ELCC: the Emergent Language Corpus Collection

Brendon Boldt
Language Technologies Insitute
Carnegie Mellon University
Pittsburgh, PA 15213
bboldt@cs.cmu.edu

David Mortensen
Language Technologies Insitute
Carnegie Mellon University
Pittsburgh, PA 15213
dmortens@cs.cmu.edu

Abstract

We introduce the Emergent Language Corpus Collection (ELCC): a collection of corpora collected from open source implementations of emergent communication systems across the literature. These systems include a variety of signalling game environments as well as more complex tasks like a social deduction game and embodied navigation. Each corpus is annotated with metadata describing the characteristics of the source system as well as a suite of analyses of the corpus (e.g., size, entropy, average message length). Currently, research studying emergent languages requires directly running different systems which takes time away from actual analyses of such languages, limits the variety of languages that are studied, and presents a barrier to entry for researchers without a background in deep learning. The availability of a substantial collection of well-documented emergent language corpora, then, will enable new directions of research which focus their purview on the properties of emergent languages themselves rather than on experimental apparatus.

1 Introduction

Emergent communication (also called *emergent language*) studies machine learning simulations that attempt to model the development of communication systems from scratch. These simulations have recently been argued to have many potential applications, both in artificial intelligence and in the scientific study of human communicative behavior [Boldt and Mortensen, 2024b]. However, this potential has been stymied, due in part to the challenges—up to this point—in comparing emergent communication systems (ECSs)[1] in a way that allows their properties and utility to be understood. There have been no standard datasets, for example, that represent output from various systems across the literature. The only recourse, for researchers desiring to compare ECSs or the languages that they generate, has been to reimplement or collect code for earlier ECSs. Getting this code to run is a significant challenge, requiring the research to chase outdated versions of Python, install CUDA drivers (again), and work with sparsely documented code. The results of such efforts, too, would lack comparability and reproducibility. To our knowledge, efforts at large-scale comparison between different outputs of different ECSs are not reported in the literature.

This paper introduces the Emergent Language Corpus Collection (ELCC), a resource that addresses this glaring lacuna in the field of emergent communication research; namely, it is a set of corpora

[1]Emergent communications systems are more commonly referred to as simply "environments"; we choose to use the term "system" in order to emphasize that what goes into producing an emergent language is more than just an environment including also the architecture of the agents, optimization procedure, datasets, and more.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

of “languages” generated by most of the prominent ECS types described in the current literature (accompanied by extensive metadata regarding the typology of the ECS, how the data were produced, and the statistical properties of the resulting corpus). With these data, it is possible for researchers, even those with limited technological expertise, to compare emergent languages, whether in their structural properties [van der Wal et al., 2020] or their utility in pretraining models for downstream NLP tasks [Yao et al., 2022a]. ELCC is published on Hugging Face Datasets at `https://huggingface.co/datasets/bboldt/elcc` with data and code licensed under the CC BY 4.0 and MIT licenses, respectively.

We discuss related work in Section 2. Section 3 lays out the design of ELCC while Section 4 describes the content of the collection. Section 5 presents some brief analyses, discussion, and future work related to ELCC. Finally, we conclude in Section 6.

Contributions The primary contribution of this paper is a first-of-its kind data resource which will enable broader engagement and new research direction within the field of emergent communication. Additionally, code published for reproducing the data resource also improve the reproducibility of existing ECS implementations in the literature, supporting further research beyond just the data resource itself.

2 Related Work

Emergent communication There is no direct precedent to this paper in the emergent communication literature that we are aware of. Perkins [2021b] introduces the TexRel dataset, but this is a dataset observations for training ECSs, not data generated by them. Some papers do provide the emergent language corpora generated from their experiments (e.g., Yao et al. [2022a]), although these papers are few in number and only include the particular ECS used in the paper. At a high level, the EGG framework [Kharitonov et al., 2021] strives to make emergent languages easily accessible, though instead of providing corpora directly, it provides a framework for implementing ECSs. Thus, while EGG is more useful for someone building new systems entirely, it is not as geared towards research projects aiming directly at analyzing emergent languages themselves.

