setups described in this chapter, and explore how to adapt generation models representing speakers against a set of resolution models representing listeners with diverse knowledge spaces.

Chapter 10

Speaker Adaptation in Visually Grounded Referential Games

The material in this chapter is based on: Ece Takmaz*, Nicolo' Brandizzi*, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernández. 2023. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4198–4217, Toronto, Canada. Association for Computational Linguistics. *Shared first authorship.

Contributions: Ece Takmaz: Implementing and running the initial models, contributing to the analysis, writing and revising the paper. Nicolo' Brandizzi: Implementing and running the final setups, writing and revising the paper. Mario Giulianelli: Creating data domains, conducting the majority of the analysis, writing and revising the paper. Sandro Pezzelle: Supervising the research, writing and revising the paper. Raquel Fernández: Supervising the research, writing and revising the paper.

10.1 Introduction

In the previous chapter, we have shown the significance of incorporating visual and linguistic contexts when generating referring utterances in multimodal dialogue. In this chapter, we investigate another phenomenon existing in dialogue: adaptation. Speakers tend to adapt their language use to the perceived knowledge, information, and linguistic abilities of their interlocutors by representing the mental states of the interlocutors via Theory of Mind (Isaacs and Clark, 1987; Clark, 1996; Pickering and Garrod, 2004; Tomasello, 2005).

In this chapter, we model a communicative situation where the interlocutors have *asymmetric language abilities*: a proficient speaker interacts with a listener characterized by limited semantic knowledge to complete a reference game, as illustrated in Figure 10.1. Our goal is to mimic a scenario in which, for example, a high school physics professor can make complex atomic models understandable to young students by using terminology familiar to them, such as culinary terminology to explain Thomson's 'plum pudding model'. We focus on the speaker's Referring Expression Generation (REG; Reiter and Dale, 1997; Krahmer and van Deemter, 2012) in a multimodal dialogue setting and use REG models equipped with visual perception to generate discriminative image descriptions within a set of related image candidates. Several psycholinguistic theories have proposed that language production is interwoven with comprehension via 'forward prediction'—i.e., producing an utterance involves predicting how a comprehender would understand it (e.g., Pickering and Garrod, 2013; Roelofs, 2020). Inspired by this idea, we equip our speaker model with a *simulator*, i.e., a module that 'simulates' whether a listener would be able to identify the target referent. Based on this predicted behavior (i.e., the expected effect of the planned utterance), the simulator modifies the generation plan on the fly to increase communicative success.

These are the main contributions of the research explained in this chapter:[1]

- We model adaptation between agents with asymmetric knowledge, using a referential task as case study, where agents communicate in natural language about realistic images (in contrast to related work using synthetic data—see Section 10.2).

- We propose a novel simulation-based approach and test it in two settings: (1) a *self-aware* setting where the speaker predicts how a generic listener (with the same knowledge as the speaker) would resolve a planned utterance, and (2) an *audience-aware* setting where the speaker learns—from the behavior of a listener with restricted semantic knowledge—to form representations of the listeners' knowledge (Clark, 1985; Isaacs and Clark, 1987) and predict their responses.

[1]Our code and models are publicly available at `https://github.com/nicofirst1/speaker-adaptation`

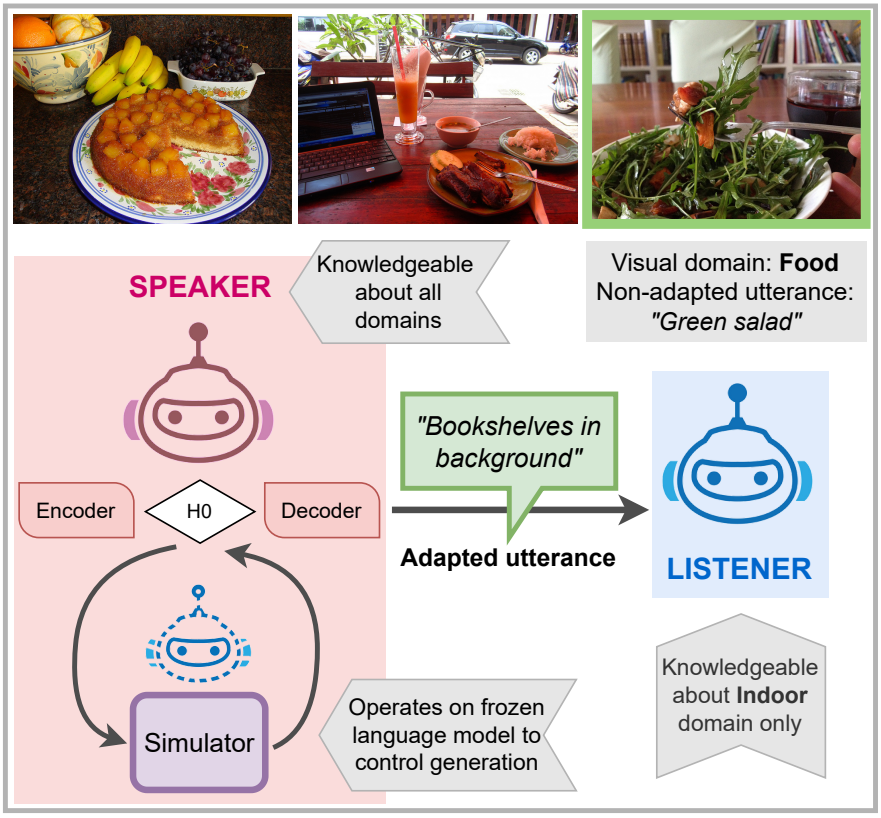


Figure 10.1: An illustration of our knowledge-asymmetric setup where an expert *Speaker* interacts with a less knowledgeable *Listener*. The *Speaker* tailors its utterance about an image from the **food** domain for a *Listener* who only knows about the **indoor** domain. The speaker's *Simulator* module inspired by Theory of Mind guides this adaptation. The adapted utterance exploits indoor terms (*'bookshelves'*) without referring to food.

- We exploit the simulator's representations in an innovative way: by leveraging a *plug-and-play* approach originally introduced for controllable text generation (Dathathri et al., 2020), which steers language production at the decoding stage without altering the underlying language model.

- We show that our approach leads to increased resolution accuracy; in particular, our audience-aware speaker is able to adapt its utterances effectively when referring to a target within a visual domain unknown to the listener.

- We provide an in-depth analysis of the patterns present in the adapted utterances and the model's production strategies underpinning our results.

10.2 Related Work

10.2.1 Pragmatic Reference Generation

Speakers tend to design their referring expressions to be pragmatically informative, i.e., discriminative from the listener's perspective. Most approaches to pragmatic reference expression generation (REG) have considered scenarios where we can assume a shared set of linguistic conventions between speakers and addressees (common domain and training data). The Rational Speech Act framework (RSA; Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Goodman and Frank, 2016) has become a popular option for characterizing such settings, with REG models that reason probabilistically about their interlocutors' interpretation via recursively defined speaker and listener models (Andreas and Klein, 2016; Monroe et al., 2017; Cohn-Gordon et al., 2018; Zarrieß and Schlangen, 2019; Fried et al., 2021), possibly taking into account information accumulated during interaction (Hawkins et al., 2020). There also exist joint speaker-listener models that are not recursive in the RSA sense. In these models, speakers can become listener-aware at inference time thanks to enhanced decoding algorithms (Vedantam et al., 2017) or they can learn to generate discriminative utterances at training time, for example via altered supervised training objectives (Mao et al., 2016) or auxiliary reinforcement learning (RL) modules (Yu et al., 2017a), including approaches where the RL rewards are determined by the reference resolution success of a listener model (Lazaridou et al., 2020).

Our model, too, produces audience-aware discriminative image descriptions through an auxiliary module that captures the listener's perspective. However, in contrast to the above studies, the setting we investigate has two distinct key features: (1) we model situations with *knowledge asymmetry* between the dialogue participants, and (2) we experiment with *plug-and-play controlled generation* methods that result in temporary updates to the speaker's language model—rather than steering generation via recursive probabilistic reasoning. We review work related to these two aspects next.

10.2.2 Knowledge Asymmetry & Referring Tasks

What if the speaker and the listener have access to differing semantic knowledge? It is well known that speakers are able to adapt to less proficient addressees (Isaacs and Clark, 1987). Janarthanam and Lemon (2010) were one of the first to address adaptation in dialogue systems with asymmetric knowledge. They modeled REG for technical domains where users may not know the jargon, using RL to learn a REG policy from a user simulation. More recently, Ohashi and Higashinaka (2022) focus on generating utterances in task-oriented dialogue with users that have limited vocabulary. They exploit the natural language understanding module of the system (representing user understanding) to set up a reward function,

which is then used to fine-tune the NLG module via RL.

In the context of visually grounded referring tasks, Bao et al. (2022) focus on a scenario where the listener has comprehension difficulties and model adaptation by reweighing the probability of candidate referring utterances as a function of their likelihood to be successfully interpreted by the listener. Similarly, Liu et al. (2023a) apply ToM-based listener modeling, where the speaker generates multiple candidate utterances and ranks them with the help of the ToM listener. Generating and ranking multiple utterances, however, is an inefficient production mechanism. For these reasons, others have tried to condition the speaker model prior to utterance generation, mainly with external modules. Corona Rodriguez et al. (2019) model interactions where the listener has an impaired perceptual system and implement this conditioning through an external policy network that takes as input listener embeddings. While Zhu et al. (2021) propose a ToM module that tracks the listener's understanding via meta-learning for few-shot coordination in a setup where listeners understand different languages. Singh et al. (2023) train an attention-based adapter layer in a reward-based manner as part of a multi-agent referential game where the speaker aims to generate utterances that would be understood by one listener, but not the other. Finally, Greco et al. (2023) have a setup that is the most similar to ours, where Expert speakers adapt to Layman listeners. But unlike our *plug-and-play* approach, the authors follow the RSA framework in developing audience-aware models that are updated through interaction.

10.2.3 Adaptive Controlled Generation

Most of the approaches to adaptation we have reviewed apply RL to the speaker model or fine-tune its language model through interaction. As a result, the speaker is not able to retain its original knowledge, which might cause catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999). With the advent of large pretrained language models, a plethora of new methods for controlled text generation have been proposed, including prefix-tuning (Li and Liang, 2021; Ben-David et al., 2022), prompting (Brown et al., 2020), adapters (Houlsby et al., 2019; Pfeiffer et al., 2020a,b), and energy-based constraints (Qin et al., 2022). Visual prefixes and prompts (Alayrac et al., 2022) have also been used to condition generation, especially without training the full language model.

We argue that this recent line of research offers promising alternative frameworks for adaptive REG. In particular, we investigate a solution to adaptation inspired by the *plug-and-play* approach to controlled text generation (PPLM; Dathathri et al., 2020; Pascual et al., 2021), which has been used to steer large pretrained language models towards generating texts with certain features (e.g., positive/negative sentiment or a given vocabulary distribution). In Dathathri et al. (2020), latent representations are updated at inference time with the help of a classifier while keeping the model parameters unchanged. Building on this

idea, we propose a modular approach to REG adaptation in asymmetric knowledge settings where a module trained to predict the listener's behavior —similar to the 'prediction net' in the machine ToM model by Rabinowitz et al. (2018)—is exploited to control generation on the fly.

10.3 Problem Formulation

We provide an abstract overview of the problem we address and our approach. Details on the data and the experimental pipeline are given in Section 10.4 and Section 10.5.

Scenario Our setup is a classic referential game: two artificial agents, a speaker and a listener, share a visual context involving multiple images. The speaker produces an utterance to refer to one of the images (the target) and the listener attempts to identify the referent given that utterance. In particular, we model a scenario with *knowledge asymmetry*, where the speaker is more knowledgeable than the listener. We hypothesize that, in such a setup, for communication to be successful, the speaker will need to adapt its utterances to the listener's representational space and language. To make this possible, we endow the speaker with a simulation module and an adaptation mechanism.

Simulation We provide the speaker with a module that simulates how a listener would process a planned utterance. We assume that, by having interacted with listeners in the past, the speaker has learned a model of certain listener types (e.g., a prototypical idea of what a 3-year-old would understand). We operationalize this by pretraining several instances of the simulator, one per listener type, to predict how a listener is likely to resolve a referring utterance. We compare three settings:

- *Baseline*: No simulation takes place.
- *Self-aware*: The simulator is trained to predict how a listener with the same knowledge as the speaker would resolve an utterance. This is equivalent to a pragmatic speaker who reasons about the effect of its utterances on a generic listener (see Section 10.2.1), but in our approach at test time the listener's interpretations are predicted rather than directly observed. Our proposal is also inspired by human production models based on 'self-monitoring' (Levelt, 1993; Roelofs, 2020).
- *Audience-aware*: The simulator is trained to predict how a listener with a subset of the speaker's knowledge, i.e., a single domain, would resolve an utterance. Thus, the speaker learns a model—a theory of mind—of a less knowledgeable listener type that allows the speaker to make predictions

about the listener's behavior. When performing the referential task, we assume that the speaker knows the type of the listener beforehand, i.e., which simulator needs to be engaged (similarly to knowing that we are addressing a 3-year-old, for example).

Adaptation Rather than fine-tuning the speaker's language model, we exploit the pretrained simulators to control utterance generation on the fly via a monitoring loop. The simulator checks whether planned utterances would be effective; if that is not the case, a loss is backpropagated to update the initial hidden state of the speaker's decoder and a new utterance is generated. Our hypothesis is that such a mechanism will lead to referring utterances that are adapted to the listener's knowledge.

10.4 Data

We use the dataset of referring utterances automatically extracted from the PhotoBook dialogues as explained in Chapter 9 (Takmaz et al., 2020a). In contrast to the previous chapters, we leave aside the dialogue context for simplicity, as our focus in this chapter is on the generation and adaptation of referring utterances in visual contexts. As the visual contexts in PhotoBook feature images from the same visual domain, this facilitates the separation of the dataset into domain-specific subsets. The images in PhotoBook belong to 30 different visual domains (e.g., *'person-umbrella'*, *'car-motorcycle'*). To model speaker adaptation to different semantic domains, we split the dataset of PhotoBook referring utterances according to the visual domain of each game. We cluster the image domains as a function of the similarity between their vocabulary vectors, constructed by counting word frequencies in the referring utterances belonging to a given domain. We obtain a set of 5 macro-domains (*appliances*, *food*, *indoor*, *outdoor*, *vehicles*), selected so that the domain vocabularies have minimal overlap. For each cluster of visual domains, we extract the corresponding referring utterance and visual context. We then randomly split these into training (70%), validation (15%), and test set (15%). We also merge the 5 domain-specific datasets into an *'all-domains'* dataset to be used to train domain-general models as described in Section 10.5. See summary in Table 10.1.

10.5 Experimental Pipeline

As described in Section 10.3, our experimental pipeline includes two agents—a speaker and a listener—implemented as a generative language model instantiating the speaker, a discriminative model instantiating the listener, and a third model, a simulator used by the speaker to assess the forward effect of its planned utterance

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	-	-

Table 10.1: Statistics of the domain-specific datasets: # of utterances (N) and % within the entire dataset (Prop), vocabulary size ($|V|$), # of unique images (Images), % of domain-specific vocabulary (Specific), and max. lexical overlap with another domain (Overlap). The max. overlap is between *outdoor* and *vehicles*. Example shared words are *'left'*, *'black'*, *'driving'*, and *'glasses'*.

on the listener. The language model and the discriminator model are adapted from those by Takmaz et al. (2020a) described in Chapter 9, and the simulator model is built on the discriminator's architecture with additional components. We train these models from scratch to have full control over the linguistic and visual knowledge of the agents and their degree of asymmetry. We use ResNet-152 to encode the images (He et al., 2016). See Appendix E.1 for more information about the training schemes and hyperparameters.

10.5.1 Generative Language Model

The speaker is a visually conditioned language model that generates an utterance describing a target image within a visual context. The model follows an encoder-decoder architecture consisting of a visual encoder that represents the visual context along with the target image, and a decoder for language generation. The decoder generates a referring utterance via nucleus sampling (Holtzman et al., 2020), also paying attention to the encoder output at every time step. See Appendix E.1.1 for more details about the model architecture.

We train the visually conditioned language model from scratch using the training set including all domains in PhotoBook and optimize the model with respect to Cross-Entropy Loss using Adam (Kingma and Ba, 2015). We select the best model based on its performance on a set of NLG metrics on the validation set. The weights of the trained speaker are then frozen and used as the core language generation model in all our experiments identically.

Performance The speaker's language model obtains reasonable scores in terms of classic NLG metrics:[2] 23.8 BLEU-2, 32.9 ROUGE, 44.1 CIDEr, and 57.7

[2]Comparable to those obtained by Takmaz et al. (2020a) with their 'Ref' model.

BERTScore F1 (Papineni et al., 2002; Lin, 2004; Vedantam et al., 2015; Zhang et al., 2020b). All scores are averages across 4 seeds on the test set. For details, see Appendix E.2.1.

10.5.2 Discriminator

Our listener is a discriminator model that receives six images in the visual context plus an utterance, and is tasked with identifying the target image that the utterance refers to. To encode the utterance, we use word embeddings trained from scratch to make sure no knowledge leaks from any pretraining. The model combines the visual context and the utterance to produce a multimodal context vector. The listener identifies the target image by comparing this multimodal context vector to the representations of each candidate image via dot-product and selecting the image with the highest score. See Appendix E.1.2 for the detailed description of the model architecture.

