Language-Goal Imagination to Foster Creative Exploration in Deep RL

Anonymous Authors[1]

Abstract

Developmental machine learning studies how artificial agents can model the way children learn open-ended repertoires of skills. Children are known to use language and its compositionality as a tool to imagine descriptions of outcomes they never experienced before and target them as goals during play. We introduce IMAGINE, an intrinsically motivated deep RL architecture that models this ability. Such imaginative agents, like children, benefit from the guidance of a social peer who provides language descriptions. To take advantage of goal imagination, agents must be able to leverage these descriptions to interpret their imagined goals. This generalization is made possible by modularity: a decomposition between learned goal-achievement reward function and policy relying on deep sets, gated attention and object-centered representations. We introduce the Playground environment and study how this form of goal imagination improves generalization and exploration over agents lacking this capacity.

1. Introduction

Building autonomous machines that can discover and learn a variety of skills is a long-standing goal in Artificial Intelligence. In this quest, we can draw inspiration from the way children learn and explore open-ended environments (Asada et al., 2009). In such exploration, *intrinsic motivations* are key to trigger spontaneous exploration, for the mere purpose of experiencing novelty, surprise or learning progress (Kaplan & Oudeyer, 2007; Kidd & Hayden, 2015).

Children also leverage the properties of language to assimilate years of experience embedded in their culture, in only a few years (Tomasello, 1999; Bruner, 1991). As they discover language, their goal-driven exploration changes. Piaget (1926) identified a form of egocentric speech where children narrate their ongoing activities. Later, Vygotsky (1978) showed that they can generate novel plans and goals by using the expressive generative properties of language.

This paper presents **I**ntrinsic **M**otivations **A**nd **G**oal **IN**vention for **E**xploration (IMAGINE): a learning architecture which leverages natural language (NL) interactions with a descriptive social partner (SP) to explore procedurally-generated scenes and interact with objects. IMAGINE discovers meaningful environment interactions through its own exploration (Figure 1a) and episode-level NL descriptions provided by SP (1b). These descriptions are turned into targetable goals by the agent (1c). It learns to represent them by jointly training a language encoder mapping NL to goal embeddings and a goal-achievement reward function (1d). The latter provides training signals for policy learning (ticks in Figure 1d-e). More importantly, IMAGINE can invent new goals by composing known ones, and trains on them autonomously via internal signals (1f).

Related work. The idea that language understanding is grounded in one's experience of the world and should not be secluded from the perceptual and motor systems has a long history in Cognitive Science (Glenberg & Kaschak, 2002; Zwaan & Madden, 2005). This vision was transposed to intelligent systems (Steels, 2006; McClelland et al., 2019), applied to human-machine interaction (Dominey, 2005; Madden et al., 2010) and recently to deep RL via frameworks such as *BabyAI* (Chevalier-Boisvert et al., 2019).

In their review of *RL algorithms informed by NL*, Luketina et al. (2019) distinguish between *language-conditional* problems where language is required to solve the task and *language-assisted* problems where language is a supplementary help. In the first category, most works propose instruction-following agents (R. K. Branavan et al., 2010; Chen & Mooney, 2011; Bahdanau et al., 2019; Co-Reyes et al., 2018; Jiang et al., 2019; Goyal et al., 2019; Cideron et al., 2019). Although our system is *language-conditioned*, it is not *language-instructed*: it is never given any instruction or reward but sets its own goals and learns its own internal reward function. Bahdanau et al. (2019) and Fu et al. (2019) also learn a reward function but require extensive expert knowledge (expert dataset and known environment dynamics respectively), whereas our agent uses experience generated by its own exploration.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

