

# You See, I Hear: Grounding Meaning Across Modalities

Naomi Pitzer

University of Southampton  
Southampton, United Kingdom  
np5n22@soton.ac.uk

Daniela Mihai

University of Southampton  
Southampton, United Kingdom  
A-D.Mihai@soton.ac.uk

## Abstract

Language is often viewed through the lens of embodied cognition and symbol grounding, where meaning arises from interaction with the environment rather than from abstract symbols alone. This work explores how divergent perceptual experiences affect the emergence of communication in artificial agents. By extending a multi-turn referential game to both unimodal and multimodal setups, we examine how sensory alignment influences the structure and semantics of emergent languages. Experiments across synthetic and environmental datasets reveal that unimodal agents, benefiting from shared perceptual grounding, exhibit lower entropy, higher referential accuracy, and more structured message encodings. In contrast, multimodal agents develop more abstract, distributed communication strategies. These findings highlight the role of perception in shaping emergent language and provide a framework for future work in adaptive, modality-aware communication systems.

## CCS Concepts

• **Computing Methodologies** → **Multi-agent reinforcement learning**.

## Keywords

Emergent communication, symbol grounding, multi-agent learning

### ACM Reference Format:

Naomi Pitzer and Daniela Mihai. 2018. You See, I Hear: Grounding Meaning Across Modalities. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (12th ACM Celebration of Women in Computing: womENCourage 2025)*. ACM, New York, NY, USA, 2 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 Introduction and Motivations

Noam Chomsky, widely regarded as the “father of modern linguistics”, once characterised spoken language as “a way of decoding noises [one] hears and converting them into a system that matches [one’s] own representations” [3]. While the biological underpinnings of language are extremely complex, this description encapsulates its cognitive role: mapping sensory information onto internal representations to achieve mutual understanding. This process is central to theories such as symbol grounding [5] and embodied cognition [6], which argue that symbols and words lack inherent

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*12th ACM Celebration of Women in Computing: womENCourage 2025, Braşov, Romania*  
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-XXXX-X/2018/06  
<https://doi.org/XXXXXXX.XXXXXXX>

meaning until they are grounded in sensory and motor experiences. According to these frameworks, language functions as a tool for encapsulating and transmitting elements of lived experience, making communication possible through the alignment of perceptual commonalities. However, when individuals perceive their environment through entirely different modalities, such as one through auditory input and the other through visual input, the absence of a common perceptual foundation complicates mutual understanding.

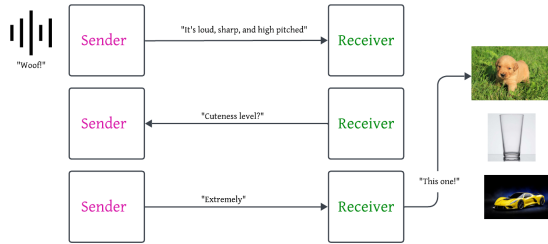
This challenge is not limited to humans. As we develop artificial agents and robots with diverse sensors and modalities, understanding how communication can emerge across perceptual boundaries becomes important. Emergent communication, where agents develop their own protocols through interaction, provides a strong framework to study this problem [1]. While prior work has shown its potential for coordination in artificial systems [2], most research assumes homogeneous perception, where agents perceive the environment through identical or overlapping modalities. This assumption neglects the reality of heterogeneous perceptual modalities, which are common in real-world scenarios and present both challenges and opportunities for advancing communication systems.

This research addresses this gap by investigating the communication strategies that emerge between agents with disjoint sensory modalities. The goal is to establish an experimental framework for future studies in multimodal emergent communication and to offer preliminary insights into how such systems adapt and develop shared language despite perceptual divergence. By exploring how agents with disjoint sensory modalities learn to communicate, this work opens new research directions in both artificial intelligence and linguistic theory. It lays a foundation for designing multi-agent systems, such as heterogeneous robotic teams operating in real-world environments, including applications like search-and-rescue missions. Moreover, it provides a framework for future studies on emergent semantics under sensory divergence, with potential relevance to communication strategies in assistive technologies and human populations with sensory impairments.