We train one listener model per domain in Table 10.1.[3] The models are optimized with Cross-Entropy Loss using the Adam optimizer. The best models are selected based on resolution accuracy on the validation set. We keep these domain-specific listener models frozen in the rest of the study. See Appendix E.1.2 for further details.

Performance We distinguish between *in-domain* (IND) accuracy—i.e., the resolution accuracy achieved on the test set of the domain on which the listener has been trained—and *out-of-domain* (OOD) accuracy—accuracy on domains the listener has not been exposed to (e.g., the accuracy on images from the *vehicles* domain of a listener exclusively trained on the *food* domain). Our listeners are truly domain specific: they are able to identify the target image with an average accuracy of 83.08% in IND, while their OOD accuracy is 19.05% on average—barely above a random baseline (16.67%). See Appendix E.2.2 for the full results broken down per domain.

10.5.3 Simulator

As explained in Section 10.3, the speaker is endowed with a simulator module. The simulator receives inputs in two parallel streams. In one stream, it receives the visual context $\mathbf{v}$ coupled with the speaker's planned utterance u_t, and in the second stream, the visual context along with the language model's initial hidden state h_0. The motivation behind this architectural choice is related to the plug-and-play approach at the core of our proposal. The first stream is inspired by previous work on ToM (e.g., Rabinowitz et al., 2018): its main input is the same

[3]We also train a general listener model on all domains which is only used to train the self-aware simulator; see Section 10.5.3.

as what a listener would receive, an utterance. However, to control generation on the fly, we need to modify the language model's internal representations. Thus, the main reason for the second stream is technical: the gradients from the simulator's loss cannot flow back to the language model's hidden states if the input to the simulator is text due to the non-differentiability of the *argmax* operation.[4] The second stream uses a combination of linear layers and standardization to compute the dot product between h_0 and $\mathbf{v}$. The outcomes of the two streams are multiplied to obtain the final representation that is compared to the candidate images.

We train one audience-aware simulator per domain-specific listener and one self-aware general simulator with Cross-Entropy Loss using the AdamW optimizer (Loshchilov and Hutter, 2019). The training set sizes of both types of simulators are the same, with the target behavior being different. In the simulation of a general listener, the simulator predicts the behavior of a listener that was exposed to all domains as the speaker, contrary to one domain in the domain-specific case. We choose the best simulator per listener type based on the simulators' prediction accuracies (more details in Appendix E.1.3). The simulators are then frozen in the rest of the pipeline.

Performance The self-aware simulator achieves an accuracy of 70% when predicting the behavior of a general listener. The audience-aware simulators predict the behavior of domain-specific listeners with an average accuracy of 78.20% for IND samples, and 72.78% for OOD samples.[5] The drop in accuracy from IND samples to OOD samples could be due to difficulties in ascertaining the reactions of a listener on OOD data. See details of the results in Appendix E.2.3.

10.6 Audience-Aware Adaptation

In our framework, adaptation takes place at inference time building on our pretrained, frozen models for the language model, the discriminators and simulators described in Section 10.5. We first explain our adaptation mechanism (Section 10.6.1) and then report the results obtained (Section 10.6.2).

10.6.1 Adaptation Mechanism

Algorithm 1 describes the adaptation mechanism sketched in Section 10.3, which exploits the simulator to iteratively monitor the generation outcomes of the speaker. Given the visual context $\mathbf{v}$, the initial hidden state of the speaker's

[4]We observed that using the Gumbel-Softmax trick (Jang et al., 2017) led to unstable behaviour.

[5]Possibly because the general knowledge space is bigger, it could also be more difficult to model a general listener than a domain-specific listener with a limited knowledge space.

decoder h_0 and the currently planned utterance u_t, the simulator makes a prediction for the listener's selection.[6] We calculate the Cross-Entropy Loss between the simulator's prediction and the true target. We use the gradients flowing back from this loss to update h_0 with the Adam optimizer. That is, adaptation is performed by backpropagating the loss to modify only the initial hidden state of the speaker's decoder. Based on the updated h_0, the language model generates a new utterance to be reviewed by the simulator. The mechanism stops when: either (1) the simulator predicts that the listener will choose the gold target image; or (2) when the maximum number of adaptation steps is reached (st_{adp}). At each step, we reset the random seed to ensure that the changes in the sampling of the words are only attributable to the updates to h_0, showing the effects of adaptation directly without being confounded by the stochastic nature of sampling.

Algorithm 1: Adaptation Mechanism

Input: st_{adp} : maximum number of adaptation steps
lr_{adp} : learning rate for adaptation
$seed$: random seed
Data: h_0 : speaker's initial hidden state
$\mathbf{v}$: visual context
t_g : true target

$i \leftarrow 0$
while $i \leq st_{adp}$ **do**
  $set\_seed(seed)$
  $u_t = Speaker(\mathbf{v}, h_0)$
  $o_{sim} = Simulator(\mathbf{v}, u_t, h_0)$
  $t_{sim} = \arg\max(o_{sim})$
  if $t_{sim} == t_g$ **then**
    break
  $loss = CrossEntropy(o_{sim}, t_g)$
  $h_0 = backprop(loss, h_0, lr_{adp})$
  $i \mathrel{+}= 1$
$t_l = Listener(\mathbf{v}, u_t)$

10.6.2 Results

We evaluate whether our approach leads to increased communicative success, quantified in terms of listener resolution accuracy. We report the results for the three settings described in Section 10.3. For each of the three modules involved

[6] To avoid excessive language drift and help regularize utterance generation, at inference time we condition h_0 with the previous gold utterance referring to the target image in the current dialogue (if it exists), as done by Takmaz et al. (2020a). This resonates with precautions taken in other plug-and-play approaches against text degeneration (Dathathri et al., 2020).

in these settings, we provide an evaluation card (Hupkes et al., 2023) to clarify the nature of our generalization tests in the Appendix E.3.

Baseline Table 10.2 provides a breakdown of resolution accuracies per type of domain-specific listener in the setting without simulation; Table 10.3 shows the averages. Not surprisingly, the results obtained with generated utterances are lower than those reported in Section 10.5.2. However the patterns are the same: when the speaker agent refers to an image within a domain known to the listener (IND), the average resolution accuracy is 52.30%; communication however breaks down in out-of-domain instances, where the average OOD score is 19.06%, close to random choice.

	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

Table 10.2: Resolution accuracy in the Baseline setting. Rows indicate the listener domain and columns the evaluation domain. Shaded cells show IND accuracy. Averages across 5 seeds. Full table with SDs in Appendix E.2.2.

Self-aware adaptation As shown in Table 10.3, with the added capability to simulate and adapt to a generic listener, we observe an increase in IND resolution accuracy (52.30% vs. 65.09%). Yet, this setting does not help to bridge the knowledge gap between speaker and listener: when the input is OOD for a domain-specific listener, adaptation with a general simulator does not lead to higher communicative success (19.06% vs. 19.11%).

	Baseline	**Self-aware**	**Audience-aware**
OOD	19.06 ± 0.47	19.11 ± 1.12	26.74 ± 1.48
IND	52.30 ± 1.10	65.09 ± 1.98	71.77 ± 2.16

Table 10.3: Average resolution accuracy for our 3 settings in OOD and IND. Results on the test set over 5 runs.

Audience-aware adaptation When the speaker adapts its utterances by predicting the behavior of a domain-specific listener, we see a significant increase in both IND and OOD (Table 10.3). This indicates that audience-aware adaptation

helps in knowledge asymmetric scenarios, including in IND situations where the agents communicate about a domain known to the listener (65.09% vs. 71.77%). More importantly, while there is certainly room for improvement, the speaker is able to generate utterances that can more often be resolved in OOD (19% vs. 26.74%).

10.7 Analysis

Our experiments show that simulation-based adaptation leads to more successful communication. In this section, we analyze the speaker model and its generated utterances to understand which neural processing mechanisms and which production strategies are behind our main results.

10.7.1 Probing for Domain Information

We begin with an analysis of the neural representations of the speaker model in the *audience-aware* setting. We focus on h_0, the first hidden state of the LSTM decoder. This is the output of the visual encoder on which the simulator module intervenes in order to adapt the speaker's utterance plan. Because h_0 is the result of encoding a target image (within a visual context), we expect it to carry information about the semantic domain of the image. If it was not able to differentiate visual domains, it would be very unlikely to successfully adapt to domain-specific listeners. We test this hypothesis using diagnostic probing (Adi et al., 2017; Conneau et al., 2018a; Hupkes et al., 2018). We train a logistic regression classifier on a 70% of hidden states h_0 collected from the speaker when at test time, and then we assess whether it can predict the image domain corresponding to the remaining 30% of the hidden states. As expected, the probing classifier is able to do so with perfect precision and recall (both equal 1.0) across the 5 visual domains. Using the same approach, we test whether the domain of the *listener* – rather than the image domain – is also encoded in h_0.[7] Our hypothesis is that this should not be the case: before the simulator kicks in, the speaker model has no information on the listener's domain-specific knowledge. Probing accuracy scores vary between 0.13 and 0.16 across domains (the random baseline is 0.17), indicating that indeed the speaker's hidden state does not carry listener information before adaptation.

As the simulator activates, the original h_0 is updated for a maximum of st_{adp} adaptation steps. We now look at the updated hidden states $h_0^1, \ldots, h_0^{st_{adp}}$ and test whether their encoding of the image and the listener domain changes with adaptation. First, we use the probing classifier previously trained to predict image domains from h_0 to test the adapted hidden states. We find that the encoding

[7] We train a logistic regression classifier on the 70% split of the h_0 but this time using as label the domain of the listener. We then evaluate whether the classifier can predict the listener domain in the 30% probing test set.

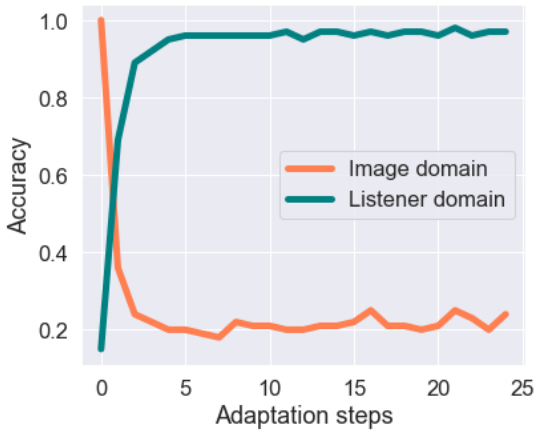


Figure 10.2: Probing accuracy for image domain and listener domain predictions over adaptation steps. The 0-th step corresponds to the non-adapted h_0.

of the image domain deteriorates with domain-specific adaptation (Figure 10.2). Then, we probe $h_0^1, h_0^2, \ldots$ for listener information and we show that the listener's domain can be predicted almost perfectly from the adapted h_0 after only three adaptation steps (Figure 10.2).[8] Taken together, these observations indicate that the neural processing mechanism that leads to more successful interaction is one by which information about the semantic domain of the visual context is replaced by information on the domain of the listener – and one which only requires a few gradient updates.

10.7.2 The Speaker's Adapted Vocabulary

We analyze macro-level properties of the corpus of adapted utterances as compared to the utterances generated in the simulator-less baseline setting. We compute type-utterance ratio and type-token ratio over adaptation steps to monitor the relative size and the variety of the vocabulary as the speaker uses its simulator module. As Figure 10.3 shows, after an initial drop for the first 1-3 adaptation steps, type-utterance ratio and type-token ratio increase substantially with respect to the non-adapted utterances (and to the gold referring utterances). The speaker vocabulary becomes much more diverse. What remains rather stable throughout adaptation, instead, is the unigram part-of-speech distribution (Figure E.2 in Appendix E.5). While, after the first adaptation step, the difference in POS usage is notable (e.g., less punctuation, more nouns), only proper nouns and determiners show substantial changes in relative proportions, with proper nouns increasing and determiners decreasing over time.

[8]For this analysis, we train and test one probing classifier for each adaptation step. Using the classifier trained on h_0 would not make sense as we showed that it is not possible to extract listener information from non-adapted representations.

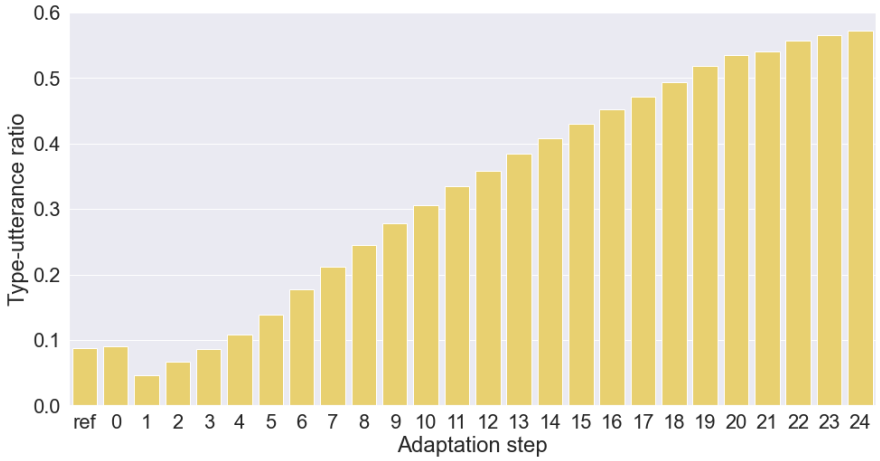


Figure 10.3: Type-utterance ratio across adaptation steps (type-token ratio in Figure E.1, Appendix E.5). Human gold utterances (*ref*) and non-adapted utterances (0) also shown.

10.7.3 Adaptation Strategies

The trends observed so far characterize the effect of adaptation across steps, but they do not differentiate between successful and unsuccessful adaptation. In Figure 10.4, we split adapted utterances (the ones actually generated by the speaker when it believed its utterance would be successful) according to whether they lead to a correct listener guess. We observe that more successful utterances contain words with a lower age of acquisition[9] (AoA, $t = -28.88$, $p < 0.001$), they show a lower rate of lexical choice from the target image vocabulary ($t = -28.76$, $p < 0.001$), and a higher rate of words from the listener vocabulary ($t = 5.88$, $p < 0.001$). The average AoA in an utterance increases with adaptation steps (see Figure E.3 in Appendix E.5), suggesting that the excessive abstractness of the descriptions may be behind the limited gains we observe with adaptation.

10.7.4 Qualitative Inspection

In Figure 10.5, we provide examples of adapted sentences from the test set to demonstrate how the audience-aware adaptation strategies affect the lexical choices made by the language model. In the top example, the image domain is 'food'; however, the listener was trained on the 'indoor' domain. We see that the speaker moves away from generating detailed mentions of food to including a word related to the listener's own domain, *bookshelves*. In the bottom example where the listener has only been exposed to the 'food' domain and the image domain is 'outdoor', the model avoids mentioning the *truck*. Instead, it produces an

[9]Age of Acquisition is a psycholinguistic measure expressing the age at which a word is typically learned. We use the ratings by Kuperman et al. (2012); they range from 0 to 25.

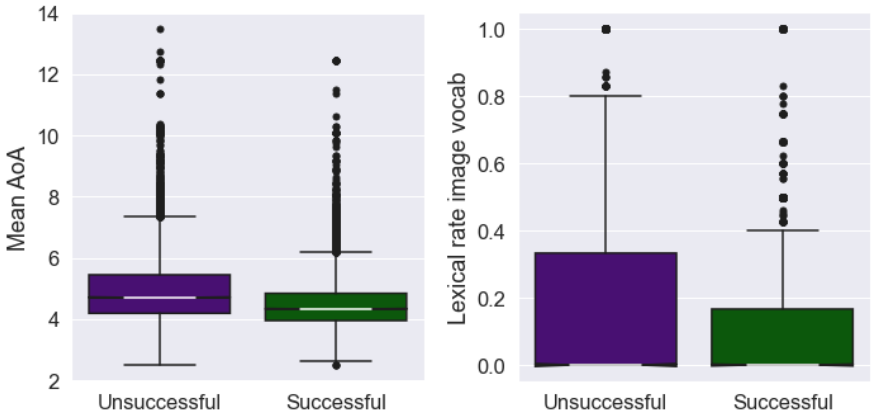


Figure 10.4: Factors affecting the success of an adapted utterance, age of acquisition (left) and % of words in an utterance belonging to the target image domain (right).

utterance containing a prominent color in the image, i.e., pink, and some visible entities that belong to the listener's domain, namely, *donuts*. These observations suggest that the model exploits various adaptation strategies.

In the whole set of adapted utterances, we observe comprehensible sentences; however, there is also a large number of less fluent, unnatural ones. As we do not use pretrained large language models, sometimes, the speaker's initial utterances themselves are not fluent. The dynamics of adaptation may further exacerbate this situation and lead the language model towards generating unnatural utterances. Such utterances may not be understood by human listeners; yet, they could make sense to artificial listeners. In order to ensure that the adapted utterances are comprehensible to humans, further precautions may be needed, such as incentivizing the generative model to keep the adapted utterances grammatical and fluent, possibly with the aid of human feedback.