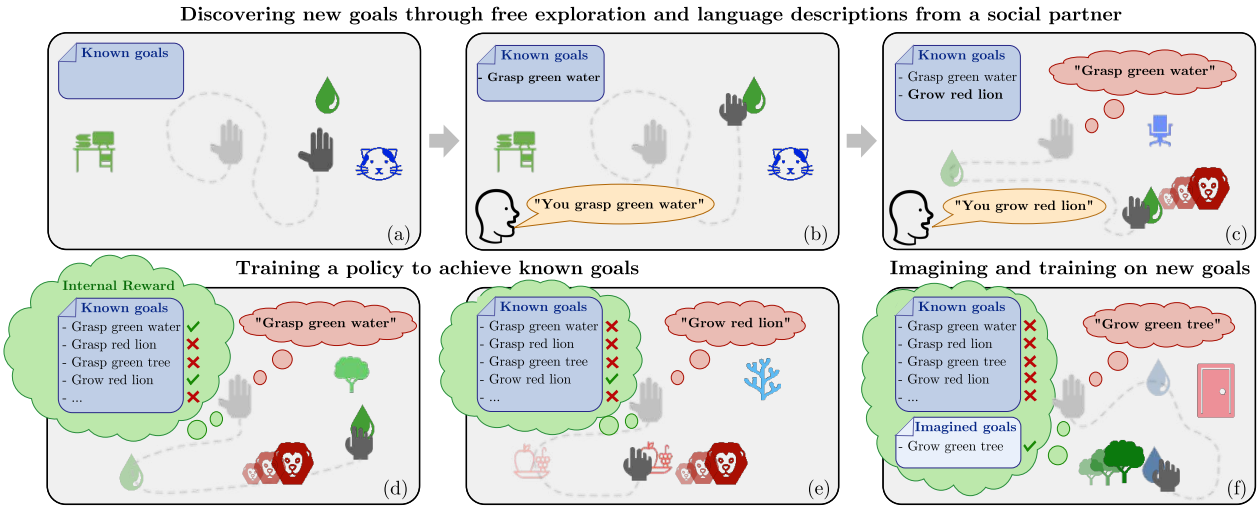


Figure 1: **IMAGINE overview**. In the *Playground* environment, the agent (hand) can move, grasp objects and grow some of them. Scenes are generated procedurally with objects of different types, colors and sizes. A social partner provides descriptive feedback (orange), that the agent converts into targetable goals (red bubbles).

Imagining goals by composing known ones only works in association with *systematic generalization* (Bahdanau et al., 2018; Hill et al., 2019): generalizations of the type *grow any animal* + *grasp any plant* → *grow any plant*. These were found to emerge in instruction-following agents, including generalizations to new combinations of motor predicates, object colors and shapes (Hermann et al., 2017; Hill et al., 2019; Bahdanau et al., 2019). Because systematic generalization can only occur when objects share common attributes (e.g. type, color), we represent objects as *single-slot object files* (Green & Quilty-Dunn, 2017): separate entities characterized by shared attributes. We introduce a new object-centered inductive bias: object-based modular architectures based on Deep Sets (Zaheer et al., 2017).

Contributions. We introduce goal imagination using language compositionality to drive exploration of intrisically motivated agents and a modular architecture to enable it.

2. Problem Definition

Open-ended learning environment. We consider a setup where agents evolve in an environment filled with objects and without prior on the set of possible interactions. An agent decides what to learn by setting its own goals, and has no access to external rewards. However, to allow the agent to learn relevant skills, a social partner (SP) watches the scene to guide the agent. Following a developmental approach (Asada et al., 2009), we propose a hard-coded surrogate SP modelling important aspects of human development:

- At the beginning of each episode, the agent chooses a goal by formulating a sentence. SP then provides agents with optimal learning opportunities by organizing the scene with: 1) the required objects to reach the goal (relevant scene) 2) procedurally-generated distracting objects (new scene and providing further discovery opportunities). This is close to a developmental scaffolding modelling the Zone of Proximal Development (ZPD) introduced by Vygotsky to describe infant-parent learning dynamics (Vygotsky, 1978).
- At the end of each episode, SP utters a set of sentences describing achieved and meaningful outcomes (except sentences from a test set). Linguistic guidance given through descriptions are a key component of how parents "teach" language to infants. It contrasts with instruction following (providing a linguistic command and then a reward), rarely seen in real parent-child interactions (Tomasello, 2009; Bornstein et al., 1992).

Learning objective. This paper investigates how goal imagination can lead agents to creatively explore their environment and discover interesting interactions with objects around. In this quest, SP guides agents towards a set of interesting outcomes by uttering NL descriptions. Through compositional recombinations of these sentences, goal imagination aims to drive creative exploration, to push agents to discover outcomes beyond the set of outcomes known by SP. We evaluate this desired behavior by two metrics: 1) the generalization of the policy to new language goals unknown to SP, (test set defined in Supp. Section 1) ; 2) goal-oriented exploration metrics. We measure generalization for each goal as the success rate SR over 30 episodes and report $\overline{\text{SR}}$ the average over goals. We evaluate exploration with the *interesting interaction count* (I2C). Given a set of interaction, I2C measures the penchant of agents to explore interactions of this set, (details in Supp. Section 2). We report means and *std* over 10 seeds and statistical significance using a two-tail Welch's t-test at level $\alpha = 0.05$ (Colas et al., 2019).