Data resources At a high level, ELCC is a collection of different subdatasets which all represent various manifestations of common phenomenon (emergent communication, in this case). On a basic level, ELCC could is somewhat analogous to any multi-lingual dataset (where “human language” is the phenomenon). Taking the notion of “phenomenon” more narrowly (i.e., of more direct scientific interest), it could be compared to Blum et al. [2023], which presents a collection of grammar snapshot pairs for 52 different languages as instances of diachronic language change. Zheng et al. [2024] present a dataset of conversations from Chatbot Arena containing where “text generated by different LLMs” is the phenomenon of interest. Insofar as ELCC documents the basic typology of different ECSs, it is similar to the World Atlas of Language Structures (WALS) Dryer and Haspelmath [2013].

3 Design

3.1 Format

ELCC is a collection of ECSs each of which has one or more associated *variants* which correspond to runs with different hyperparameter settings (e.g., different random seed, message length, dataset). Each variant has metadata along with the corpus generated from its settings. Each ECS has its own metadata as well and code to generate the corpus and metadata of each variant. The structure of ELCC is illustrated in Figure 1.

ECS metadata Environment metadata provides a basic snapshot of a given system and where it falls in the taxonomy of ECSs. As the collection grows, this makes it easier to ascertain the contents of the collection and easily find the most relevant corpora for a given purpose. This metadata will also serve as the foundation for future analyses of the corpora looking how the characteristics of an ECS relate to the properties of its output.

```
data/ ..................................................... top-level directory
  ecs-1/...................................... directory for a particular ECS
    metadta.yml.................................... metadata about the ECS
    code/......................... directory containing files to produce the data
    data/........................ directory containing corpus and metadata files
      hparams-1/.............. directory for run with specific hyperparameters
        corpus.jsonl ........................................ corpus data
        metadata.json ............ metadata specific for corpus (e.g., metrics)
      hparams-2/.............................................. as above
      hparams-n/.............................................. as above
  ecs-2/........................................................ as above
  ecs-n/........................................................ as above
```

Figure 1: The file structure of ELCC.

source The URL of where the code for producing the data is from.

upstream_source The URL of the original repo if **source** is a fork.

paper The URL of the paper documenting the ECS (if any).

game_type The high level category of the game implemented in the ECS; currently one of *signalling*, *conversation*, or *navigation*.

game_subtype A finer-grained categorization of the game, if applicable.

observation_type The type of observation that the agents make; currently either *vector* or *image* (i.e., an image embedding).

observation_continuous Whether or not the observation is continuous as opposed to discrete (e.g., image embeddings versus concatenated one-hot vectors).

data_source Whether the data being communicated about is from a natural source (e.g., pictures), synthetic, or does not apply (e.g., in a social deduction game).

variants A dictionary where each entry corresponds to one of the variants of the particular ECS. Each entry in the dictionary contains any relevant hyperparameters that distinguish it from the other variants.

seeding_available Whether or not the ECS implements seeding the random elements of the system.

multi_step Whether or not the ECS has multiple steps per episode.

symmetric_agents Whether or not agents both send and receive messages.

multi_utterance Whether or not multiple utterances are included per line in the dataset.

more_than_2_agents Whether or not the ECS has a population of >2 agents.

These metadata are stored as YAML files in each ECS directory. A Python script is provided to validate these entries against a schema. See Appendix A for an example of such a metadata file.

Corpus Each *corpus* comprises a list of *lines* each of which is, itself, an array of *tokens* represented as integers. Each line corresponds to a single episode or round in the particular ECS. In the case of multi-step or multi-agent systems, this might comprise multiple individual utterances which are then concatenated together to form the line (no separation tokens are added). Each corpus is generated from a single run of the ECS; that is, they are never aggregated from distinct runs of the ECS.

Concretely, a *corpus* is formatted as a JSON lines (JSONL) file where each *line* is a JSON array of integer *tokens*. This is visualized in Figure 2. There are a few advantages of JSONL: (1) it is JSON-based meaning it is standardized and has wide support across programming languages, (2) it is easy to interpret and work with for those without a computer science background, and (3) it is

```
[3,14,1,59,26,5,35]
[89,7,93,2,38,4,62]
[6,43,3,8]
```

Figure 2: Example of a three-line corpus in the JSON lines format.