10.8 Conclusion

We focused on a standard reference game—a speaker produces an utterance, and a listener uses it to pick the referent from a visual context. However, our setup is *asymmetric*—the speaker has general semantic knowledge, while the listener has little knowledge of all domains but one (e.g., *food*). Such a setting is a perfect scenario for studying adaptation, i.e., the common process in human communication by which a speaker tunes its language to that of a listener to achieve communicative success. We modeled this mechanism using a *plug-and-play* approach to controllable text generation: the speaker's output is conditioned on the effect of the planned utterance on the listener, as predicted by an internal simulator. Our results show that *speaking the language of a listener* increases communicative suc-

Target:FOOD. Listener domain:INDOOR
Gold (not adapted): *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?*
Adapted: *must bookshelves in the salad?*

Target:OUTDOOR. Listener domain:FOOD
Gold (not adapted): *I have the pink food truck again ... white shirt lady*
Generated (not adapted): *girl at black phone, red truck, brown hair, pink*
Adapted: *pink donuts*

Figure 10.5: Examples showing how audience-aware adaptation changes the generated utterances. For simplicity, we only show the target images and not the whole visual contexts. We report the final adapted utterances when the adaptation mechanism stops because the simulator predicts that the listener will select the correct image.

cess. Through adaptation, the speaker's language becomes less tied to the input domain and more tied to the listener's vocabulary, revealing that audience-aware adaptation can be realized without irreversible changes to generation models.

Our approach and findings pave the way for pragmatic models that can account for different communicative scenarios. Future work may study adaptation to other dimensions such as age group or sociocultural background. Moreover, adaptation could be explored in multiple 'directions'—in our setup, only the speaker adapts. We also simplify the setup by abstracting away the online process that leads to the simulation of the listeners. It would be beneficial to allow the simulators to learn to predict listener behavior during interaction in an online manner. Finally, our approach could be applied to other and possibly more complex communicative tasks, perhaps in conjunction with a mechanism leveraging human feedback via reinforcement learning.

Chapter 11

Conclusion

Humans interact with the world and other fellow human beings through multi-faceted visuo-linguistic processes, producing and receiving a vast array of signals. In this thesis, I have focused on a select subset of such processes to explore the modeling of the interplay between vision and language in neural networks. I have examined these processes in two parts. In Part One, I have investigated how to model the process of generating verbal descriptions for images, while taking human eye movements into account. Then, I looked into the variation in human signals that occurs in visuo-linguistic processes. Additionally, I have explored how to model the reading process, predicting eye movements over text. In Part Two, I have moved on to visually grounded dialogue, and explored the modeling, quantification, and adaptation of referring utterances. I summarize the findings of these parts in the next section.

11.1 Findings

Part One The central theme of this part was human gaze in linguistic processes. I explored whether incorporating gaze improves image description generation models (Chapter 4), whether pretrained encoders can account for the variation in visuo-linguistic human signals (Chapter 5), and whether pretrained multilingual models can predict human reading behavior in the form of eye movements (Chapter 6).

The results in this part have shown that human gaze constitutes an integral part of vision-language tasks, and that incorporating eye movements into state-of-the-art models for image description generation is beneficial. In this way, the models generate descriptions that are sequentially and semantically in line with what humans uttered. This finding affirms the contribution of the information relayed by human gaze to models of language production conditioned on visual stimuli, at the same time, shedding light on the inner workings of visuo-linguistic processes in humans.

I have also shown that the complex relationship between vision and language leads to intriguing patterns of variation in human signals. Furthermore, I have found that such variation is encoded by pretrained multimodal models to a limited extent, with potential room for collecting more human signals and taking them into consideration while developing such models. As discussed throughout the thesis, these outcomes bear significance both for AI and cognitive science, unveiling the variation in human signals and the power of pretrained models in capturing such variations. Such findings can be informative about human expectations, the complexity of stimuli, and uncertainty. I have also shown that, with minor extensions, multilingual models can account for human reading behavior, suggesting that similar approaches could be implemented for multimodal and multilingual models for predicting other cognitive signals.

Part Two In this part, I have focused on referring utterances in visually grounded dialogue. I have explored the listener-speaker dynamics in such settings, and how the utterances evolve as common ground is built, with a focus on the strategies adopted by humans.

In Chapter 8, using a pretrained multimodal model, I have quantified referring utterances in terms of their descriptiveness and discriminativeness in reference games. The results have yielded patterns resonating with human strategies as a conversation unfolds: humans produce shorter utterances that are less descriptive over time while maintaining discriminativeness. Such strategies involve leveraging the dialogue history and distilling discriminative words for the sake of task performance and communicative success.

I have modeled the process of generating and resolving referring utterances in conversational and visual contexts in Chapter 9, and found that dialogue history serves as a memory of the common ground, and guides processes pertaining to referring utterances. The model that generates the first and subsequent references to an image by keeping track of the dialogue history produces effective and human-like utterances.

Finally, in Chapter 10, I have investigated how to adapt referring utterance generation models when they encounter listeners with knowledge gaps. To tackle this issue, I have proposed modules that simulate listeners, following Theory of Mind. By developing a plug-and-play methodology to adapt the generation models, I have enabled these models to perform audience-aware adaptation, which led to increased performance in communication during reference games.

All the results in this thesis substantiate the importance of accounting for *human* visuo-linguistic processes when developing, evaluating, and utilizing models performing *artificial* visuo-linguistic processes. In this thesis, I have taken steps in this direction, embedding multimodal processes in human cognition into computational models.

11.2 Limitations and Future Work

Part One A limitation of the studies detailed in this part lies in the relatively small size of the datasets, given the scarcity of datasets involving human signals in the literature. Therefore, I mainly opted for starting from models pretrained with larger datasets. In Chapter 4, I trained an initial model on a large machine-translated dataset, and in Chapter 6, I used pretrained multilingual models and only trained lightweight adapters. In Chapter 5, I resorted to approaches that circumvent training.

Another data-related limitation was that, in Chapters 4 and 5, I only conducted experiments for a single language. Since the cross-modal interaction between vision and language could show some variation based on the properties of the languages (i.e., word order and morphological constraints), leading to variation in visual attention and structural choices (Norcliffe and Konopka, 2015; Myachykov et al., 2011), the findings might differ based on the languages of the datasets and the pretrained models. For future work, I recommend collecting larger datasets containing eye-tracking data to model gaze in multilingual setups. Such data is lacking, in particular, for multimodal settings as opposed to reading eye-tracking datasets, which are more widespread.

Regarding Chapter 5, which explored the variation in human signals, it would also be informative to explore other models and tasks, as well as explicit, discrete features that would contribute to predicting visuo-linguistic variation. Regarding the data, there could be noise in human signals, and the preprocessing steps I took could also partly affect the findings. Investigating the variation in gaze before/after speech onset with participant-specific analyses could also reveal interesting dynamics. As the dataset contains descriptions from 45 participants, with an average of 15 participants describing each image, a different pool of participants (in particular, of a different size) may produce disparate results. A larger corpus may also allow for the training and fine-tuning of models. This is a line of work I have only partially explored in that chapter. A probing approach where I trained lightweight layers on top of image representations yielded even lower correlation coefficients and higher losses than the similarity-based approach reported in the chapter.

When training and fine-tuning models for visuo-linguistic processes, a weakness of the gaze models described in Chapter 4 was that in the absence of gaze data for an image, these models could not generate a caption accurately. Hence, instead of feeding gaze data into the models as input, in future work, it would be advantageous to equip the models with the ability to predict the next gaze sequentially along with the next word. Similarly, when modeling multilingual reading behavior in Chapter 6, unidirectional predictive models could also be utilized to compare against bidirectional models to investigate the cognitive plausibility of these architectures.

Part Two A limitation of this part is that, although I use data from dialogues, I do not model interactive, collaborative reference using the full dialogues within PhotoBook games. Given the novelty of the approaches I investigated, I made certain assumptions and simplified the setups. Instead of using full dialogues, I used automatically extracted referring utterances, which may have induced some level of noise in chapters 9 and 10. Although a human-annotated portion of PhotoBook exists, as used in Chapter 8, this subset does not cover the whole dataset. Future work can collect complete human annotations with referent-utterance pairs in visually grounded dialogue, ideally including other phenomena such as dialogue acts. In this way, the interaction between language and vision can be analyzed more deeply.

As I do not utilize full dialogues, in Chapter 9, I model the generation and resolution of a single referring utterance. Similarly, in Chapter 10, I do not model continual mutual adaptation during a dialogue. Instead, I focus on the speaker's adaptation to the listener in a single turn. Future work should ideally model the whole dialogue building on a more interactive adaptation setup than the plug-and-play approach, which still requires training simulators per listener type.

I train the models from scratch using the PhotoBook data, and do not use state-of-the-art large pretrained vision-and-language models that are commonly based on transformers nowadays. I opted for this setup since it is more aligned with my research questions. This setup enables controlling the domain-specificity of the linguistic knowledge of the models. I also acknowledge the imbalance in the set sizes of the domains, as well as the possible lexical and visual overlaps in the samples across domains. The overlaps may facilitate the adaptation of certain sentences from one domain to another (asymmetry is not controlled in a fine-grained manner). This is not uncommon in human communication, so ideal models should be able to perform well even in such a scenario.

Future work should study a broader set of aspects of human communication in setups that are more faithful to authentic dialogue, where an interlocutor can be both a listener and a speaker. Every interlocutor should be able to continually observe the others, and adapt their language production and understanding accordingly during the whole conversation incrementally, which is a line of work I did not delve into due to time and compute constraints.

11.3 Final Remarks

While describing an image and participating in visually grounded dialogue, multiple stages of actions need to be performed accurately. Current state-of-the-art models come short in some instances and sometimes behave unexpectedly, exposing the inconsistencies and the chasm between human and machine intelligence when processing multimodal inputs.

The aim of this thesis has been to explore various facets of visuo-linguistic

processes in humans and to model them within the framework of deep learning. The results have revealed the importance of modeling multimodality in deep neural networks in light of findings from psycholinguistics and cognitive science. I have found that models of image description generation and visually grounded dialogue can benefit from a deeper understanding of human cognition.

Although current trends in AI seem to indicate that increasing the scale of data and models would be the best way forward, I believe that investigating the complex interaction between vision and language in humans can inform how best we can integrate them into deep neural networks. To this end, I suggest collecting more multimodal data from humans across languages. Although such datasets may not amount to a scale that is enough to train or fine-tune deep learning models, they can provide insights into the intricacies of human cognition in multimodal settings, leading to the development of better models.

Regarding conversational settings, incorporating more aspects of the listener-speaker dynamics in novel ways could prove beneficial. Contemporary conversational models appear to be generating quite good outputs already; yet, there is still room for improvement. This is particularly the case for developing generative models that interact with a diverse array of users.

I hope this thesis motivates further work in the direction of exploring computational approaches leveraging human signals, which could simultaneously benefit the development of better AI models and provide insights into human cognition itself.

Appendix A

Appendix to Chapter 4

A.1 Data Preprocessing

This appendix provides details on the pipeline used to time-align audio and descriptions. After processing a transcribed caption, we insert it as a grammar rule into a Java Speech Grammar Format (JSGF) file to be fed into CMUSphinx. As CMUSphinx supports English by default, we incorporated into the tool the phonetic and language models and the dictionary for Dutch as provided by the developers of CMUSphinx.[1]

Some words in our JSGF files were not in the VoxForge Dutch phonetic dictionary of CMUSphinx, which lists lexical items and their corresponding pronunciations in a format similar to ARPABET, adapted for Dutch.[2] To overcome this problem, we used eSpeak[3] to obtain the International Phonetic Alphabet (IPA) transcriptions of such out-of-vocabulary words. We obtained the set of IPA symbols existing in the transcriptions of out-of-vocabulary words and the set of ARPABET symbols in the dictionary. Then, a native speaker of Dutch, who is also a linguist, manually produced a mapping from these IPA symbols to ARPABET symbols of Dutch phonemes.[4] Given this mapping, we automatically converted out-of-vocabulary tokens into the required format and appended them to the dictionary. A similar approach was also followed for numbers in numeric notation and certain English words.

For some audio-caption pairs, the tool could not find an alignment matching the grammar. We turned off the noise reduction and silence removal parameters, and experimented with parameters related to beam decoding in CMUSphinx to

[1] `https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/Dutch/`

[2] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[3] `http://espeak.sourceforge.net/`

[4] The mapping from IPA symbols to ARPABET symbols is provided in our GitHub repository.

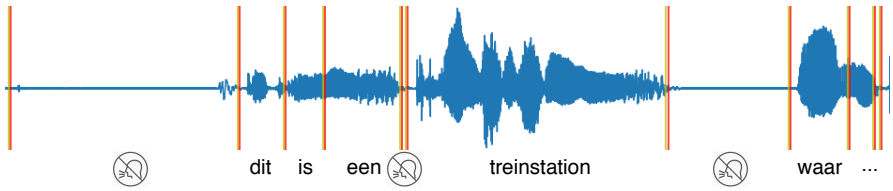


Figure A.1: Temporal alignment of words in a transcribed caption and the corresponding audio file.

allow for a maximal number of complete alignments. However, we had to exclude some captions with unintelligible words, particularly at the beginning or in the middle of the audio, since such an issue disrupts the alignment procedure.

Considering possible inter-participant differences in terms of pronunciation, the quality of audio files, and possible noise in the background of recordings, we assume that the time intervals of the words we obtained after these preprocessing steps are approximate indicators. Although there might be a few cases where the alignment is not quite accurate, we find this way of obtaining utterance timestamps reliable in general. An example audio-caption alignment is shown in Figure A.1.

A.2 SSD: Further Details

Type	Description	SSD	*gr*	*rg*
R	een dubbeldekker bus	1.41	2.82	0.00
G	een dubbeldekker bus in een stad			
R	een dubbeldekkerbus in uh in engeland	2.64	2.72	2.55
G	een dubbeldekker bus in een stad			
R	een rustige straat met een bus tegemoetkomend naar <unk> nummer 43	5.87	4.31	7.43
G	een dubbeldekker bus die op een weg rijdt			
R	een bus met lijn 43 die aan het rijden is waarvan uh de bus uit twee <unk> bestaat	8.62	0.43	16.81
G	een dubbeldekker bus			

Table A.1: Examples of SSD scores for several descriptions generated (G) by GAZE-2SEQ compared to the reference description (R). *gr* and *rg* indicate the direction of the calculation. Lower SSD scores are better.

SSD is the average of two terms, *gr* and *rg*, which quantify the overall distance between a generated sentence (G) and a reference sentence (R). Eq. A.1 (identical to Eq. 4.1 in Section 4.4) shows the calculation from G to R and Eq. A.2 from R to G:

$$gr = \sum_{i=1}^{N} cos(G_i, R_s(i)) + pos(G_i, R_s(i)) \quad [A.1]$$

$$rg = \sum_{j=1}^{M} cos(R_{\mathrm{j}}, G_s(j)) + pos(R_{\mathrm{j}}, G_s(j)) \quad [A.2]$$

N and M refer to the number of tokens in G and R, respectively. Cosine and positional distances are computed between the i_{th} element of G and another token, which is the most semantically similar word to G_i in R. $R_s(i)$ is the most semantically similar word to G_i and $G_s(j)$ is the most semantically similar word to R_j:

$$R_s(i) = \arg\min_{j}(cos(G_{\mathrm{i}}, R_{\mathrm{j}})) \quad [A.3]$$

$$G_s(j) = \arg\min_{i}(cos(R_{\mathrm{j}}, G_{\mathrm{i}})) \quad [A.4]$$

Table A.1 shows some example descriptions generated by the GAZE-2SEQ model and corresponding references for a single image. We report the overall SSD scores along with gr and rg values separately.

A.3 Data Split Statistics

Table A.2 lists the number of images belonging to each split after we divide the DIDEC corpus (description-view partition) with respect to the images. In addition, the total number of captions in each split is provided.

	train	**val**	**test**	**total**
Images	247	30	30	307
Captions	3658	444	446	4548

Table A.2: Number of images and captions.

The number of human descriptions per image varies in DIDEC, and as we also removed some captions during preprocessing, images do not have an equal number of captions. Therefore, we report the average number of captions per image for each split, as well as their range, in Table A.3.

	train	**val**	**test**	**overall**
Avg	14.81	14.80	14.87	14.81
Min	11	12	13	11
Max	16	16	16	16

Table A.3: Number of captions per image.

A.4 Reproducibility

We implemented and trained our models in Python version 3.6[5] and PyTorch version 0.4.1.[6] All models were run on a computer cluster with Debian Linux OS. Each model used a single GPU GeForce 1080Ti, 11GB GDDR5X, with NVIDIA driver version 418.56 and CUDA version 10.1.

Pretraining with the translated MS COCO dataset took approximately 5 days. No-Gaze and Gaze-Agg took around 1.5 hours, and Gaze-Seq and Gaze-2Seq models took 2 hours to fine-tune over the pretrained model.

Since the pretrained model and the fine-tuned No-Gaze, Gaze-Agg and Gaze-Seq models use essentially the same architecture, they have an equal number of parameters: 85 million. Gaze-2Seq has more parameters due to the addition of the Gaze LSTM: 100 million.

In all the models, the biases in linear layers were set to 0, and the weights were uniformly sampled from the range (-0.1, 0.1). Embedding weights were initialized uniformly in the range (-0.1, 0.1). LSTM hidden states were initialized to 0.

Below, we give details regarding the manually-tuned hyperparameters.

A.4.1 Hyperparameters for Pretraining

We experimented with learning rate (0.001, 0.0001), dimensions for the word embeddings and hidden representations (512, 1024), and batch size (64, 128). The best pretrained model is selected based on its CIDEr score on the validation split of our translated MS COCO dataset, with an early-stopping patience of 20 epochs. We use a learning rate of 0.0001, optimizing the Cross-Entropy Loss with the Adam optimizer. The batch size is 128. The image features have 2048 dimensions and the hidden representations 1024. The generations for the validation set are obtained through beam search with a beam width of 5.

