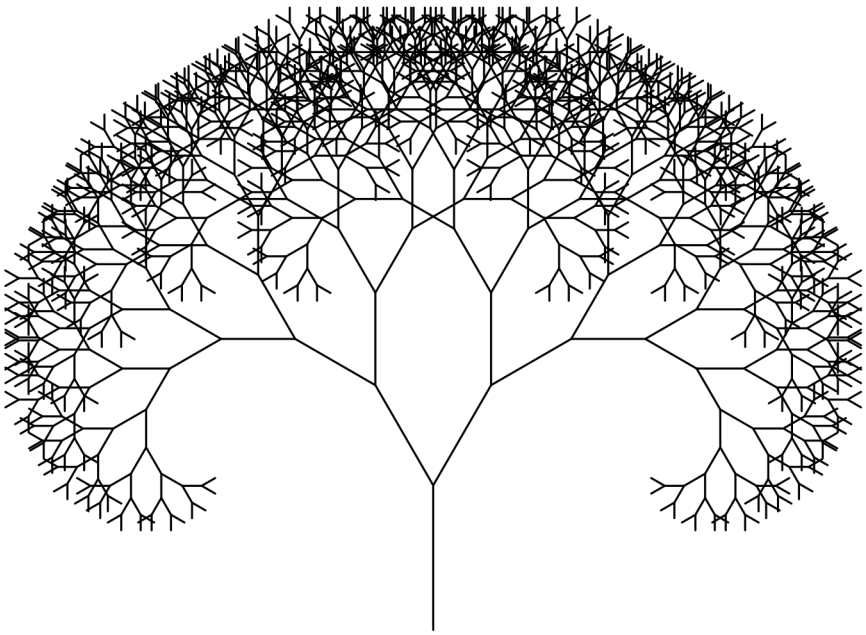
Proceedings

AI in Language Evolution



Edited by: Bart de Boer

AI in evolang workshop

Bart de Boer[*1], editor

[*]Corresponding Author: bart@ai.vub.ac.be
[1]AI-lab, Vrije Universiteit Brussel, Brussels

Since the last few editions of the Evolang conference, Artificial Intelligence (AI) has had an increasing impact on scientific research, and this has influenced research on language evolution as well. Evolang has a long tradition of using AI techniques (some of the papers at the first Evolang in 1996 were based on AI) but most recently, this was reflected by a workshop on machine learning at the Kanazawa conference (2022) and by a workshop on large-scale computational approaches at the Madison conference (2024). Since this time, AI methods have become even easier to use.

AI techniques can for instance be used in the form of large language models in corpus analysis, as chatbots to interact with human participants in experiments, as part of agent-based models in simulation and as tools for sound- and image analysis when analyzing animal behavior. Different applications require different techniques, and perhaps more subtly, have different constraints on how existing AI models can be applied without introducing biases and artefacts. Exactly because modern AI tools are so easy to use, there is a definite risk that they may be applied incorrectly. However, there are as yet no textbooks that help researchers decide on how to use AI techniques appropriately for their research.

This workshop therefore proposes to bring together recent and ongoing research that uses AI techniques. The focus of the workshop is methodological: it is not so much about the linguistic/evolutionary questions that the work addresses but is meant to help exchange techniques and ideas on how to use AI techniques productively and correctly in evolution of language research.

Thanks to all contributors to the workshop, and to all those who contributed to the review process of the abstracts contained in this volume.

Table of Contents

1 Invited lecture: Willem Zuidema
Learnability and evolvability of grammar: What the AI revolution teaches us about language evolution

3 Axel Ekström & Runhui Song
Learning from a hundred thousand tube models: Exploring the role of vocal tract size in shaping phonetic output

7 Rodrigo Manríquez, Sonja Kotz, Andrea Ravignani & Bart de Boer
Spiking Neural Networks trained on Acoustic Data

11 Andres Karjus, Sophie-Marie Ertelt & Muhammad Okky Ibrohim
Evolutionary idea dynamics: drift and diffusion in LLM-annotated historical corpora

15 Bach Phan-Tat, Kris Heylen, Dirk Geeraerts, Stefano De Pascale & Dirk Speelman
From Parsers to Prompts: Combining AI and Linguistic for Interpretable Language Evolution

21 Jonas Gillain, Katrien Beuls & Paul Van Eecke
Exploring Open-ended Neural Concept Representation for Language Emergence

25 Nathaniel Imel & Noga Zaslavsky
Evolution and compression in LLMs: On the emergence of human-aligned categorization

29 Anna Jon-And & Jérôme Michaud
Modeling Cognition with Minimal AI

33 Yuqing Zhang, Ecesu Ürker, Tessa Verhoef, Gemma Boleda & Arianna Bisazza
NeLLCom-Lex: A Neural-agent Framework to Study and Simulate Lexical Semantic Change

37 Roman Miletitch & Limor Raviv
Swarm Robotics: Embodiment and Spatial Constraints in Emergent Communication Models

Learnability and evolvability of gramma: What the AI revolution teaches us about language evolution

Willem Zuidema[*,1]

[*]Corresponding Author: zuidema@mailbox.org
[1]Institute for Logic, Language and Computation,
University of Amsterdam, the Netherlands

The last few years have witnessed a tsunami of news, hype and critical reflections about AI, much of which has been fuelled by advances with Large Language Models. In that endless stream of claims and counter-claims about the capabilities of this new technology, it has generally been difficult to separate chaff from grain. For linguists, specifically, it has been difficult to determine what the consequences are – if any – of the surprising fluency in natural language of these models for theories of language use, learning and evolution.

In my talk, I will focus on classic arguments in debates on how human language may have evolved. In particular, I go over claims about gradualism (and the improbability of "macromutations"), about complexity (e.g., the relevance of classes in the Chomsky Hierarchy), and about domain-specificty and modularity (e.g., the need to account for language-specific computations and constraints on variation).

I will try to reevaluate these claims in light of results with Transformer-based language models, trained end-to-end with a gradual learning algorithm. I will discuss results showing successful generalization behavior of such models not only in the language domain, but also in logical and mathematical reasoning, image processing, and speech recognition and generation. While acknowledging fundamental differences between the architectures of these models and the human brain, and between the nature of the training data both receive, I will make the argument that many of the classic arguments have lost their force.

Learning from a hundred thousand tube models: Exploring the role of vocal tract size in shaping phonetic output

Axel Ekström[1,2*] and Runhui Song[2,3]

[*]Corresponding Author: axel.ekstrom@su.se
[1]Centre for Cultural Evolution, Department of Psychology, Stockholm University, Sweden
[2]Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden
[3]Department of Linguistics and Philology, Uppsala University

Simulating the vocal tract as an elongated acoustic resonator or tube has a long history in speech research (Fant, 1971; Carré, Divenyi, & Mrayati, 2017). Here, we explored the full potential of deep neural networks (DNN) – a machine learning model with multiple hidden layers that allows it to learn complex patterns in data – to analyse the behaviors of such resonators. We conducted these experiments to determine whether relationships between stricture and changes to predicted resonance frequency would differ between resonators of differing lenghts – of relevance to, for example, speech production in developing humans. To simulate an acoustic tube, we used the algorithmic approach developed by Liljencrants and Fant (1975)[1], which recursively computes a determinant through tube segments, each defined by their length and area. In this manner, the software allows for predicting resonances[2] for any arbitrary tube sequence.

XGBoost (Chen & Guestrin, 2016), an integrated learning algorithm based on Gradient Boosting (Friedman, 2001), as a baseline model (Song, 2025; Song, Sjons, & Ekström, in press). From the dataset of 50,000 random permutations, 80% were for training, and the remain 20% for testing (Song et al., in press). The main assumptions underpinning XGBoost are the construction of a series of decision trees, with each tree correcting the residuals of prior models; in this manner, the target output is gradually approximated. For the DNN model, we utilized an input layer composed from 16 elements, with two hidden layers of 64 and 32 neurons, respectively, and output layer neuron corresponding to the prediction of the *N*th formant where $N \in [1, 3]$. Inputs were standardized

[1]The algorithm for deriving resonants is only briefly summarized here; readers are referred to the original publication.

[2]In practice, while the two terms are often used interchangeably, *resonant properties* of an acoustic resonator (where energy resounds more strongly) do not have a 1:1 relationship to *formants* – spectral energy peaks. Here, however, for the sake of illustration the two are treated as functionally equivalent. For additional information on this topic, readers are referred to other sources (Fant, 1971).

with `StandardScaler`; training used the Adam optimizer and 500 epochs for the 16-segment setting, reflecting a trade-off between expressiveness and interpretability. Finally, we repeated the procedure for an 8cm tube sequence, with length and area values scaled down: segment length was fixed at 0.5cm, and area was varied 0.1cm^2 and 5cm^2.

Model performance was evaluated with standard regression metrics: R^2, MAE, MAPE, MSE, and RMSE. In the pipeline work, reliability was increased via k-fold cross-validation (five-fold for the 16-segment setting), and interpretability was obtained using SHAP, including global summary plots and local waterfall plots; SHAP values were computed using `KernelExplainer`. We treat XGBoost as a classic nonlinear baseline and compare it to the MLP surrogate; in our pipeline experiments, XGBoost was consistently outperformed by the MLP for smooth tube-to-formant mappings, motivating the neural model in the present study.

Controlling for size revealed that as tubes become smaller, the volatility – i.e., the predictability of behavior of resonances – increases (Tab. 1). That is, for smaller tubes, smaller changes to tube shape often result in significantly greater shifts in resonance frequency. These results (Fig. 1) replicate the sensitivity functions derived for an acoustic tube in exhaustive analyses performed by Carré and colleagues (Mrayati, Carré, & Guérin, 1988; Carré et al., 2017), effectively serving as a sanity check on our methodology. Our models reveal systematic scaling effects with potential implications for vocal tract evolution and development. On average, equal perturbations to tube shapes result in smaller (for longer) and larger (for shorter tubes) shifts in resonance frequences. These results provide intriguing avenues for future work in comparative acoustics, and developmental phonetics, where vocal tract length may play an determinant role in structuring vocal communication.

