Collaborative meaning-making: the emergence of novel languages in humans, machines, and human-machine interactions

Kouwenhoven, T.

7

Shaping Shared Languages

Languages are shaped by the inductive biases of their users. Using a classical referential game, we investigate how artificial languages evolve when optimised for inductive biases in humans and large language models (LLMs) via Human-Human, LLM-LLM and Human-LLM experiments. We show that referentially grounded vocabularies emerge that enable reliable communication in all conditions, even when humans and LLMs collaborate. Comparisons between conditions reveal that languages optimised for LLMs subtly differ from those optimised for humans. Interestingly, interactions between humans and LLMs alleviate these differences and result in vocabularies that are more human-like than LLM-like. These findings advance our understanding of how inductive biases in LLMs play a role in the dynamic nature of human language and contribute to maintaining alignment in human and machine communication. In particular, our work highlights the need to develop new methods that incorporate human interaction into the training processes of LLMs, and demonstrates that using communicative success as a reward signal can be a fruitful and novel direction.

Originally published as: Kouwenhoven, T., Peeperkorn, M., de Kleijn, R.E. and Verhoef, T. (2025). Shaping Shared Languages: Human and Large Language Models' Inductive Biases in Emergent Communication. In Kwok, J., editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI-25, International Joint Conferences on Artificial Intelligence Organization. Human-Centred AI

7.1 Introduction

Languages adapt to how they are learned and used. The primary reason is the continuous influence of individuals' (learning) biases and pressures that slowly shape languages to become more structured, easier to learn and communicatively efficient (Smith, 2022). Although a wealth of experiments in the field of language evolution have contributed to this knowledge (i.a. Kirby et al., 2014, 2015; Raviv et al., 2019a), only relatively recently have we started investigating whether these principles can be applied to large language models as well (Galke et al., 2024). For instance, more systematic and structured languages are typically easier for humans to learn when asked to learn novel artificial languages (Raviv et al., 2021). Recent work by Galke et al. (2024) revealed that the same is true for recurrent neural networks and transformer-based LLMs. Moreover, transmission of initially unstructured language systems over generations of human learners (i.e., iterated learning) increases structure and learnability in these languages (Kirby et al., 2015). To investigate whether this process leads to a similar outcome with LLMs, the work in Chapter 6 created a setting in which LLMs learned an initially holistic, unstructured artificial language and then repeatedly used it to communicate in a referential game. This showed that the linguistic structure of these languages increased, which enabled more successful communication between LLM agents, again mirroring observations from human experiments (Kirby et al., 2015).

With AI systems being increasingly incorporated into our daily lives (Brinkmann et al., 2023), it is argued that repeated interactions with machines become increasingly important to maintain alignment (Mikolov et al., 2018; Beuls and Van Eecke, 2024) and referential grounding (Chapter 1). In the case of humans, these repeated interactions cause languages to evolve in a way that accommodates the specific abilities and preferences of minority individuals at the group level (Josserand et al., 2024). Since the seemingly similar ways that languages adapt and optimise as a result of learning and use in both Human-Human and LLM-LLM interactions, the question that arises is whether these processes can also be used to experimentally evolve a language that is optimised for humans *and* LLMs. In other words, can humans and LLMs collaboratively shape a language that is easy to learn for both and allows for successful communication? If so, what do these languages look like?

This is investigated here by extending Chapter 6. Firstly, we provide experimental data of humans playing the same referential game used with LLMs in Chapter 6, allowing comparisons between languages evolved through LLM-LLM interactions with those resulting from Human-Human interactions. Secondly, we run experiments where humans collaborate with an LLM (Figure 7.1)[1]. While it is unclear how human and LLM abilities exactly differ, this allows us to test whether an artificial holistic language can be optimised for the inductive biases of two different types of language learners. If shared vocabularies of signals and meanings emerge that allow for successful communication, one could argue that there has been some form of

[1]This study was approved by the ethics department of Leiden University (2024-03-11-R.E. de Kleijn-V1-5354)

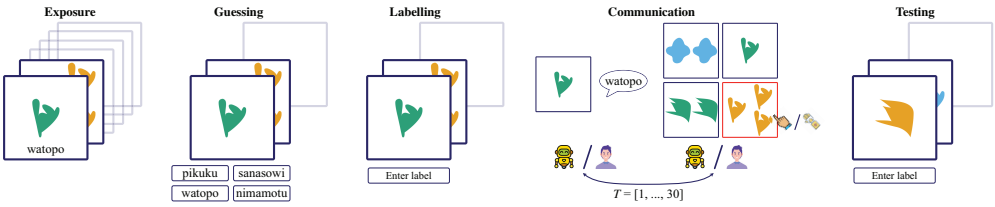


Figure 7.1: The experimental blocks in our experiment. Participants go through the exposure and guessing block twice before labelling each of the 15 training stimuli in the labelling block. The communication block is performed for 4 rounds each consisting of 30 tasks T, where participants alternate speaker-listener roles for each stimulus once. Participants label 27 (15 original and 12 novel) stimuli in the testing block. Image is adapted from Chapter 6. Icons obtained from flaticon.com

referential grounding, a prerequisite for successful communication (Clark and Brennan, 1991). Finally, Human-LLM collaboration allows investigating if and how the evolved language differs from languages that evolve within Human-Human and LLM-LLM interactions.