Source	Type	Data source	Multi-agent	Multi-step	n corp.
Kharitonov et al. [2021]	signalling	synthetic	No	No	15
Yao et al. [2022a]	signalling	natural	No	No	2
Mu and Goodman [2021b]	signalling	both	No	No	6
Chaabouni et al. [2022]	signalling	natural	Yes	No	5
Unger and Bruni [2020]	navigation	synthetic	No	Yes	18
Boldt and Mortensen [2022]	navigation	syntehtic	No	Yes	20
Brandizzi et al. [2022]	conversation	—	Yes	Yes	7

Table 1: Taxonomic summary the contents of ELCC.

line-based meaning it is easy to process with command line tools.[2] Corpora are also available as single JSON objects (i.e., and array of arrays), accessible via the Croissant ecosystem.

Corpus analysis For each corpus in ELCC we run a suite of analyses to produce a quantitative snapshot. This suite metrics is intended not only to paint a robust a picture of the corpus but also to serve as jumping-off point for future analyses on the corpora. Specifically, we apply the following to each corpus: token count, unique tokens, line count, unique lines, tokens per line, tokens per line stand deviation, 1-gram entropy, normalized 1-gram entropy, entropy per line, 2-gram entropy, 2-gram conditional entropy, EoS token present, and EoS padding. *Normalized* 1-*gram entropy* is computed as 1-*gram entropy* divided by the maximum entropy given the number of unique tokens in that corpus.

We consider an EoS (end-of-sentence) token to be present when: (1) every line ends with token consistent across the entire corpora, and (2) the first occurrence of this token in a line is only ever followed by more the same token. For example, `0` could be an EoS token in the corpus `[[1,2,0],[1,0,0]]` but not `[[1,2,0],[0,1,0]]`. EoS padding is defined as a corpus having an EoS token, all lines being the same length, and the EoS token occurs more than once in a line at least once in the corpus.

Additionally, each corpus also has a small amount of metadata copied directly from the output of the ECS; for example, this might include the success rate in a signalling game environment. We do not standardize this because it can vary widely from ECS to ECS, though it can still be useful for comparison to other results among variants within an ECS.

Reproducibility ELCC is designed with reproducibility in mind. With each ECS, code is included to reproduce the corpora and analysis metadata. Not only does this make ELCC reproducible, but it sometimes helps the reproducibility of the underlying implementation insofar as it fixes bugs, specifies Python environments, and provides examples of how to run an experiment with a certain set of hyperparameters. Nevertheless, in this code, we have tried to keep as close to the original implementations as possible. When the underlying implementation supports it, we set the random seed (or keep the default) for the sake of consistency, although many systems do not offer ways to easily set this.

	min	25%	50%	75%	max
Token Count	48616	67248	110000	1061520	42977805
Line Count	999	5765	10000	10000	2865187
Tokens per Line	5.87	7.00	11.00	33.53	7212.72
Tokens per Line SD	0.00	0.00	2.31	13.81	445.84
Unique Tokens	2	7	10	20	902
Unique Lines	18	1253	2440	4911	309405
1-gram Entropy	0.36	2.12	2.80	3.37	6.60
1-gram Normalized Entropy	0.16	0.71	0.82	0.90	1.00
2-gram Entropy	0.42	3.16	4.11	5.88	12.88
2-gram Conditional Entropy	0.06	0.85	1.41	2.54	6.29
Entropy per Line	4.38	21.23	30.80	71.85	30233.52

Table 2: Five-number summar of the analyses across corpora of ELCC. Entropy in bits.

4 Content

ELCC contains 73 corpora across 8 ECSs taken from the literature for which free and open source implementations were available. With our selection we sought to capture variation across a three distinct dimensions:

1. Variation across ECSs generally, including elements like game types, message structure, data sources, implementation details.
2. Variation among different hyperparameter settings within an ECS, including message length, vocabulary size, dataset, and game difficulty.
3. Variation within a particular hyperparameter setting that comes from inherent stochasticity in the system; this is useful for gauging the stability or convergence of an ECS.

Table 1 shows an overview of the taxonomy of ELCC based on the ECS-level metadata. In addition to this, Table 2 provides a quantitative summary of the corpus-level metrics described in Section 3.1 (full results in Appendix C). The following sections describe the individual ECSs in more detail.