Table 1. Model outcomes for first, second, and third formants (F_1, F_2, F_3) for 8cm and 16cm models. The coefficient of determination (R^2) estimates variance in the target variable (the Nth formant) explained by the model. Mean absolute error (MAE) denotes absolute differences between predicted and actual values in the dataset. Mean absolute percentage error (MAPE) is the average of absolute relative errors between predicted and true values. Mean Square error (MSE) is the average of squared differences between observed and predicted values. Root mean square error (RMSE) is the average difference between observed and predicted values. Finally, mean absolute percentage error (MAPE) refers to the average of absolute percentage differences between predicted and observed values

Measurements	**8cm**			**16cm**		
	F_1	F_2	F_3	F_1	F_2	F_3
Mean R^2	.98	0.97	.95	.98	.9677	.94
Mean MAE (Hz)	18.72	66.91	145.21	10.16	37.38	79.51
Mean MSE	612.98	8226.53	39318.42	192.81	2644.04	12834.46
Mean RMSE (Hz)	24.70	90.68	198.14	13.87	51.40	113.23
Mean MAPE	2.65 %	3.09 %	3.80%	3.11%	3.59%	4.23%

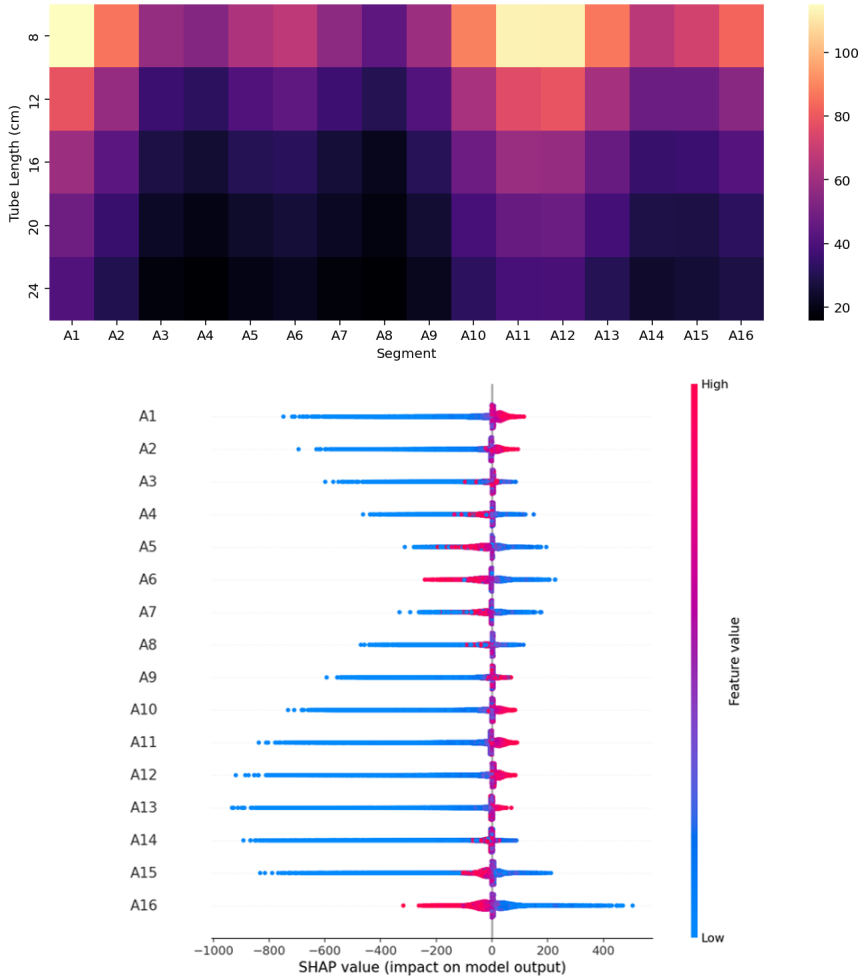


Figure 1. Top: The segment-wise SHapley Additive exPlanations (SHAP) heatmaps illustrate how strongly each part of the tube (segments A1–A16) influences a given formant (here the second formant, or F_2) for each tube segment. Taken together, these heatmaps provide a visualization sensitivity for any one formant, across the vocal tract model, indicating for each formant which regions are the most influential. Note that the segments are arranged from front to back, such that A_1 corresponds to the opening of the tube (a stand-in for lips and mouth), and A_{16} corresponds to its far end (a stand in for the glottis). For illustrative pruposes, we here show results of similar analyses for three additional VT lengths (12cm, 20cm, and 24cm) subjected to the same analyses as above. Bottom: All SHAP values for predicted second formant. Here, color coding denotes changes to predicted formants, such that *blue* indicates a decrease in segment area, and *red* indicates an increase; negative values correspond to lower predicted formants, while positive SHAP values indicate higher formant frequencies.

Acknowledgements

AE was funded through the Swedish Research Council (2025–00209).

References

Carré, R., Divenyi, P., & Mrayati, M. (2017). *Speech: A dynamic process.* Walter de Gruyter GmbH & Co KG.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Fant, G. (1971). *Acoustic theory of speech production: With calculations based on x-ray studies of russian articulations.* Mouton.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Liljencrants, J., & Fant, G. (1975). Computer program for vt-resonance frequency calculations. *STL-QPSR*, *16*, 15–21.

Mrayati, M., Carré, R., & Guérin, B. (1988). Distinctive regions and modes: a new theory of speech production. *Speech Communication*, *7*(3), 257–286.

Song, R. (2025). *Revisiting the third formant: Computational analysis of vocal tract constrictions using tube models and neural networks.* Unpublished master's thesis, Uppsala University.

Song, R., Sjons, J., & Ekström, A. (in press). Neural network-assisted analysis of tube vocal tract models. In *Proceedings of the 15th language resources and evaluation conference.* Mallorca, Spain: European Language Resources Association.

Spiking Neural Networks trained on Acoustic Data

Rodrigo Manriquez P.[*1,2], Sonja A. Kotz[2,3], Andrea Ravignani[4,5], Bart de Boer[1]

[*]Corresponding Author: rodrigo.manriquez@vub.be

[1]Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussel, Belgium;

[2]Department of Neuropsychology and Psychopharmacology, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands;

[3]Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany;

[4]Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy;

[5]Center for Music in the Brain, Department of Clinical Medicine, Aarhus University & The Royal Academy of Music Aarhus/Aalborg, Denmark.

Spiking Neural Networks (SNNs) communicate information through precisely timed spikes. They have stronger biological plausibility compared to other AI networks, which makes them well suited for testing cognitive/neural hypotheses. However, their performance remains inferior to non-spiking neural networks, as they are more challenging to train. This limits their adoption, particularly for auditory or language-related tasks.

Due to recent advances in training methods, SNNs have achieved noticeable progress in speech command recognition (Bittar & Garner, 2022; Sun et al., 2023). Moreover, SNNs have been used to explore neural oscillations in speech perception (Bittar & Garner, 2024), highlighting their potential for testing hypotheses about dynamic neural mechanisms in speech. Despite recent progress, analysis methods to study how SNNs process information remain inconsistent and are often adapted to the specificity of each task. Borrowing methods from neuroscience to interpret the dynamics of real neurons, we propose a framework to investigate an SNN trained on acoustic data.

We consider a network composed of Leaky Integrate-and-Fire (LIF) neurons (Abbott, 1999), a widely used simplified model in the SNN literature (Eshraighian et al., 2023). In a LIF neuron, each input contributes to a state variable known as the membrane potential. When the membrane potential

reaches a certain threshold, the neuron emits a spike at the output and resets its state (Figure 1). A main advantage of LIF neurons is that they enable efficient learning using a variant of gradient descent (Neftci et al., 2019), and they have also been shown to learn time-dependent information (Yu et al., 2025).

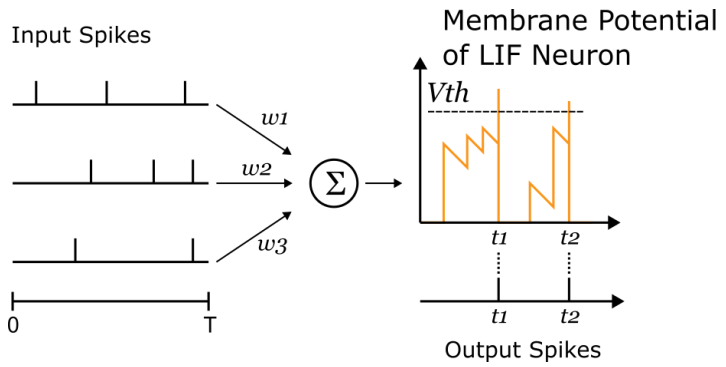


Figure 1. Scheme of a LIF neuron. In our model, a set of binary spike trains is considered as input. At each timestep, the inputs are weighted and summed, contributing to the neuron's membrane potential, a variable that reflects its internal state. The membrane potential increases with the arrival of an input spike and naturally decreases (or "leaks") at a fixed rate. When the membrane potential reaches a certain threshold, the neuron emits an output spike and resets its state by subtracting the threshold value.

In our approach, we consider a three-layer spiking autoencoder with self-recurrence (i.e., the output of each neuron is also fed back as input). The output layer consists of a single LI neuron with its potential threshold set to a very high value to prevent spiking. In this framework, the membrane potential of the output neuron is used to reconstruct the audio waveform of the spike-encoded input sound. This design enables the network to be trained as an autoencoder, allowing it to recover relevant auditory features from the spike trains. This approach differs from others that focus on classifying inputs into discrete categories (e.g. different vowels). Figure 2 illustrates the training paradigm used for this task.

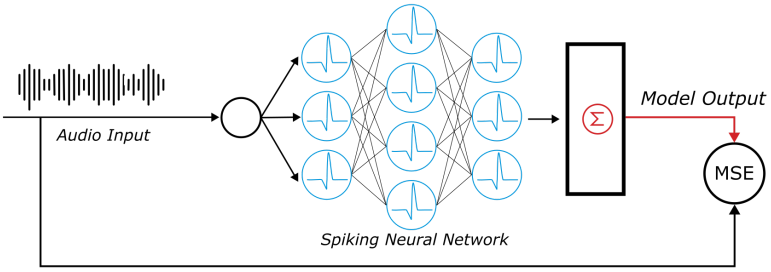


Figure 2. Spiking Neural Network autoencoder framework. Here, an auditory input is reconstructed in the output, which is recovered from the membrane potential of an output neuron that does not fire. Mean square error is computed between the difference of the waveform of the auditory input and recovered output.