A.4.2 Hyperparameters for Fine-tuning

We experimented with the same set of hyperparameters as in pretraining. The details of the hyperparameters for the selected models were given in the main text. We select the models separately based on CIDEr scores and SSD scores. We train each model type with their selected configuration with 5 different random seeds to set the random behavior of PyTorch and NumPy. We also turn off the cuDNN benchmark and also set cuDNN to deterministic.

[5] `https://www.python.org/downloads/release/python-360/`
[6] `https://pytorch.org/`

Appendix B

Appendix to Chapter 5

B.1 Data Preprocessing

We use spaCy to extract the first noun of each description. The numbers of errors in terms of lemmatization and POS-tagging are as follows when using the small, medium, and large spaCy models for Dutch, respectively: 33, 32, 23 mistakes in the full descriptions, and 3, 2, 2 for the first nouns. As the utterances sometimes contain incomplete sentences and disfluencies, POS-tagging may not be reliable in such cases, especially in the later parts of the utterances. However, the large model was reliable both for full descriptions and the first nouns. Hence, we chose to use the data processed by the large model. The model was not able to tag any nouns in 7 descriptions; for those, we use the <unk> token as a placeholder starting point. We also skipped nouns such as 'photo' ('a photo of a car'), 'number' (as in 'a number of cats'), 'couple' (as in a couple of kids).

B.2 Distribution of Speech Onsets

The histograms of the mean speech onsets and their standard deviations reveal non-normal distributions, as illustrated in Figure B.1.

B.3 Participant-Based Correlation Analysis

To have a better understanding of speaker-specific dynamics, in addition to calculating statistics per image, we also look into per-participant statistics. Each participant describes around 100 images, each with a possibly different speech onset. We calculate the correlation between a participant's speech onsets and the BLEU-2-based linguistic variation score of the corresponding images. In 24 out of 45 participants, we find significant moderate negative correlations. All 45 participants have negative correlation coefficients, indicating that all participants tend

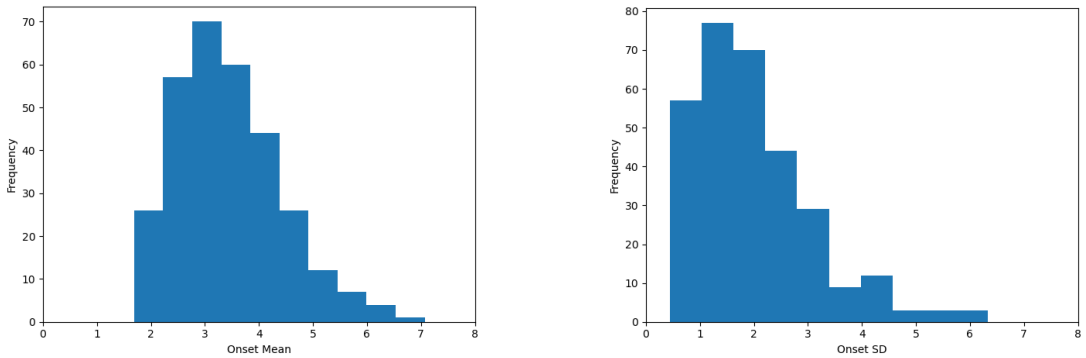


Figure B.1: Distributions of onset means and standard deviations (SD) for the images in the whole dataset.

to start describing an image earlier if that image elicits less linguistic variation across speakers. This suggests that although there can be speaker-specific and contextual factors, the features of an image can also have an overarching effect on the behavioral responses across speakers, and may allow for the prediction of such responses.

B.4 BERTje-based Variation in Descriptions

We inspect linguistic variation by comparing the representations of the descriptions extracted using a Dutch BERT model (BERTje; de Vries et al., 2019). To calculate variation based on BERTje, we utilize the last hidden state corresponding to the [CLS] token for each description as the representation. Then, for each image, we calculate the pairwise cosine similarities between these representations. The average of these similarities is assigned as the variation found in the descriptions of an image. This method yields scores in the narrow range of $0.69 - 0.86$, which indicates semantically quite similar descriptions. Since most descriptions have semantics suitable for the corresponding image, the variation in the semantic space is not substantial. Between BERTje-based variation and speech onsets, we reveal a slight negative correlation (Spearman's $\rho = -0.212, p < 0.01$). The standard deviation of speech onsets is even less correlated with BERTje-based variation (Spearman's $\rho = -0.151, p < 0.01$).

B.5 More Analyses on Linguistic Variation Metrics

We also combine BERTje- and BLEU-2-based variation scores by taking their mean. This metric yields correlations comparable to the ones achieved by the

BLEU-2 version, with a moderate increase in the correlation to the starting point variation and mean onset, yet a decrease in the correlation to gaze variation. For the sake of simplicity, we opt for the BLEU-2 version.

We also compare the BLEU-2-based metric against human evaluations for a different dataset provided by Jas and Parikh (2015), which achieves a significant correlation (Spearman's $\rho = -0.40, p < .001$), albeit to a moderate extent. Jas and Parikh (2015) propose a metric that achieves a stronger correlation ($\rho = 0.72$). Note that the provided human annotations were obtained through 3 annotators evaluating sentence similarities without looking at the images (comparing only 2 sentences at a time). In our dataset, using our metric, we compare 1 description against 14. As a result, the procedure for human annotations may not be well-aligned with our method (i.e., our metric compares 1 sentence against 4 for their dataset, as each image has 5 descriptions).

B.6 Correlation between Human Signals of Variation

We illustrate the correlation between the mean onset and the BLEU-2 scores of full descriptions in Figure B.2.

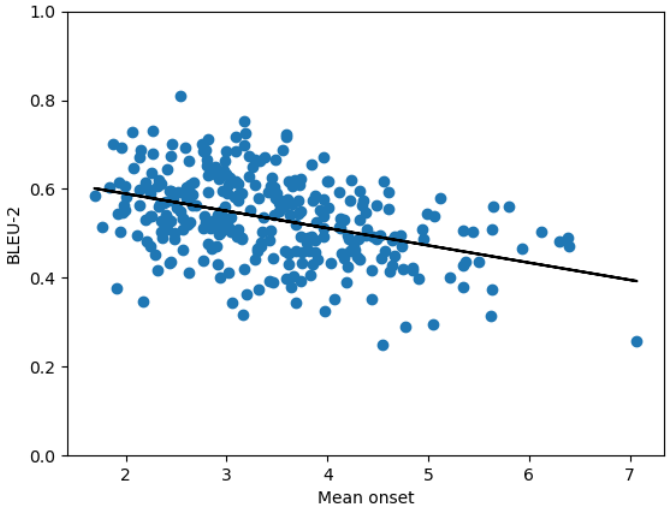


Figure B.2: Correlation between mean onset and BLEU-2.

Appendix C

Appendix to Chapter 6

C.1 Data Preprocessing

We use the XLM-RoBERTa tokenizer containing 250002 tokens. When converting the words into IDs, the tokenizer maintains the cases of the words, which could provide crucial information regarding human reading behavior. However, the way the tokens were presented to the readers differs from how the tokenizer would partition a given sentence. For instance, in the data, we see full stop appended to the last word or '(1917-1919)' as a single entry. For such cases, the tokenizer yields multiple wordpieces per token. We assign the eye-tracking feature values of the full entry to each of its wordpieces, and during training and validation, we include them in the loss separately. For the test set predictions, we calculate the average of the predictions for the wordpieces and assign it as a single prediction for the whole entry.

We combine the token entries with the same sentence ID into a single sentence. Since the sentences do not include start- and end-of-sentence tokens, we also add such special tokens where necessary. In addition, we pad or truncate the input to maintain a total wordpiece length of 200. For all special tokens, we assign '-1' as the dummy eye-tracking feature value.

C.2 Reproducibility

We use AdapterHub version 2.2.0 based on HuggingFace Transformers version 4.11.3.[1] We implement and train our models in Python version 3.7.11 and PyTorch version 1.10.1.[2] All models were run on a computer cluster running Debian Linux OS, with 4 NVIDIA GeForce GTX 1080 Ti GPUs with driver version 470.103.01 and CUDA version 11.4. Below, we detail the hyperparameters.

[1] `https://huggingface.co/docs/transformers/`
[2] `https://pytorch.org/`

C.2.1 Hyperparameters

For each model, we have performed hyperparameter search for learning rate (0.001, 0.0001, 0.00001, 0.00002) and batch size (4, 8, 16, 32). All the models were trained up to 50 epochs. [3] We saved the best model based on the validation MAE per epoch and ran random initializations of the best model with 4 different seeds. The adapters were optimized using the AdamW optimizer (Loshchilov and Hutter, 2019) with respect to MSELoss following a linear learning rate schedule. In Table C.1, we provide the hyperparameters of our best models for Subtask 1 and Subtask 2.

Model	LR	Batch size	Seed
EN stack	0.0001	4	42
ZH stack	0.001	4	8
DE stack	0.001	8	42
HI stack	0.001	4	42
RU stack	0.001	4	8
NL 1 new	0.001	4	42

Table C.1: Hyperparameters for our best submission for Subtask 1 (Language-specific-stack). The DE stack model is also used to obtain our best results for Subtask 2. LR: Learning rate.

C.3 Additional Results

RoBERTa + NER Our first submission to Subtask 1 was built on RoBERTa-base (Liu et al., 2019),[4] with a Named Entity Recognition (NER) adapter trained on the CoNLL2003 dataset[5] (Poth et al., 2021; Tjong Kim Sang and De Meulder, 2003). We used the NER adapter as we noticed a lot of named entities in the data. In this setup, we remove the NER token classification head and create a token-level regression head. The head is trained from scratch, and the NER adapter is fine-tuned. The results revealed that this setup already improves over the mean baseline across all features (MAE = 4.0317, our first submission). Although RoBERTa is monolingual (English) and its vocabulary is much smaller than XLM-R's vocabulary (50265, also its tokenizer converts non-Latin scripts into unintelligible wordpieces), this model seemed to work quite well. However,

[3] It is possible that a higher epoch cap could produce better results; however, in most cases, we observed declining performance as the number of epochs approached 50.

[4] `https://huggingface.co/docs/transformers/model_doc/roberta`

[5] `https://adapterhub.ml/adapters/AdapterHub/roberta-base-pf-conll2003/`

we wanted to make sure that the wordpieces work properly and that the underlying frozen PLM was exposed to multilingual data, which is why we switched to XLM-RoBERTa.

Language breakdown The details of the language-specific-stack models for Subtask 1 are provided in Table C.2. The majority of these models outperform the corresponding mean baselines computed with respect to the language-specific means (except for the Dutch setup, which does not include a pretrained language-specific adapter).

Model setup	FFDAvg	FFDStd	TRTAvg	TRTStd	MAE	Baseline MAE
EN stack	3.2360	1.9582	6.8383	4.9501	4.2456	5.2736
EN large stack	3.0390	1.9921	6.1242	4.8968	4.0130	
ZH stack	3.1586	3.3608	6.8213	6.6955	5.0091	5.4616
ZH large stack	3.1571	3.4448	7.3876	6.5892	5.1447	
DE stack	0.4304	0.4346	3.7796	2.8918	1.8841	2.8679
HI stack	2.5493	2.7178	5.7471	5.5693	4.1459	4.5668
RU stack	2.6062	2.6443	8.3637	5.5609	4.7938	4.9007
NL 1 new	1.8772	1.5720	3.3467	2.9443	2.4351	2.4176
NL 2 new stack	1.8904	1.5911	3.2836	3.0673	2.4581	

Table C.2: Test set results for Subtask 1 for the XLM-R language-specific models with stacking, broken down into languages. Baseline MAE is calculated with respect to the means of the language-specific data. EN: English, ZH: Chinese, DE: German, HI: Hindi, RU: Russian, NL: Dutch.

Dutch-specific models For Dutch, we only employed a single adapter as we did not have a Dutch-specific adapter pretrained on Wikipedia articles. As a result, we also tried stacking 2 new adapters. This setup yielded slightly worse scores than the former setup. Therefore, we opted to keep the single-adapter model in our submissions.

Large models We also use the large version of XLM-RoBERTa.[6] At the time of the project, only English and Chinese Wikipedia MLM adapters were available on AdapterHub (Pfeiffer et al., 2020b, 2021).[7] For English, the utility of the large model was not substantially high, and for Chinese, the large model caused a decrease in accuracy. These findings suggest that the adapters are able to

[6] `https://huggingface.co/xlm-roberta-large`

[7] EN: `https://adapterhub.ml/adapters/ukp/xlm-roberta-large-en-wiki_pfeiffer/`, ZH: `https://adapterhub.ml/adapters/ukp/xlm-roberta-large-zh-wiki_pfeiffer/`

capture the patterns in eye-tracking features, without the need to resort to larger language models. However, more hyperparameter tuning could be beneficial to explore the capacity of the large models.

C.4 R^2 Scores

In Table C.3, we provide the R^2 (coefficient of determination) scores as reported by the shared task system. The top half lists the results for Subtask 1, and the bottom half for Subtask 2.

Model	**FFDAvg**	**FFDStd**	**TRTAvg**	**TRTStd**	R^2
RoBERTa + NER	0.6963	0.3437	0.3293	0.2677	0.4093
Language-specific-stack	0.7581	0.3689	0.4868	0.3517	0.4914
First wordpiece-only	0.7506	0.3564	0.4836	0.3362	0.4817
Translate train	-13.5708	-3.1490	-6.1914	-5.4032	-7.0786
Translate test - EN (without Provo)	-1.0249	-2.3468	-0.8361	-0.7824	-1.2475
Translate test - DE	-1.2176	-0.1296	-0.4203	-0.4929	-0.5651

Table C.3: R^2 scores for the submissions to Subtask 1 and 2.

Appendix D

Appendix to Chapter 9

D.1 Reference Chain Extraction

We extract reference chains of single referring utterances from the PhotoBook dataset (Haber et al., 2019). Given a dialogue and a target image, a reference chain is comprised of utterances—maximum one per round—that refer to the target image in that dialogue. Due to the size of the PhotoBook dataset, we perform this procedure automatically, with a three-step heuristic method described in the following sections. The chain extraction code is available at `https://dmg-photobook.github.io`.

Extracting dialogue segments The goal of segment extraction is to identify all utterances that may include a description of a given target image. To identify relevant segments, we leverage the participants' recorded actions, i.e., selecting an image as common or different (more details on the available metadata in Haber et al., 2019). When an image is selected by a participant as *common* in a dialogue round, we extract all utterances up to that point in the round as candidate referring expressions. We collect referring expressions for a given image in a dialogue starting from the round when *both* speakers observe it. The speakers are then more likely to have established a conceptual pact (see Section 9.1).

Scoring referring utterances In this second step, we assign a score to each utterance in the extracted segments, indicating how likely it is for that utterance to be a description of a given image. To produce these scores, we use as reference the MS COCO image captioning dataset (Lin et al., 2014) and the Visual Genome dataset of scene graphs (Krishna et al., 2017). All 360 pictures in PhotoBook are taken from MS COCO, so we have access to at least 5 captions for each target image. Instead, the Visual Genome dataset provides detailed scene graphs for 37% of the PhotoBook images.

	Chains	Utterances	Unique utterances	Target images	Image domains	Chain length	Utterance length
Train	11540	28933	27288	360	30	2.51(0.85)	8.71(4.66)
Validation	2503	6252	6009	360	30	2.50 (0.85)	8.82 (4.67)
Test	2482	6155	5876	360	30	2.48 (0.86)	8.77(4.68)
Extracted-20	327	824	807	199	24	2.52 (0.85)	9.50 (4.75)
Gold-20	327	756	740	199	24	2.31 (0.94)	9.47 (4.77)

Table D.1: Descriptive statistics of all portions of the extracted dataset of reference utterance chains. Gold-20 is a set of 20 hand-annotated PhotoBook dialogues, with referent labels linking utterances to the target image they describe (see Section 9.3), whereas Extracted-20 are the reference chains extracted from the same 20 dialogues, as if they were not annotated. Duplicate utterances are due to chance: PhotoBook participants have uttered them in different dialogues, potentially to describe the same target image. Image domains refers to the number of MS COCO image categories covered by a dataset portion; the 360 PhotoBook images come from a total of 30 domains.

To measure the similarity of a candidate utterance to a reference MS COCO caption, we use the BERTScore (Zhang et al., 2020b). We experiment with BERTScore Precision, Recall, F1, and select BERTScore F1. As, in our dialogue setting, utterances often contain lexical material that is not part of a referring expression, we filter out stopwords from both the captions and the utterances. We use spaCy's stop-word list for English, from which we remove numerals and prepositions that encode spatial information.[1] Furthermore, to capture dyad-specific variation in referring language, we add the utterance with the highest BERTScore in a round to the reference set, and use it as an additional caption for the following rounds.

To take into account visual attributes and relationships, for each image we collect attribute tokens $T_A(i)$ (e.g. *leafy*, *tree* from *leafy(tree)*) and relationship tokens $T_R(i)$ (e.g. *man*, *playing*, *frisbee* from *playing(man, frisbee)*) from the Visual Genome dataset of scene graphs. We only consider the intersection $T_{VG}(i) = T_A(i) \cap T_R(i)$ between the sets of attribute and relationship tokens to retain only the most relevant tokens. The set difference $T_{VG}(i*) \setminus \bigcup_{i=1,i\neq i*}^{12}$ between the Visual Genome tokens of the target image and the tokens of the 11 distractors is then used as a reference set. To score an utterance, we compute its METEOR score (Banerjee and Lavie, 2005) with respect to this reference set. For all images annotated in the Visual Genome dataset, the final utterance score is the sum of BERTScore and METEOR.[2]

[1]The English stop-word list is available at `https://github.com/explosion/spaCy/blob/master/spacy/lang/en/stop_words.py` and our edits at `https://dmg-photobook.github.io`.