4.1 Scope

The scope of the contents of ELCC is largely the same as discussed in reviews such as Lazaridou and Baroni [2020] and Boldt and Mortensen [2024b, Section 1.2]. This comprises agent-based models for simulating the formation of “natural” language from scratch using deep neural networks. Importantly, *from scratch* means that the models are not pretrained or tuned on human language. Typically, such simulations make use of reinforcement learning to train the neural networks, though this is not a requirement in principle.

One criterion that we do use to filter ECSs for inclusion is its suitability for generating corpora as described above. This requires that the communication channel is discrete, analogous to the distinct words/morphemes which for the units of human language. This excludes a small number of emergent communication papers have approached emergent communication through constrained continuous channels like sketching [Mihai and Hare, 2021b] or acoustic-like signals [Eloff et al., 2023]. Other systems use discrete communication but are in essence one token per episode (e.g., Tucker et al. [2021b]) which would not form a suitable corpus for addressing most research questions.

4.2 Signalling games

The *signalling game* (or *reference game*) [Lewis, 1969] represents a plurality, if not majority, of the systems present in the literature. A brief, non-exhaustive review of the literature yielded 43 papers

[2]E.g., Creating a 100-line random sample of a dataset could be done with `shuf dataset.jsonl | head -n 100 > sample.jsonl`

which use minor variations of the signalling game, a large number considering the modest body of emergent communication literature (see Appendix B). The basic format of the signalling game is a single round of *sender* agent making an observation, passing a message to the *receiver* agent, and the receiver performing an action based on the information from the message. The popularity of this game is, in large part, because of its simplicity in both concept and implementation. Experimental variables can be manipulated easily while introducing minimal confounding factors. Furthermore, the implementations can entirely avoid the difficulties of reinforcement learning by treating the sender and receiver agents as a single neural network, resulting in autoencoder with a discrete bottleneck which can be trained with backpropagation and supervised learning.

The two major subtypes of the signalling game are the *discrimination game* and the *reconstruction game*. In the discrimination game, the receiver must answer a multiple-choice question, that is, select the correct observation from among incorrect "distractors". In the reconstruction game, the receiver must recreate the input directly, similar to the decoder of an autoencoder.

Vanilla For the most basic form of the signalling game, which we term "vanilla", we use the implementation provided in the Emergent of LanGuage in Games (EGG) framework [Kharitonov et al., 2021, MIT license]. It is vanilla insofar as it comprises the signalling game with the simplest possible observations (synthetic, concatenated one-hot vectors), a standard agent architecture (i.e., RNNs) and no additional dynamics or variations on the game. Both the discrimination game and the reconstruction game are included. This system provides a good point of comparison for other ECSs which introduce variations on the signalling game. The simplicity of the system additionally makes it easier to vary hyperparameters: for example, the size of the dataset can be scaled arbitrarily and there is no reliance on pretrained embedding models.

Natural images "Linking emergent and natural languages via corpus transfer" [Yao et al., 2022a, MIT license] presents a variant of the signalling game which uses embeddings of natural images as the observations. In particular, the system uses embedded images from the MS-COCO and Conceptual Captions datasets consisting of pictures of everyday scenes. Compared to the uniformly sampled one-hot vectors in the vanilla setting, natural image embeddings are real-valued with a generally smooth probability distribution rather than being binary or categorical. Furthermore natural data distributions are not uniform and instead have concentrations of probability mass on particular elements; this non-uniform distribution responsible for various features of human language (e.g., human languages' bias towards describing warm colors [Gibson et al., 2017, Zaslavsky et al., 2019]).

Differing observations "Emergent communication of generalizations" [Mu and Goodman, 2021b, MIT license] presents a variant of the discrimination signalling game which they term the *concept game*. The concept game changes the way that the sender's observation corresponds with the receiver's observations. In the vanilla discrimination game, the observation the sender sees is exactly the same as the correct observation that the sender sees. In the concept game, the sender instead observes a set of inputs which share a particular concept (e.g., red triangle and red circle are both red), and the correct observation (among distractors) shown to the receiver contains the same concept (i.e., red) while not being identical to those observed by the sender. The rationale for this system is that the differing observations will encourage the sender to communicate about abstract concepts rather than low-level details about the observation. This ECS also presents the vanilla discrimination game as well as the *setref game*, which is similar to the reference game except that the whole object is consistent (e.g., different sizes and locations of a red triangle).