By accessing internal state variables such as membrane potentials or firing rates derived from available spike trains, we examined oscillatory behaviour within the network. Further, by applying existing dimensionality reduction methods including PCA, UMAP, or tSNE, we obtained neural trajectories and observed how they align with a specific task or evolve over time. We propose that investigating these trajectories during an acoustic related task can be used to test hypotheses about language evolution. In particular, we focused on how neural trajectories in latent spiking layers represent combinatorial structure that is known to be present in the training data that are presented to the network. If such structure is clearly represented in terms of attractors and transitions between them, even though the network is not specifically designed to deal with combinatorial structure, it would imply that combinatorial structure may emerge naturally from the dynamics of the system, suggesting that its learnability may require fewer specialized mechanisms than previously assumed. This could make it easier to provide a transition from what has been called superficial combinatorial structure (i.e., the structure being present, without the language user making use of it (see Zuidema & de Boer 2009) to productive use of combinatorial structure (i.e., the language user making use of the structure to learn, perceive, and produce utterances). In this way, our framework provides a novel approach for probing the internal dynamics of SNNs, advancing their use as computational tools for investigating how combinatorial structure may emerge in the evolution of language.

References

Abbott, L.F. (1999). Lapicque's introduction of the integrate-and-fire model neuron (1907). *Brain Research Bulletin, 50(5-6), 303-304.*

Bittar, A., & Garner, P. N. (2022). A surrogate gradient spiking baseline for speech command recognition. *Frontiers in Neuroscience*, *16*.

Bittar, A., & Garner, P. N. (2024). Exploring neural oscillations during speech perception via surrogate gradient spiking neural networks. *Frontiers in neuroscience*, *18*, 1449181.

Eshraghian, J.K., Ward M., Neftci E.O., Wang X., Lenz G., Dwivedi G., et al. (2023). Training Spiking Neural Networks Using Lessons From Deep Learning. *Proceedings of the IEEE*,111(9):1016–54.

Neftci E.O., Mostafa H., Zenke F. (2019). Surrogate Gradient Learning in Spiking Neural Networks: Bringing the Power of Gradient-Based Optimization to Spiking Neural Networks. *IEEE Signal Processing Magazine*, 36(6):51–63.

Perich, M. G., Narain, D., & Gallego, J. A. (2025). A neural manifold view of the brain. *Nature Neuroscience*, *28*(8), 1582–1597.

Sun, P., Chua, Y., Devos, P., & Botteldooren, D. (2023). Learnable axonal delay in spiking neural networks improves spoken word recognition. *Frontiers in Neuroscience*, *17*.

Yu Z., Sun P., Goodman D. F. M. (2009). Beyond Rate Coding: Surrogate Gradients Enable Spike Timing Learning in Spiking Neural Networks. *arXiv*.

Zuidema, W., & de Boer, B. . The evolution of combinatorial phonology. *Journal of Phonetics, 37*(2), 125–144.

Evolutionary idea dynamics: drift and diffusion in LLM-annotated historical corpora

Andres Karjus[*1,3,4], Sophie-Marie Ertelt[1,2], and Muhammad Okky Ibrohim[1]

[*]Corresponding Author: akarjus@tlu.ee
[1]University of Tartu, Institute for Social Studies
[2]Örebro University, School of Business
[3]Estonian Business School
[4]Tallinn University, School of Humanities

Language and cultural evolution research has fruitfully used large text corpora and lexical databases in various scenarios, like modeling selection and drift in linguistic change and word frequencies (Karjus, Blythe, Kirby, & Smith, 2020; Newberry, Ahern, Clark, & Plotkin, 2017; Montero, Karjus, Smith, & Blythe, 2023), the evolution of word and meanings (Turney & Mohammad, 2019-01; Xu, Malt, & Srinivasan, 2017-08) and semantic shifts (Hamilton, Leskovec, & Jurafsky, 2016; Tahmasebi, Borin, Jatowt, Xu, & Hengchen, 2021-06; Ramiro, Srinivasan, Malt, & Xu, 2018). These approaches have typically operated with clearly defined competing (often binary) variants, database entries, or simple units like words. Cultural evolution research has made use of experimental observations (Ravignani, Thompson, Grossi, Delgado, & Kirby, 2018-03), and various metadata and demographic databases (Hahn & Bentley, 2003; Youngblood, Baraghith, & Savage, 2021-11-01; Tinits & Sobchuk, 2020).

With recent advances in large language models (LLMs), it has become possible (or at least significantly easier, compared to supervised learning approaches) to also detect and annotate more complex phenomena and semantic patterns in free-running text and other unstructured data (Ren, Caputo, & Jones, 2024-08; Karjus, 2025; Ziems et al., 2023-12-15). These may include semantic constructs, ideas, cultural expressions — that do not correspond to single lexemes but instead emerge from distributed patterns of phrasing. Methodologically, this is zero–few-shot text classification (instruction-following LLMs) as an alternative to supervised classifiers and dictionary methods. The main advantages are scalability (allowing consistent annotation across decades and countries where hand coding is infeasible), sensitivity to paraphrase and implicit mentions, and rapid iteration over category schemes; disadvantages include possible modern-language biases when analyzing historical texts, temporal domain shift, sensitivity to prompt framing, and opaque error modes.

Here we use LLMs to identify ideational traits related to environment, sci-

ence and technology in historical newspaper corpora of G20 countries in 1900-2025, as suggested in the Deep Transitions framework (Kanger et al., 2022-01-01), to explore the global evolution of industrial modernity. Industrial modernity denotes the dominant 20th-century worldview built on faith in science, technology, economic growth, and the subordination of nature to human progress. Articles sourced from the proprietary Proquest database containing relevant keywords are segmented into local contexts and annotated for the presence or absence of ideational traits, e.g., views of nature as a resource, techno-solutionism, and limitless progress. We use GPT-5.x-family models (via the data provider's API implementation; a limitation of using this dataset), with a prompt that provides theory-driven task instructions, a trait scheme with brief definitions and examples, and then a single article segment to analyze at a time. Output is constrained to a fixed schema to support scripted, large-scale annotation.

This yields, for each country and year, a vector of trait prevalences, which can be modeled individually as time series, or jointly as country profiles which "move" through the high-dimensional cultural trait space; random-walk baselines offer a natural null model in such cases (Cocho, Flores, Gershenson, Pineda, & Sánchez, 2015-04-07; Feltgen, Fagard, & Nadal, 2017). Departures from neutral diffusion, assessed via displacement magnitudes, preferred directions, and correlating changes across countries, provide a way to detect systematic shifts in ideational structure.

This does, however, require a discussion on taking into account LLM (or any classifier) error rates, heterogeneity in misclassification across traits, changes in classifier bias when terminology shifts historically, and the difficulty of distinguishing genuine ideational drift from fluctuations caused by annotation noise, OCR errors, and shifting media practices. We benchmark performance on a small hand-annotated gold set stratified across historical periods to test for temporal domain shift, and estimate misclassification rates that are then propagated to yearly prevalence estimates via bootstrap uncertainty intervals.

References

Cocho, G., Flores, J., Gershenson, C., Pineda, C., & Sánchez, S. (2015-04-07). Rank Diversity of Languages: Generic Behavior in Computational Linguistics. *10*(4), e0121898.

Feltgen, Q., Fagard, B., & Nadal, J.-P. (2017). Frequency patterns of semantic change: Corpus-based evidence of a near-Critical dynamics in language change. *4*(11).

Hahn, M. W., & Bentley, R. A. (2003). Drift as a mechanism for cultural change: An example from baby names. *270*, S120-S123.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. *2016*, 2116–2121.

Kanger, L., Tinits, P., Pahker, A.-K., Orru, K., Tiwari, A. K., Sillak, S., Šeļa, A., & Vaik, K. (2022-01-01). Deep Transitions: Towards a comprehensive framework for mapping major continuities and ruptures in industrial modernity. *72*, 102447.

Karjus, A. (2025). Machine-assisted quantitizing designs: Augmenting humanities and social sciences with artificial intelligence. *12*(277).

Karjus, A., Blythe, R. A., Kirby, S., & Smith, K. (2020). Quantifying the dynamics of topical fluctuations in language. *10*(1), 86–125.

Montero, J. G., Karjus, A., Smith, K., & Blythe, R. A. (2023). Reliable detection and quantification of selective forces in language change. *21*(1), 31–73.

Newberry, M. G., Ahern, C. A., Clark, R., & Plotkin, J. B. (2017). Detecting evolutionary forces in language change. *551*(7679), 223–226.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *115*(10), 2323–2328.

Ravignani, A., Thompson, B., Grossi, T., Delgado, T., & Kirby, S. (2018-03). Evolving building blocks of rhythm: How human cognition creates music via cultural transmission.

Ren, Z., Caputo, A., & Jones, G. (2024-08). A Few-shot Learning Approach for Lexical Semantic Change Detection Using GPT-4. In N. Tahmasebi, S. Montariol, A. Kutuzov, D. Alfter, F. Periti, P. Cassotti, & N. Huebscher (Eds.), *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change* (pp. 187–192). Association for Computational Linguistics.

Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., & Hengchen, S. (2021-06). *Computational approaches to semantic change.* Language Science Press.

Tinits, P., & Sobchuk, O. (2020). Open-ended cumulative cultural evolution of Hollywood film crews. *2*, e26.

Turney, P. D., & Mohammad, S. M. (2019-01). The natural selection of words: Finding the features of fitness. *14*(1), 1–20.

Xu, Y., Malt, B. C., & Srinivasan, M. (2017-08). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *96*, 41–53.

Youngblood, M., Baraghith, K., & Savage, P. E. (2021-11-01). Phylogenetic reconstruction of the cultural evolution of electronic music via dynamic community detection (1975–1999). *42*(6), 573–582.

Ziems, C., Shaikh, O., Zhang, Z., Held, W., Chen, J., & Yang, D. (2023-12-15). Can Large Language Models Transform Computational Social Science? 1–53.

From Parsers to Prompts: Combining AI and Linguistic for Interpretable Language Evolution

Bach Phan-Tat[*1], Kris Heylen[1], Dirk Geeraerts[1], Stefano De Pascale[1], and Dirk Speelman[1]

[*]Corresponding Author: ttbach.phan@kuleuven.be
[1]Department of Linguistics, KU Leuven, Leuven, Belgium

1. Introduction

Lexical semantic change detection is a well-established field in computational linguistics and natural language processing, with several shared tasks and curated datasets (e.g.,Basile, Caputo, Caselli, Cassotti, & Varvara, 2020; Kutuzov & Pivovarova, 2021; Schlechtweg, McGillivray, Hengchen, Dubossarsky, & Tahmasebi, 2020). Modern approaches in the study of semantic change often consist of applications of vector space modelling (VSM) (Tahmasebi, Borin, Jatowt, Xu, & Hengchen, 2021; Tahmasebi & Dubossarsky, 2023), which converts raw text into numerical vector representations. Semantic change is then quantified by applying clustering methods or distance measures (e.g., cosine distance) to those vectors.

