# Conversational Grounding as Natural Language Supervision – the need for divergent agent data

**Oliver Lemon**
Interaction Lab, Dept. of Computer Science
Heriot-Watt University
Edinburgh, Scotland
o.lemon@hw.ac.uk

## Abstract

We explore how Conversational Grounding messages in Natural Language can provide general and detailed feedback mechanisms for learning. We first present the fine-grained and *targeted* feedback signals provided by Conversational Grounding and discuss their potential advantages in models of language and task learning. We argue that a key factor holding back research in this area is lack of appropriate data on tasks with *divergent* agents, which can *resolve disagreements and errors*, and we propose requirements and methods for new data collections enabling such work.

## 1 Introduction

This paper argues for a perspective which explores the use of 'conversational grounding' signals by interactive learning agents, so that cooperative agents can learn how to find common-ground (Dafoe et al., 2021). We discuss the potential advantages of this view, and propose new data collections that we will need to enable it, focussed on agent *divergence*.

In human communication, conversational 'grounding' dialogue acts, like repair ('Not that bowl, the white one') and clarification ('Is it white or brown?) are used to establish mutually agreed *common ground* when engaged in a collaborative task (Brennan and Clark, 1996). Recently, a number of position and survey papers (Schlangen, 2019; Chandu et al., 2021; Benotti and Blackburn, 2021) highlight the importance of these types of *targeted* communicative feedback signals in human task coordination and language learning. Such detailed signals are also likely to be useful for AI models of learning. See brief examples below:

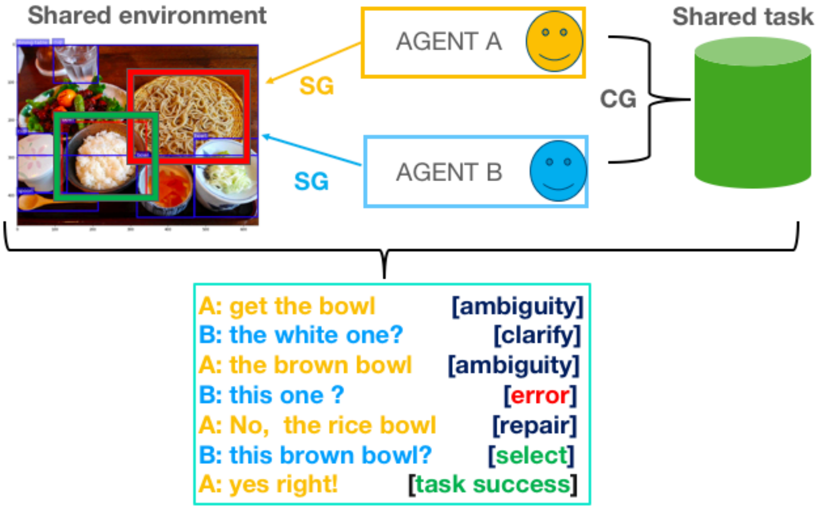
1. confirmation ("is it X?")
2. clarification ("is it X or Y?")
3. other-repair ("no not X but Y")
4. self-repair ("I mean X not Y")
5. underspecified repair ("that's not X")
6. implicit repair "the *bowl* is X."

As well as being human-understandable, conversational grounding signals could also lead to better task completion, better learned representations (for example in terms of disentanglment (Suglia et al., 2020)), and may also help to counter language drift in agent communities. We will discuss such potential advantages, as well as requirements on new data collections needed to train such systems. The paper closes with some future research directions.

## 2 Grounding in conversation versus symbol grounding

Current methods for developing collaborative and communicating AI agents which are situated in the real world focus on *symbol grounding* – see the image in figure 1. Here, language learning is often modelled as agents learning how to agree on targets in a visual scene (for example the brown bowl of rice), in tasks such as GuessWhat?! (De Vries et al., 2017). To achieve this, an omniscient Teacher agent instructs a Learner about the one true state of the scene and how to properly describe it in a formal or Natural Language (a form of Supervised Learning). But as well as being slow and requiring large amounts of data, this approach does not yet account for the ways in which *collaborative meaning is contextual and dynamic, local to a task, co-constructed, and negotiated.*

Consider for example the dialogue in figure 1. As this example illustrates, humans are able to learn and adapt meanings for specific tasks, so what makes conversation vital for learning is that it allows *divergent* agents to be flexible and adaptive enough to *rapidly reach agreement that is fit for current purposes*, to *adapt* their language when confronted with other agents and other tasks, and



A: get the bowl [ambiguity]
B: the white one? [clarify]
A: the brown bowl [ambiguity]
B: this one ? [error]
A: No, the rice bowl [repair]
B: this brown bowl? [select]
A: yes right! [task success]

Figure 1: Collaborative NL supervision: Symbol Grounding (SG) + Conversational Grounding (CG).

to *learn* new concepts. We call these processes, studied in Cognitive Science and models of conversation, *conversational grounding* (CG), rather than *symbol grounding* (SG). They have been argued to be universal in human languages and a foundation of cooperative communication (Brennan and Clark, 1996; Dingemanse et al., 2015; Healey et al., 2018; Benotti and Blackburn, 2021). Note that CG is independent of SG, and not built upon it, since CG happens also in non-visual tasks (such as buying an airline ticket, agreeing on a restaurant etc).