Our results show that structured and referentially grounded languages can emerge when humans and LLMs interact repeatedly in our experiment. The emergent languages from these interactions tend to be more human-like than LLM-like, suggesting that the LLMs are flexible towards the strong human preferences that shape the languages. Finally, languages optimised for LLMs result in less variation and are more degenerate than those optimised for humans.

7.2 Background

In this section, we discuss relevant work in the field of language evolution, the role of inductive biases in language evolution, and why this is relevant for LLM research.

7.2.1 Language evolution

Language allows us to communicate successfully because of the vocabulary we share, but also due to its open-ended nature, which enables the possibility of expressing novel meanings through compositional semantics. This defining feature of human language means that the meaning of any phrase is derived from the meanings of its individual components and the rules by which they are combined (Hockett, 1960). The evolution of compositionality has been investigated abundantly in the field of language evolution through human experiments (e.g. Kirby et al., 2008; Raviv et al., 2019a) and computational simulations (e.g. de Boer, 2006; Steels and Loetzsch, 2012; Lazaridou and Baroni, 2020). These experiments typically involve learning artificial languages or playing a signalling game. Here, *learning* artificial languages imposes a constraint for which it is believed to lead to more compressible and structured languages (Kirby et al., 2015). *Communication* in signalling games imposes a pressure for expressivity, requiring

participants to develop a vocabulary of signal-meaning mappings that allows them to communicate about novel stimuli. In this case, some form of referential grounding must be established through the process of repeated interactions. Participants—human or machine—generally establish novel signal-meaning mappings quickly, which enables successful communication.

7.2.2 Inductive biases

An important aspect of this chapter is the notion of biases. Here, we do not focus primarily on behavioural biases observed in humans (e.g., the confirmation bias), but rather are interested in implicit inductive biases that may result in biased language learning. This is relevant since seemingly arbitrary aspects of linguistic structure may actually result from general learning and processing biases deriving from the structure of thought processes, perceptuo-motor factors, cognitive limitations, and pragmatics (Christiansen and Chater, 2008). Especially so since fundamental predispositions influence how humans and artificial systems learn and process information (i.e., language). At a population level, these biases may *manifest* themselves as preferences for compressibility, simplicity, and efficiency—cognitive tendencies (Kirby et al., 2015; Tamariz and Kirby, 2015; Gibson et al., 2019) that naturally influence language evolution. For example, in the case of human systems (e.g., language) that are culturally transmitted, a memory constraint can enforce systems to be easy to learn and simple, because the hard-to-learn elements are less likely to be transmitted. Furthermore, the sound systems of human languages seem to be optimised for criteria such as acoustic distinctiveness or articulatory ease (Liljencrants and Lindblom, 1972; Lindblom and Maddieson, 1988) through a process of self-organisation (De Boer, 2000). Some even argue that humans' cognitive limitations may be *beneficial* for language acquisition (DeCaro et al., 2008; Poletiek et al., 2018).

Human constraints like these could well have evolved differently and are inherently different between humans and computational language learners, such as reinforcement learning agents and LLMs (although this is an ongoing debate (Kozachkov et al., 2023)). In the case of simulations of (reinforcement learning) agents, inductive biases typically do not match those present in humans. As such, they are often induced artificially by incorporating biases to guide learning dynamics as a means to recover human-like properties (for a review see Galke and Raviv, 2025). In the case of LLMs, which are fundamentally different from humans, we focus on increasingly apparent inductive biases of the Transformer architecture (Futrell and Mahowald, 2025) that may influence how languages evolve in the context of our experiment.

One example is a bias for simplicity. Rende et al. (2024) carefully cloned training data such that texts only contained between-token interactions up to a certain degree. Revealing that Transformers first learn low-degree between-token interactions, and only later learn high-degree interactions. Similarly, LLMs pick up grammar as the simplest explanation for data early during training. Only shortly thereafter, general linguistic capabilities arise (Chen et al., 2024). Moreover, Transformers seem to have an inductive bias favouring structure in (natural) language.