However, one major conceptual problem of the above shared tasks and models is that they operationalise 'semantic change' as change in the distribution of senses of words (between 2 time periods). For example, in the subtask 1 of SemEval 2020 shared task, each target lemma is classified as 1 or 0, depending on whether it has gained or lost a sense, while in subtask 2 of the same shared task, the change score of each lemma is computed based on the shift of the distribution of its senses. Additionally, these approaches tend to suffer from the sensitivity to corpus size (Antoniak & Mimno, 2018; Sahlgren & Lenci, 2016) as they often require a large amount of data to generate useful vector representation, and the lack of interpretability (Lenci, Sahlgren, Jeuniaux, Cuba Gyllensten, & Miliani, 2022). All these issues make it difficult to connect their outputs to explicit hypotheses and theories about language change and language evolution.

In practice, many types of change do not map neatly onto such clear-cut sense distinctions. Subtle shifts in typical collocates, prototypical structures (Geeraerts, 1997), argument-structure (Christiansen & Joseph, 2016), or conceptual frame (Traugott, 2017) may leave a coarse-grained sense inventory intact while still reflecting substantial reorganization in how a word is used and conceptualised. For instance, *meeting* retains the general ORGANISED GATHERING meaning, yet its usage profile develops a digitally mediated default, with collocates

like *Zoom/Teams, virtual, hybrid* from only a primarily physical setting with typical collocates such as *in-person, conference room*. The sense remains stable, but the prototypical scenario and implied practices change. Another example is *babysit* (Christiansen & Joseph, 2016), whose earlier framing treats the beneficiary as an adjunct (e.g., *I babysat for the Smiths*) but now promotes the child to direct object position (e.g., *I babysat the kids*). A coarse sense inventory can still gloss it as LOOK AFTER CHILDREN TEMPORARILY, even though the constructions change what is foregrounded (a service to parents versus direct care for children). A strictly sense-based operationalisation therefore risks underestimating these more graded, profile-level developments, which motivates approaches that track changes in usage profiles directly rather than inferring them only through sense annotations, such as grammatical profiling (Kutuzov, Pivovarova, & Giulianelli, 2021) and usage-fluctuation analysis (McEnery, Brezina, & Baker, 2019).

We present an approach that 1) is conceptually simpler, 2) requires less data than VSM methods, 3) is more direct and theory-driven in interpreting the changes across different dimensions of words' usages and meanings.

2. Methodology

Given a concept/lexeme, we 1) annotate the different (theory-driven) dimensions of word usages using available tools such as automatic parsers or Large Language Models (LLMs), 2) measure change in these dimensions, 3) interpret the changing dimensions using LLM-based prompting and verify it with the theoretical narratives.

To demonstrate the feasibility of this approach, we report 3 case studies: 2 on the SemEval 2020 Task 1 (Schlechtweg et al., 2020) and one on three chemical concepts (air, water, acid) during the scientific evolution.

Our first type of dimension is syntactic dependency relations (Evert, 2008; Seretan, 2011) (slots) annotated using Stanza (Qi, Zhang, Zhang, Bolton, & Manning, 2020). For example, for the lemma *president*, we could extract all adjectives modifying president, all verbs for which president is the subject, all verbs for which president is the object, and so on. Each of the mentioned slots constitutes one dimension (i.e., how *president* is described, what *president* does, what happens to *president*). This yields a set of slot-fillers of each slot for each period. The decision of using syntactic collocates instead of surface collocates like what McEnery et al. (McEnery et al., 2019) did is to avoid accidental co-occurrence (Evert, 2008) and to separate the dimensions. We then quantified changes in each slot by computing the Jensen–Shannon Divergence (JSD) (Menéndez, Pardo, Pardo, & Pardo, 1997) between the slot-filler distributions of consecutive periods, yielding a JSD score per slot indicating how strongly each syntactic dimension of usage has shifted over time.

Our second type of dimension is semantic frames. We first frame-parsed the data using an open-sourced transformer model (Chanin, 2023), then collected all

the frames that the target lemmas participate in. We then quantified the distributional change of the frames using JSD. This indicate how strongly the frames linked to the target lemmas have shifted over time.

3. Case studies

In the 2 case studies with SemEval dataset (Phan-Tat, Heylen, Geeraerts, Pascale, & Speelman, 2026b, 2026a), we demonstrate that it can be used for lexical semantic change detection and can outperform a number of VSM approaches in many cases while remaining lightweight and transparent. Our qualitative analysis further indicates that the model captures semantically relevant signals and supports interpretable explanations of the observed changes.

In Case Study 3, we applied the framework to three chemical concepts from late eighteenth- to early nineteenth-century chemistry to detect the linguistic imprint of a major paradigm shift—here, the emergence of oxygen chemistry. We worked with a small and highly imbalanced subset of the Royal Society Corpus 6.0 Open (Fischer, Knappen, Menzel, & Teich, 2020), which is challenging for many VSM approaches. For each concept, we computed JSD over multiple syntactic slots and decomposed the slot-level divergence into contributions from individual fillers to identify which items drove the change. Following Davies (2025), we used an AI-assisted interpretation workflow. The quantitative outputs consist of (i) JSD values measuring how each slot's filler distribution of each concept changes across periods, and (ii) a decomposition of each slot-level JSD into per-filler contributions, which identifies the contributions of different fillers to the observed divergence. ChatGPT was prompted with these outputs and asked to propose candidate historical–semantic explanations. We then verify the output of ChatGPT by (a) confirming in-corpus evidence via concordance inspection and (b) checking whether the proposed historical claims align with established sources: peer-reviewed historiography and philosophy-of-science scholarship on the Chemical Revolution, complemented by contemporaneous scientific writings when relevant (e.g., Blumenthal & Ladyman, 2017; Cavendish, 1766; Chang, 2011; Stewart, 2012). Although ChatGPT sometimes produced unsupported claims, it was often useful, especially with its recent 'Thinking' mode. The results and interpretations showed that this approach successfully captures the linguistic imprint of the phlogiston-oxygen paradigm shift during the Chemical Revolution.

4. Proposal for language evolution research

From the perspective of language evolution, our approach is most naturally situated within cultural evolutionary accounts, which treat language as a culturally transmitted system that changes over time as a consequence of pressures arising during transmission and use. In this view, language evolution does not refer exclusively to the biological emergence of a language faculty, but also to the cultural processes that shape linguistic structure on historical timescales. This relationship

between culture and language has long been discussed in the literature (Smith, 2006; Steels, 2011; Smith, 2014; Degaetano-Ortlieb & Teich, 2022; Hoijer, 1948; Witherspoon, 1980). Semantic change literature often highlights how conceptual changes are encoded in lexical behaviour (Geeraerts, 1997). Diachronic semantic changes in historical corpora can therefore be treated as cultural micro-evolutions.

The generalisation of the framework is as follows (AI components are **bold**): Given a corpus and a target input, we posit a set of theoretically motivated dimensions; Each dimension is operationalised as a categorical distribution over observable annotations (e.g., semantic frames, thematic roles, constructional patterns, or register/genre labels), which can easily be done using **automatic parsers or LLMs**; Changes are now quantified as the divergence between these distributions across time; **Modern LLMs** can be used to assist in qualitative interpretation. When prompted with the quantitative outputs (e.g., dimension-level divergences, top contributing items, and representative contexts), they can propose candidate explanations of the shifting dimensions, as in Case Study 3. We could further use **retrieval-augmented generation (RAG)** to ground these interpretations in relevant external sources so that candidate explanations are accompanied by traceable citations and can be more reliably evaluated.

Because the method is lightweight and relies only on off-the-shelf AI/NLP tools, it can be applied to diverse diachronic corpora and domains. In this way, it complements experimental paradigms and agent-based simulations of cultural language evolution by adding an empirical, corpus-based layer that quantifies and localises change in attested historical usage. Concretely, it identifies which dimensions shift, when the shifts occur, and which lexical items contribute most to the observed divergence, thereby providing testable targets and constraints for modelling assumptions and simulations.

Acknowledgements

This work was supported by the Horizon Europe MSCA Doctoral Network project CASCADE (101119511) (www.horizoncascade.net).

References

Antoniak, M., & Mimno, D. (2018). Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Linguistics*, *6*, 107–119.

Basile, P., Caputo, A., Caselli, T., Cassotti, P., & Varvara, R. (2020). DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In V. Basile, D. Croce, M. Maro, & L. C. Passaro (Eds.), *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020* (pp. 411–419). Torino: Accademia University Press.

Blumenthal, G., & Ladyman, J. (2017). The development of problems within

the phlogiston theories, 1766–1791. *Foundations of Chemistry*, *19*(3), 241–280.

Cavendish, H. (1766). XIX. Three papers, containing experiments on factitious air. *Philosophical Transactions of the Royal Society of London*, *56*, 141–184. (_eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1766.0019)

Chang, H. (2011). The Persistence of Epistemic Objects Through Scientific Change. *Erkenntnis*, *75*(3), 413–429. (Publisher: Springer Science and Business Media LLC)

Chanin, D. (2023). *Open-source frame semantic parsing.*

Christiansen, B. J., & Joseph, B. D. (2016). On the relationship between argument structure change and semantic change. *Proceedings of the Linguistic Society of America*, *1*, 26.

Davies, M. (2025). *AI/LLM integration with the corpora from English-Corpora.org.*

Degaetano-Ortlieb, S., & Teich, E. (2022). Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*, *18*(1), 175–207.

Evert, S. (2008). Corpora and collocations. In *Corpus Linguistics. An International Handbook.*

Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020). The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study.

Geeraerts, D. (1997). *Diachronic Prototype Semantics: A Contribution Historical Lexicology.* Oxford University Press.

Hoijer, H. (1948). Linguistic and Cultural Change. *Language*.

Kutuzov, A., & Pivovarova, L. (2021). RuShiftEval: A Shared Task on Semantic Shift Detection for Russian. In *Computational linguistics and intellectual technologies.*

Kutuzov, A., Pivovarova, L., & Giulianelli, M. (2021). Grammatical Profiling for Semantic Change Detection. In *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 423–434). Online: Association for Computational Linguistics.

Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., & Miliani, M. (2022). A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*, *56*(4), 1269–1313.

McEnery, T., Brezina, V., & Baker, H. (2019). Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse. *International Journal of Corpus Linguistics*, 413–444.