Multi-agent population "Emergent communication at scale" [Chaabouni et al., 2022, Apache 2.0-license] presents a signalling game system with populations of agents instead of the standard fixed pair of sender and receiver. For each round of the game, then, a random sender is paired with a random receiver. This adds a degree of realism to the system, as natural human languages are ways developed within a population and not just between two speakers (cf. idioglossia). More specifically, language developing among a population of agents prevents some degree "overfitting" between sender and receiver; in this context, having a population of agents functions as an ensembling approach to regularization.

4.3 Other games

Considering that the signalling game is close to the simplest possible game for an ECS, moving beyond the signalling game generally entails an increase in complexity. There is no limit to the theoretical diversity of games, although some of the most common games that we see in the literature are conversation-based games (e.g., negotiation, social deduction) and navigation games. These games often introduce new aspects to agent interactions like: multi-step episodes, multi-agent interactions, non-linguistic actions, and embodiment.

These kinds of systems, as a whole, are somewhat less popular in the literature. On a practical level, more complex systems are more complex to implement and even harder to get to converge reliably—many higher-level behaviors, such as planning or inferring other agent's knowledge, are difficult problems for reinforcement learning in general, let alone with discrete multi-agent emergent communication. On a methodological level, more complexity in the ECS makes it harder to formally analyze the system as well as eliminate confounding factors in empirical investigation. With so many moving parts, it can be difficult to prove that some observed effect is not just a result of some seemingly innocent hyperparameter choice (e.g., learning rate, samples in the rollout buffer) [Boldt and Mortensen, 2022]. Nevertheless, we have reason to believe that these complexities are critical to understanding and learning human language as a whole [Bisk et al., 2020], meaning that the difficulties of more complex systems are worth overcoming as they are part of the process of creating more human-like, and therefore more useful, emergent languages.

Grid-world navigation "Generalizing Emergent Communication" [Unger and Bruni, 2020, BSD-3-clause license] introduces an ECS which takes some of the basic structure of the signalling game and applies it to a navigation-based system derived from the synthetic Minigrid/BabyAI environment [Chevalier-Boisvert et al., 2018, 2023]. A sender with a bird's-eye view of the environment sends messages to a receiver with a limited view who has to navigate to a goal location. Beyond navigation, some environments present a locked door for which the receiver must first pick up a key in order to open.

What distinguishes this system most from the signalling game is that it is multi-step and embodied such that the utterances within an episodes are dependent on each other. Among other things, this changes the distribution properties of the utterances. For example, if the receiver is in Room A at timestep T, it is more likely to be in Room A at timestep $T + 1$; thus if utterances are describing what room the receiver is in, this means that an utterance at $T + 1$ has less uncertainty given the content of an utterance at T. Practically speaking, the multiple utterances in a given episode are concatenated together to form a single line in the corpus in order to maintain the dependence of later utterances on previous ones.

Continuous navigation "Mathematically Modeling the Lexicon Entropy of Emergent Language" [Boldt and Mortensen, 2022, GPL-3.0 license] introduces a simple navigation-based ECS which is situated in a continuous environment. A "blind" receiver is randomly initialized in an obstacle-free environment and must navigate toward a goal zone guided by messages from the center which observes the position of the receiver relative to the goal. The sender sends a single discrete token at each timestep, and a line in the dataset consists of the utterances from each timestep concatenated together. This system shares the time-dependence between utterances of the grid-world navigation system although with no additional complexity of navigating around obstacle, opening doors, etc. On the other hand, the continuous nature of this environment provides built-in stochasticity since there are (theoretically) infinitely many distinct arrangements of the environment that are possible, allowing for more natural variability in the resulting language.

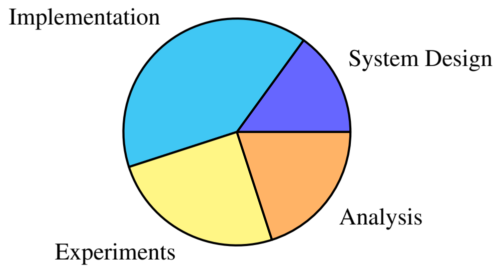

Figure 3: Typical allocation of resources to an emergent communication paper (time-wise and page count-wise): analysis is a relatively small component. ELCC enables research which can focus primarily on analysis.