3. Methods

3.1. The *Playground* environment

We introduce *Playground*, a simple environment designed to study the impact of goal imagination on exploration and generalization by disentangling it from the problems of perception and fully-blown NL understanding. *Playground* is a continuous 2D world, with procedurally-generated scenes containing N = 3 objects, from 32 different object types (*e.g. dog, cactus, sofa, water, etc.*), organized into 5 categories (*animals, furniture, plants, etc*), see Figure 1.

Agent perception and embodiment. Agents have access to state vectors describing the scene: the agent's body and the objects. Each object is represented by a set of features describing its type, position, color, size and whether it is grasped. The agent can perform bounded translations in the 2D plane, grasp and release objects with its gripper. It can make animals and plants grow by bringing them the right supply (food or water for animals, water for plants).

Goal and description A simple grammar generates the descriptions of the 256 achievable goals, such as *go bottom left, grasp red cat* or *grow green animal*. Supp. Section 1 completely details the environment and the grammar

3.2. The IMAGINE Architecture

IMAGINE agents build a repertoire of goals and train two internal models: 1) a goal-achievement reward function $\mathcal{R}$ to predict whether a given description matches a behavioral trajectory; 2) a policy Π to achieve behavioral trajectories matching descriptions. The architecture is presented in Figure 2 and follows this logic:

1. The *Goal Generator* samples a target goal g_{target} from known and imagined goals ($\mathcal{G}_{\text{known}} \cup \mathcal{G}_{\text{im}}$).
2. The agent (*RL Agent*) interacts with the environment using its policy Π conditioned on g_{target}.
3. State-action trajectories are stored in *mem*(Π).
4. SP's descriptions of the last state are considered as potential goals $\mathcal{G}_{\text{SP}}(\mathbf{s}_T) = \mathcal{D}_{\text{SP}}(\mathbf{s}_T)$.
5. *mem*($\mathcal{R}$) stores positive pairs ($\mathbf{s}_T$, $\mathcal{G}_{\text{SP}}(\mathbf{s}_T)$) and infers negative pairs ($\mathbf{s}_T$, $\mathcal{G}_{\text{known}} \setminus \mathcal{G}_{\text{SP}}(\mathbf{s}_T)$).
6. The agent then updates:
   - *Goal Gen.*: $\mathcal{G}_{\text{known}} \leftarrow \mathcal{G}_{\text{known}} \cup \mathcal{G}_{\text{SP}}(\mathbf{s}_T)$ and $\mathcal{G}_{im} \leftarrow \text{Imagination}(\mathcal{G}_{\text{known}})$.
   - *Language Encoder* (L_e) and *Reward Function* ($\mathcal{R}$) are updated using data from *mem*($\mathcal{R}$).
   - *RL agent*: We sample a batch of state-action transitions ($\mathbf{s}$, $\mathbf{a}$, $\mathbf{s}'$) from *mem*(Π). Then, we use *Hindsight Replay* and $\mathcal{R}$ to bias the selection of substitute goals to train on (g_{s}) and compute the associated rewards ($\mathbf{s}$, $\mathbf{a}$, $\mathbf{s}'$, g_{s}, r). Finally, the policy and critic are trained via RL.

Goal generator. It is a generative model of NL goals. It generates target goals g_{target} for data collection and substitute goals g_{s} for hindsight replay. When goal imagination is disabled, the goal generator samples uniformly from the set of known goals $\mathcal{G}_{\text{known}}$, sampling random vectors if empty. When enabled, it samples with equal probability from $\mathcal{G}_{\text{known}}$ and $\mathcal{G}_{\text{im}}$ (set of imagined goals). $\mathcal{G}_{\text{im}}$ is generated using a mechanism grounded in construction grammar that leverages the compositionality of language to imagine new goals from $\mathcal{G}_{\text{known}}$. The heuristic consists in computing sets of *equivalent words*: words that appear in two sentences that only differ by one word. For example, from *grasp red lion* and *grow red lion*, *grasp* and *grow* can be considered *equivalent* and from *grasp green tree* one can imagine a new goal *grow green tree* (see Figure 1f). Among them, some are meaningless, some are syntactically correct but infeasible (e.g. *grow red lamp*) and some belong to $\mathcal{G}^{\text{test}}$, or even to $\mathcal{G}^{\text{train}}$ before they are encountered by the agent and described by SP. The pseudo-code and all imaginable goals are provided in Supp. Section 3.