Menéndez, M., Pardo, J., Pardo, L., & Pardo, M. (1997). The Jensen-Shannon divergence. *Journal of the Franklin Institute*, *334*(2), 307–318.

Phan-Tat, B., Heylen, K., Geeraerts, D., Pascale, S. D., & Speelman, D. (2026a). *Reframe or remain: Unsupervised lexical semantic change detection with*

frame semantics.

Phan-Tat, B., Heylen, K., Geeraerts, D., Pascale, S. D., & Speelman, D. (2026b). *Transparent semantic change detection with dependency-based profiles.*

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020, April). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.* arXiv. (arXiv:2003.07082 [cs] Read_Status: New Read_Status_Date: 2025-08-01T12:54:50.822Z)

Sahlgren, M., & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 975–980). Austin, Texas: Association for Computational Linguistics.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1–23). Barcelona (online): International Committee for Computational Linguistics.

Seretan, V. (2011). *Syntax-Based Collocation Extraction.* Springer Netherlands.

Smith, A. D. (2014). Models of language evolution and change. *WIREs Cognitive Science*, *5*(3), 281–293.

Smith, K. (2006). Cultural Evolution of Language. In K. Brown (Ed.), *Encyclopedia of Language & Linguistics (Second Edition)* (Second Edition ed., pp. 315–322). Oxford: Elsevier.

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, *8*(4), 339–356.

Stewart, J. (2012). The Reality of Phlogiston in Great Britain. *HYLE–International Journal for Philosophy of Chemistry*, *18*.

Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., & Hengchen, S. (2021). Computational approaches to semantic change.

Tahmasebi, N., & Dubossarsky, H. (2023). *Computational modeling of semantic change.* (_eprint: 2304.06337)

Traugott, E. C. (2017, March). *Semantic Change.* Oxford University Press.

Witherspoon, G. (1980). Language in Culture and Culture in Language. *International Journal of American Linguistics*, *46*(1), 1–13.

Exploring Open-ended Neural Concept Representation for Language Emergence

Jonas Gillain*[1], Katrien Beuls[1], and Paul Van Eecke[2]

*Corresponding Author: jonas.gillain@unamur.be
[1]Faculté d'informatique, Université de Namur, Belgium
[2]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Belgium

Humans can acquire concepts and skills in a modular and, to a certain extent, independent way. For example, we can learn to recognise penguins and later learn to draw them. If we want to do the same with monkeys, we will not need to learn drawing from scratch. This modularity has been central in the opposition of different methodologies in the field of language emergence. The language game paradigm has historically proposed experiments in which categories – or concepts – are created in an open-ended and explainable way (Steels, 1995; Steels & Belpaeme, 2005). On the other hand, the past decade has seen a wave of experiments using neural techniques to tackle complex problems in the real world (Rita et al., 2025), like puzzle-solving (Foerster et al., 2016) or reference (Lazaridou et al., 2017), but most architectures proposed end-to-end systems, where skills and concepts are indistinguishable. Our goal is to progress towards modular and transferable concept representations by proposing an open-ended architecture that also leverages the complexity that neural networks can capture.

This abstract presents our methodology and reports on preliminary results from a reference game experiment [1]. Our goal is that the agents construct their linguistic inventory and the associated concepts from the ground up. We equip them with two elementary templates to initialise small neural networks: *concept networks*, and a *fitness network*. Every agent possesses a single fitness network, while many concept networks eventually coexist in his inventory, each associated with a symbolic form that can be uttered in an interaction. A concept network can be connected to the fitness network to form a pipeline, taking as input an image and outputting a value between 0 and 1, which will come to reflect the *fitness* of this concept relative to the image (see Figure 1).

While all networks initially perform poorly, they will be updated as the agents give each other feedback after they interact, following the typical language game

[1]The code is available at
`https://gitlab.unamur.be/beehaif/modular-concepts/`

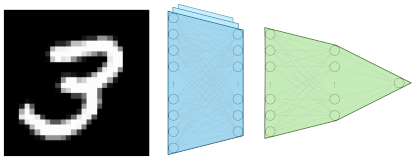


Figure 1. A concept network (in blue) - fitness network (in green) pipeline. After some, the ouptut value will come to reflect how well the concept fits the image. Multiple concept networks coexist in an agent's inventory.

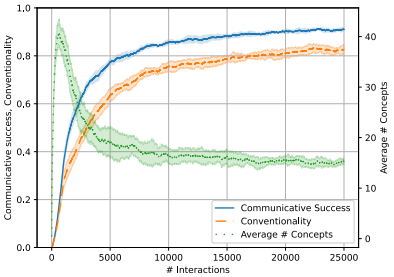


Figure 2. Evolutionary dynamics during training for a population of 5 agents - average $\pm$ 1 standard deviation over 10 runs.

methodology (Van Eecke et al., 2022; Botoko Ekila et al., 2024):

- a new concept will be initialised to discriminate a topic from the scene that is currently presented to the agent,
- in the case of a successful interaction (the speaker attracted the attention of the listener to the correct object in the scene with his utterance), both participating agents will update the concept they used,
- in the case of a failed interaction, the listener will update the concept he used to try to shift it towards the representation of the speaker.

A score is kept for each concept, which will be increased in the case of successful communication and decreased in the case of failure. A successful concept's competitors will also be punished proportionally to how close they are to its representation. This score will be taken into account when selecting the best concept to discriminate a topic from a scene.

In practice, a concept network consists of a 784-neuron input layer, fully-connected to a 128-neuron output layer, with batch normalisation and ReLU activation (784 corresponds to 28×28, the size of the input images). The fitness network is also fully connected, with a 128-neuron input layer, a 64-neuron hidden layer – each with batch normalisation and ReLU activation – and a single-neuron output layer, with sigmoïd activation. The different modules are implemented using PyTorch (Paszke et al., 2019), and trained using the Binary Cross Entropy loss function, and Adam optimiser with a learning rate of 0.015 (Kingma & Ba, 2014). This network architecture has been selected as a compromise between complexity and size. Indeed, because a concept network will initially be trained on a very limited amount of data (one scene, with one positive example – the topic – and the rest being negative examples), it is important that the number of parameters is low enough so that there is a chance that this single training pass will push the

network in the right direction, and allow it to be discriminative the next time it encounters a topic of the same class.

We found that this architecture allows for the emergence of an efficient convention. We evaluated our methodology on a population of 5 agents, with every scene consisting of one instance of each class of the MNIST dataset – 0 to 9 (LeCun et al., 1998). After 25,000 interactions with training images, we froze the networks and measured the success on unique scenes from the test images. The evolutionary dynamics during training are shown in Figure 2. The population reached communicative success in 91% of the interactions and conventionality in 82% (an interaction is said to be conventional when the listener would have produced the same utterance as the speaker). We evaluated our method on a similarly structured dataset, FashionMNIST (Xiao et al., 2017), without further parameter tuning, reaching communicative success in 86% and conventionality in 75% of the interactions.

With this research, we contribute to the current methods for language emergence in populations of simulated agents by introducing the use of neural networks for concept representation. We exploit the complexity that they can capture as well as their versatility to provide an open-ended approach applicable to different image datasets. We believe that the modularity of this architecture will allow us to progress towards the transferability of concepts across different tasks.

References

Botoko Ekila, J., Nevens, J., Verheyen, L., Beuls, K., & Van Eecke, P. (2024). Decentralised emergence of robust and adaptive linguistic conventions in populations of autonomous agents grounded in continuous worlds. *arXiv preprint*, arXiv:2401.08461.

Foerster, J., Assael, Y., de Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, *29*, 2137–2145.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint*, arXiv:1412.6980.

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint*, arXiv:1612.07182.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2324.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*, 8024–8035.

Rita, M., Michel, P., Chaabouni, R., Pietquin, O., Dupoux, E., & Strub, F. (2025). Language evolution with deep learning. *The Oxford Handbook of Approaches to Language Evolution*, 335–370.

Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, *2*, 319–332.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*, 469–529.

Van Eecke, P., Beuls, K., Botoko Ekila, J., & Rădulescu, R. (2022). Language games meet multi-agent reinforcement learning: A case study for the naming game. *Journal of Language Evolution*, *7*, 213–223.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint*, arXiv:1708.07747.

Evolution and compression in LLMs: On the emergence of human-aligned categorization

Nathaniel Imel[*1] and Noga Zaslavsky[1*]

[*]Corresponding Authors: ni2128@nyu.edu, nogaz@nyu.edu
[1]Department of Psychology, New York University, NY, United States

It has been shown that human semantic systems achieve near-optimal compression via the Information Bottleneck (IB) complexity-accuracy tradeoff (Zaslavsky et al., 2018, 2019, 2021), and that cultural transmission via iterated learning can drive initially random systems towards IB efficiency (Imel et al., 2025; Carlsson et al., 2024). Large language models (LLMs) are not trained for this objective, which raises the question: are LLMs capable of evolving efficient human-aligned semantic systems? To address this, we focus on color — a key testbed for theories of categorization with uniquely rich human data — and replicate with LLMs two influential human studies.

First, we conducted an English color naming experiment designed to assess the efficiency and human-alignment of the color naming systems of LLMs (Figure 1B). Each LLM was prompted to label the 330 color chip stimuli (Figure 1A) from the World Color Survey (WCS; Cook et al., 2005) and then evaluated with respect to (a) The English color naming data of Lindsey and Brown (2014) and (b) the IB theoretical bound of Zaslavsky et al. (2018). We find that LLMs vary widely in their complexity and English-alignment. While many open-weight models struggle to align with English speakers, the larger instruction-tuned models generally achieve better human-alignment and IB-efficiency (Figure 1D,E).

Second, we conducted an iterated language learning (ILL) experiment (based on Xu et al., 2013), designed to test whether LLMs simply mimic patterns in their training data or actually exhibit a human-like inductive bias toward IB-efficiency. To this end, we propose a method which we refer to as Iterated in-Context Language Learning (IICLL, Figure 1C; building on Zhu and Griffiths (2024)). At each generation t, an LLM is prompted with a small dataset, d_{t-1}, consisting of colors with pseudo labels sampled from the previous generation's language, L_{t-1}. With d_{t-1} in context, the LLM performs the naming task for the full meaning space.