Related work on 'emergent communication' investigates agents which develop their own communication protocols which are generally not human-understandable (Kottur et al., 2017; Lazaridou and Baroni, 2020). Moreover, emergent communication work does not deal with fine-grained targeted messages such as clarification and repair which are essential to human collaboration, and which should also be beneficial for computational models of language learning (see below). This issue is partly related to the restricted nature of the data and tasks used, which are generally simple reference games such as GuessWhat?! using *shared* images (De Vries et al., 2017; Suglia et al., 2021), where *agents are not divergent* and phenomena such as repair do not occur (as we discuss in section 4), meaning that models for CG tasks cannot yet be trained.

### 2.1 Conversation, Collaboration, and Divergence

Current work in collaborative AI (Dafoe et al., 2021) uses a variety of *pre-trained Deep Learning models* with general vision and language capabilities, which are then fine-tuned for specific tasks. But such agents will not agree on everything, and currently they cannot detect or recover from disagreement or errors. This is a critical limitation, which prevents the latest neural models from being useful for developing collaborative learning agents. Sources of potential *divergence* are many – agents' perceptions and language may have been fine-tuned differently, or may have diverged due to continual learning, or they may have different visual perspectives on the current task, simply due to different physical locations, and they may have different plans. If one of the collaborating agents is human, divergences might also be individual (e.g. due to disability), idiosyncratic / personal, or cultural in nature.

We argue that we need to develop mechanisms for divergent agents to learn how to rapidly reach agreement on shared tasks – with each other and with humans – when they may have different perspectives, perceptions, language, and plans. But we already have such mechanisms in everyday global use: conversational grounding in Natural Language. Therefore, we propose a new focus on divergent agent tasks, which combine symbol grounding and conversational grounding: SG+CG

## 3 Potential advantages

Firstly, divergent agents which have no way to reach agreement are unable to collaborate – thus the fundamental potential advantage of SG+CG is that divergent agents will be able to complete a much wider range of shared tasks than current state-of-the-art agents – which are able to learn visually grounded language but cannot perform conversational collaborative grounding.

Secondly, conversational grounding might also be also more computationally effective than symbol grounding alone, as it provides new fine-grained feedback signals for machine learning, rapid adaptation, and optimisation.

Thirdly, the more detailed feedback signals provided by CG might lead to better quality of learned representations, for example perhaps enhancing aspects such as disentanglement, and compositionality of learned language. This is because, for example, specific properties of objects can be clarified and repaired, rather than whole class labels, in principle leading to more detailed model updates. Finally, other issues in emergent communication, such as language drift in communities of agents, might be impacted by CG abilities, as they provide agents with more targeted abilities to correct language use, concepts, and plans of others.

In summary, we could expect quantitative performance gains in terms of 1) ability to coordinate successfully (i.e. task success), 2) speed of learning and adaptation to new tasks (task efficiency), 3) the quality of learned neural representations, for example as determined using the CompGuessWhat!? multi-task evaluation framework (Suglia et al., 2020), and 4) properties of language learning in communities such as reduced language drift.

## 4 Data collection requirements for SG+CG

Training data for conversational grounding phenomena such as clarification and repair is missing from all the large-scale vision-and-language learning datasets (Schlangen, 2019; Benotti and Blackburn, 2021; Chandu et al., 2021)[1]. Current setups either 1) do not collect *any data at all* on collaborative *grounding phenomena* such as clarification and repair with *divergent agents*, or else 2) do not collect sufficient *volume* of such data; or 3) do not collect *'ecologically-valid'* data in scenarios which are close to real-world tasks. New datasets are needed.

Recent years have seen many different shared tasks and associated datasets for visually grounded language learning (to name a few: GuessWhat?! (De Vries et al., 2017), BURCHAK (Yu et al., 2017b), Minecraft (Narayan-Chen et al., 2019), CUPS (Loáiciga et al., 2021), CerealBar (Suhr et al., 2019), IGLU (Kiseleva et al., 2021), TEACh (Padmakumar et al., 2021)). However, most such tasks fail to meet the requirements of SG+CG since they do not focus on conversational grounding for divergent agents, and/or use only abstract shapes and images (Zarrieß et al., 2016; Yu et al., 2017b; Narayan-Chen et al., 2019; Suhr et al., 2019; Kiseleva et al., 2021) rather than real images or 3D scenes. Note that the 'Visual Dialog' work of (Das et al., 2017) does in fact use divergent agents (one can see the image, one cannot) but there is no shared task, making the whole dataset problematic (Agarwal et al., 2020; Massiceti et al., 2019).