[2]We implement BERTScore and use NLTK's code for METEOR (`https://www.nltk.org/api/nltk.translate.html`). We set METEOR's alignment penalty to 0 as our references are unordered collections of tokens.

Selecting referring utterances The last step, utterance selection, produces reference chains consisting of single utterances—maximum one per round. As PhotoBook dialogues are made up of five rounds, reference chains will have a minimum length of 1 and a maximum possible length of 5. First, given an extracted dialogue segment, we discard all utterances produced by speakers who do not have that image in their visual context. Then, for each target image in the corresponding dialogue round, we collect a ranked candidate list of n top-scoring utterances. As an utterance can be selected as a candidate for multiple images in the same round, we discard a candidate *(utterance, image)* pair if its score is lower than that of any other *(utterance', image)* pair in the same round. Finally, we pick the utterance with the highest score among the remaining candidates. For some images, all of the n top-scoring utterances are assigned to other images, and with higher scores. This causes a slight decrease in the number of utterances in the extracted dataset. We set $n = 4$ to minimize the number of discarded utterances. Table D.1 reports relevant statistics for the dataset splits of our extracted reference utterance chains.

D.2 Data Processing for Models

We further process the dataset of automatically extracted utterance chains. Every utterance is uniquely identified by the game ID, round number, message number, and the ID of the image that they refer to. From these utterances and their contexts, we build the data we feed into our models.

While providing the 6 candidate images to the reference resolution models, we also keep track of the respective histories of candidates (the last utterance up to that time in the game).

As the distribution of the 6 images and the position of the target are not uniform for each target-context pair, this may constitute a bias in the reference resolution model. Therefore, to overcome this, we shuffle the images in the context for all splits at the beginning of each epoch. In the generation models, this shuffling is done once at the beginning of training for all splits.

D.2.1 BERT Representations

Since utilizing pretrained BERT models and representations has proven to be beneficial to many NLP tasks (Devlin et al., 2019), we also decided to use BERT to encode the linguistic input in the reference resolution models. For this purpose, we use the BERT-base-uncased model and the tokenizer as provided in the HuggingFace's Transformers library (Wolf et al., 2020). The utterances are first encoded into the correct format for BERT models. Afterwards, they go through the BERT model to produce the hidden states that correspond to the representations of each of the input wordpieces. Finally, all utterances are fed into the

reference resolution model in the form of a set of BERT representations.

We also experimented with using BERT-large-uncased model as well as extracting hidden states from multiple layers and aggregating them. Neither option provided further improvements on the results we obtained with the final hidden states from the BERT-base-uncased model. Hence, we opted to use the base model's outputs, where each hidden state is of size 768.

D.2.2 Embeddings from Scratch

For the generation models where we do not use BERT representations, we create a vocabulary of tokens from the training set with the help of TweetTokenizer from the NLTK library.[3] We then map the words that occurred only once in the training split to '<unk>'. This results in a vocabulary of size 2816 (including <pad>, <unk>, <sos>, and <eos>). In addition to these special tokens, we also add <nohs> to point out that there was **no his**tory (no previous utterance) for the target image at that point in the game. This token is utilized in the models that base their generation on the previous utterance. An input of <nohs> means that what the generation model is expected to produce is the very first utterance for that image in the game.

The tokens in all 3 splits are converted to indices using this final vocabulary. For the copy model, we need to keep track of what the actual form of an <unk> token is. For this purpose, we build a full vocabulary from the whole dataset to have access to every word in all splits in their actual surface forms. This vocabulary is of size 5793 (including all 5 special tokens mentioned above).

Since we do not want the generation model to output the <nohs> token, the search space of the decoder does not include this token. The Copy model needs to keep track of unknown tokens in the previous utterance and map the previous utterance using an extended vocabulary so that the decoder would be able to 'copy' from the input itself, rather than only generating words from the reduced vocabulary. Mapped expected next utterance is used in calculating the loss. Actual inputs to the encoder and the decoder still contain unknown words, as we do not maintain special embeddings for the surface forms of each of the unknown tokens.

D.3 Evaluation Metrics

For the evaluation of the reference resolution models, we use accuracy and mean reciprocal rank (MRR) implemented by us. Accuracy is a stricter measure as it is either 0 or 1 for a given instance.

For the generation models, we use the *compute_metrics* function provided in the library at `https://github.com/Maluuba/nlg-eval` to obtain corpus-level

[3] `https://www.nltk.org/api/nltk.tokenize.html`

BLEU, ROUGE, and CIDEr.

We also report BERTScore (Zhang et al., 2020b) for the generation models. To obtain this score, we use the library provided by the authors at `https://github.com/Tiiiger/bert_score` and import the *score* function in our evaluation scripts. We use the BERT-uncased-model, we do not apply rescaling to baseline or importance weighting. The hash code for BERTScore that we used in evaluation is 'bert-base-uncased_L9_no-idf_version=0.3.2(hug_trans=2.6.0)'. We obtain precision, recall and F1 variants of BERTScore.

In the generation models, we apply teacher forcing during training; therefore, a token embedding at timestep t is the embedding of the expected token from the ground-truth utterance. During validation, the models use the embedding of the word they generated in the previous timestep.

D.4 Reproducibility

The models are implemented in Python 3.7.5[4] and PyTorch 1.4.1[5]. In training our models, we use the Adam optimizer (Kingma and Ba, 2015) to minimize the Cross-Entropy Loss with sum reduction.[6]

We experimented with learning rate (0.001, 0.0001, 0.00001), dimensions for the embeddings (512, 1024), hidden and attention dimensions (512, 1024), batch size (16, 32) and dropout probability (0.0, 0.3, 0.5). We selected the best configurations per model type via manual tuning.

We train each model type with their selected configuration with 5 different random seeds, setting the random behavior of PyTorch and NumPy. We also turn off the cuDNN benchmark and also set cuDNN to deterministic.

In all the models, the biases in linear layers were set to 0 and the weights were uniformly sampled from the range (-0.1, 0.1). In the models that learn embeddings from scratch, embedding weights were initialized uniformly in the range (-0.1, 0.1). The hidden and cell states of the LSTMs were initialized with task-related input at the first timestep.

Computing infrastructure The models were trained and evaluated on a computer cluster with Debian Linux OS. No parallelization was implemented; each model used a single GPU GeForce 1080Ti, 11GB GDDR5X, with NVIDIA driver version 418.56 and CUDA version 10.1.

Average runtimes Please see Tables D.2 and D.3. These durations indicate the total approximate runtime of training. The best models are reached in a

[4] `https://www.python.org/downloads/release/python-375/`

[5] `https://pytorch.org/`

[6] Copy model, in fact, uses the Negative Log-Likelihood Loss that receives log-softmax probabilities. This is equivalent to Cross-Entropy Loss with logits.

Model	Runtime
Baseline	1 hour
Proposed	5.5 hours
Ablation	2.8 hours

Table D.2: Resolution: approximate training runtimes.

Model	Runtime
Ref	6.5 hours
ReRef	7.5 hours
Copy	14 hours

Table D.3: Generation: approximate training runtimes.

shorter amount of time.

Number of parameters in each model Please see Tables D.4 and D.5.

Model	Parameters
Baseline	182K
Proposed	8.9M
Ablation	8.5M

Table D.4: Resolution models: number of parameters.

Model	Parameters
Ref	16.1M
ReRef	24.9M
Copy	24.0M

Table D.5: Generation models: number of parameters.

D.4.1 Configurations of the Reference Resolution Models

We select the reference resolution models based on their performance in accurately predicting the correct target among 6 images. We also report MRR, as it also provides further information in terms of the ranking of the correct image among the distractors.

After hyperparameter search, we decided on a batch size of 32, a learning rate of 0.0001, attention and hidden dimensions both set to 512, and a dropout probability of 0.5 for the proposed reference resolution model. We trained the ablation model with the same settings.

D.4.2 Configurations of the Generation Models

Best-performing generation models for each model type were selected based on their performance with respect to the F1 component of BERTScore. We also performed hyperparameter search for the beam width used in decoding, after which we decided to use a beam width of 3. The best-performing model for each model type outperformed the other models in its own category over all metrics.

As revealed by hyperparameter search, all reported generation models use 1024 dimensions for embeddings and 512 dimensions for hidden and attention layers. They all use a learning rate of 0.0001. Ref and Copy models use a batch size of 32 and the ReRef model, 16. Ref and ReRef models use a dropout probability of 0.3, whereas the Copy model yielded better results without dropout.

D.5 Results on the Validation Set

For each model we report in the main text, we also provide the validation set performances in Table D.6 for the generation and Table D.7 for the resolution models.

Model	BLEU-2	ROUGE	CIDEr	BERT-F1	ACC	MRR
Ref	22.40 (1.22)	31.29 (1.56)	41.26 (3.18)	55.24 (1.38)	59.69 (3.48)	74.41 (2.21)
ReRef	45.41 (0.89)	51.14 (0.42)	127.08 (4.17)	67.94 (0.23)	91.70 (1.09)	95.32 (0.70)
Copy	36.44 (0.31)	43.00 (0.35)	104.27 (1.16)	62.93 (0.21)	83.28 (0.77)	90.07 (0.49)

Table D.6: Average metric scores of the 3 generation models on the validation set. We report the average of 5 runs and standard deviations in parentheses. ACC is the reference resolution accuracy of the sentences generated by the generation models, and MRR is their mean reciprocal rank as obtained through our best reference resolution model.

Subset	ACC	MRR	Instances
First	81.85 (0.45)	88.88 (0.29)	2503
Later	88.51 (0.19)	93.33 (0.12)	3749
Overall	85.85 (0.10)	91.55 (0.07)	6252

Table D.7: Validation set scores of the reference resolution model: averages of 5 runs with the best configuration, with the standard deviations in parentheses.

D.6 Linguistic Measures

The linguistic measures used were chosen to quantitatively explore whether artifacts of the compression, reuse, and grounding present in the human utterances,

as well as other human-like linguistic patterns, can be seen in the generated utterances. We compare the performance of the generation models with regards to the similarity of their generated sentences to human traits, namely a) whether there is a change in token use between the first and last mention (Table D.8) and b) whether this relative distance or the values in the first mention differ significantly between human and model references (Table D.9).

In the case of givenness markers, we measure this as the proportion of tokens that correspond to definite (*the*), indefinite (*some, a, an*), and other markers of the existence of shared context (*again, before, one, same, also*) which occur in the utterance. In the case of compression, we measure the lengths of the utterances in terms of tokens and content tokens, i.e., tokens that are not in the stopword list from NLTK version 3.4.5 (Loper and Bird, 2002). We also measure the proportion of content words in an utterance that correspond to nouns, verbs, and adjectives. Finally, for entrainment, examining only later utterances (not the first referent to an image), we measure firstly what proportion of the utterance in question consists of reused unigrams and bigrams from the previous utterance. We also measure within the reused tokens, the proportion of which is made up of nouns, adjectives, and verbs, to discover their relative importance in terms of reuse. These measures can all be found in Tables D.8 and D.9. For these analyses, we compared the generated output from the best seed for each model variant. These were seeds 1, 1, and 24 for the Ref, Copy, and ReRef models, respectively. We report both effect size (d) as measured by Cohen's d, and p-value ($^{*}p < 0.05$, $^{**}p < 0.005$, $^{***}p < 0.001$) for each comparison. We use the SciPy stats package (SciPy version 1.3.3) *ttest_ind* to perform the independent t-test, and our own implementation to calculate Cohen's d effect size.

Additionally, to check general fluency, we evaluate the coherence and vocabulary use of the models in comparison to humans. We measure *Type Token Ratio (TTR)*, the proportion of unique tokens in an utterance. This can capture ungrammatical repetition patterns in the generation, and, if following human trends, should increase in subsequent mentions. Although both models have significantly lower TTR than the human data, ReRef, unlike Copy, shows a significant increase in subsequent mentions, with much higher TTR than Copy, even though both models show similar average utterance length for later utterances *(ReRef: 7.22, Copy: 7.79)*. In terms of vocabulary, for the generated outputs, ReRef has a much smaller *(first: 492, later: 705)* vocabulary than Copy *(first: 1098, later: 1469)*, although these are both much lower than Human vocabulary size *(first: 1836, later: 1727)* and show an increase rather than a decrease in later mentions.

Overall, Tables D.8 and D.9 show that both of our context-aware speaker models, ReRef and Copy, are able to generate referring utterances, which make use of the dialogue history in a manner akin to humans with respect to multiple aspects of language style.

Comparing the context-aware models, ReRef shows a stronger degree of shortening than Copy, with very similar levels of bigram reuse to humans while Copy

	Human			*ReRef*			*Copy*			*Ref*		
	first	later	*d*	first	later	*d*	first	later	*d*	first	later	*d*
Givenness												
givenness	0.05	0.08	-0.36*	0.02	0.10	-0.89*	0.04	0.09	-0.53*	0.05	0.05	-0.03
definite	0.03	0.05	-0.27*	0.01	0.08	-0.85*	0.03	0.06	-0.48*	0.04	0.05	-0.04
seen	0.01	0.03	-0.26*	0.00	0.02	-0.43*	0.01	0.03	-0.29*	0.00	0.00	0.03
indefinite	0.07	0.02	0.77*	0.15	0.01	1.88*	0.10	0.01	1.14*	0.15	0.15	0.03
Compression												
length_c	11.29	8.28	0.63*	11.32	7.22	1.15*	10.77	7.79	0.65*	13.66	13.59	0.00
prop content	0.53	0.57	-0.20*	0.41	0.54	-0.70*	0.50	0.58	-0.39*	0.40	0.39	0.01
prop noun	0.37	0.41	-0.29*	0.30	0.44	-0.86*	0.37	0.43	-0.37*	0.28	0.28	-0.01
prop adj	0.09	0.10	-0.02	0.06	0.07	-0.14*	0.08	0.09	-0.10*	0.08	0.08	0.04
prop verb	0.13	0.11	0.12*	0.19	0.11	0.76*	0.13	0.12	0.12*	0.17	0.17	0.01

Table D.8: Trends in Subsequent mentions across humans, ReRef, Copy and Ref. The presence of * indicates significant differences between first and later means, with $p < 0.001$. d shows effect size measured by Cohen's d.

Human		*ReRef*			*Copy*			*Ref*		
	mean	mean	d	p	mean	d	p	mean	d	p
Lexical Entrainment:										
reuse prop within mention:										
–reuse_c	0.562	0.660	-0.334	***	0.612	-0.168	***	0.320	0.868	***
–reuse_bigrams_c	0.325	0.304	0.050	*	0.283	0.103	***	0.091	0.682	***
reuse prop within reused:										
–noun	0.701	0.746	-0.161	***	0.716	-0.050	*	0.740	-0.124	***
–adj	0.158	0.146	0.054	*	0.146	0.057	*	0.180	-0.079	**
–verb	0.095	0.066	0.165	***	0.097	-0.011	0.653	0.063	0.172	***
–NN bigrams	0.064	0.051	0.069	**	0.056	0.043	0.064	0.013	0.328	***

Table D.9: Human comparison with ReRef, Copy and Ref for givenness markers and Compression. The presence of * indicates a significant difference between the human mean and that of the model. *(***: $p < 0.001$, **: $p < 0.005$, *: $p < 0.01$)*

shows more similar traits to humans in terms of the proportion of markers and POS tags (as revealed by smaller effect sizes). In general, both models are successful at generating human-like utterances as we measure them; however, it seems that while Copy does generate utterances with the most similar proportional similarities to humans and exhibits similar proportions of unigram reuse, it does so at the expense of coherence. In terms of content bigram reuse, Copy seems to be less selective in what it repeats from previous referring utterances than ReRef, most likely due to the increased overall level of repetition in the generation. ReRef, on the other hand, shows amplified versions of the human trends, yet very similar content bigram and noun-noun bigram reuse proportion to humans, while maintaining low levels of same content word repetition as well as a high TTR, which indicates that coherence is also maintained.

Appendix E

Appendix to Chapter 10

E.1 Training Details

We provide the details of the setups of the generative language model in Section E.1.1, the discriminators in Section E.1.2, the simulators in Section E.1.3, and the adaptation mechanism in Section E.1.4. We use Python version 3.9.0 and PyTorch version 1.11.0 in the development and testing of all our models. In Table E.1, we report the hyperparameters used in the training of our final models.

	LM	Disc	Simulator
Learning rate	0.0001	0.0001	0.0004
Batch size	3	64	32
Dropout	0.3	0.2	0
Attention dim	512	512	1024
Embed dim	1024	768	1024
Hidden dim	512	512	1024
Patience	30	30	5

Table E.1: Hyperparameters used for training the generative language model (LM), discriminator (Disc), and simulator models.

E.1.1 Generative Language Model

In addition to the main hyperparameters listed in Table E.1, the language model requires several additional parameters. In nucleus sampling, we set the p value for *top-p* to 0.9 and sample from a vocabulary that consists of the words in the training splits of all 5 domains. The maximum length of the generated utterances is set to 30. The model is initialized and trained with 4 different seeds, which

yield similar performances. We use an early stopping patience of 30 epochs based on the validation set scores.[1]

Regarding the architectural details of the visually conditioned language model, in the visual encoder, we feed both the standardized target image vector and the concatenation of the six images in the full visual context into a linear layer followed by the ReLU non-linearity. We then concatenate the ensuing representations of the target image with the visual context and once more apply a linear layer followed by a ReLU non-linearity to obtain the final visual context, $\mathbf{v}$. This visual context is used to initialize a bidirectional LSTM encoder that takes as input the previous utterance referring to the target image in the current dialogue, if exists (see footnote 6 in Section 10.6.1), otherwise a special token indicating the absence of such an utterance. The final forward and backward hidden states of this encoder are concatenated, go through a linear layer and *tanh* non-linearity. The output is then set as the initial hidden state h_0 of the LSTM decoder (Hochreiter and Schmidhuber, 1997).