We analyzed the resulting LLM trajectories through the lens of IB, following Imel et al. (2025)'s analysis of the human ILL data. We find that akin to humans, LLMs iteratively restructure initially random systems towards greater IB-efficiency and similarity to WCS languages (Figure 1F,G,H), with Gemini being

the only model able to recapitulate the wide range of near-optimal IB-tradeoffs observed in humans (Figure 1I). Taken together, our findings demonstrate how human-aligned semantic categories can emerge in LLMs via the same fundamental principle that underlies semantic efficiency in humans.

An extended version of this work, available at `https://arxiv.org/abs/2509.08093`, will appear in ICLR 2026.

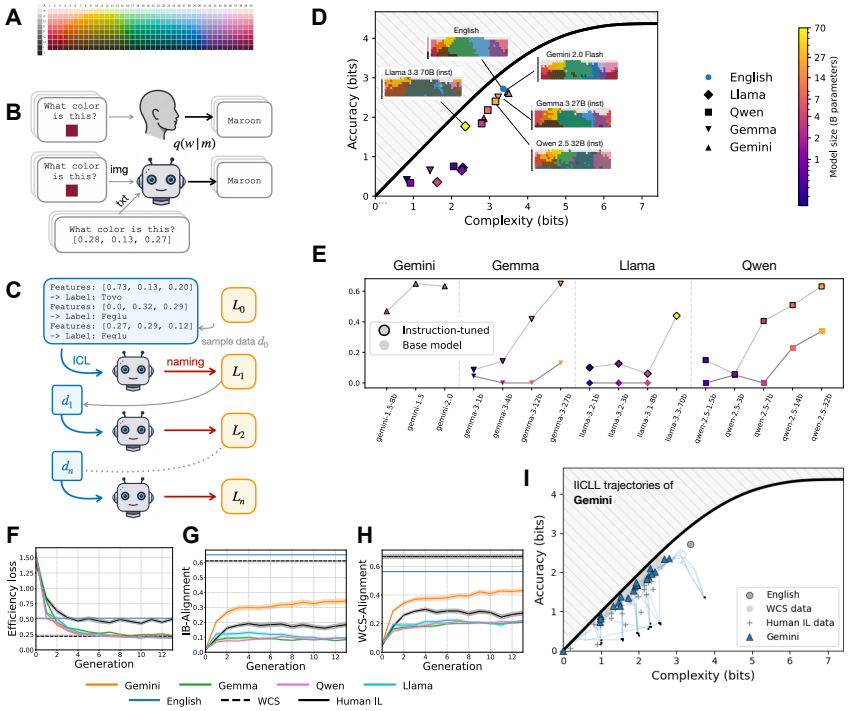


Figure 1.: **A.** The WCS color naming grid. **B.** Color naming task with humans (top) and LLMs (bottom). **C.** The IICLL paradigm (see main text). **D.** IB complexity-accuracy tradeoffs achieved by LLMs, plotted with respect to the English tradeoff (blue circle) and IB theoretical bound (black curve; from Zaslavsky et al., 2018). The English and LLM color naming system are plotted against the WCS grid, where each chip is colored by the color-centroid of its modal category. **E.** English-alignment, measured by the max-Normalized Mutual Information (AMI_{max}; Vinh et al., 2010) between the English and LLM systems; markers are the same as in (A). Across model families, size and instruction-tuning are associated with higher complexity and better alignment to English. **F-H**. Over generations of IICLL, LLM systems become more efficient (F), more aligned to IB optima (G), and more aligned to WCS languages (H). **I.** Gemini 2.0 converges to near-optimal solutions within the same range of human IL chains (Imel et al., 2025) and WCS languages (Zaslavsky et al., 2018). Small black dots correspond to random initialization of chains with varying number of categories, $k \in \{2, 3, 4, 5, 6, 14\}$. Thin blue lines correspond to IICLL trajectories.

References

Carlsson, E., Dubhashi, D., & Regier, T. (2024). Cultural evolution via iterated learning and communication explains efficient color naming systems. *Jour-*

nal of Language Evolution, *9*(1-2), 49–66.

Cook, R. S., Kay, P., & Regier, T. (2005). The World Color Survey Database: History and use. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (pp. 223–241). Oxford: Elsevier Science Ltd.

Imel, N., Culbertson, J., Kirby, S., & Zaslavsky, N. (2025). Iterated language learning is shaped by a drive for optimizing lossy compression. In *Proceedings of the 47th annual meeting of the cognitive science society.*

Lindsey, D. T., & Brown, A. M. (2014). The color lexicon of American English. *Journal of Vision*, *14*(2), 17.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, *11*(95), 2837–2854.

Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1758), 20123073.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-Optimal compression. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society.*

Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. In *41st Annual Meeting of the Cognitive Science Society.*

Zhu, J.-Q., & Griffiths, T. L. (2024). *Eliciting the Priors of Large Language Models using Iterated In-Context Learning* (No. arXiv:2406.01860). arXiv.

Modeling Cognition with Minimal AI

Anna Jon-And[*1,2] and Jérôme Michaud[1,3]

[*]Corresponding Author:anna.jon-and@su.se
[1]Centre for cultural evolution, Department of Psychology, Stockholm University, Stockholm, Sweden
[2]Department of Romance Studies and Classics, Stockholm University, Stockholm, Sweden
[3]Department of Mathematics and Physics, Mälardalen University, Västerås, Sweden

1. Introduction

Artificial intelligence plays an increasingly central role in research on language evolution, yet many widely adopted AI tools rest on assumptions that are poorly aligned with human cognition. Modern neural approaches, including large language models, offer impressive performance but obscure the mechanisms that give rise to linguistic structure, rely on unrealistic memory and data resources, limiting our ability to evaluate hypotheses about the cognitive prerequisites of language (Cuskley, Woods, & Flaherty, 2024; Mahowald et al., 2024). In this contribution, we argue that if we use AI with the aim of testing hypotheses about cognitive capacities required for language, assumptions need to be cognitively plausible and the methods used need to be interpretable. We illustrate this by outlining a research program that employs minimal and transparent AI architectures to probe which domain-general mechanisms may be *necessary and sufficient* for the emergence of grammar.

Our approach builds on three methodological principles:

(1) Minimalism. To identify prerequisites rather than reproduce full linguistic competence, we construct learning systems with *as little built-in structure as possible*, following the tradition of cognitive architecture research that seeks to explain diverse cognitive phenomena with few general mechanisms (Newell, 1994). The aim is not to maximize performance but to determine which minimal cognitive assumptions allow linguistic abstractions to emerge.

(2) Analyzability. Instead of neural networks, we rely on equation-based reinforcement learning (Rescorla & Wagner, 1972; Sutton & Barto, 2018). This design enables precise inspection of internal decision processes, allowing us to trace how specific mechanisms contribute to hierarchical structure, chunk reuse, and the emergence of grammatical abstraction (Jon-And & Michaud, 2024a, 2024b).

(3) Cognitive plausibility. The architecture only incorporates learning and memory mechanisms that are strongly supported across species, such as associative learning and generalization (Hull, 1943; Ghirlanda & Enquist, 2003;

Ghirlanda, Lind, & Enquist, 2020), or that are well-studied in humans, like chunking (Bybee, 2002; Christiansen & Arnon, 2017; Tomasello, 2003), and limited working memory (Cowan, 2001; Miller, 1956). Crucially, we model **faithful sequence representation**, a cognitive capacity proposed to be uniquely human and central to language and culture (Ghirlanda, Lind, & Enquist, 2017; Lind, Vinken, Jonsson, Ghirlanda, & Enquist, 2023; Enquist, Ghirlanda, & Lind, 2023; Jon-And, Jonsson, Lind, Ghirlanda, & Enquist, 2023; Lind & Jon-And, 2025). We also implement an *evolutionary type system* in which grammatical categories arise dynamically from the learner's experience, as a generalization over sequential order and chunking, rather than being predefined (Lambek, 1958; Jon-And & Michaud, 2024a).

Using these principles, we investigate how a learner with a short and flexible sequence memory can segment a word stream into meningful macro-units (sentences) and develop increasingly structured representations. The learner decides incrementally whether to chunk elements into hierarchical structures or place boundaries, while generalizing over recurring combinations give rise to the emergent type system. Across several models, we observe the spontaneous emergence of hierarchical grouping, recurrent chunks, and eventually functional grammatical categories such as noun-like and verb-like types (Jon-And & Michaud, 2024a, 2024b). Importantly, these structures arise under cognitive constraints: when memory is limited, the learner must reorganize information into reusable combinatorial units, whereas systems with very large or unbounded memory — such as many neural AI models — can succeed without developing grammar-like abstractions. This suggests that grammar may emerge as a solution to the combinatorial challenges imposed by human cognitive limitations, rather than as an optimization target per se.

Methodologically, this work highlights both the potential and the pitfalls of using AI to study language evolution. Minimal interpretable architectures allow researchers to test hypotheses about uniquely human cognitive properties, such as faithful sequence representation and chunk-based processing. At the same time, many off-the-shelf AI tools can inadvertently smuggle in assumptions — about memory capacity, optimization goals, or implicit structure — that make them unsuitable models of language learning and therefore unsuitable for evolutionary inference. Careful attention to cognitive constraints, transparency, and mechanistic interpretability is essential when employing AI in this domain.

By presenting a case study in cognitively grounded AI modeling, this work contributes to the workshop's methodological goal: discussing how to use AI productively and appropriately in language-evolution research. Rather than treating AI as a black box or a source of ready-made predictions, we propose using minimalist AI systems as *theoretical instruments* for uncovering the cognitive conditions under which linguistic structure can arise. This perspective complements existing neural and agent-based approaches and underscores the importance of

aligning AI methods with the scientific questions that drive research on the evolution of language.

Acknowledgements

This work was supported by the Swedish Research Council (VR 2022-02737).

References

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in second language acquisition*, *24*(2), 215–221.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in cognitive science*, *9*(3), 542–551.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87–114.

Cuskley, C., Woods, R., & Flaherty, M. (2024). The limitations of large language models for understanding human language and cognition. *Open Mind*, *8*, 1058–1083.

Enquist, M., Ghirlanda, S., & Lind, J. (2023). *The human evolutionary transition: From animal intelligence to culture.* Princeton University Press.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36.

Ghirlanda, S., Lind, J., & Enquist, M. (2017). Memory for stimulus sequences: a divide between humans and other animals? *Open Science*, *4*(6), 161011.

Ghirlanda, S., Lind, J., & Enquist, M. (2020). A-learning: A new formulation of associative learning theory. *Psychonomic Bulletin & Review*, *27*, 1166–1194.

Hull, C. L. (1943). Principles of behavior: an introduction to behavior theory.