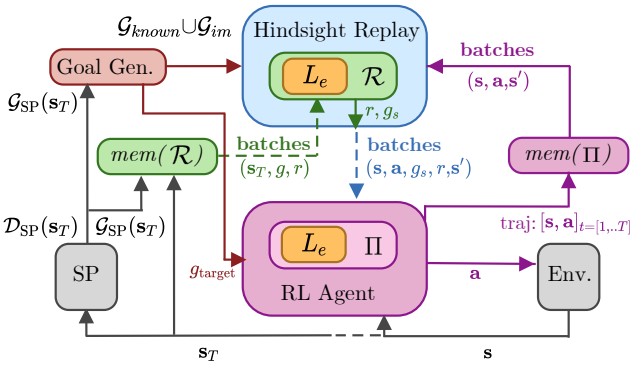
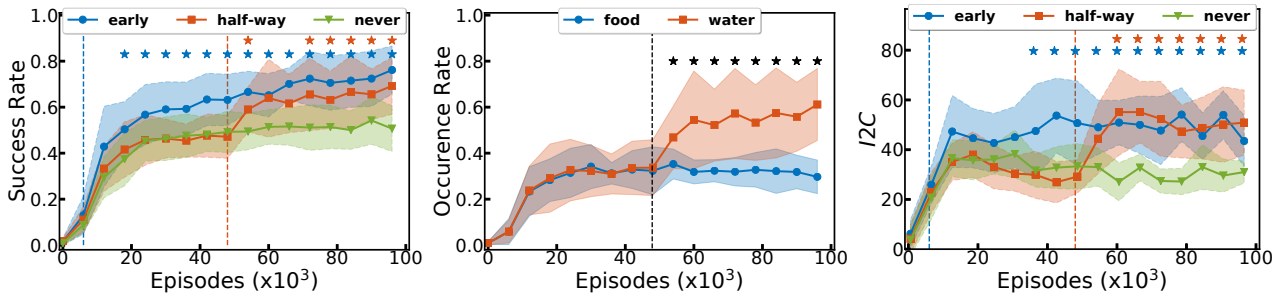


Figure 2: **IMAGINE architecture.** Colored boxes show the different modules of IMAGINE. Lines represent update signals (dashed) and function outputs (plain). The language encoder L_e is shared. See Supp. Section 5.1

Object-centered modular architectures. The goal-achievement reward function, policy and critic leverage novel *modular-attention* (MA) architectures based on Deep Sets (Zaheer et al., 2017), gated attention mechanisms (Chaplot et al., 2017) and object-centered representations. The idea is to ensure efficient skill transfer between objects, no matter their position in the state vector. This is done through the combined use of a shared neural network that encodes object-specific features and a permutation-invariant function to aggregate the resulting latent encodings. The shared network independently encodes, for each object, an affordance between this object (object observations), the agent (body observations) and its current goal. The goal embedding, generated by the language encoder L_e, is first cast into an attention vector in $[0, 1]$, then fused with the concatenation of object and body features via an Hadamard product (gated-attention (Chaplot et al., 2017)). The resulting object-



(a) Testing set performance (b) Behavioral adaptation (c) Exploration metrics on testing set

Figure 3: **Goal imagination drives exploration and generalization.** Vertical dashed lines mark the onset of goal imagination. (a) $\overline{\text{SR}}$ on testing set. (b) Behavioral adaptation, empirical probabilities that the agent brings supplies to a plant when trying to grow it. (c) I2C computed on the testing set. Stars indicate significance (a and c are tested against *never*).

specific encodings are aggregated by a permutation-invariant function and mapped to the desired output via a final network (e.g. into actions or action-values). Supp. Section 5.1 provides visual representations and details about the reward function, policy and critic based on this architecture.