Very often the data collection environments do not contain different agent perspectives or perceptions, nor multiple objects of the same type (so clarification is not needed for task success), and/or the tasks do not allow agents to express repairs or clarification (see GuessWhat?!). The few datasets which do meet most of the SG+CG requirements (e.g. CUPS, TEACh) fail to contain sufficient examples of miscommunication, repairs, semantic coordination etc required for model training.

What is needed is a new focus on CG for divergent agents in ecologically-valid environments such as AI2-THOR and VirtualHome (Kolve et al., 2019; Huang et al., 2022) – where we can create new large-scale collections of conversational grounding phenomena. New online data-collection tools will also be useful here, such as SLURK (Götze et al., 2022) – which allows multiple humans to communicate about controlled task environments.

Most recently, the TEACh (Padmakumar et al., 2021) dataset used in the 2022 Alexa SimBot challenge provides a set of 3000 human-human dialogues about shared tasks in a simulated 3D home (using the AI2-THOR simulator (Kolve et al., 2019)), where a Commander interacts with a Fol-

[1] Note that there are some useful datasets containing some non-visual CG phenomena, such as SMCalFlow – a large dialogue dataset about tasks involving calendars, weather, places, and people, which includes repair phenomena (Andreas et al., 2020)

lower. The Commander has egoentric 3D views from both agents, and oracle access to a map and task details. Crucially, the Follower can make mistakes which need correction, and the agents have different perspectives, leading to a limited form of divergence. This setup and environment is closer to what is needed for work on SG+CG

### 4.1 Methods for creating divergence, disagreement, and resolution

We will need to use data-collection techniques deploying divergent agents which can *resolve disagreements* – i.e. which can coordinate different perceptions, language, and knowledge, as was done in the seminal work on Maptask (Anderson et al., 1991). We can use similar ideas to create datasets where agents need to detect and correct ambiguity and disagreements via interaction. As (Schlangen, 2019) notes, current data sets such as GuessWhat?! are not useful here because for the human crowdworkers looking at the images *" ... the perceptual task being so easy for them, a need for dealing with miscommunication never arose ... and hence no such strategies can be learned from that data."* He proposes the MeetUp game (Ilinykh et al., 2019), where 2 players have to coordinate to meet in a particular place which they must agree on as being the target, and a variant MatchIt (without navigation) where the aim is for 2 agents to decide if they are looking at the same image. (Shekhar et al., 2017) create a version of an image captioning dataset (FOIL-COCO) where mistakes (perceptual divergence) are introduced into the original caption (and they showed that state-of-the-art vision and language models could not detect and correct such mistakes). The BURCHAK dataset (Yu et al., 2017b) collected repairs and corrections in an abstract teaching game, where Reinforcement Learning was used to train a grounded language learner for colours and shapes (Yu et al., 2017a). (Healey et al., 2018) use a chat tool (DiET) which allows edits and manipulations of dialogue turns, to collect data on handling of disagreements. Each of these efforts collects important collaborative phenomena around agent divergence, but has not collected sufficient volumes of data for model training.

Several other data collections such as REX and PentoRef (Tokunaga et al., 2012; Zarrieß et al., 2016) also focus on controlled settings where images are of abstract shapes rather than real-world, but nevertheless collected some data on phenomena such as repair in a joint task setting. The recent IGLU task (Kiseleva et al., 2021) also collects some clarification data, but also focusses on collaborative building of minecraft-style block structures, rather than realistic tasks. In terms of more realistic environments, in the CUPS corpus (Loáiciga et al., 2021) two agents have different views on a simulated tabletop scene, but where some different cups have been removed from each participant's view. Again, this setup elicits some repair and clarification phenomena of the type we target. All of these designs and experiences are useful for collecting SG+CG data, as they exemplify methods with which the data of interest can be collected.

### 4.2 Proposal: divergence in future SG+CG data collections

We propose to create new tasks similar to TEACh (using the open source AI2-THOR) but where we carefully control each agent's/human's *divergent perspectives, perceptions, language, and plans* relevant to a shared task. The agent *perspectives* will mean that different objects may be in view for each agent (this happens naturally when Leader and Follower are in different positions in a 3D scene). Divergent *perceptions and language* will involve agents using different class and property labels, and agents may also be given different *plans* which will require negotiation. Finally some scenes must also contain distractors (e.g. several different bowls, cups etc) - all of which will lead to a greater volume of repairs, clarification, and coordination. We can then create specific data-collections which focus on particular SG+CG tasks, for instance clarification and repair of object class/properties, repair of reference, and plan repair. Given sufficient data collected in this way, we can then train models for SG+CG in a multi-task fashion.

