LT2318: Artificial Intelligence: Cognitive Systems (AICS)
Reading list

Simon Dobnik
CLASP, University of Gothenburg
simon.dobnik@gu.se

October 21, 2021

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2017. Don't just assume; look and answer: Overcoming priors for visual question answering. *arXiv*, arXiv:1712.00377 [cs.CV]:1–15.

Saad Ullah Akram. 2012. Visual recognition of isolated swedish sign language signs. Master's thesis, School of Computer Science and Communication, Control and Robotics, Royal Institute of Technolog, Stockholm, Sweden.

Peter Anderson. 2018. *Vision and Language Learning: From Image Captioning and Visual Question Answering towards Embodied Agents*. A thesis submitted for the degree of doctor of philosophy, The Australian National University.

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 visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2015. Deep compositional question answering with neural module networks. *arXiv*, arXiv:1511.02799 [cs.CV].

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *Proceedings of NAACL-HLT 2016*, pages 1545–1554, San Diego, California. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Yoav Artzi. 2020. Robust control in situated instruction following. Invited talk, Cornell University, ACL 2020 Workshop on Advances in Language and Vision Research (ALVR).

Amelie Åstbom. 2017. How function of objects affects geometry of spatial descriptions. A study of Swedish and Japanese. C-uppsats (bachelor's thesis/extended essay), Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik, opponent: Linnea Strand, examiner: Christine Howes.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46.

Lawrence W. Barsalou. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–609.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59:617–645.

John A. Bateman, Mihai Pomarlan, and Gayane Kazhoyan. 2019. Embodied contextualization: Towards a multistratal ontological treatment. *Applied Ontology*, Pre-press:1–35.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. 2013. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332.

Bert Baumgaertner, Raquel Fernández, and Matthew Stone. 2012. Towards a flexible semantics: Colour terms in collaborative reference tasks. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 80–84, Montréal, Canada. 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.

Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc.

Cynthia Breazeal, Matt Berlin, Andrew Brooks, Jesse Gray, and Andrea L. Thomaz. 2006. Using perspective taking to learn from ambiguous demonstrations. *Robotics and Autonomous Systems*, 54(5):385–393.

Jason Brownlee. 2019. A gentle introduction to generative adversarial networks (GANs). Blog-post article, Machine Learning Mastery.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(1–47).

Donna K Byron. 2003. Understanding referring expressions in situated language some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction*, pages 39–47.

Remi Cadene, Corentin Dancette, Matthieu Cord, and Devi Parikh. 2019. RUBi: Reducing unimodal biases for visual question answering. In *NeurIPS*, pages 841–852.

Ozan Arkan Can, Pedro Zuidberg Dos Martires, Andreas Persson, Julian Gaal, Amy Loutfi, Luc De Raedt, Deniz Yuret, and Alessandro Saffiotti. 2019. Learning from implicit information in natural language instructions for robotic manipulations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 29–39, Minneapolis, Minnesota. Association for Computational Linguistics.

A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, L. Fadiga, B. Wrede, K. Rohlfing, E. Tuci, K. Dautenhahn, J. Saunders, and A. Zeschel. 2010. Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3):167–195.

José Miguel Cano Santín. 2019. Fast visual grounding in interaction: bringing few-shot learning with neural networks to an interactive robot. Masters in language technology (mlt), 30 hec, Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik and Mehdi Ghanimifard, examiner: Aarne Ranta.

José Miguel Cano Santín, Simon Dobnik, and Mehdi Ghanimifard. 2020. Fast visual grounding in interaction: bringing few-shot learning with neural networks to an interactive robot. In *Proceedings of Conference on Probability and Meaning (PaM-2020), Gothenburg, Sweden*, pages 1–9, Gothenburg, Sweden. Association for Computational Linguistics (ACL), Special Interest Group on Computational Semantics (SIGSEM).

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent communication through negotiation. *arXiv*, arXiv:1804.03980 [cs.AI].

Joyce Y Chai, Maya Cakmak, and Candace Sidner. 2018a. Teaching robots new tasks through natural interaction. In K. A. Cluck and J. E. Laird, editors, *Interactive Task Learning: Agents, Robots, and Humans Acquiring New Tasks through Natural Interactions*, Strungmann Forum Reports, chapter 9. MIT press.