4. Experiments and Results

As IMAGINE achieve near perfect generalizations on training goals: $\overline{\text{SR}} = 0.95 \pm 0.05$, we focus on the testing set performance. Supp. Sections 2 to 4 provide more results.

Global generalization performance. Figure 3a shows $\overline{\text{SR}}$ on the set of testing goals, when the agent starts imagining new goals early (after $6 \cdot 10^3$ episodes), half-way (after $48 \cdot 10^3$ episodes) or when not allowed to do so. Imagining goals leads to significant improvements in generalization.

A particular generalization: growing plants. Agents learn to grow animals from SP's descriptions, but are never told they could grow plants. When evaluated offline on the *growing-plants* goals before goal imagination, agents' policies perform a sensible zero-shot generalization and bring them water or food with equal probability, as they would do for animals (Figure 3b, left). As they start imagining these goals, their behavior adapts (Figure 3b, right). A reward function with good zero-shot abilities only provides positive rewards when the agent brings water. The policy slowly adapts to this internal reward signal and pushes agents to bring more water. We name this: *behavioral adaptation*.

Exploration. Figure 3c presents the I2C metric computed on the set of interactions related to $\mathcal{G}^{\text{test}}$ and demonstrates the exploration boost triggered by goal imagination. Supp. Section 2 additional results on exploration.

Modularity is Essential to Goal Imagination. Flat architectures (FA) consider the whole scene at once. We compared policies based on MA and in Table 4. Both used MA reward function as the FA reward function showed poor performance on $\mathcal{G}^{\text{train}}$. MA shows stronger generalization and is the only architecture allowing an additional boost with goal imagination. Only MA policy can leverage the novel reward signals coming from imagined goals and show *behavioral adaptation*.

Figure 4: Policy architectures performance. $\overline{\text{SR}}_{\text{test}}$ at convergence.

	MA *	FA
Im.	0.76 ± 0.1	0.15 ± 0.05
No Im.	0.51 ± 0.1	0.17 ± 0.04
p-val	4.8e-5	0.66

5. Discussion and Conclusion

IMAGINE is an autonomous learning architecture that leverages NL interactions with a social partner. IMAGINE sets its own goals and builds behavioral repertoires without external rewards. As such, it is distinct from traditional instruction-following RL agents. This is done through the joint training of a language encoder for goal representation and a goal-achievement reward function to generate internal rewards. Our proposed modular architectures with gated-attention enable efficient out-of-distribution generalization of the reward function and policy. The ability to imagine new goals by composing known ones leads to further improvements over initial generalization abilities and fosters exploration beyond the set of interactions relevant to SP. Our agent even tries to grow pieces of furniture with supplies, a behavior that can echo the way a child may try to feed his doll.

IMAGINE does not need externally-provided rewards but learns which behaviors are *interesting* from language-based interactions with SP. In contrast with hand-crafted reward functions, NL descriptions provide an easy way to guide machines towards relevant interactions. In addition, Section 4 shows that agents can learn to achieve goals from a relatively small number of descriptions, paving the way towards human-provided descriptions.

References

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., and Yoshida, C. Cognitive developmental robotics: A survey. *IEEE transactions on autonomous mental development*, 1(1):12–34, 2009.

Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T. H., de Vries, H., and Courville, A. Systematic generalization: What is required and can it be learned?, 2018.

Bahdanau, D., Hill, F., Leike, J., Hughes, E., Kohli, P., and Grefenstette, E. Learning to Understand Goal Specifications by Modelling Reward. In *International Conference on Learning Representations*, jun 2019.

Bornstein, M. H., Tamis-LeMonda, C. S., Tal, J., Ludemann, P., Toda, S., Rahn, C. W., Pêcheux, M.-G., Azuma, H., and Vardi, D. Maternal responsiveness to infants in three societies: The united states, france, and japan. *Child development*, 63(4):808–821, 1992.

Bruner, J. The Narrative Construction of Reality. *Critical Inquiry*, 18(1):1–21, oct 1991. ISSN 0093-1896. doi: 10.1086/448619.

Chaplot, D. S., Sathyendra, K. M., Pasumarthi, R. K., Rajagopal, D., and Salakhutdinov, R. Gated-attention architectures for task-oriented language grounding, 2017.

Chen, D. L. and Mooney, R. J. Learning to Interpret Natural Language Navigation Instructions from Observations. In *AAAI Conference on Artificial Intelligence (AAAI), 2011*, 2011.