For example, GPT-2 models struggle more to learn impossible languages (e.g., languages lacking hierarchical structure or having unnatural or irreversible word orders) compared to English (Kallini et al., 2024), indicating that structure aids language learning. Additionally, the ability to generalise to novel stimuli increases when LLMs learn from more structured artificial languages (Galke et al., 2024). Recent work also revealed a primacy and recency bias in LLMs. They handle information better when it appears either at the beginning or towards the end of a prompt (Liu et al., 2024; Mina et al., 2025). Finally, LLMs have an inductive preference for verbose answers (Zheng et al., 2023; Saito et al., 2023), while humans prefer short, efficient answers (Gibson et al., 2019).

Although the underlying mechanisms of these biases differ between humans and machines, we find substantial overlap in terms of their behavioural effects. As such, we hypothesise that the aforementioned effects of continuous learning and use of language will also come into play when humans and machines collaborate, resulting in a language optimised for the preferences of both entities.

7.2.3 Why is this relevant for LLMs?

It is increasingly assumed that LLMs can be used as models of language (Millière, 2024) and that classical approaches from emergent communication can inform more human-like language learning in machines (Beuls and Van Eecke, 2024; Galke and Raviv, 2025). Moreover, language modelling and linguistics should complement each other (Futrell and Mahowald, 2025) as comparing LLMs to human language users, can help answer cognitive and typological questions (Warstadt and Bowman, 2022; van Dijk et al., 2023a). Vice-versa, methods from psychology can help to quantify inductive biases of LLMs (Griffiths et al., 2024; Galke and Raviv, 2025) or vision-and-language models (e.g. Chapter 4; Kouwenhoven et al., 2025) and compare them to known biases in humans.

For instance, the process of iterated learning, in which the transmitted information will ultimately come to mirror the minds of the learners (Griffiths and Kalish, 2007a), has been used to discover inherent LLM biases. Ren et al. (2024) showed that iterated learning causes subtle biases in LLM priors to be gradually amplified, Chapter 6 concluded that artificial languages can be optimised for LLM-augmented agents with iterated learning, and Shumailov et al. (2024) argue that generative models converge on uninterpretable junk when they are trained on AI-generated data. While the latter is typically seen as drift, crucially, we argue that what this shows is that the generated content is slowly shaped to be optimised for model preferences, *not* for human preferences. To prevent what Shumailov et al. (2024) refers to as model collapse, they argue that genuine human interactions with systems will become increasingly important. Similarly, Smith et al. (2024) responded that, like in human language transmission, the need to be expressive may prevent both the convergence on a few frequent uninformative sentences and the emergence of a long tail of uninterpretable junk.

These findings advance our understanding of internal LLM representations. This thereby contributes to maintaining alignment and mutual understandability between humans and machines in interaction. We address this by examining how adaptation processes unfold when humans and machines interact and develop a novel artificial language together.

7.3 Methodology

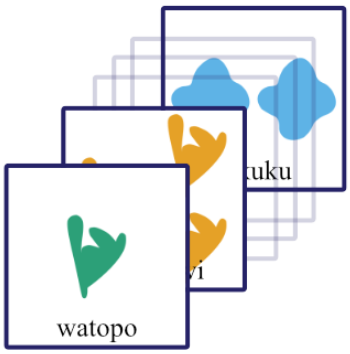
This experiment revolves around the classical referential Lewis game as implemented in Chapter 6, which is based on previous work in emergent communication (Raviv et al., 2021; Kirby et al., 2015, e.g.). We extend this setup to incorporate humans. In total, 45 participants participated in the experiment, 30 of whom formed 15 Human-Human pairs, and the remaining 15 interacted with an LLM in a Human-LLM setup. This allows us to directly compare languages adapted for human preferences to those adapted for the LLM-LLM simulations. But perhaps most interestingly, the Human-LLM condition provides an opportunity to investigate whether languages can be optimised for entities with different mechanisms and cognitive capacities (e.g., memory). If so, we can unravel what these look like.

During the experiment, participants first learn an artificial language and then use it to communicate with each other. The artificial language comprises a meaning space consisting of three attributes (shape, colour, and amount) that each can have three values, totalling to 27 unique stimuli. The corresponding labels are initialised following the design of Kirby et al. (2008), creating a holistic artificial language without structure (e.g., ”watopo”, ”sanasowi”, ”pikuku”) that contained a limited set of characters to prevent participants from writing English words. Participants first individually learn 15 random signal-meaning pairs through the exposure, guessing, and labelling blocks. Hereafter, participants are tasked to use the newly acquired language to communicate in a referential game. In this game, participants alternate between a speaker and listener role, where the speaker observes a target stimulus and labels it. Using this label as a signal, the listener is then tasked with identifying the correct target among three distractors. Cooperation is successful when the listeners' guess matches the target stimulus. After the communication block, there is a testing block in which participants individually label 27 meanings, including 12 unseen meanings, to assess how well they generalise to novel inputs. The duration of the entire experiment is roughly 70 minutes. An experiment overview is provided in Figure 7.1. [2]

7.3.1 LLMs as participants

Human participants learn the language by going through the exposure and guessing blocks twice. They iteratively go over the 15 training stimuli and may extract some apparent, but not present, patterns or consistencies. They are then tasked to label the stimuli, before moving on to

[2] All code, materials, and data are available on OSF: `https://osf.io/52yar/`.