## 2 Multi-Modal Referential Game

To investigate, we implement a variant of the referential game, with our version based on the setup introduced by Evtimova et al.[4]. In standard referential games, a sender agent encodes an input stimulus, such as an image, into a message, which is then transmitted to a receiver agent [7]. The receiver, upon receiving the message along with a set of candidate images, including distractor images, must infer which image the sender intended to refer to. In our adapted version of the game, communication is extended to multiple time-steps: the receiver is allowed to send messages back to the sender before making a final decision. This bidirectional

interaction ensures that the emergent language is shaped collaboratively by both agents. Crucially, in our multimodal setting, the sender and receiver do not share the same perceptual modality. One agent perceives an image representation of the object, while the other perceives an audio version. To evaluate the impact of perceptual divergence, we implement both this multimodal variant and a unimodal counterpart in which both agents share the same sensory input. This comparison enables a more detailed analysis of how modality affects the structure and semantics of emergent language. The messages exchanged between agents in our system are binary strings, composed of sequences of 0s and 1s. These discrete, fixed-length messages are learned from scratch and not pre-defined, enabling the agents to develop their own communication protocol during training. The system was evaluated on both a simple synthetic dataset and a more complex, environmentally grounded dataset to assess its performance across varying levels of perceptual complexity.



**Figure 1: Example play of the multi-step multi-modal referential game proposed by Evtimova et al. [4]**

### 3 Experiments and Results

Our experiments revealed notable differences between unimodal and multimodal agents in how they develop and structure emergent communication protocols. These differences spanned accuracy, message structure, and representational strategy, reflecting the influence of modality alignment on communication efficiency. Unimodal agents slightly outperformed multimodal agents in referential accuracy and exhibited significantly lower receiver entropy, indicating reduced uncertainty and more deterministic communication. In contrast, multimodal agents, facing perceptual mismatch, tended to need longer and denser messages to compensate for the lack of shared modality.

Message similarity analysis showed higher within-class consistency in unimodal agents, reflecting more stable and deterministic encoding. However, variable bits (those that frequently flipped between 0 and 1) carried greater discriminatory weight in these systems than they did in multimodal systems, and perturbing them led to performance drops. Flipping constant bits (those that were 0 or 1 most of the time) had a weaker impact than it did on multimodal systems, suggesting variable bits encoded more fine-grained information about the input. In contrast, multimodal agents were largely unaffected by changes to variable bits but were highly sensitive to constant bit perturbations. This indicates a more abstract,

distributed encoding strategy, where meaning depended on stable bit patterns and surrounding context. While unimodal messages encoded lower-level features more directly, multimodal messages relied on more diffuse, relational structures.

Importantly, feature influence analysis showed that sender messages corresponding to low-frequency sounds formed distinct clusters in the embedding space, especially in the unimodal setup. While less pronounced in the multimodal case, this pattern suggests that emergent messages still reflect meaningful properties of the input and are not entirely abstract. These structural patterns were also present, though somewhat attenuated, in experiments on the more complex environmental dataset.

These findings demonstrate that perceptual divergence leads to more abstract, distributed communication protocols, while also highlighting the adaptability of emergent language systems in bridging heterogeneous experiences. Beyond characterising agent strategies, this research also highlights the importance of input embedding generation, methods in which sensory data is represented before being processed by agents. In both synthetic and environmentally grounded settings, the clarity of emergent messages was closely tied to how perceptual inputs were encoded. As such, achieving real-world applicability will require further research into more effective embedding methods. Overall, this work not only uncovers preliminary communication strategies for heterogeneous agents but also lays the groundwork for future studies into perception-aware communication, robust message encoding, and representation learning in multimodal environments.

### 4 Conclusion

Beyond its immediate findings, this study advocates a broader reconceptualisation of emergent communication. It encourages that we approach it not as a product of shared representations, but as an adaptive process shaped by perceptual context. Future work could build on this framework to support real-world coordination between heterogeneous agents, such as audio-driven and vision-driven robots, or inform linguistic research into how meaning emerges between individuals with different sensory or cognitive experiences.

### References

- [1] Brendon Boldt and David Mortensen. 2024. A Review of the Applications of Deep Learning-Based Emergent Communication. arXiv:2407.03302 [cs.CL] <https://arxiv.org/abs/2407.03302>
- [2] Liqun Chen and Siaw-Lynn Ng. 2021. Securing emergent behaviour in swarm robotics. arXiv:2102.03148 [cs.RO] <https://arxiv.org/abs/2102.03148>
- [3] Noam Chomsky. 1989. The Concept of Language - Upon Reflection - University of Washington Interview. <https://www.youtube.com/watch?v=hdUblwHRkY>
- [4] Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho. 2018. Emergent Communication in a Multi-Modal, Multi-Step Referential Game. arXiv:1705.10369 [cs.LG] <https://arxiv.org/abs/1705.10369>
- [5] Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 1 (1990), 335–346. doi:10.1016/0167-2789(90)90087-6
- [6] George Lakoff and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and its Challenges to Western Thought*. University of Chicago Press, Chicago. <https://www.degruyter.com/database/COGBIB/entry/cogbib.7098/html> 2010.
- [7] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-Agent Cooperation and the Emergence of (Natural) Language. arXiv:1612.07182 [cs.CL] <https://arxiv.org/abs/1612.07182>