Joyce Y. Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018b. Language to action: Towards interactive task learning with physical agents. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2–9. International Joint Conferences on Artificial Intelligence Organization.

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. *arXiv*, arXiv:1708.05122 [cs.HC]:1–9.

Yejin Choi. 2020. Intuitive reasoning as (un)supervised language generation. Seminar, Paul G. Allen School of Computer Science and Enginneering, University of Washington and Allen Institute for Artificial Intelligence, MIT Embodied Intelligence Seminar.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2017. Acquiring common sense spatial knowledge through implicit spatial templates. *arXiv*, arXiv:1711.06821 [cs.AI]:1–8.

Guillem Collell and Marie-Francine Moens. 2018a. Do neural network cross-modal mappings really bridge modalities? *arXiv*, arXiv:1805.07616 [stat.ML]:1–13.

Guillem Collell and Marie-Francine Moens. 2018b. Learning representations specialized in spatial knowledge: Leveraging language and vision. *Transactions of the Association for Computational Linguistics*, 6:133–144.

Robin Cooper. in prep. From perception to communication: An analysis of meaning and action using a theory of types with records (TTR). Draft at `https://sites.google.com/site/typetheorywithrecords/drafts`.

Kenny Coventry and Simon Garrod. 2005. Spatial prepositions and the functional geometric framework. towards a classification of extra-geometric influences. In Laura Anne Carlson and Emile van der Zee, editors, *Functional features in language and space: insights from perception, categorization, and development*, volume 2, pages 149–162. Oxford University Press.

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, volume 3343 of *Lecture Notes in Computer Science*, pages 98–110. Springer Berlin Heidelberg.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063.

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.

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.

Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. 2017. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *The IEEE International Conference on Computer Vision (ICCV)*.

Vittorio Di Tomaso and Vecenzo Lombardo. 1998. A computational model for the interpretation of static locative expressions. In Patrick Olivier and Klaus-Peter Gapp, editors, *Representation and Processing of Spatial Expressions*, pages 73–90. Lawrence Erlbaum Associates, Mahwah, N.J.

M. W. M. G Dissanayake, P. M. Newman, H. F. Durrant-Whyte, S. Clark, and M. Csorba. 2001. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotic and Automation*, 17(3):229–241.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik and Amelie Åstbom. 2017. (Perceptual) grounding as interaction. In *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–26, Saarbrücken, Germany.

Simon Dobnik and Mehdi Ghanimifard. 2020. Spatial descriptions on a functional-geometric spectrum: the location of objects. In *Spatial Cognition XII: Proceedings of the 12th International Conference, Spatial Cognition 2020, Riga, Latvia*, pages 219–234, Cham, Switzerland. Springer International Publishing.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Simon Dobnik and Erik de Graaf. 2017. KILLE: a framework for situated agents for learning language through interaction. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 162–171, Gothenburg, Sweden. Northern European Association for Language Technology (NEALT), Association for Computational Linguistics.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSci 2013 Production of referring expressions – bridging the gap between cognitive and computational approaches to reference at CogSci*, pages 1–6, Berlin, Germany.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language at COLING*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Simon Dobnik, John D. Kelleher, and Mehdi Ghanimifard. 2019. Language, action and perception (APL-ESSLLI): Lecture notes of a course in language and computation. lecture notes, ESSLLI 2019, 31 European Summer School on Logic, Language and Information, University of Latvia, Riga, Latvia.

Simon Dobnik, John D. Kelleher, and Christine Howes. 2020. Local alignment of frame of reference assignment in English and Swedish dialogue. In *Spatial Cognition XII: Proceedings of the 12th International Conference, Spatial Cognition 2020, Riga, Latvia*, pages 251–267, Cham, Switzerland. Springer International Publishing.

Simon Dobnik, John D. Kelleher, and Christos Koniaris. 2014. Priming and alignment of frame of reference in situated conversation. In *Proceedings of DialWatt – Semdial 2014: The 18th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–52, Edinburgh.

Simon Dobnik and Sharid Loáiciga. 2019. On visual coreference chains resolution. In *Proceedings of LondonLogue – Semdial 2019: The 23rd Workshop on the Semantics and Pragmatics of Dialogue*, pages 1–3, London, UK. Queen Mary University of London.