Jon-And, A., Jonsson, M., Lind, J., Ghirlanda, S., & Enquist, M. (2023). Sequence representation as an early step in the evolution of language. *PLOS Computational Biology*, *19*(12), e1011702.

Jon-And, A., & Michaud, J. (2024a). Emergent grammar from a minimal cognitive architecture. In J. Nölle, L. Raviv, K. E. Graham, S. Hartmann, Y. Jadoul, & e. a. Josserand, M. (Eds.), *The Evolution of Language: Proceedings of the 15th International Conference (EvoLangXV).* Madison.

Jon-And, A., & Michaud, J. (2024b). Usage-based grammar induction from minimal cognitive principles. *Computational Linguistics*, *50*(4), 1375–1414.

Lambek, J. (1958). The mathematics of sentence structure. *The American Mathematical Monthly*, *65*(3), 154–170.

Lind, J., & Jon-And, A. (2025). A sequence bottleneck for animal intelligence and language? *Trends in cognitive sciences*, *29*(3), 242–254.

Swarm Robotics: Embodiment and Spatial Constraints in Emergent Communication Models

Roman Miletitch[*1] and Limor Raviv[1,2]

[1]Language Evolution and Adaptation in Diverse Situations (LEADS) group, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]Donders Center for Cognition, Radboud University, Nijmegen, the Netherlands

Swarm robotics is a powerful approach to multi-agent systems inspired by the collective behavior of social insects (Bonabeau, Dorigo, & Theraulaz, 1999; Sahin, 2005). It is used to study how repeated local interactions in large populations of simple, embodied robots can generate complex group-level behaviors, such as foraging, aggregation, and communication (Brambilla, Ferrante, Birattari, & Dorigo, 2013). Specifically, the goal is to demonstrate emergent complexity: bridging the substantial gap between the simple cognition of each individual robot, and the rich, cumulative behavior produced by the swarm. Swarm robotics is thus a form of agent-based modeling, in which multiple agents have simple cognitive capacities (e.g., limited to no memory) and communication protocols. In addition, robots are embodied, possessing physical bodies (real or simulated) with local perception, and interact not only with each other but also with objects and obstacles in their environment.

Embodiment is the defining feature of swarm robotics (Brooks, 1991): agents are not immaterial, abstract entities, but have a tangible body equipped with specific sensors and actuators, operating within a physical environment. This physical body occupies space, leading to collisions and occlusions, as well as allowing direct manipulation of the environment. Sensors and actuators are grounded in physics: their placement on the robot introduces blind spots, and their behavior is affected by factors such as noise or mechanical inertia and latency. Swarm robotics experiments can be conducted with both real and simulated robots, often using the same code, and ideally yielding comparable results. An arena and real robots (e.g., e-puck, footbots, kilobots) are required for live experiments, while simulations require realistic 2D or 3D physics engines and virtual robots that are modeled after real ones in terms of their morphology and movement dynamics. This physical layer introduces a finer granularity in spatial organization. In particular, constraining where robots move, when they encounter one another, and what they perceive at any given moment can effectively shape contact patterns, coordination demands, and information flow within the swarm, thereby directly

influencing the emergence and dynamics of social behavior.

This setup is therefore more ecologically valid, and enables the modeling of more complex behaviors, population structures, and environments, in less abstract and/or symbolic ways. With respect to language evolution research, most classic computational work in the field has either explored the interactions of two embodied robots with realistic starting conditions (Steels, Spranger, Van Trijp, Höfer, & Hild, 2012; Steels, 2001; Spranger, 2016), or the interactions of large communities of simple agents without embodiment (Griffiths & Kalish, 2007; Smith, 2009), but not a combination of the two, that is, larger populations of embodied robots (Cambier et al., 2020; Cambier & Miletitch, 2025). In addition, while most classic agent-based models focus only on simulating the target behavior (i.e., communication) in isolation, here agents are first and foremost engaged in realistic tasks such as foraging or navigation, and simultaneously also communicate - enabling deeper insights into the relationship between communication and other aspects of daily life (Miletitch, Reina, Dorigo, & Trianni, 2019; Cambier, Albani, Frémont, Trianni, & Ferrante, 2021). As such, this method is most appropriate when the group dynamics of interest are spatially grounded, when ecological conditions are directly tested, and/or when the phenomenon under study depends on embodied interactions and environmental aspects (whether because realism is required for emergence, or because one wishes to test a hypothesis using more realistic models in which agents are engaged in concrete tasks). For example, swarm robotics is a good platform to test the potential relationship between language evolution in large communities and ecological factors, such as when testing the emergence of structure and/or linguistic diversity in groups that need to survive in harsh vs. abundant environments (modeled with different spread and availability of resources). Conversely, swarm robotics is less suitable for behaviors that rely on complex symbolic representations that are not grounded in the environment, when centralized control is required, or when a standard multi-agent model is sufficient. In particular, it offers limited advantages in domains where the environment is largely irrelevant, where physical interactions and embodiment do not shape behavior, and when agent coordination does not depend on spatial constraints, communication limits, or real-world dynamics.

In practice, designing a swarm robotics experiment begins with selecting the collective behaviors of interest, which in turn determines the type of robot required, its sensor range, actuators, size, cognitive architecture, communication bandwidth, morphology, battery capacity, and, pragmatically, its cost. For example, robots' cognition can include generational/social learning cognition, and can be coded through simple decision trees, Bayesian rules, or neural networks, depending on the goals of the study. A corresponding environment must then be constructed, with items and structures represented in ways that align with the robots' sensing and acting capabilities. Finally, individual behaviors must be coded, from which the targeted swarm behavior is expected to emerge.

To make this process more concrete and accessible, we are organizing a workshop at EvoLang 2026 that will combine a brief introduction to swarm dynamics and the ARGoS simulator (Pinciroli et al., 2012), a live coding demo and a hands-on try-it-yourself session using a pre-prepared simulation environment (with documentation and starter code). Participants will have the opportunity to create their own robotic behaviors and test different experimental setups. All material for the workshop is at `https://romamile.com/swlang`.

As with all agent-based models, swarm designs are highly sensitive to parameter choices, often in ways that are not immediately apparent: arena size, robot speed, and swarm density can all substantially influence the resulting dynamics. Physical constraints can also introduce unintended biases and artifacts, particularly in navigation (e.g., systematic drift, loss of traction, pushing of items in the environment), which might artificially favor specific areas of the arena. In simulation, further biases may arise given that aspects that occur naturally with real robots must be explicitly coded (e.g., scripted updates of environmental states, changes to items triggered by robot actions, or spawning of resources). As in any model, interpreting the simulation's results demands caution: simplified models can produce compelling patterns, but not all conclusions generalize to natural systems. Assessing how much can legitimately be extrapolated from such experiments therefore depends on the realism and scope of the model. For example, we would need to ensure that robots' behavior remains faithful to naturalistic patterns and constraints (e.g., gravity), and that we are explicit about which aspects of the target behavior are captured and which are deliberately omitted or simplified. Ultimately, our goal with using swarm robotics is to model a specific behavior or to test a hypothesis, rather than to engineer an efficient solution to a task.

References

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm intelligence: From natural to artificial systems.* Oxford University Press.

Brambilla, M., Ferrante, E., Birattari, M., & Dorigo, M. (2013). Swarm robotics: a review from the swarm engineering perspective. *Swarm Intelligence*, *7*(1), 1–41.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1–3), 139–159.

Cambier, N., Albani, D., Frémont, V., Trianni, V., & Ferrante, E. (2021). Cultural evolution of probabilistic aggregation in synthetic swarms. *Applied Soft Computing*, *113*, 108010.

Cambier, N., & Miletitch, R. (2025). Task-driven language evolution with swarm robotics. In L. Raviv & C. Boeckx (Eds.), *The oxford handbook of approaches to language evolution* (pp. 371–389). The Oxford University Press.

Cambier, N., Miletitch, R., Frémont, V., Dorigo, M., Ferrante, E., & Trianni, V.

(2020). Language evolution in swarm robotics: A perspective. *Frontiers in Robotics and AI*, *7*, 12.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, *31*(3), 441–480.

Miletitch, R., Reina, A., Dorigo, M., & Trianni, V. (2019). Emergent naming of resources in a foraging robot swarm. *arXiv preprint arXiv:1910.02274*.

Pinciroli, C., Trianni, V., O´Grady, R., Pini, G., Brutschy, A., Brambilla, M., Mathews, N., Ferrante, E., Di Caro, G., Ducatelle, F., et al.. (2012). Argos: a modular, parallel, multi-engine simulator for multi-robot systems. *Swarm intelligence*, *6*(4), 271–295.

Sahin, E. (2005). Swarm robotics: From sources of inspiration to domains of application. *Swarm Robotics*, 10–20.

Smith, K. (2009). Iterated learning in populations of bayesian agents. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 31). Amsterdam, The Netherlands.

Spranger, M. (2016). *The evolution of grounded spatial language.* Language Science Press.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent systems*, *16*(5), 16–22.

Steels, L., Spranger, M., Van Trijp, R., Höfer, S., & Hild, M. (2012). Emergent action language on real robots. In *Language grounding in robots* (pp. 255–276). Springer.

Lind, J., Vinken, V., Jonsson, M., Ghirlanda, S., & Enquist, M. (2023). A test of memory for stimulus sequences in great apes. *Plos one*, *18*(9), e0290546.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in cognitive sciences*, *28*(6), 517–540.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Newell, A. (1994). *Unified theories of cognition.* Harvard University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement [incollection]. In *Classical conditioning: current research and theory* (p. 64-69). Appleton-Century-Crofts.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Harvard University Press.