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T. H., and Bengio, Y. Baby{AI}: First Steps Towards Grounded Language Learning With a Human In the Loop. In *International Conference on Learning Representations*, 2019.

Cideron, G., Seurin, M., Strub, F., and Pietquin, O. Self-educated language agent with hindsight experience replay for instruction following. *arXiv preprint arXiv:1910.09451*, 2019.

Co-Reyes, J. D., Gupta, A., Sanjeev, S., Altieri, N., Andreas, J., DeNero, J., Abbeel, P., and Levine, S. Guiding policies with language via meta-learning, 2018.

Colas, C., Sigaud, O., and Oudeyer, P.-Y. A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms. *arXiv preprint arXiv:1904.06979*, 2019.

Dominey, P. F. Emergence of grammatical constructions: evidence from simulation and grounded agent experiments. *Connection Science*, 17(3-4):289–306, sep 2005. ISSN 0954-0091. doi: 10.1080/09540090500270714.

Fu, J., Korattikara, A., Levine, S., and Guadarrama, S. From Language to Goals: Inverse Reinforcement Learning for Vision-Based Instruction Following. In *International Conference on Learning Representations*, 2019.

Glenberg, A. M. and Kaschak, M. P. Grounding language in action. *Psychonomic Bulletin & Review*, 9(3):558–565, sep 2002. ISSN 1069-9384. doi: 10.3758/BF03196313.

Goyal, P., Niekum, S., and Mooney, R. J. Using Natural Language for Reward Shaping in Reinforcement Learning. In *IJCAI 2019*, mar 2019. URL http://arxiv.org/abs/1903.02020.

Green, E. J. and Quilty-Dunn, J. What is an object file? *The British Journal for the Philosophy of Science*, 2017.

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., Wainwright, M., Apps, C., Hassabis, D., and Blunsom, P. Grounded Language Learning in a Simulated 3D World. jun 2017. URL http://arxiv.org/abs/1706.06551.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., and Santoro, A. Emergent systematic generalization in a situated agent, 2019.

Jiang, Y., Gu, S., Murphy, K., and Finn, C. Language as an Abstraction for Hierarchical Deep Reinforcement Learning. In *Workshop on "Structure & Priors in Reinforcement Learning" at ICLR 2019*, jun 2019. URL http://arxiv.org/abs/1906.07343.

Kaplan, F. and Oudeyer, P.-Y. In search of the neural circuits of intrinsic motivation. *Frontiers in neuroscience*, 1:17, 2007.

Kidd, C. and Hayden, B. Y. The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460, 2015.

Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., Whiteson, S., and Rocktäschel, T. A Survey of Reinforcement Learning Informed by Natural Language. *IJCAI'19*, jun 2019. URL http://arxiv.org/abs/1906.03926.

Madden, C., Hoen, M., and Dominey, P. F. A cognitive neuroscience perspective on embodied language for human–robot cooperation. *Brain and Language*, 112(3):180–188, mar 2010. ISSN 0093-934X. doi: 10.1016/J.BANDL.2009.07.001.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., and Schütze, H. Extending machine language models toward human-level language understanding. *arXiv preprint arXiv:1912.05877*, 2019.

Piaget, J. *The language and thought of the child*. Routledge, 1926. ISBN 0415267501.

R. K. Branavan, S., S. Zettlemoyer, L., and Barzilay, R. Reading Between the Lines: Learning to Map High-level Instructions to Commands. In *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 1268–1277, 2010.

Steels, L. Semiotic dynamics for embodied agents. *IEEE Intelligent Systems*, 21(3):32–38, 2006.

Tomasello, M. *The cultural origins of human cognition*. Harvard University Press, 1999. ISBN 9780674005822.

Tomasello, M. *Constructing a language*. Harvard university press, 2009.

Vygotsky, L. S. Tool and Symbol in Child Development. In *Mind in Society*, chapter Tool and Symbol in Child Development, pp. 19–30. Harvard University Press, 1978. ISBN 0674576292. doi: 10.2307/j.ctvjf9vz4.6.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In *Advances in neural information processing systems*, pp. 3391–3401, 2017.

Zwaan, R. and Madden, C. Embodied sentence comprehension. *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, pp. 224–245, 2005. doi: 10.1017/CBO9780511499968.010.