{'shape':1,'colour':'green','amount':1,'word':'watopo'}
{'shape':3,'colour':'orange','amount':3,'word':'sanasowi'}
.
.
.
{'shape':2,'colour':'blue','amount':2,'word':'pikuku'}

Figure 7.2: Left: Humans learn the language by being exposed to stimuli and the corresponding signals in the exposure block. Right: LLMs learn the same vocabulary by virtue of in-context learning. A JSON-like structure containing the signal-meaning mappings is prepended to each prompt to serve as learning stimuli.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a
 language learner who has to learn an artificial language with words
and their corresponding features. Your task is to complete the
vocabulary by generating a word that describes the last item. Only
respond with the word.<|eot_id|><|start_header_id|>user<|
end_header_id|>\n
{'shape':2,'colour':'orange','amount':1,'word':'giniwite'}
{'shape':1,'colour':'green','amount':3,'word':'hanosa'}
    :
{'shape':3,'colour':'blue','amount':2,'word':'tusetetu'}
{'shape':1,'colour':'green','amount':3,'word':'<|eot_id|><|
start_header_id|>assistant<|end_header_id|>[comp/prefill]
```

Prompt 7.3.1: A prompt snippet used for labelling and guessing. During communication, we add a `communicativeSuccess` attribute, update the system prompt to inform about the communicative task, and instruct that 'Communicative success is important'.

the communication block in which they interact with another human or a LLM. In either case, they were told that they interacted with a human. The LLM agents, however, are not updated and receive instructions to learn the languages by virtue of in-context learning. Specifically, the stimuli are presented in a structured, JSON-like format (Figure 7.2) that has proven to be effective in Galke et al. (2024) and Chapter 6. As such, we assume that these signal-meaning mappings in the context of a prompt provide enough (distributional) information for a LLM to learn a mapping between the attributes of the stimuli and signal syllables (Prompt 7.3.1). Although the prompt structure 'invites' the LLM to infer a signal from the stimulus attributes, we are agnostic about how exactly and what kind of mapping is deduced, but are interested in

the resulting behaviours. In the experiments, we use the instruction-tuned variant of Llama-3 70B (Llama Team, 2024) with greedy sampling. We opt for an instruction-tuned model since this allows us to specify the need for communicative success. This potentially affects how the model's inductive biases are expressed, but we leave this for future work.

One of essentially two tasks is performed throughout the experiment: labelling or guessing. The labelling block and speaking in the communication block involve labelling, and the guessing block and discrimination during communication involve guessing. Generating signals is achieved through prompt completion. Guessing is done by prefilling the prompt with distractor stimuli or labels and selecting the item with the highest probability. This alleviates LLMs' inconsistent behaviour in answering multiple-choice questions (Khatun and Brown, 2024), and follows recommendations from computational linguistics (Hendrycks et al., 2021; Wang et al., 2024). During communication, we add a `communicativeSuccess` attribute set to 1 if the previous interaction for this stimulus was successful and zero otherwise. This attribute functions as a memory between interactions and acts as a pressure for expressivity. In human language evolution, such pressure plays an important role since it prevents languages from becoming degenerate (Kirby et al., 2015). Importantly, the agents observe the training vocabulary in their context *with* the current stimulus in the guessing and labelling block, rendering them as simple look-up tasks. We do, however, *not* include the current stimulus during communication and testing, requiring the agents to extract an appropriate mapping and generalise to new stimuli. Akin to standard practice in older simulations (e.g. Steels and Loetzsch, 2012), the agent vocabularies are updated when labels are generated after the labelling block and during the communication block. This allows the vocabularies of signal-meaning mappings to evolve over the course of the simulation. As such, prompts are slightly different after each interaction. Moreover, given the primacy and recency bias in LLMs (Liu et al., 2024; Mina et al., 2025), we shuffle the vocabulary before creating prompts to account for unwanted ordering effects. Some example prompts used in the Human-LLM condition are displayed in Section D.1.

7.3.2 Metrics

Besides comparing the percentage of communicative success (*PercCom*), the primary goal of this work is to understand what a language looks like when optimised for different entities. Specifically, we investigate whether the languages display some degree of structure in the form of compositionality. In this experiment, this means that attribute values are denoted with label parts that are reused to describe other similar stimuli. Capturing this is not at all trivial, especially provided the freedom given to participants when they label stimuli. A common metric that gauges whether similar meanings map to similar signals is Topographic Similarity (*TopSim*; Brighton and Kirby, 2006). While providing a good indication of compositional language use, it does not account for variability in language, such as word-order freedom. It could therefore show an incomplete picture (i.e., a low *TopSim*) as languages can still be compositional