NeLLCom-Lex: A Neural-agent Framework to Study and Simulate Lexical Semantic Change

Yuqing Zhang[*1], Ecesu Ürker[2], Tessa Verhoef[3], Gemma Boleda[2,4], and Arianna Bisazza[1]

[*]Corresponding Author: yuqing.zhang@rug.nl
[1]Center for Language and Cognition, University of Groningen
[2]Department of Translation and Language Sciences, Universitat Pompeu Fabra
[3]Leiden Institute of Advanced Computer Science, Leiden University
[4]Catalan Institution for Research and Advanced Studies (ICREA)

Language use, particularly the referential context of communication, is widely recognized as a key factor shaping linguistic systems (Campbell, 2013; Goldberg, 1995). Yet the mechanisms through which it exerts this influence remain difficult to study, since their effects typically unfold over long timescales (Winters et al., 2015; Hawkins, 2004). To address this challenge, computational simulations have been used to investigate the emergence and change of lexical systems (Steels, 1997; De Boer, 2006). Color systems are among the most extensively studied and well-documented lexical domains (Berlin & Kay, 1991; Cook, Kay, & Regier, 2005; Gärdenfors, 2000, 2014), and have therefore long served as a testbed for agent-based computational models. Simulation studies with artificial agents have shown that agents can develop color categories from scratch through language games (Steels et al., 2005; Steels, 1997; Steels & McIntyre, 1998; De Boer, 2006), and that incorporating simple perceptual constraints can yield human-like color systems (Loreto et al., 2012). More recent work has employed neural-network-based agents, combining mechanistic control with larger-scale experimentation in which machines, rather than humans, interact via reinforcement learning and develop human-like lexicons (Carlsson et al., 2024; Kågebäck et al., 2020; Chaabouni et al., 2021; Lazaridou et al., 2018; Lazaridou & Baroni, 2020; Kharitonov et al., 2019; Lian et al., 2023). Such models allow controlled manipulation of contextual variables and enable the study of language emergence and change across different timescales.

Building on this approach, the present study investigates the relationship between language use and the structure of lexical color systems within a single generation using our neural agent framework, **NeLLCom-Lex**[1]. Unlike prior simulations, where agents develop color systems from scratch that may be uninterpretable to humans, our agents are first trained on English color terms, ensuring

[1]Code and materials are available at `https://github.com/yuqing0304/NeLLCom-Lex`.

that the resulting lexicon remains human-interpretable. Using this framework, we examine how variation in communicative context affects the granularity of lexical distinctions (fine- vs. coarse-grained). More specifically, we test whether agents exhibit human-like pragmatic behavior in their naming choices, using more specific terms when the context requires finer discriminations (Figure 1)

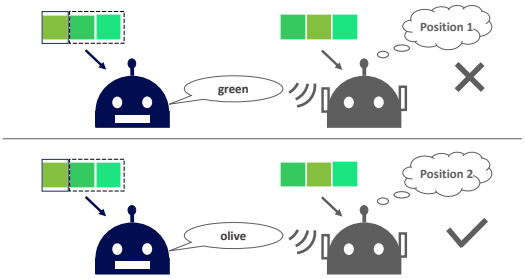


Figure 1.: Dyadic reference game. Reproduced from Figure 1c in Zhang et al. (2025)

NeLLCom-Lex extends NeLLCom (Lian et al., 2023), an agent communication framework in which a speaker agent and a listener agent are implemented as feed-forward neural networks (FNNs). Agents first acquire a predefined language through supervised learning (SL), learning from labeled input–output pairs, and then engage in reinforcement learning (RL), learning through trial-and-error interaction to maximize a shared communicative reward. To model how language use shapes the color lexicon, we trained agents in the SL phase on English color terms using the English version of the Colors dataset (Gualdoni & Boleda, 2024). This dataset consists of human dyadic reference-game interactions in which a speaker describes a target color chip so that a listener can identify it among two distractors (see Figure 1). Targets appear in one of three contextual conditions: **FAR** (both distractors clearly distinct), **SPLIT** (one distinct and one similar distractor), and **CLOSE** (both distractors similar). For the RL phase, we generated a large set of color triplets for each context condition using the same sampling procedure as in the original dataset.

In this framework, both speaker and listener receive the target chip (and, when present, its distractors) as CIELAB color vectors, which are projected into embeddings via FNNs. Color terms are mapped to representations using an embedding layer. The **speaker** is provided with a referential context consisting of a target and two distractors. Each color is encoded independently through separate feed-forward blocks, after which the resulting embeddings are concatenated and passed through an additional feed-forward block to produce a joint representation. A final linear classifier then predicts a single-symbol message. To

manipulate the speaker's access to contextual information, in the *context-aware* setting all colors are encoded normally, in other words the speaker knows the distractor colors, whereas in the *context-unaware* setting, distractor embeddings are zeroed out, such that the speaker has no access to the distractor colors. The **listener** receives the speaker's message along with the referential context, where the three candidate chips are presented in randomized order. Each chip is independently projected into a hidden representation, while the color term is mapped to its embedding. The listener then computes similarity scores between the term embedding and each candidate chip embedding, and transforms these scores into a probability distribution via a softmax layer. In other words, based on the embeddings of the color chips and the color term, and their similarity, the listener predicts which chip is most likely to be the target. The listener is trained to select the intended referent. In the SL phase, speaker and listener agents are trained separately whereas they are updated together during RL to optimize communicative success.

We conducted two sets of analyses: 1) how access to contextual information during training affects agent behavior, and 2) how overall communicative need influences agent behavior. For the first analysis, we examined agents that received access to contextual information at different stages of training in order to determine how the timing of context introduction shapes their behavior. Specifically, we tested five conditions: (i) SL without context, (ii) SL with context, (iii) SL and RL without context, (iv) SL without context followed by RL with context, and (v) SL and RL with context. The results showed that, regardless of when context was introduced, agents exposed to contextual information eventually developed human-like pragmatic behavior: that is, they used more informative words (measured using the informativeness metric from Gualdoni and Boleda (2024)) in close contexts (the context illustrarted in Figure 1) compared to far contexts. Moreover, agents without access to context developed inefficient and overly informative lexicons, with many words referring to very narrow sets of referents. As a result, they required larger vocabularies to achieve comparable communicative accuracy, yet still performed worse than agents with access to context. Notably, access to context during SL alone did not lead to human-like lexicons; these agents continued to produce overinformative and inefficient systems, highlighting the necessity of exposure to context during RL for developing human-like lexical structure.

For the second analysis, we manipulated the overall distribution of contexts during RL training and evaluated the behavior of agents that had access to contextual information in both SL and RL. Agents were trained under three conditions: (i) **AllFar**, in which they were only exposed to the far contexts, (ii) **HalfHalf**, they were exposed to equal amounts of far and close contexts, and (iii) **AllClose**, they were only exposed to close contexts. They were then tested on an unseen, balanced evaluation set consisting of 50% far and 50% close contexts. The results showed that all agents exhibited pragmatic adaptation; however, agents trained in

far contexts were less sensitive to contextual variation, in other words they did not change their choice of word depending on the distractors, as AllFar training provided little pressure to modulate informativeness based on contextual difficulty. Although overall lexical diversity was comparable across agents, the AllClose agents developed the most human-like and efficient lexicons. These findings indicate that exposure to increased contextual pressures during training promotes more flexible pragmatic behavior and more adaptive lexical systems.

We are currently using NeLLCom-Lex to simulate the meaning narrowing process, which has been widely observed in historical linguistic change. For example, the Old English word *dēor* once denoted any wild animal but gradually came to refer specifically to deers only, a shift likely shaped by repeated use in contexts such as hunting, where *deer* were "the favorite animal of the chase" (Online Etymology Dictionary, n.d.). We adopted the SL and RL with context pipeline and increased the frequency of RL training data drawn from a targeted, narrower color region (e.g., olive), analogous to the disproportionate exposure to deer in relevant communicative contexts. Preliminary results show that increased exposure to a narrower subregion during RL does not reliably induce denotational narrowing. Changes in denotational structure, as measured by informativeness and prototype shift, are small and not systematically aligned with predictions of semantic narrowing. These findings suggest that frequency skew alone does not lead to stable denotational narrowing in color semantics. One possible explanation is the highly continuous and overlapping structure of color meaning space, where category boundaries are graded and prototypes are unstable. Unlike discrete taxonomic domains (e.g., animal categories), color categories may resist narrowing under frequency pressure due to semantic overlap. These findings highlight domain-specific semantic structure constraining the extent to which frequency-based communicative pressures can reshape lexical meaning, but further work is needed for more conclusive findings.

References

Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution.* University of California Press, Berkeley.

Campbell, L. (2013). *Historical linguistics.* Edinburgh University Press.

Carlsson, E., Dubhashi, D., & Regier, T. (2024). Cultural evolution via iterated learning and communication explains efficient color naming systems. *Journal of Language Evolution*, *9*(1-2), 49–66.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, *118*(12), e2016569118.

Cook, R. S., Kay, P., & Regier, T. (2005). The world color survey database. In *Handbook of categorization in cognitive science* (pp. 223–241). Elsevier.

De Boer, B. (2006). Computer modelling as a tool for understanding language

evolution. In *Evolutionary epistemology, language and culture: A non-adaptationist, systems theoretical approach* (pp. 381–406). Springer.

Gärdenfors, P. (2000). Conceptual spaces: The geometry of thought.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces.* MIT Press.

Goldberg, A. E. (1995). *Construction grammar: a construction grammar approach to argument structure.* University of Chicago Press.

Gualdoni, E., & Boleda, G. (2024). Why do objects have many names? a study on word informativeness in language use and lexical systems. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 18150–18163). Miami, Florida, USA: Association for Computational Linguistics.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars.* OUP Oxford.

Kharitonov, E., Chaabouni, R., Bouchacourt, D., & Baroni, M. (2019). EGG: a toolkit for research on emergence of lanGuage in games. In S. Padó & R. Huang (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp): System demonstrations* (pp. 55–60). Hong Kong, China: Association for Computational Linguistics.

Kågebäck, M., Carlsson, E., Dubhashi, D., & Sayeed, A. (2020). A reinforcement-learning approach to efficient communication. *PLOS ONE*, *15*(7), 1-26.

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. In *International conference on learning representations.* Vancouver, Canada.

Lian, Y., Bisazza, A., & Verhoef, T. (2023). Communication Drives the Emergence of Language Universals in Neural Agents: Evidence from the Word-order/Case-marking Trade-off. *Transactions of the Association for Computational Linguistics*, *11*, 1033-1047.

Loreto, V., Mukherjee, A., & Tria, F. (2012). On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, *109*(18), 6819–6824.

Online Etymology Dictionary. (n.d.). *Deer. In* Online Etymology Dictionary. (Retrieved May 19, 2025, from https://www.etymonline.com/search?q=deer)

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of communication*, *1*(1), 1–34.

Steels, L., Belpaeme, T., et al.. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, *28*(4), 469–488.

Steels, L., & McIntyre, A. (1998). Spatially distributed naming games. *Advances in complex systems*, *1*(04), 301–323.

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(3), 415–449.

Zhang, Y., Ürker, E., Verhoef, T., Boleda, G., & Bisazza, A. (2025). Nellcom-lex: A neural-agent framework to study the interplay between lexical systems and language use. *arXiv preprint arXiv:2509.22479*.