Simon Dobnik and Vera Silfversparre. 2021. The red cup on the left: Reference, coreference and attention in visual dialogue. In *Proceedings of PotsDial - Semdial 2021: The 25th Workshop on the Semantics and Pragmatics of Dialogue*, Proceedings (SemDial), pages 50–60, Potsdam, Germany.

Fethiye Irmak Doğan, Sinan Kalkan, and Iolanda Leite. 2019. Learning to generate unambiguous spatial referring expressions for real-world environments. *arXiv*, arXiv:1904.07165 [cs.RO]:1–8.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. *CoRR*, abs/1605.06676.

Sébastien Forestier and Pierre-Yves Oudeyer. 2017. A unified model of speech and tool use early development. In *39th Annual Conference of the Cognitive Science Society (CogSci 2017)*, pages 2013–2018.

Mehdi Ghanimifard. 2020. *Why the pond is not outside the frog? Grounding in contextual representations by neural language models*. Doctoral thesis, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Gothenburg, Sweden.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics*, pages 1–12, Montpellier, France. Association for Computational Linguistics.

Mehdi Ghanimifard and Simon Dobnik. 2018. Knowing when to look for what and where: Evaluating generation of spatial descriptions with adaptive attention. In *Computer Vision – ECCV 2018 Workshops. ECCV 2018*, volume 11132 of *Lecture Notes in Computer Science (LNCS)*, pages 1–9, Proceedings of the Workshop on Shortcomings in Vision and Language (SiVL), ECCV 2018, Munich, Germany. Springer, Cham.

Mehdi Ghanimifard and Simon Dobnik. 2019a. What a neural language model tells us about spatial relations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 71–81, Minneapolis, Minnesota, USA. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Association for Computational Linguistics.

Mehdi Ghanimifard and Simon Dobnik. 2019b. What goes into a word: generating image descriptions with top-down spatial knowledge. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG-2019)*, pages 1–15, Tokyo, Japan. Association for Computational Linguistics.

Arthur M. Glenberg. 1997. What memory is for. *The Behavioral and brain sciences*, 20:1–55.

Patrick Goebel. 2013. *ROS by example*. Lulu.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. *arXiv*, arXiv:1712.03316 [cs.CV].

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–11.

Erik de Graaf. 2016. Learning objects and spatial relations with Kinect. Master's thesis, Department of Philosophy, Linguistics and Theory of Science. University of Gothenburg,

Gothenburg, Sweden, June, 8th. Supervisor: Simon Dobnik, examiner: Richard Johansson, opponent: Lorena Llozhi.

Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. 2019. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, Florence, Italy. Association for Computational Linguistics.

Antonia F de C Hamilton, Klaus Kessler, and Sarah H Creem-Regehr. 2014. Perspective taking: building a neurocognitive framework for integrating the "social"and the "spatial". *Frontiers in Human Neuroscience*, 8(403):1–3.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1–3):335–346.

Felix Hill, Karl Moritz Hermann, Phil Blunsom, and Stephen Clark. 2017. Understanding grounded language learning agents. *arXiv*, arXiv:1710.09867 [cs.CL].

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. 2020. Grounded language learning fast and slow. *arXiv*, arXiv:2009.01719 [cs.CL]:1–17.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Nikolai Ilinykh and Simon Dobnik. 2020. When an image tells a story: The role of visual and semantic information for generating paragraph descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.

Nikolai Ilinykh and Simon Dobnik. 2021. How vision affects language: Comparing masked self-attention in uni-modal and multi-modal transformer. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 45–55, Groningen, Netherlands (Online). Association for Computational Linguistics.

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.

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.

Andrej Karpathy and Li Fei-Fei. 2015a. Automated image captioning with convnets and recurrent nets. Technical report, The Vision Lab, Stanford University.

Andrej Karpathy and Li Fei-Fei. 2015b. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

John D. Kelleher and Fintan J. Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

John D. Kelleher, Fintan J. Costello, and Josef van Genabith. 2005. Dynamically structuring updating and interrelating representations of visual and linguistic discourse. *Artificial Intelligence*, 167(1):62–102.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. InproTKs: a toolkit for incremental situated processing. *Proceedings of SIGdial 2014: Short Papers*.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China. Association for Computational Linguistics.

Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: an interactive 3d environment for visual AI. *arXiv*, arXiv:1712.05474 [cs.CV].

Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3345.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael 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.