despite having multiple word orders, the existence of synonyms, or homonyms (Conklin and Smith, 2023). Hence, we report multiple metrics in addition to *TopSim* that together indicate the degree of compositionality. Specifically, we report on *synonymy* (one-to-many mappings), *homonymy* (many-to-one), and word order freedom (*Freedom*), for which Conklin and Smith (2023) proposed entropy-based metrics. A language where each attribute value is encoded by a single character in a position has low entropy, and thus a low *synonymy*. Languages with a uniform distribution over all characters to refer to an attribute value have high synonymy. *Homonymy* is similar; it looks at how many attribute-values a character in a position can refer to, i.e., when *homonymy* ≈ 1 characters can map to multiple attribute-values. Finally, we compute word-order freedom (*freedom*) to account for variability in the order by which labels are composed. It assesses whether each value of a specific attribute is encoded in a specific position of the label, i.e., there is little freedom, or whether attribute values can be encoded in any position of the label, i.e., displaying a high degree of word order freedom (*Freedom* ≈ 1).

Systematic generalisation to novel stimuli is assessed through the generalisation score *GenScore* from Raviv et al. (2021). It gauges whether the labels produced for unknown (i.e., testing) stimuli are labelled in consistent ways to labels produced for similar known (i.e., training) stimuli. In addition to character-based metrics, we assess whether participants reuse parts of labels in different labels by computing the *Ngram* diversity (Li et al., 2016) over all the produced labels in a block. *Ngram* diversity is the average ratio of unique vs. total *Ngrams* for $N \in \{1, 2, 3, 4\}$ in all labels. Low *Ngram* diversity implies that labels are composed of reused parts, and high diversity means that labels do not share many *Ngrams*, thus are very different. The percentage of unique labels captures the degree of degeneracy (*RatioUniLabels*). Finally, we measure whether a pressure for communicative success, known to drive efficiency in human experiments (Smith and Culbertson, 2020), results in shorter labels using *WordLength*.

7.4 Evaluation

We use linear mixed effect models to analyse our results. Specifically, we fit $PercCom \sim Metric + (1|RoundId)$ where $Metric$ can either be *TopSim*, *RatioUniLabels* and is the average value of two players in a round. To measure effects across conditions, we use $PercCom \sim Metric + Metric * Condition + (1|RoundId)$. The slope $\hat{\beta}$ determines the direction of the effect and the rate of change. Additionally, we use conditional R^2_c, and marginal R^2_m (Nakagawa and Schielzeth, 2013). The former considers fixed and random effects to show how much variance can be explained by the model. Higher values of R^2_c indicate that the model captures more variance and that correlations are stronger. R^2_m describes how much variance can be explained by the fixed effects. We report Pearson's R to describe the relationship between *TopSim* and *GenScore*, and use a paired T-test, or Welch's test when assumptions of normality and variance are not met, to assess whether the metrics differ significantly.

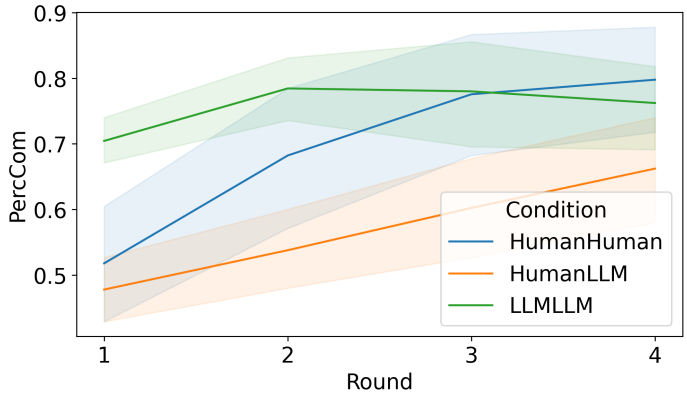


Figure 7.3: The average communicative performance (*PecrCom*) per round across the conditions. Communication steadily increases over rounds except for the LLM-LLM condition, in which coordination happens in the first round but does not increase afterwards.

7.5 Results

Human artificial language learning happened in a way that was expected based on earlier work (Kirby et al., 2015; Raviv et al., 2021). The results of these 15 Human-Human ($n = 30$) experiments act as a benchmark of human behaviour in our setup. We find that learning artificial languages is not a trivial task. After two rounds of exposure, labels were correctly guessed approximately half the time ($47.0\% \pm 49.9$). Freely labelling stimuli was done correctly only in $10.4\% \pm 30.6$ of the labels. Nevertheless, reliable communication protocols emerged during the communicating block; interactions were significantly more successful in the final round compared to the first round (Figure 7.3, $t(14) = -6.30, p < .001, d = 1.63, PercCom_{r1} = .518 \pm .176, PercCom_{r4} = .798 \pm .169$). *TopSim* positively influenced communication (*PercCom*) ($\hat{\beta} = .087 \pm .009, R^2_c = .731, R^2_m = .714, p < .001$) and generalisation to new stimuli was more consistent when the languages that evolved during communication displayed more *TopSim* ($r = .826, p < .001$).