E.1.2 Discriminators

In these models instantiating the listeners, the word embeddings go through a dropout layer and a linear layer followed by the Leaky-ReLU non-linearity, after which standardization is applied. The visual context is processed in the same way as in the generative language model. Each word representation is concatenated with the representation of the visual context. The resulting vectors go through a linear layer and ReLU. Finally, we apply attention over these vectors to obtain the attention-weighted multimodal context vector. It is this context vector that is compared to the representations of candidate images via dot product.

We use the same set of hyperparameters for each domain as shown in Table E.1. The domain-specific listener models were selected based on their accuracy on the in-domain validation set. We report accuracy and MRR on the in- and out-of-domain test sets in Table E.3.

OOD word masking Our listeners are initialized with the same vocabulary comprising all the words in the training data. However, the domain-specific listeners only learn the words that exist in their own training sets. Therefore, if the speaker generates an OOD word for a domain-specific listener, in order not to further confound the effects of adaptation on the listeners, we mask the word with the <unk> vector. This vector is the same across all domains.

[1] We use the 'nlg-eval' library at `https://github.com/Maluuba/nlg-eval` to obtain scores for the common NLG metrics and also use BERTScore version 0.3.11 provided at `https://github.com/Tiiiger/bert_score`.

E.1.3 Simulator

We select the simulator models based on their accuracy in predicting the behavior of the listener models on the validation set. The simulator models are trained using the AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 0.0001, and a plateau learning scheduler with a patience of 2, a factor of 0.5, a threshold of 0.5.

E.1.4 Adaptation Mechanism

We optimize the values of the number of adaptation steps and the learning rate for the adaptation mechanism. We perform 2 hyperparameter sweeps using the Weight & Biases (WandB) platform (Biewald, 2020), evaluating a range of values. We find a positive correlation between both hyperparameters and adaptation accuracy, with Pearson's correlation coefficients of 0.71 for the learning rate, and 0.66 for the number of steps.

E.2 Additional Results

Here, we provide additional results yielded by our models for the speaker in Section E.2.1, the listener in Section E.2.2, the simulator in Section E.2.3 and for the adaptation mechanism in Section E.2.4.

E.2.1 Speaker Results

We provide the detailed results of the speaker model on the test set in Table E.2 with the averages and standard deviations over 4 runs.

BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**Rouge**
40.06 ± 1.60	23.81 ± 1.51	14.09 ± 1.20	8.46 ± 0.89	32.92 ± 0.93
CIDEr	**BertScore - Recall**	**BertScore-Recall - F1**	**BertScore - Precision**	
44.07 ± 1.68	58.91 ± 0.19	57.7 ± 0.12	57.9 ± 0.16	

Table E.2: Speaker results on the test set as measured by common NLG evaluation metrics.

E.2.2 Listener Results

Table E.3 reports the domain-specific listener performances on IND and OOD gold data. We observe that the domain-specific listeners perform well in in-domain settings and perform close to the random baseline in OOD settings.

Table E.4 presents the domain-specific listener accuracies on speaker-generated input. Especially in IND settings, we see lower scores as compared to the use of the gold data, presumably because the listener models were trained on gold data.

Domain	Epoch	IND		OOD	
		Accuracy	**MRR**	**Accuracy**	**MRR**
Appliances	23	84.12 ± 0.33	90.27 ± 0.10	20.28 ± 0.23	44.07 ± 0.11
Food	21	85.40 ± 0.28	91.20 ± 0.20	17.72 ± 0.18	42.42 ± 0.06
Indoor	14	82.94 ± 0.13	89.32 ± 0.09	19.14 ± 0.09	43.46 ± 0.06
Outdoor	19	83.96 ± 0.23	90.01 ± 0.14	19.64 ± 0.07	43.52 ± 0.06
Vehicles	17	78.99 ± 0.35	86.81 ± 0.14	18.46 ± 0.28	42.36 ± 0.20

Table E.3: Listener performance on gold utterances. Accuracy and MRR for the in-domain (IND) and out-of-domain (OOD) samples given to listeners trained on specific domains (indicated under the 'Domain' column).

Listener domain	Data domain				
	Appliances	**Food**	**Indoor**	**Outdoor**	**Vehicles**
Appliances	57.61 ± 1.38	20.10 ± 0.63	19.92 ± 0.47	21.27 ± 0.83	15.98 ± 0.82
Food	19.11 ± 1.70	54.29 ± 1.06	18.60 ± 0.84	18.85 ± 0.49	18.85 ± 0.49
Indoor	22.71 ± 1.30	19.65 ± 1.77	53.62 ± 0.79	20.82 ± 1.05	16.77 ± 0.79
Outdoor	15.08 ± 1.04	21.46 ± 0.70	19.62 ± 0.69	52.93 ± 1.11	17.69 ± 0.97
Vehicles	16.36 ± 1.55	16.17 ± 0.81	17.41 ± 0.64	20.13 ± 0.59	43.08 ± 1.16

Table E.4: Listener accuracies on speaker-generated data. Each row indicates the domain a listener was trained on, and the columns indicate the domain of the input samples. Results over 5 seeds.

E.2.3 Simulator Results

The detailed outcomes of the simulator models are reported in Table E.5. Here, we also report the results for the subset where the *listener* made a correct prediction (Pos) vs. it made an incorrect prediction (Neg). The simulators are better able to capture the correct listener behavior, possibly because during the training of simulators, in-domain data provides a clear picture of the listener's correct behavior.

Simulator	Setting	Avg	Pos	Neg
All domains	–	69.97 ± 0.79	85.15 ± 1.39	54.73 ± 0.76
Domain-specific	IND	78.20 ± 1.26	88.09 ± 1.98	67.36 ± 2.96
	OOD	72.78 ± 0.56	73.67 ± 1.69	72.58 ± 0.71

Table E.5: Simulator's accuracy in predicting the behavior of a listener knowledgeable about *all domains* (as the speaker) and a listener with *domain-specific* knowledge for IND and OOD samples. 'Avg' is the overall accuracy, 'Pos' and 'Neg' are the percentages of correct predictions for the samples where the listener picked the correct (Pos) and the incorrect image (Neg).

	OOD			IND		
	Golden	**Speaker**	**Adapted**	**Golden**	**Speaker**	**Adapted**
Appliances	20.21	19.30	27.74	84.21	57.21	74.28
Indoor	18.50	19.53	28.34	83.22	52.94	69.62
Food	17.06	18.31	26.26	85.61	55.54	78.15
Outdoor	18.89	18.54	26.21	84.38	52.83	73.04
Vehicles	18.25	17.35	25.16	78.67	42.09	63.75

Table E.6: Test results for the audience-aware adaptation pipeline, 5 seeds for each domain.

E.2.4 Adaptation Results

In Table E.6, we provide the test set results of the adaptation pipeline, broken down into domains and for IND and OOD inputs separately. The outcomes show that adaptation has effects in both IND and OOD settings, increasing resolution accuracies over speaker-generated utterances.

E.3 Evaluation Cards

For each of the three main modules in our experiments, we provide an evaluation card to clarify the nature of our generalization tests.[2] See Table E.7 for the generator, Table E.8 for the simulator, and Table E.9 for the listener. We also register our work in the GenBench evolving survey of generalization in NLP (Hupkes et al., 2023).[3]

E.4 Additional Experiments

Here, we provide details on additional experiments we performed in our adaptation pipeline.

In our adaptation mechanism, one of the stopping conditions is that the simulator predicts that the listener will be able to guess the referent. We also explored continuing adaptation until the *listener* itself correctly guesses the referent. We report the results in Table E.10, which reveal that using this stopping condition would yield higher results since the utterances are adapted until the actual listener makes a correct guess, mimicking an online interaction setup.

[2] `https://genbench.org/eval_cards`
[3] `https://genbench.org/references`

Motivation			
Practical □ △ ◯	*Cognitive* □ △ ◯	*Intrinsic*	*Fairness*
Generalization type			
Compo- sitional	*Structural*	*Cross Task*	*Cross Language*	*Cross Domain* □ △ ◯	*Robustness*
Shift type			
Covariate	*Label* □ △ ◯	*Full*	*No shift*
Shift source			
Naturally occurring	*Partitioned natural*	*Generated shift*	*Fully generated* □ △ ◯
Shift locus			
Train–test	*Fine-tune train–test*	*Pretrain–train*	*Pretrain–test* □ △ ◯

Table E.7: **Generator**'s evaluation card for the three main setups: baseline □, self-aware adaptation △, and audience-aware adaptation ◯

Motivation			
Practical △ ◯	*Cognitive* △ ◯	*Intrinsic*	*Fairness*
Generalization type			
Compo- sitional	*Structural*	*Cross Task*	*Cross Language*	*Cross Domain* △ ◯	*Robustness*
Shift type			
Covariate	*Label* △	*Full*	*No shift* ◯
Shift source			
Naturally occurring	*Partitioned natural*	*Generated shift*	*Fully generated* △ ◯
Shift locus			
Train–test	*Fine-tune train–test*	*Pretrain–train*	*Pretrain–test* △ ◯

Table E.8: **Simulator**'s evaluation card for the two setups in which it is used (i.e., baseline setup excluded): self-aware adaptation △, and audience-aware adaptation ◯.

Motivation			
Practical □ △ ◯	*Cognitive* □ △ ◯	*Intrinsic*	*Fairness*
Generalization type			
Compo- sitional	*Structural*	*Cross Task*	*Cross Language*	*Cross Domain* □ △ ◯	*Robustness*
Shift type			
Covariate □ △ ◯	*Label*	*Full*	*No shift* □ △ ◯
Shift source			
Naturally occurring	*Partitioned natural* □ △ ◯	*Generated shift*	*Fully generated* □ △ ◯
Shift locus			
Train–test	*Fine-tune train–test*	*Pretrain–train*	*Pretrain–test* □ △ ◯

Table E.9: **Listener**'s evaluation card for the three main setups: baseline □, self-aware adaptation △, and audience-aware adaptation ◯. In out-of-domain settings (OOD), the type of shift is covariate. In in-domain settings (IND), there is no shift between the training and test.

Target domain	**Golden**	**Speaker**	**Adapted**
Appliances	16.85	20.04	**38.89**
Food	85.57	55.26	**91.74**
Indoor	18.69	18.47	**39.49**
Outdoor	19.03	18.33	**37.96**
Vehicles	13.75	16.63	**35.43**

Table E.10: Listener accuracy using the listener stopping condition in the adaptation mechanism.

E.5 Additional Analyses

We note that we measure the type-utterance ratio for each step (i.e., the vocabulary size divided by the number of utterances available for that step), rather than the vocabulary size, because different steps correspond to different numbers of utterances: adaptation stops when the simulator module predicts the target image.

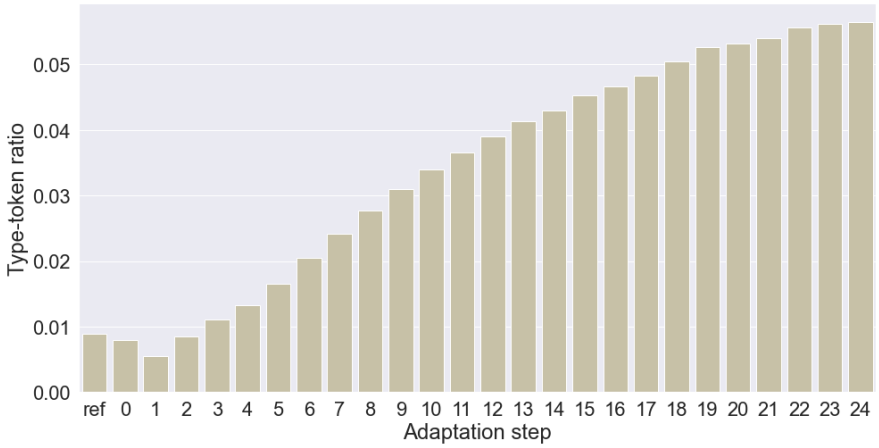


Figure E.1: Type-token ratio across adaptation steps. Human gold utterances (*ref*) and non-adapted utterances (0) are also shown.

Figure E.2 shows unigram part-of-speech distribution across adaptation steps for in- and out-of-domain conditions.

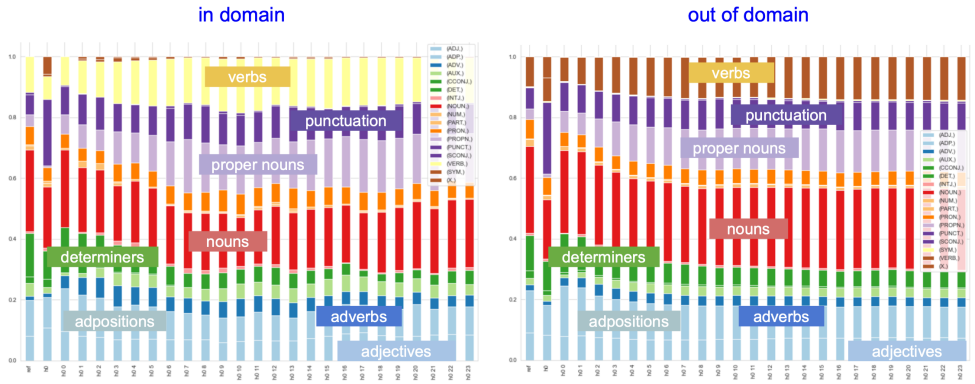


Figure E.2: Unigram POS distribution across adaptation steps.

We also measure the domain-specificity of utterances over steps, both in terms of the target image domain and the listener domain, as the percentage of domain-specific words in an utterance. We consider as domain-specific the words that appear *only* in interactions about a certain domain. The speaker, throughout

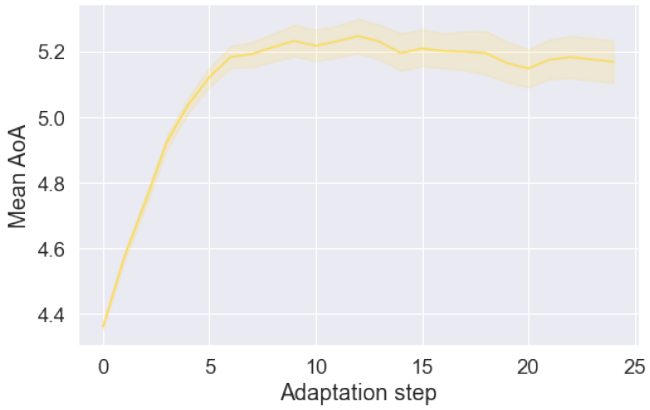


Figure E.3: Mean utterance Age of Acquisition over adaptation steps. Step 0 corresponds to the non-adapted utterance.

adaptation, produces more words belonging to both the image and the listener domain (Figure E.4) and thus fewer domain-agnostic words. We saw that, over adaptation steps, the decoder hidden state forgets image domain information in favor of the listener domain. This does not translate into no longer producing words from the image domain, suggesting that the speaker may be focusing more on the specific image than on its semantic domain.

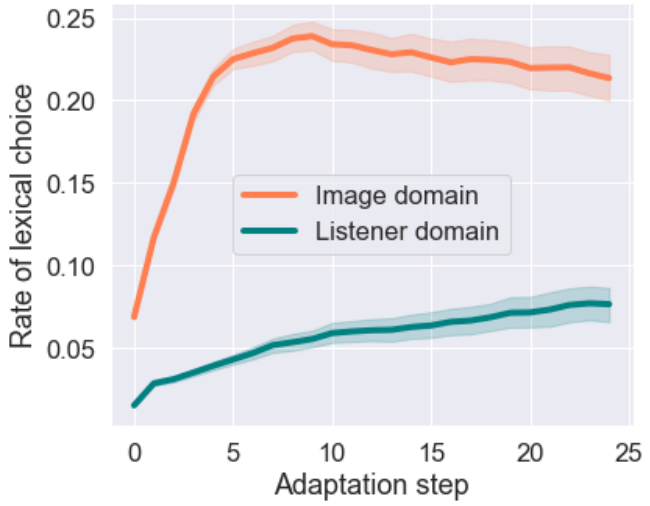


Figure E.4: Rate of lexical choice from image and listener domain-specific vocabularies.

Figure E.3 shows mean utterance age of acquisition rating (Kuperman et al., 2012) over steps.

Bibliography

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating clip: Towards characterization of broader capabilities and downstream implications.

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

Özge Alacam, Eugen Ruppert, Sina Zarrieß, Ganeshan Malhotra, Chris Biemann, and Sina Zarrieß. 2022. Modeling referential gaze in task-oriented settings of varying referential complexity. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 197–210, Online only. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj

Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Andrew James Anderson, Benjamin D. Zinszer, and Rajeev D.S. Raizada. 2016a. Representational similarity encoding for fmri: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage*, 128:44–53.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016b. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and VQA. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

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, Austin, Texas. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Michael Argyle, Mark Cook, and Duncan Cramer. 1994. Gaze and mutual gaze. *The British Journal of Psychiatry*, 165(6):848–850.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Pushmeet Kohli, and Edward Grefenstette. 2018. Jointly learning "what" and "how" from instructions and goal-states. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Adrian Bangerter. 2004. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6):415–419. PMID: 15147496.