In summary, learning from previous work such as MapTask, Cups, FOIL, MeetUp, and TEACh, we propose to create new *divergent* Human-Human and Human-Agent data collections with greatly increased environmental pressure (Choi et al., 2018) to perform conversational grounding as a form of Natural Language supervision. The new tasks and data collections must involve distractors, ambiguity, vagueness, and different agent perspectives and plans leading to disagreement and coordination on a suitable common-ground (Schlangen, 2019; Benotti and Blackburn, 2021; Anderson et al., 1991). Finally, the simulated tasks should ideally have physical counterparts, so that communication trained from data collected in simulation can ultimately be tested in real-world scenarios. Approaches such as RoboTHOR (Deitke et al., 2020), built on AI2-THOR, are well-suited to this requirement.

## 5 Research directions

The main problem we have discussed is that current vision-and-language data sets do not allow us to learn conversational grounding Natural Language supervision signals. The new data collections outlined above will then enable several other important research directions:

- Developing and training new multi-task models of learning which are capable of understanding and generating conversational grounding inputs and outputs.
- Developing tools for the analysis of conversational grounding behaviours and policies (i.e. in what circumstances does grounding occur? What sequences of actions are effective?)
- Evaluating the benefits of SG+CG in real-world tasks where teams of agents (sometimes including humans) need to coordinate on shared tasks, for instance via RoboTHOR.

### Acknowledgements

Thanks to Alessandro Suglia and Ioannis Konstas for helpful discussions. This work is partially supported by the European Commission under the Horizon 2020 framework programme for Research and Innovation (H2020-ICT-2019-2, GA no. 871245), SPRING project, http://spring-h2020.eu

## References

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.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-Oriented Dialogue as Dataflow Synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Luciana Benotti and Patrick Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online. Association for Computational Linguistics.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22:482–1493.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. 2018. Compositional obverter communication learning from raw visual input.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground. *Nature*, 593.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4466–4475.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. 2020. Robothor: An open simulation-to-real embodied AI platform. *CoRR*, abs/2004.06799.

Mark Dingemanse, Seán Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa Gisladottir, Kobin Kendrick, Stephen Levinson, Elizabeth Manrique, Giovanni Rossi, and N. Enfield. 2015. Universal principles in the repair of communication problems. *PLOS ONE*, 10:e0136100.

Jana Götze, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann, and David Schlangen. 2022. The slurk Interaction Server Framework: Better Data for Better Dialog Models.

Patrick G. T. Healey, Gregory J. Mills, Arash Eshghi, and Christine Howes. 2018. Running Repairs: Coordinating Meaning in Dialogue. *Topics in Cognitive Science (topiCS)*, 10(2).

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment.

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, Arthur Szlam, Yuxuan Sun, Katja Hofmann, Michel Galley, and Ahmed Awadallah. 2021. Neurips 2021 competition iglu: Interactive grounded language understanding in a collaborative environment.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2019. AI2-THOR: An Interactive 3D Environment for Visual AI.

Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. *CoRR*, abs/1706.08502.

Angeliki Lazaridou and Marco Baroni. 2020. Emergent multi-agent communication in the deep learning era.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021. 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.

Daniela Massiceti, Puneet K. Dokania, N. Siddharth, and Philip H. S. Torr. 2019. Visual dialogue without vision or dialogue.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Teach: Task-driven embodied agents that chat.

David Schlangen. 2019. Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings.

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. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Alessandro Suglia, Yonatan Bisk, Ioannis Konstas, Antonio Vergari, Emanuele Bastianelli, Andrea Vanzo, and Oliver Lemon. 2021. An empirical study on the generalization power of neural representations learned via visual guessing games. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2135–2144, Online. Association for Computational Linguistics.

Alessandro Suglia, Ioannis Konstas, Andrea Vanzo, Emanuele Bastianelli, Desmond Elliott, Stella Frank, and Oliver Lemon. 2020. Compguesswhat?!: A multi-task evaluation framework for grounded language learning. In *Proceedings of ACL*.

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.

Takenobu Tokunaga, Ryu Iida, Asuka Terai, and Naoko Kuriyama. 2012. The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 422–429, Istanbul, Turkey. European Language Resources Association (ELRA).

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2017a. Learning how to learn: An adaptive dialogue agent for incrementally learning visually grounded word meanings. *Proceedings of the First Workshop on Language Grounding for Robotics*.

Yanchao Yu, Arash Eshghi, Gregory Mills, and Oliver Lemon. 2017b. The burchak corpus: a challenge data set for interactive learning of visually grounded word meanings. *Proceedings of the Sixth Workshop on Vision and Language*.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A corpus of spoken references in task-oriented dialogues. In

*Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131, Portorož, Slovenia. European Language Resources Association (ELRA).