A qualitative inspection revealed that, during communication, participants quickly replaced the labels they learned before, with only parts of labels 'surviving' this cut. Moreover, the number of shapes displayed (1, 2, or 3) was sometimes encoded by repeating the shape and colour labels several times, e.g., "pufepufe" was used to indicate two green shapes. Although expressive, this solution does not generalise to larger numbers and is therefore arguably not compositional.

Artificial language learning in LLMs was assessed in Chapter 6. Here, we briefly discuss the results of LLM-LLM ($n = 15$) simulations. LLMs guessed labels correctly for $97.3\% \pm 16.1$,

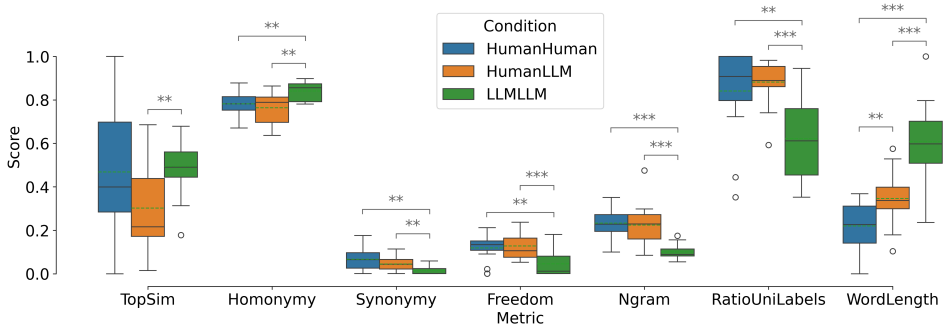


Figure 7.4: An overview of structure metrics used to measure compositional structure in the languages produced in the testing block. Generally, languages optimised for LLMs differ from those optimised for humans. Languages optimised for both mediate these differences. The asterisks indicate whether an independent Welch's t-test reveals a significant difference between the conditions where $* \; p < .05$, $** \; p < .01$, $*** \; p < .001$. *TopSim* and *WordLength* are normalised to values between 0 and 1 for visualisation purposes.

and the produced labels in the labelling block exactly matched the initial labels for $45.3\% \pm .498$. Indicating that learning the language is easier for LLMs than for humans. This is not surprising since the target stimulus is present in the prompt context in these blocks, and there is virtually no memory constraint. Communication happens reliably as well (Figure 7.3), however, communication can—but does not always—result in degenerate vocabularies with few uniquely used labels ($RatioUniLabels = .621 \pm .198$), significantly differing from human diversity in labels ($RatioUniLabels = .841 \pm .201, t(28) = -3.01, p = .005$, Figure 7.4). Interestingly, this happens even though the ratio of unique labels is modestly related to *PercCom* during communication ($\hat{\beta} = .300 \pm .134, R^2_c = .180, R^2_m = .092, p = .025$), suggesting that expressiveness is beneficial. A tentative explanation could be that aligning vocabularies happens much faster than in the other conditions. While humans would optimise languages whilst retaining expressiveness, LLMs start producing more duplicate and longer labels. We also ran additional simulations with other smaller models (i.e. Llama-3-8B (Llama Team, 2024), OLMo-2 7B, and OLMo-2 13B (Walsh et al., 2025)). The results are presented in Section D.3. In general, learning the artificial languages was comparable to Llama-3 70B for all models, but communication proved more difficult for the smaller models. Here, agents comprised of OLMo-2 7B models struggled the most and were unable to communicate robustly above chance levels.

What about Human-LLM communication? Our main contribution comes from the Human-LLM condition in which participants ($n = 15$) collaborated with LLMs. Successful communication necessitates that both entities adopt their behaviours and that a reliable referentially grounded vocabulary emerges. This is especially interesting since we observed that the ability to learn artificial languages differs between humans and LLMs, and that the optimised

languages differ in their use of homonyms, i.e., duplicate labels for different meanings. Despite these differences, communication *is* still possible ($PercCom = 0.662 \pm 0.161$, Figure 7.3). The final performance was lower than the other conditions, but the data suggest that prolonged interactions may result in higher communicative success. This exciting result shows that, even though the learning mechanisms of both entities may initially learn different signal-meaning relations, communication is possible. As such, the process of repeated learning and use of these artificial languages can overcome initial differences and indeed shape languages to be optimised for humans *and* LLMs. Out of 15 participants, 9 believed their partner was another human, despite communicating with an LLM. Performance did not change significantly as a result of this.

What do these languages look like? Having established that participants can communicate, we now examine if and how languages differ across different experimental conditions, focusing on the language metrics. Generally, we find that languages optimised for LLMs differ from those optimised for humans, and that languages optimised for *both* are more human-like than LLM-like (Figure 7.4). This is most notably visible in the ratio of uniquely produced labels, their respective lengths, and the *Ngram* diversity. Languages optimised for humans contain more unique labels, have higher *Ngram* diversity, and are shorter when compared to languages optimised for LLMs. The latter of which corroborates the well-known human preference for efficient communication (Smith and Culbertson, 2020). The length, number of unique labels, and diversity of label parts resulting from human-LLM collaboration seem to adhere more to human preferences than to LLM preferences. A similar pattern is visible for the compositionality metrics that allow for variation. There is more *homonymy* in LLM-optimised languages than in human-LLM languages, suggesting that the meanings of these words should be disambiguated by the context (i.e., the distractor stimuli) in which they appeared. This strategy is not straightforward and perhaps requires more cognitive capacity than available for humans, which could result in lower performance in the collaborative condition. It also seems that humans introduce synonymy into languages, i.e., they use more than one character to refer to specific attribute values. This introduces variability that can explain why the ratio of unique labels and *Ngram* diversity is higher in the collaborative condition. Finally, the word order of messages is somewhat flexible for humans, whereas LLMs tend to converge on a more fixed word order. The languages shaped by both seem to have human-like word order freedom. While these variations may introduce difficulties for LLMs to decode the meanings during communication, we do not find that *PercCom* is affected by the degree of *homonymy*, *synonymy*, or *freedom*.

The canonical *TopSim* metric suggests structure is lower when humans and LLMs collaborate compared to LLM simulations. This makes intuitive sense, given our observation that it was also more difficult to establish successful communication. Nevertheless, a linear mixed effects model fitted to predict *PercCom* with *TopSim*, the experimental condition, the interaction

between them, and a random effect for round revealed that *TopSim* strongly affects *PercCom* ($\hat{\beta} = .092 \pm .008, R^2_c = .580, R^2_m = .580, p < .001$). This means that irrespective of the experimental condition, a higher degree of structure in the produced labels was beneficial for communication. Moreover, generalisation to novel stimuli happened more consistently when the languages in the last round of communication displayed more structure. Again, confirmed by the mixed effects model predicting *GenScore* with *TopSim*, the condition, their interaction and round as random effect ($\hat{\beta} = .047 \pm .006, R^2_c = .784, R^2_m = .673, p < .001$, Figure D.1).

Generally, we find that while remarkably human-like, the languages shaped by intrinsic LLM constraints are in fact subtly different from those shaped by humans. Thereby providing a more nuanced view of what was observed in Chapter 6 as it only looked at *TopSim* and *NGram* diversity. Returning to the question of what languages optimised for entities with different inductive biases look like, they seem to be shaped in such a way as to conform more to human pressures than those present in LLMs.

7.6 Discussion

The primary goal of our work was to investigate if and how artificial languages differ when they are optimised for human *and* artificially intelligent language users. To do so, we extended Chapter 6, suggesting that LLMs can shape and use languages in referential communication. That setup was adapted to allow participants to interact with other human participants and with LLMs. This enabled controlled comparisons between the languages that evolve under different conditions. Our findings showed that human pairs, LLM pairs, and Human-LLM pairs can learn and successfully use languages in a referential game. This suggests that mechanisms that influence how language evolves for humans, specifically, learning and using a language repeatedly (e.g. Smith, 2022), also apply to computational and collaborative Human-LLM settings. In all conditions, successful communication was achieved by optimising an initially holistic, unstructured vocabulary to fit better with the inductive biases of the language users. We compared the languages across conditions and revealed that 1) while very human-like, LLM languages tend to be more strict (i.e., there is little variation), and that 2) languages adapted for human-LLM pairs tend to be more human-like than LLM-like (i.e., they are more diverse and have variation). Overall, our findings corroborate earlier claims that interactions between humans and machines are beneficial to establishing some form of referential grounding (Mikolov et al., 2018; Beuls and Van Eecke, 2024).

On the level of vocabulary, the ratio of uniquely produced labels by LLMs revealed that vocabularies can become degenerate. While this is also observed in human experiments when there is no pressure against it (Kirby et al., 2008), communicative success as a pressure is typically enough to prevent this (Kirby et al., 2015, e.g.). In contrast, even though the LLMs in our experiments were tested in a communicative setting, this did not prevent the languages from becoming underspecified. Possibly, this happened because the instruction to achieve

communicative success was not explicit enough and did not induce sufficient pressure for expressivity. Alternatively, it may be the case that the LLMs solved the problem in a non-humanlike manner and employed some kind of shortcut learning (Schwartz and Stanovsky, 2022; Mitchell and Krakauer, 2023). It could, for example, be that the distractors did not require the labels to be very specific, but instead allowed using underlying concurrences that were picked up by LLMs but not by humans. This would also explain the wide range of scores on this *RatioUniLabels*. On a character level, we observe related patterns in the form of high levels of *homonymy*, meaning that attribute values could be associated with multiple label characters, and that context was necessary to disentangle the correct meaning. While the duplicate labels can explain these scores for LLM simulations, this is not the case for humans. Here, the surprising behaviour of repeating label parts to indicate the *amount* attribute can explain the *homonymy* values.