Yuwei Bao, Sayan Ghosh, and Joyce Chai. 2022. Learning to mediate disparities towards pragmatic communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2829–2842, Dublin, Ireland. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume*

2: Short Papers), pages 579–584, Berlin, Germany. Association for Computational Linguistics.

Janet Beavin Bavelas and Nicole Chovil. 2000. Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology*, 19(2):163–194.

Lisa Beinborn and Nora Hollenstein. 2024. *Cognitive Plausibility in Natural Language Processing*, 1st edition. Springer Nature.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. PADA: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433.

Uri Berger, Lea Frermann, Gabriel Stanovsky, and Omri Abend. 2023. A large-scale multilingual study of visual constraints on linguistic selection of descriptions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2285–2299, Dubrovnik, Croatia. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. 2023. Towards language models that can see: Computer vision through the lens of natural language.

Yves Bestgen. 2021. LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 90–96, Online. Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Kathryn Bock, David E. Irwin, and Douglas J. Davidson. 2004. Putting first things first. *The interface of language, vision, and action: Eye movements and the visual world*, pages 249–278.

Margaret A. Boden. 2003. *The Creative Mind: Myths and Mechanisms*. Routledge.

Lena Bolliger, David Reich, Patrick Haller, Deborah Jakobi, Paul Prasse, and Lena Jäger. 2023. ScanDL: A diffusion model for generating synthetic scanpaths on texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15513–15538, Singapore. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Susan E. Brennan. 2005. How conversation is shaped by visual and spoken evidence. In J. Trueswell M. Tanenhaus, editor, *Approaches to studying world-situated language use: Bridging the language-as-product and language-action traditions*, pages 95–129. MIT Press, Cambridge, MA.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

Carsten Brockmann, Amy Isard, Jon Oberlander, and Michael White. 2005. Modelling alignment for affective dialogue. In *Proceedings of the Workshop on adapting the interaction style to affective factors at the 10th international conference on user modeling (UM-05)*.

Geert Brône and Bert Oben. 2018. *Eye-tracking in Interaction: Studies on the role of eye gaze in dialogue*. Advances in Interaction Studies. John Benjamins Publishing Company.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sarah Brown-Schmidt, Si On Yoon, and Rachel Anna Ryskin. 2015. People as contexts in conversation. In *Psychology of Learning and Motivation*, volume 62, chapter 3, pages 59–99. Elsevier.

Andrea Bruera and Massimo Poesio. 2023. Family lexicon: using language models to encode memories of personally familiar and famous people and places in the brain. *bioRxiv*.

Andrea Bruera, Yuan Tao, Andrew Anderson, Derya Çokal, Janosch Haber, and Massimo Poesio. 2023. Modeling brain representations of words' concreteness in context using gpt-2 and human ratings. *Cognitive Science*, 47(12):e13388.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Timothy J. Buschman and Earl K. Miller. 2007. Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, 315(5820):1860–1862.

Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. 2009. An alignment-capable microplanner for Natural Language Generation. In *Proceedings of the*

12th European Workshop on Natural Language Generation (ENLG 2009), pages 82–89, Athens, Greece. Association for Computational Linguistics.

Guy Thomas Buswell. 1935. *How people look at pictures: A study of the psychology and perception in art*. University of Chicago Press.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2021. What vision-language models 'see' when they see scenes. *ArXiv*, abs/2109.07301.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *ECCV Spotlight*.

Monica S. Castelhano, Michael L. Mack, and John M. Henderson. 2009. Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3):6–6.

Monica S. Castelhano and Carrick C. Williams. 2021. *Scene Perception*. Elements in Perception. Cambridge University Press.

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-AI games. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, Fuming Ma, and Qi Ju. 2019. Improving image captioning with conditional generative adversarial nets. In *AAAI Conference on Artificial Intelligence*.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18030–18040.

Shi Chen and Qi Zhao. 2018. Boosted attention: Leveraging human attention for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84.

Xianyu Chen, Ming Jiang, and Qi Zhao. 2021. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10876–10885.

Xinyi Chen, Raquel Fernández, and Sandro Pezzelle. 2023. The BLA benchmark: Investigating basic language abilities of pre-trained multimodal models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5817–5830, Singapore. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Herbert H. Clark. 1985. Language use and language users. *Handbook of social psychology (3rd ed.)*, pages 179–231.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Herbert H. Clark. 2003. Pointing and placing. In *Pointing: Where Language, Culture, and Cognition Meet*, pages 243–268. Psychology Press.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62 – 81.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1 – 39.

Moreno I. Coco and Frank Keller. 2010. Scan patterns on visual scenes predict sentence production. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, pages 1934–1939.

Moreno I. Coco and Frank Keller. 2012. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, 36(7):1204–1223.

Moreno I. Coco and Frank Keller. 2014. Classification of visual and linguistic tasks using eye-movement features. *Journal of Vision*, 14(3):1–18.

Moreno I. Coco and Frank Keller. 2015a. Integrating mechanisms of visual guidance in naturalistic language production. *Cognitive processing*, 16(2):131—150.

Moreno I. Coco and Frank Keller. 2015b. The interaction of visual and linguistic saliency during syntactic ambiguity resolution. *Quarterly Journal of Experimental Psychology*, 68(1):46–74.

Michael A. Cohen, Daniel C. Dennett, and Nancy Kanwisher. 2016. What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences*, 20(5):324–335.

Reuben Cohn-Gordon, Noah 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, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Elizabeth Coppock, Danielle Dionne, Nathanial Graham, Elias Ganem, Shijie Zhao, Shawn Lin, Wenxing Liu, and Derry Wijaya. 2020. Informativity in image captions vs. referring expressions. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 104–108, Gothenburg. Association for Computational Linguistics.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018a. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(2).

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018b. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154.

Rodolfo Corona Rodriguez, Stephan Alaniz, and Zeynep Akata. 2019. Modeling Conceptual Understanding in Image Reference Games. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fintan J. Costello and Mark T Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, 24(2):299–349.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Shuwen Deng, David R. Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A. Jäger. 2023. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. *Proc. ACM Hum.-Comput. Interact.*, 7(ETRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Ding, Bowen Chen, Li Du, Bing Qin, and Ting Liu. 2022. CogBERT: Cognition-guided pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3210–3225, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sibo Dong, Justin Goldstein, and Grace Hui Yang. 2022. Gazby: Gaze-based bert model to incorporate human attention in neural information retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '22, page 182–192, New York, NY, USA. Association for Computing Machinery.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Pamela Downing. 1977. On the creation and use of english compound nouns. *Language*, 53(4):810–842.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Ondřej Dušek and Filip Jurčíček. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 185–190, Los Angeles. Association for Computational Linguistics.

Maria K. Eckstein, Belén Guerra-Carrillo, Alison T. Miller Singley, and Silvia A. Bunge. 2017. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, 25:69 – 91. Sensitive periods across development.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, Elena Lloret, Elena-Simona Apostol, Ciprian-Octavian Truică, Branislava Šandrih, Sanda Martinčić-Ipšić, Gábor Berend, Albert Gatt, and Grăzina Korvel. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73.

Yulia Esaulova, Martina Penke, and Sarah Dolscheid. 2019. Describing events: Changes in eye movements and language production due to visual and conceptual properties of scenes. *Frontiers in Psychology*, 10.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4120–4129.

Fernanda Ferreira and Gwendolyn Rehrig. 2019. Linearisation during language production: evidence from scene meaning and saliency maps. *Language, Cognition and Neuroscience*, 34(9):1129–1139.

Kurt Feyaerts, Geert Brône, and Bert Oben. 2017. *Multimodality in Interaction*, Cambridge Handbooks in Language and Linguistics, page 135–156. Cambridge University Press.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Daniel Fried, Justin Chiu, and Dan Klein. 2021. Reference-centric models for grounded collaborative dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2130–2147, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181 – 218.

Albert Gatt, Emiel Krahmer, Kees van Deemter, and Roger P.G. van Gompel. 2017. Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive Science*, 41(S6):1457–1492.

Spandana Gella and Frank Keller. 2018. An evaluation of image-based verb prediction models against human eye-tracking data. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 758–763, New Orleans, Louisiana. Association for Computational Linguistics.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lila R. Gleitman, David January, Rebecca Nappa, and John C. Trueswell. 2007. On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, 57(4):544–569.

Cristina Gómez-Alonso and Aida Valls. 2008. A similarity measure for sequences of categorical data based on the ordering of common elements. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 134–145. Springer.

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

Noah D. Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Elizabeth R. Grant and Michael J. Spivey. 2003. Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, 14(5):462–466.

Claudio Greco, Diksha Bagade, Dieu-Thu Le, and Raffaella Bernardi. 2023. She adapts to her student: An expert pragmatic speaker tailoring her referring expressions to the layman listener. *Frontiers in Artificial Intelligence*, 6.

H. Paul Grice. 1975. Logic and conversation. In D. Davidson and G. Harman, editors, *The Logic of Grammar*, pages 64–75. Dickenson, Encino, California.

Zenzi M. Griffin. 2004. Why look? reasons for eye movements related to language production. In J. M. Henderson & F. Ferreira, editor, *The interface of language, vision, and action: Eye movements and the visual world*, chapter 7, pages 213–247. Psychology Press, New York.

Zenzi M. Griffin and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological Science*, 11:274–9.

Eleonora Gualdoni, Thomas Brochhagen, Andreas Mädebach, and Gemma Boleda. 2023. What's in a name? a large-scale computational study on how competition between names affects naming variation. *Journal of Memory and Language*, 133:104459.

Surabhi Gupta and Amanda Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, pages 1–6.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 85–95, Austin, Texas. Association for Computational Linguistics.

Joy E. Hanna and Susan E. Brennan. 2007. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57(4):596 – 615. Language-Vision Interaction.

Joy E. Hanna and Michael K. Tanenhaus. 2004. Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, 28(1):105 – 115.

Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual adaptation for efficient machine communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 408–419, Online. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sen He, Hamed R. Tavakoli, Ali Borji, and Nicolas Pugeault. 2019. Human attention in image captioning: Dataset and analysis. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8528–8537.

John M. Henderson. 2017. Gaze control as prediction. *Trends in Cognitive Sciences*, 21(1):15–23.

John M. Henderson and Fernanda Ferreira. 2013. *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. Taylor & Francis.

John M. Henderson, Taylor R. Hayes, Gwendolyn Rehrig, and Fernanda Ferreira. 2018. Meaning guides attention during real-world scene description. *Scientific reports*, 8(1):1–9.

John M. Henderson, Svetlana V. Shinkareva, Jing Wang, Steven G. Luke, and Jenn Olejarczyk. 2013. Predicting cognitive state from eye movements. *PLOS ONE*, 8(5):1–6.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Nora Hollenstein. 2021. *Leveraging Cognitive Processing Signals for Natural Language Understanding*. Ph.D. thesis, ETH Zurich.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Nora Hollenstein, Emmanuele Chersoni, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2022. CMCL 2022 shared task on multilingual and crosslingual prediction of human reading behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 121–129, Dublin, Ireland. Association for Computational Linguistics.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021a. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78, Online. Association for Computational Linguistics.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021b. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2011. *Eye Tracking : A Comprehensive Guide to Methods and Measures*. Oxford University Press, United Kingdom.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51(6):118:1–118:36.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Zhichao Hu, Gabrielle Halberg, Carolynn R. Jimenez, and Marilyn A. Walker. 2016. Entrainment in pedestrian direction giving: How many kinds of entrain-

ment? In *Situated Dialog in Speech-Based Human-Computer Interaction*, pages 151–164. Springer.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

Ellen A. Isaacs and Herbert H. Clark. 1987. References in conversation between experts and novices. *Journal of experimental psychology: general*, 116(1):26.

Laurent Itti and Christof Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506.

Laurent Itti and Christof Koch. 2001. Computational modeling of visual attention. *Nature reviews. Neuroscience*, 2:194–203.

Laurent Itti, Christof Koch, and Erns Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

T. Florian Jaeger and Roger Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Andrew Jaegle, Vahid Mehrpour, and Nicole Rust. 2019. Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *Current Opinion in Neurobiology*, 58:167–174. Computational Neuroscience.

Srinivasan Janarthanam and Oliver Lemon. 2010. Learning to adapt to unknown users: Referring expression generation in spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 69–78, Uppsala, Sweden. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Mainak Jas and Devi Parikh. 2015. Image specificity. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2727–2736.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*.

Pamela W. Jordan and Marilyn A. Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Daniel Kahneman. 2012. *Thinking, Fast and Slow*. Penguin Books.

Yuki Kamide, Gerry T.M Altmann, and Sarah L Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49(1):133–156.

Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. 2017. Gaze embeddings for zero-shot image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6412–6421.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 317–324, Barcelona, Spain. Association for Computational Linguistics.

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

Casey Kennington, Ryu Iida, Takenobu Tokunaga, and David Schlangen. 2015. Incrementally tracking reference in human/human dialogue using linguistic and extra-linguistic information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 272–282, Denver, Colorado. Association for Computational Linguistics.

Boaz Keysar. 2007. Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, 4(1):71–84.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 199–209. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Sigrid Klerke and Barbara Plank. 2019. At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Robert M. Krauss and Sidney Weinheimer. 1967. Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning & Verbal Behavior*, 6(3):359–363.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. 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 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

Willem J. M. Levelt. 1981. The speaker's linearization problem. *Philosophical Transactions of the Royal Society B*, 295:305–315.

Willem J. M. Levelt. 1993. *Speaking: From Intention to Articulation*. The MIT Press.

Bai Li and Frank Rudzicz. 2021. TorontoCL at CMCL 2021 shared task: RoBERTa with multi-stage fine-tuning for eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 85–89, Online. Association for Computational Linguistics.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023a. Multimodal foundation models: From specialists to general-purpose assistants.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020a. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision – ECCV 2020*, pages 121–137, Cham. Springer International Publishing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, and Graham Neubig. 2023a. Computational language acquisition with theory of mind. In *The Eleventh International Conference on Learning Representations*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical

study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. In *NeurIPS*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Simon P. Liversedge, Denis Drieghe, Xin Li, Guoli Yan, Xuejun Bai, and Jukka Hyönä. 2016. Universality in eye movements and reading: A trilingual investigation. *Cognition*, 147:1–20.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31(1):87–112.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Steven G. Luke and Kiel Christianson. 2018. The Provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2):826–833.

Ruotian Luo, Brian L. Price, Scott D. Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Brian MacWhinney. 1977. Starting points. *Language*, 53(1):152–168.

Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2023. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2523–2548, Dubrovnik, Croatia. Association for Computational Linguistics.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan 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.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.

Daniela Massiceti, Puneet K. Dokania, N. Siddharth, and Philip H. S. Torr. 2018. Visual dialogue without vision or dialogue. In *NeurIPS Workshop On Critiquing And Correcting Trends In Machine Learning*.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharya. 2020. A survey on using gaze behaviour for natural language processing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4907–4913. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

David D. McDonald. 1978. Subsequent reference: Syntactic and rhetorical constraints. In *Theoretical Issues in Natural Language Processing-2*.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

Charles Metzing and Susan E. Brennan. 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213.

Antje S. Meyer. 2004. The use of eye tracking in studies of sentence generation. In J. M. Henderson and F. Ferreira, editors, *The interface of language, vision,*

and action: Eye movements and the visual world, chapter 6, pages 191–212. Psychology Press, New York.

Antje S. Meyer and Femke van der Meulen. 2000. Phonological priming effects on speech onset latencies and viewing times in object naming. *Psychonomic Bulletin & Review*, 7:314–319.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Abhijit Mishra and Pushpak Bhattacharyya. 2018. *Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-Tracking*, 1st edition. Springer Publishing Company, Incorporated.

Margaret Mitchell, Ehud Reiter, and Kees Van Deemter. 2013a. Typicality and object reference. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013b. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1184, Atlanta, Georgia. Association for Computational Linguistics.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

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.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *Computer Vision – ECCV 2020*, pages 336–352, Cham. Springer International Publishing.

Andriy Myachykov, Dominic Thompson, Christoph Scheepers, and Simon Garrod. 2011. Visual attention and structural choice in sentence production across languages. *Language and Linguistics Compass*, 5(2):95–107.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, Madison, WI, USA. Omnipress.

Elisabeth Norcliffe and Agnieszka E. Konopka. 2015. *Vision and Language in Cross-Linguistic Research on Sentence Production*, pages 77–96. Springer India, New Delhi.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Marcus Nyström and Kenneth Holmqvist. 2010. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods*, 42:188–204.

Byung-Doh Oh. 2021. Team Ohio State at CMCL 2021 shared task: Fine-tuned RoBERTa for eye-tracking data prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 97–101, Online. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023b. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Atsumoto Ohashi and Ryuichiro Higashinaka. 2022. Adaptive natural language generation for task-oriented dialogue via reinforcement learning. In *Proceedings*

of the 29th International Conference on Computational Linguistics, pages 242–252, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aude Oliva. 2005. Gist of the scene. In *Neurobiology of attention*, pages 251–256. Elsevier.

Aude Oliva and Antonio Torralba. 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36.

OpenAI. 2023. Gpt-4 technical report.

Jiefu Ou, Benno Krojer, and Daniel Fried. 2023. Pragmatic inference with a CLIP listener for contrastive captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1904–1917, Toronto, Canada. Association for Computational Linguistics.

Anka Slana Ozimič and Grega Repovš. 2020. Visual working memory capacity is limited by two systems that change across lifespan. *Journal of Memory and Language*, 112:104090.

Sofia Paneri and Georgia G. Gregoriou. 2017. Top-down control of visual attention by the prefrontal cortex. functional specialization and long-range interactions. *Frontiers in neuroscience*, 11:545.