Geert-Jan M. Kruijff, John D. Kelleher, and Nick Hawes. 2006. Information fusion for visual reference resolution in dynamic situated dialogue. In Elisabeth André, Laila Dybkjær, Wolfgang Minker, Heiko Neumann, and Michael Weber, editors, *Perception and Interactive Technologies. International Tutorial and Research Workshop, PIT 2006 Kloster Irsee, Germany*, pages 117–128. Springer, Berlin, Heidelberg.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):125–138. Special issue on human and robot interactive communication.

G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *CVPR 2011*, pages 1601–1608.

Simon Lang. 2011. Sign language recognition with kinect. Bachelorarbeit, Institut für Informatik, Freie Universität Berlin, Berlin, Germany.

Simon Lang, Marco Block, and Raúl Rojas. 2012. Sign language recognition using kinect. In *Artificial Intelligence and Soft Computing: 11th International Conference, ICAISC 2012,*

Zakopane, Poland, April 29-May 3, 2012, Proceedings, Part I, pages 394–402, Berlin, Heidelberg. Springer Berlin Heidelberg.

Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, Johan Bos, and Ewan Klein. 2001. Training personal robots using natural language instruction. *IEEE Intelligent Systems*, 16:38–45.

Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, and Ewan Klein. 2002. Mobile robot programming using natural language. *Robotics and Autonomous Systems*, 38(3–4):171–181.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv*, arXiv:1612.07182v2 [cs.CL]:1–11.

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.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021a. Reference and coreference in situated dialogue. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 39–44, Online. Association for Computational Linguistics.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021b. Reference and coreference in situated dialogue. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 39–44, Online. Association for Computational Linguistics.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. IEEE.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. ArXiv:1612.01887 [cs.CV].

Mateusz Malinowski. 2017. *Towards holistic machines : From visual recognition to question answering about real-world images*. Doctor of engineering science, Doctor of Engineering Science, Saarbrücken, Germany.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9.

Matthew Marge. 2019. Towards natural dialogue with robots. Presentation in the clasp seminar, Army Research Lab, Gothenburg, Sweden.

Matthew Marge and Alexander I. Rudnicky. 2019. Miscommunication detection and recovery in situated human–robot dialogue. *ACM Trans. Interact. Intell. Syst.*, 9(1):3:1–3:40.

David Marr. 2010. *Vision: A computational approach*. MIT Press Scholarship Online.

Arild Matsson. 2018. Implementing perceptual semantics in type theory with records (ttr). Masters in language technology (mlt), 30 hec, Masters in Language Technology (MLT), Department of Philosophy, Linguistics and Theory of Science. University of Gothenburg, Gothenburg, Sweden, September 24. Examiner: Peter Ljunglöf; supervisors: Simon Dobnik and Staffan Larsson; opponent: Axel Almqvist.

Arild Matsson, Simon Dobnik, and Staffan Larsson. 2019. ImageTTR: Grounding Type Theory with Records in image classification for visual question answering. In *Proceedings of the IWCS 2019 Workshop on Computing Semantics with Types, Frames and Related Structures*, pages 55–64, Gothenburg, Sweden. Association for Computational Linguistics.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012a. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2012b. Learning to parse natural language commands to a robot control system. In *Proceedings of the 13th International Symposium on Experimental Robotics (ISER)*.

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

Samuel McNerney. 2011. A brief guide to embodied cognition: Why you are not your brain. Guest blog November 4, Scientific American.

Hongyuan Mei, Mohit Bansal, and Matthew R Walter. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *AAAI*, pages 2772–2778.

Yousuf Ali Mohammed. 2020. Guesswhat?! from what we answered before: Improving the vqa task in goal-oriented games using the previous context of dialogue improving vqa task in goal-oriented games using the previous context of the dialogue. Masters in language technology (mlt), 30 hec, Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik and Mehdi Ghanimifard, examiner: Staffan Larsson.

Will Monroe. 2018. *Learning in the rational speech acts model*. Doctor of philosophy, Department of Computer Science, Stanford University.

Will Monroe, Noah D. Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. *CoRR*, abs/1606.03821.

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.

Marius Muja and David G Lowe. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331–340):2.

Amitabha Mukerjee. 1998. Neat versus scruffy: a review of computational models of spatial expressions. In Patrick Olivier and Klaus-Peter Gapp, editors, *Representation and Processing of Spatial Expressions*, pages 1–36. Lawrence Erlbaum Associates, Mahwah, N.J.