The process of tokenisation plays an important role in these simulations. One could argue that it may help in learning some mapping between tokens and meanings. The meaning attributes and their values are common English words, while the initialised artificial languages consist of non-words that are tokenised into separate tokens. Meaning that the LLM is presented with a parsed set of attribute meanings and chunks of labels (i.e., the tokens). All that is left is to attend to a specific token given a particular meaning, which is precisely what a Transformer model is made for. Nevertheless, this does not undermine that these models *indeed* attribute attention correctly and that this produces human-like languages.

Similarly, our observation that a shared referential communicative system can be established is quite remarkable. Humans and LLMs may well use entirely different mechanisms and learn different relations between meanings and signals. Yet, their vocabularies become referentially grounded and are pragmatically understood by both humans and LLMs. This confirms that even though LLMs are not trained for this task, they can be used as relatively unbiased language learners (Wilcox et al., 2023), thereby providing a concrete example of how a pragmatic view of understanding, as argued for by van Dijk et al. (2023a), can be beneficial for collaborative tasks. This work also underscores the point made by Millière and Rathkopf (2024) that how LLMs or other AI models solve a cognitive task cannot be used as an argument against particular cognitive competences or language understanding, as long as the solution generalises.

Our results concretely corroborate the idea that insights from emergent communication literature can inform and improve language learning in language models (Smith et al., 2024; Beuls and Van Eecke, 2024; Galke and Raviv, 2025). We observed that just as languages accommodate for specific abilities and preferences in humans (Josserand et al., 2024), Human-LLM languages also adapt to the abilities and preferences of their users in that they are more human-like than LLM-like. Specifically, human preferences for simplicity and efficiency (Kirby et al., 2015; Gibson et al., 2019) likely drove vocabulary diversity while reducing lengths to human-like levels. This indicates that, in this experiment, LLMs are more flexible communicative partners than humans. These findings reinforce the idea that repeated interactions with humans are

crucial to maintaining referentially grounded human-like vocabularies instead of training only on recursively generated data (e.g. Shumailov et al., 2024) or using AI-augmented optimisation algorithms (e.g. Lee et al., 2024). In particular, this underscores the need for new methods that incorporate human interaction into the training processes of LLMs and shows that using communicative success as a reward signal can be a fruitful and novel approach.

Finally, we acknowledge that our results depend on methodological considerations, including the use of in-context learning, the model, the prompt format, and the sampling method. However, the primary goal was to extend previous work by investigating if languages optimised for human *and* LLM preferences can evolve. As such, we stayed close to well-established experimental methods in the field of language emergence and used prompts developed in Chapter 6. Importantly, we did not optimise for communicative success, human-like results, or compositional vocabularies.

7.7 Conclusion

Given the growing presence of contemporary LLMs in everyday life, there is an increasing need to understand their inductive biases to maintain alignment with humans. We tested whether general mechanisms of language learning and use have similar effects in an artificial language learning experiment conducted with Human-Human, LLM-LLM, and Human-LLM pairs. We show that referentially grounded vocabularies emerge in all conditions, indicating that initially unstructured artificial languages can be optimised for inductive biases of different language users. Comparisons across conditions revealed that, while similar to human vocabularies, LLM languages are subtly different. Interestingly, these differences are alleviated when humans and LLMs collaborate. This underscores that to achieve successful interactions between humans and machines, it is essential to optimise for communicative success. Overall, these findings advance our understanding of how LLMs may adapt to the dynamic nature of human language, contribute to its evolution, and maintain alignment with human understanding of language. While our setup only uses simple stimuli and basic languages, achieving this for human-level languages is a key research direction towards more natural language learning in LLMs.