Dim P. Papadopoulos, Alasdair D. F. Clarke, Frank Keller, and Vittorio Ferrari. 2014. Training object class detectors from eye tracking data. In *Proceedings of 13th European Conference on Computer Vision*, pages 361–376.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), pages 4465–4470, Online. Association for Computational Linguistics.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

Marc Pomplun, Helge Ritter, and Boris Velichkovsky. 1996. Disambiguating complex visual information: towards communication of personal views of a scene. *Perception*, 25 8:931–48.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charlotte Pouw, Nora Hollenstein, and Lisa Beinborn. 2023. Cross-lingual transfer of cognitive processing complexity. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 655–669, Dubrovnik, Croatia. Association for Computational Linguistics.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Susan K. Prion and Katie Anne Haerling. 2014. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clinical Simulation in Nursing*, 10:535–536.

Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Shaolin Qu and Joyce Chai. 2008. Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 244–253.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Marlou Rasenberg. 2023. *Mutual understanding from a multimodal and interactional perspective*. Ph.D. thesis, Radboud University Nijmegen.

Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. 2020. Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, 44(11):e12911.

Keith Rayner. 1977. Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition*, 5:443–448.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124 3:372–422.

Erik D. Reichle, Alexander Pollatsek, Donald L. Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review*, 105 1:125–57.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015a. Exploring models and data for image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2953–2961, Cambridge, MA, USA. MIT Press.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015b. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 91–99.

Yuqi Ren and Deyi Xiong. 2021. CogAlign: Learning to align textual neural representations to cognitive language processing signals. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3758–3769, Online. Association for Computational Linguistics.

Daniel C. Richardson and Rick Dale. 2005. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science*, 29 6:1045–60.

Daniel C. Richardson, Rick Dale, and Natasha Z. Kirkham. 2007. The art of conversation is coordination. *Psychological Science*, 18(5):407–413.

Ardi Roelofs. 2020. Self-monitoring in speaking: In defense of a comprehension-based account. *Journal of Cognition*, 3(1).

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Subhro Roy, Michael Noseworthy, Rohan Paul, Daehyung Park, and Nicholas Roy. 2019. Leveraging past references for robust language grounding. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 430–440, Hong Kong, China. Association for Computational Linguistics.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David E. Rumelhart, James L. McClelland, and PDP Research Group. 1986. Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations. The MIT Press.

Jennifer D Ryan and Kelly Shen. 2020. The eyes are a window into memory. *Current Opinion in Behavioral Sciences*, 32:1 – 6.

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, page 71–78, New York, NY, USA. Association for Computing Machinery.

Naomi Saphra, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. 2023. First tragedy, then parse: History repeats itself in the new era of large language models.

Matthew Saxton. 2009. The inevitability of child directed speech. In *Language acquisition*, pages 62–86. Springer.

Jürgen Schmidhuber. 2010. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247.

Simeon Schüz, Ting Han, and Sina Zarrieß. 2021. Diversity as a by-product: Goal-oriented language generation leads to linguistic variation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 411–422, Singapore and Online. Association for Computational Linguistics.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Komal Sharan, Ashwinkumar Ganesan, and Tim Oates. 2019. Improving visual reasoning with attention alignment. In *Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I*, page 219–230, Berlin, Heidelberg. Springer-Verlag.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*.

Todd Shore and Gabriel Skantze. 2018. Using lexical alignment and referring ability to address data sparsity in situated dialog reference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2288–2297, Brussels, Belgium. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, pages 1–14. Computational and Biological Learning Society.

Aaditya K Singh, David Ding, Andrew Saxe, Felix Hill, and Andrew Lampinen. 2023. Know your audience: specializing grounded language models with listener subtraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3884–3911, Dubrovnik, Croatia. Association for Computational Linguistics.

Dan I. Slobin. 2003. Language and Thought Online: Cognitive Consequences of Linguistic Relativity. In *Language in Mind: Advances in the Study of Language and Thought*, pages 157–192. The MIT Press.

Vidya Somashekarappa, Christine Howes, and Asad Sayeed. 2020. An annotation approach for social and referential gaze in dialogue. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 759–765, Marseille, France. European Language Resources Association.

Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2023. Multimodal integration of human-like attention in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2647–2657.

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. 2021. VQA-MHUG: A gaze dataset to study multimodal neural attention in visual question answering. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 27–43, Online. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Com-*

putational Natural Language Learning, pages 12–25, Online. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 6327–6341. Curran Associates, Inc.

Elizabeth S. Spelke. 1990. Principles of object perception. *Cognitive Science*, 14(1):29 – 56.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Maria Staudte, Matthew W. Crocker, Alexis Héloir, and Michael Kipp. 2014. The influence of speaker gaze on listener comprehension: Contrasting visual versus intentional accounts. *Cognition*, 133(1):317–328.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2023. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):539–559.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 81–88, Sydney, Australia. Association for Computational Linguistics.

Svetlana Stoyanchev and Amanda Stent. 2009. Lexical and syntactic adaptation and their impact in deployed spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 189–192, Boulder, Colorado. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning.

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.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020a. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020b. Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4664–4677, Online. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2487–2496.

Jan Theeuwes, Artem Belopolsky, and Christian N.L. Olivers. 2009. Interactions between working memory, attention and eye movements. *Acta Psychologica*, 132(2):106 – 114. Spatial working memory and imagery: From eye movements to grounded cognition.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Michael Tomasello. 1999. *The Cultural Origins of Human Cognition*. Harvard University Press.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Antonio Torralba, Aude Oliva, Monica Castelhano, and John Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review*, 113:766–86.

John C. Trueswell, Yi Lin, Benjamin F. Armstrong, Erica A. Cartmill, and Lila R. Gleitman. 2016. Perceiving referential intent: Dynamics of reference in natural parent–child interactions. *Cognition*, 148:117–135.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*.

Stéphan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised Dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4130–4136, Portorož, Slovenia. European Language Resources Association (ELRA).

Ahmet Üstün, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6650–6662, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Preethi Vaidyanathan, Emily Prud'hommeaux, Cecilia O. Alm, Jeff B. Pelz, and Anne R. Haake. 2015. Alignment of eye movements and spoken language for semantic image understanding. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 76–81, London, UK. Association for Computational Linguistics.

Preethi Vaidyanathan, Emily T. Prud'hommeaux, Jeff B. Pelz, and Cecilia O. Alm. 2018. SNAG: Spoken narratives and gaze dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 132–137, Melbourne, Australia. Association for Computational Linguistics.

Femke F. van der Meulen, Antje S. Meyer, and Willem J. M. Levelt. 2001. Eye movements during the production of nouns and pronouns. *Memory & Cognition*, 29:512–521.

Emiel van Miltenburg. 2023. Evaluating nlg systems: A brief introduction. Originally published on the website of the International Conference on Natural Language Generation (INLG) 2023.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018a. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emiel van Miltenburg, Ákos Kádár, Ruud Koolen, and Emiel Krahmer. 2018b. DIDEC: The Dutch Image Description and Eye-tracking Corpus. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3658–3669. Association for Computational Linguistics.

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018c. Varying image description tasks: spoken versus written descriptions. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 88–100. Association for Computational Linguistics.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2018. Object referring in videos with language and human gaze. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4129–4138.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Peter Vickers, Rosa Wainwright, Harish Tayyar Madabushi, and Aline Villavicencio. 2021. CogNLP-Sheffield at CMCL 2021 shared task: Blending cognitively inspired features with transformer-based language models for predicting eye tracking patterns. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 125–133, Online. Association for Computational Linguistics.

Jette Viethen, Robert Dale, and Markus Guhe. 2011. Generating subsequent reference in shared visual scenes: Computation vs re-use. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1158–1167, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Jorrig Vogels, Emiel Krahmer, and Alfons Maes. 2013. Who is where referred to how, and why? the influence of visual saliency on referent accessibility in spoken language production. *Language and Cognitive Processes*, 28(9):1323–1349.

P.C. Wason and J.ST.B.T. Evans. 1974. Dual processes in reasoning? *Cognition*, 3(2):141–154.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shih-Lun Wu, Yi-Hui Chou, and Liangze Li. 2023. Listener model for the PhotoBook referential game with CLIPScores as implicit reference chain. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1420–1432, Toronto, Canada. Association for Computational Linguistics.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *32nd International Conference on Machine Learning (ICML)*, pages 2048–2057.

Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. 2020. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alfred L. Yarbus. 1967. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. 2017a. A joint speaker-listener-reinforcer model for referring expressions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3521–3529.

Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. 2017b. Supervising neural attention models for video captioning by human gaze data. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6119–6127.

Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. 2013. Studying relationships between human gaze, description, and computer vision. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746.

Sina Zarrieß and David Schlangen. 2019. Know what you don't know: Modeling a pragmatic speaker that refers to objects of unknown categories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 654–659, Florence, Italy. Association for Computational Linguistics.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588.

Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. 2020a. Human gaze assisted artificial intelligence: A review. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4951–4958. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12901–12911. PMLR.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Rolf A. Zwaan and Carol J. Madden. 2005. Embodied sentence comprehension. *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, page 224–245.

Samenvatting

Wanneer mensen een afbeelding beschrijven, zijn er complexe visuele en talige processen aan het werk. Sprekers hebben bijvoorbeeld de neiging om naar een object te kijken net voordat ze het benoemen, maar doen dat niet altijd. Ook kunnen sprekers tijdens een gesprek meerdere keren naar een entiteit verwijzen, waarbij ze uitdrukkingen gebruiken die in hun gedeelde kennis ontstaan en doorontwikkelen. In dit proefschrift ontwerp ik computationele modellen van zulke visuele en linguïstische processen, waarbij ik inspiratie haal uit theorieën en bevindingen uit de cognitiewetenschap en de psycholinguïstiek. Dit werk, waarin ik de ingewikkelde relatie tussen taal en buitentalige modaliteiten binnen diepe kunstmatige neurale netwerken wil vastleggen, draagt bij aan de onderzoekslijn naar multimodale natuurlijke taalverwerking. Dit proefschrift bestaat uit twee delen: (1) het modelleren van de menselijke blik in taalgebruik (productie en begrip), en (2) het modelleren van communicatiestrategieën in referentiële taken in visueel gebaseerde dialogen. In het eerste deel verdiep ik me in het verbeteren van modellen voor het beschrijven van afbeeldingen met behulp van oogbewegingsgegevens; het evalueren van de variatie in menselijke signalen tijdens het beschrijven van afbeeldingen; en het voorspellen van menselijk leesgedrag in de vorm van oogbewegingen. In het tweede deel bouw ik modellen voor het kwantificeren, genereren, oplossen en aanpassen van uitingen in referentiële taken die zich binnen visuele en conversationele contexten bevinden. De resultaten bevorderen ons begrip van menselijke visuo-linguïstische processen door de ingewikkelde strategieën te onthullen die bij dergelijke processen een rol spelen, en wijzen op het belang om hiermee rekening te houden bij het ontwikkelen en gebruiken van multimodale modellen. De bevindingen werpen licht op hoe de vooruitgang op het gebied van kunstmatige intelligentie zou kunnen bijdragen aan het bevorderen van het onderzoek naar crossmodale processen bij mensen en vice versa.

Abstract

When people describe an image, there are complex visual and linguistic processes at work. For instance, speakers tend to look at an object right before mentioning it, but not every time. Similarly, during a conversation, speakers can refer to an entity multiple times, using expressions evolving in the common ground. In this thesis, I develop computational models of such visual and linguistic processes, drawing inspiration from theories and findings from cognitive science and psycholinguistics. This work, where I aim to capture the intricate relationship between non-linguistic modalities and language within deep artificial neural networks, contributes to the line of research into multimodal Natural Language Processing. This thesis consists of two parts: (1) modeling human gaze in language use (production and comprehension), and (2) modeling communication strategies in referential tasks in visually grounded dialogue. In the first part, I delve into enhancing image description generation models using eye-tracking data; evaluating the variation in human signals while describing images; and predicting human reading behavior in the form of eye movements. In the second part, I build models quantifying, generating, resolving, and adapting utterances in referential tasks situated within visual and conversational contexts. The outcomes advance our understanding of human visuo-linguistic processes by revealing intricate strategies at play in such processes, and point to the importance of accounting for them when developing and utilizing multimodal models. The findings shed light on how the advancements in artificial intelligence could contribute to advancing the research on crossmodal processes in humans and vice versa.

Titles in the ILLC Dissertation Series:

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing

ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning

ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them

ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing

ILLC DS-2020-10: **Arnold Kochari**
Perceiving and communicating magnitudes: Behavioral and electrophysiological studies

ILLC DS-2020-11: **Marco Del Tredici**
Linguistic Variation in Online Communities: A Computational Perspective

ILLC DS-2020-12: **Bastiaan van der Weij**
Experienced listeners: Modeling the influence of long-term musical exposure on rhythm perception

ILLC DS-2020-13: **Thom van Gessel**
Questions in Context

ILLC DS-2020-14: **Gianluca Grilletti**
Questions & Quantification: A study of first order inquisitive logic

ILLC DS-2020-15: **Tom Schoonen**
Tales of Similarity and Imagination. A modest epistemology of possibility

ILLC DS-2020-16: **Ilaria Canavotto**
Where Responsibility Takes You: Logics of Agency, Counterfactuals and Norms

ILLC DS-2020-17: **Francesca Zaffora Blando**
Patterns and Probabilities: A Study in Algorithmic Randomness and Computable Learning

ILLC DS-2021-01: **Yfke Dulek**
Delegated and Distributed Quantum Computation

ILLC DS-2021-02: **Elbert J. Booij**
The Things Before Us: On What it Is to Be an Object

ILLC DS-2021-03: **Seyyed Hadi Hashemi**
Modeling Users Interacting with Smart Devices

ILLC DS-2021-04: **Sophie Arnoult**
Adjunction in Hierarchical Phrase-Based Translation

ILLC DS-2021-05: **Cian Guilfoyle Chartier**
A Pragmatic Defense of Logical Pluralism

ILLC DS-2021-06: **Zoi Terzopoulou**
Collective Decisions with Incomplete Individual Opinions

ILLC DS-2021-07: **Anthia Solaki**
Logical Models for Bounded Reasoners

ILLC DS-2021-08: **Michael Sejr Schlichtkrull**
Incorporating Structure into Neural Models for Language Processing

ILLC DS-2021-09: **Taichi Uemura**
Abstract and Concrete Type Theories

ILLC DS-2021-10: **Levin Hornischer**
Dynamical Systems via Domains: Toward a Unified Foundation of Symbolic and Non-symbolic Computation

ILLC DS-2021-11: **Sirin Botan**
Strategyproof Social Choice for Restricted Domains

ILLC DS-2021-12: **Michael Cohen**
Dynamic Introspection

ILLC DS-2021-13: **Dazhu Li**
Formal Threads in the Social Fabric: Studies in the Logical Dynamics of Multi-Agent Interaction

ILLC DS-2021-14: **Álvaro Piedrafita**
On Span Programs and Quantum Algorithms

ILLC DS-2022-01: **Anna Bellomo**
Sums, Numbers and Infinity: Collections in Bolzano's Mathematics and Philosophy

ILLC DS-2022-02: **Jan Czajkowski**
Post-Quantum Security of Hash Functions

ILLC DS-2022-03: **Sonia Ramotowska**
Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences

ILLC DS-2022-04: **Ruben Brokkelkamp**
How Close Does It Get?: From Near-Optimal Network Algorithms to Suboptimal Equilibrium Outcomes

ILLC DS-2022-05: **Lwenn Bussière-Carae**
No means No! Speech Acts in Conflict

ILLC DS-2022-06: **Emma Mojet**
Observing Disciplines: Data Practices In and Between Disciplines in the 19th and Early 20th Centuries

ILLC DS-2022-07: **Freek Gerrit Witteveen**
Quantum information theory and many-body physics

ILLC DS-2023-01: **Subhasree Patro**
Quantum Fine-Grained Complexity

ILLC DS-2023-02: **Arjan Cornelissen**
Quantum multivariate estimation and span program algorithms

ILLC DS-2023-03: **Robert Paßmann**
Logical Structure of Constructive Set Theories

ILLC DS-2023-04: **Samira Abnar**
Inductive Biases for Learning Natural Language

ILLC DS-2023-05: **Dean McHugh**
Causation and Modality: Models and Meanings

ILLC DS-2023-06: **Jialiang Yan**
Monotonicity in Intensional Contexts: Weakening and: Pragmatic Effects under Modals and Attitudes

ILLC DS-2023-07: **Yiyan Wang**
Collective Agency: From Philosophical and Logical Perspectives

ILLC DS-2023-08: **Lei Li**
Games, Boards and Play: A Logical Perspective

ILLC DS-2023-09: **Simon Rey**
Variations on Participatory Budgeting

ILLC DS-2023-10: **Mario Giulianelli**
Neural Models of Language Use: Studies of Language Comprehension and Production in Context

ILLC DS-2023-11: **Guillermo Menéndez Turata**
Cyclic Proof Systems for Modal Fixpoint Logics

ILLC DS-2023-12: **Ned J.H. Wontner**
Views From a Peak: Generalisations and Descriptive Set Theory

ILLC DS-2024-01: **Jan Rooduijn**
Fragments and Frame Classes: Towards a Uniform Proof Theory for Modal Fixed Point Logics

ILLC DS-2024-02: **Bas Cornelissen**
Measuring musics: Notes on modes, motifs, and melodies

ILLC DS-2024-03: **Nicola De Cao**
Entity Centric Neural Models for Natural Language Processing