Edon Mustafa and Konstantinos Dimopoulos. 2014. Sign language recognition using kinect. In *Dautov, R., Gkasis, P., & Karama-nos, A. et al.(2014). Proceedings of the 9th South East EuropeanDoctoral Student Conference*, pages 271–285. Thessaloniki: SEERC.

Jason M. O'Kane. 2013. *A Gentle Introduction to ROS*. CreateSpace Independent Publishing Platform.

Patrick Olivier and Klaus-Peter Gapp, editors. 1998. *Representation and Processing of Spatial Expressions*. Lawrence Erlbaum Associates, Mahwah, N.J.

Sharon Oviatt, Björn Schuller, Philip R. Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger. 2017. *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 1*, volume 1. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA.

Sandro Pezzelle, Ionut-Teodor Sorodoc, and Raffaella Bernardi. 2018. Comparatives, quantifiers, proportions: A multi-task model for the learning of quantities from vision. *arXiv*, arXiv:1804.05018 [cs.CV]:1–12.

Giovanni Pezzulo, Lawrence Barsalou, Angelo Cangelosi, Martin Fischer, Michael Spivey, and Ken McRae. 2011. The mechanics of embodiment: A dialog on embodiment and computational modeling. *Frontiers in Psychology*, 2:5.

James Pustejovsky and Nikhil Krishnaswamy. 2020. Situated meaning in multimodal dialogue: Human-robot and human-computer interactions. Journal article manuscript, Department of Computer Science, Brandeis University.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Lisbon, Portugal. Association for Computational Linguistics.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's Mechanical Turk*, Los Angeles, CA. North American Chapter of the Association for Computational Linguistics (NAACL).

Terry Regier. 1996. *The human semantic potential: spatial language and constrained connectionism*. MIT Press, Cambridge, Massachusetts, London, England.

Terry Regier and Laura A. Carlson. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Gudula Retz-Schmidt. 1988. Various views on spatial prepositions. *AI magazine*, 9(2):95–95.

Benjamin Saul Rosman. 2014. *Learning domain abstractions for long lived robots*. Doctor of philosophy, Institute of Perception, Action and Behaviour, School of Informatics, The University of Edinburgh, Edinburgh.

Brandon Cain Roy. 2013. *The birth of a word*. Doctor of philosophy in media arts and sciences, Program in Media Arts and Sciences, School of Architecture and Planning, Massachusetts Institute of Technology.

Deb Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer speech and language*, 16(3):353–385.

Deb Roy. 2005. Semiotic schemas: a framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

Deb Roy. 2011. The birth of a word. Talk, TED: Ideas worth spreading.

Stuart J Russell, Peter Norvig, and Ernest Davis. 2016. *Artificial Intelligence: A Modern Approach*. Prentice Hall series in Artificial Intelligence. Pearson Education M.U.A.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A platform for embodied ai research. *arXiv*, arXiv:1904.01201 [cs.CV]:1–16.

Matthias Scheutz, Rehj Cantrell, and Paul Schermerhorn. 2011. Toward humanlike task-based dialogue processing for human robot interaction. *AI Magazine*, 32(4):77–84.

Thomas Scialom, Patrick Bordes, Paul-Alexis Dray, Jacopo Staiano, and Patrick Gallinari. 2020. What BERT sees: Cross-modal transfer for visual question generation. *arXiv*, arXiv:2002.10832 [cs.CL]:1–11.

Ravi Shekhar, Ece Takmaz, Raquel Fernández, and Raffaella Bernardi. 2019. Evaluating the representational hub of language and vision models. *arXiv*, arXiv:1904.06038 [cs.CL].

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Gabriel Skantze. 2016. Real-time coordination in human-robot interaction using face and voice. *AI Magazine*, 37(4):19–31.

Gabriel Skantze and Samer Al Moubayed. 2012. IrisTK: a statechart-based toolkit for multiparty face-to-face interaction. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 69–76. ACM.

Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication*, 65:50–66.

Danijel Skočaj, Miroslav Janiček, Matej Kristan, Geert-Jan M. Kruijff, Aleš Leonardis, Pierre Lison, Alen Vrečko, and Michael Zillich. 2010. A basic cognitive system for interactive continuous learning of visual concepts. In *ICRA 2010 workshop ICAIR - Interactive Communication for Autonomous Intelligent Robots*, pages 30–36, Anchorage, AK, USA.

Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. 2011. A system for interactive learning in dialogue with a tutor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2011*, San Francisco, CA, USA.

Lucia Specia. 2020. Challenges in evaluating vision and language tasks. Keynote talk at advances in language and vision workshop at acl 2020, Imperial College London, Online.

Luc Steels. 2012. *Experiments in cultural language evolution*. Advances in interaction studies. John Benjamins Pub. Co, Amsterdam and Philadelphia.

Luc Steels and Jean-Christophe Baillie. 2003. Shared grounding of event descriptions by autonomous robots. *Robotics and Autonomous Systems*, 43(2–3):163–173.

Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–489.

Luc Steels and Martin Loetzsch. 2009. Perspective alignment in spatial language. In Kenny R. Coventry, Thora Tenbrink, and John. A. Bateman, editors, *Spatial Language and Dialogue*. Oxford University Press.

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.

Axel Storckenfeldt. 2018. Categorisation of conversational games in free dialogue referring to spatial scenes. C-uppsats (bachelor's thesis/extended essay), Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik, examiner: Ylva Byrman.

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.

Stefanie Tellex. 2020. Towards complex language in partially observed environments. Mit csail embodied intelligence seminar, Brown University.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1).

Joshua B. Tenenbaum. 2020. Cognitive and computational building blocks for morehuman-like language in machines. Acl 2020 keynote, Center for Brains, Minds and Machines, MIT.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Jesse Thomason. 2019. Vision-and-dialog navigation (talk). Microsoft research talks, University of Washington.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*.

Andrea Lockerd Thomaz, Matt Berlin, and Cynthia Breazeal. 2005. An embodied computational model of social referencing. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 591–598. IEEE.

Ronja Utescher. 2019. Visual TTR - modelling visual question answering in type theory with records. In *Proceedings of the 13th International Conference on Computational Semantics - Student Papers*, pages 9–14, Gothenburg, Sweden. Association for Computational Linguistics.

Andrea Vedaldi. 2016. Convolutional networks for computer vision applications. IV&L summer school on vision and language, Malta. http://www.robots.ox.ac.uk/∼vedaldi/assets/teach/vedaldi16deepcv.pdf.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67. Association for Computational Linguistics.

Jette Viethen, Robert Dale, and Markus Guhe. 2011a. Generating subsequent reference in shared visual scenes: Computation vs. re-use. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1158–1167. Association for Computational Linguistics.

Jette Viethen, Robert Dale, and Markus Guhe. 2011b. The impact of visual context on the content of referring expressions. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 44–52. Association for Computational Linguistics.

Sida I Wang, Percy Liang, and Christopher D Manning. 2016. Learning language games through interaction. *arXiv preprint arXiv:1606.02447*.

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

Terry Winograd. 1976. *Understanding Natural Language*. Edinburgh University Press.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv*, arXiv:1502.03044 [cs.LG]:1–22.

Xiaofeng Xu, Ivor W. Tsang, and Chuancai Liu. 2018. Zero-shot learning with complementary attributes. *arXiv*, arXiv:1804.06505 [cs.CV].

Hee-Deok Yang. 2014. Sign language recognition with the kinect sensor based on conditional random fields. *Sensors*, 15(1):135–147.

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

Lei Yuan, David Uttal, and Steven Franconeri. 2016. Are categorical spatial relations encoded by shifting visual attention between objects? *PloS one*, 11(10):1–22.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.

Noga Zaslavsky, Terry Regier, Naftali Tishby, and Charles Kemp. 2019. Semantic categories of artifacts and animals reflect efficient coding. *arXiv*, arXiv:1905.04562 [cs.CL].

H. Zender, M. Janicek, and Geert-Jan Kruijff. 2012. Situated communication for joint activity in human-robot teams. *IEEE Intelligent Systems*, 27(2):27–35.

Hendrik Zender, Óscar Martínez-Mozos, Patric Jensfelt, Geert-Jan M. Kruijff, and Wolfram Burgard. 2008. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502. Special issue "From sensors to human spatial concepts".

Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. 2019. Large-scale visual relationship understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9185–9194.

Z. Zhang, Y. Wang, Q. Wu, and F. Chen. 2019. Visual relationship attention for image captioning. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.