Social deduction "RLupus: Cooperation through the emergent communication in The Werewolf social deduction game" [Brandizzi et al., 2022, GPL-3.0 license] introduces an ECS based on the social deduction game *Werewolf* (a.k.a., *Mafia*) where, through successive rounds of voting and discussion, the "werewolves" try to eliminate the "villagers" before the villagers figure out who the werewolves are. In a given round, the discussion takes the form of all agents broadcasting a message to all other agents after which a vote is taken on whom to eliminate. As there are multiple rounds in a given game, this system introduces multi-step as well as multi-speaker dynamics into the language. Furthermore, the messages also influence distinct actions in the system (i.e., voting). These additional features in the system add the potential for communication strategies that are shaped by a variety of heterogeneous factors rather than simply the distribution of observations (as in the signalling game).

5 Discussion

Work enabled by ELCC In the typical emergent communication paper, only a small amount of time and page count is allocated to analysis with the lion's share being taken up by designing the ECS, implementing it, and running experiments (see Figure 3). Even if one reuses an existing implementation, a significant portion of work still goes towards designing and running the experiments, and the analysis is still limited to that single system. ELCC, on the other hand, enables research which can not only be dedicated wholly to analysis but analysis across a variety of system. Furthermore, removing the necessity of implementing and/or running experiments allows researchers without machine learning backgrounds to contribute to emergent communication research from more linguistic angles that otherwise would not be possible.

In particular, ELCC enable work that focuses on the lexical properties of emergent communication, looking at the statical properties and patterns of the surface forms of a given language (e.g., Zipf's law). Ueda et al. [2023] is a prime example of this; this paper investigates whether or not emergent languages obey Harris' Articulation Schema relating conditional entropy to the presence of word boundaries. The paper finds mixed evidence for HAS in emergent languages but only evaluated a handful of settings in a single ECS; ELCC could be used in such a case to radically extend the scope emergent languages evaluated. Additionally, ELCC can similarly extend the range of emergent languages evaluated in the context of machine learning, such as Yao et al. [2022a], Boldt and Mortensen [2024a] which look at emergent language's suitability for deep transfer learning to downstream NLP tasks.

ECS implementations and reproducibility In the process of compiling ELCC, we observed a handful of trends in the implementations of emergent communication systems. A significant proportion of papers do not publish the implementations of experiments, severely limiting the ease reproducing the results or including such work in a project such as ELCC, considering that a large amount of the work in creating an ECS is not in the design but in the details of implementation. Even when a free and open source implementation is available, many projects suffer from underspecified Python dependencies (i.e., no indication of versions) which can be difficult to reproduce if the project is older than a few years. Furthermore, some projects also fail to specify the particular hyperparameter settings or commands to run the experiments presented in the paper; while these can often be recovered with some investigation, this and the above issue prove to be obstacles which could easily be avoided. For an exemplar of a well-documented, easy-to-run implementation of an ECS and its experiments, see Mu and Goodman [2021b] at `https://github.com/jayelm/emergent-generalization/` which not only provides dependencies with version and documentation how to download the data but also a complete shell script which executes the commands to reproduce the experiments.

Future of ELCC While ELCC is a complete resource as presented in this paper, ELCC is intended to be an ongoing project which incorporates further ECSs, analyses, and taxonomic features as the body of emergent communication literature and free and open source implementations continues to grow. This approach involves the community not only publishing well-documented implementation of their ECSs but also directly contributing to ELCC in the spirit of scientific collaboration and free and open source software. ELCC, then, is intended to become a hub for a variety of stakeholders in the emergent communication research community, namely a place for: ECS developers to contribute and publicize their work, EC researchers to stay up-to-date on new ECSs, and EC-adjacent researchers to find emergent languages which they can analyze or otherwise use for their own research.

Limitations While ELCC provides a representative sample of the ECSs present in the literature, it is not comprehensive collection of all of the open source implementations let alone all ECSs in the literature. This limitation is especially salient in the case foundational works in EC which have no open source implementations (e.g., Mordatch and Abbeel [2018]). Additionally, ELCC only provides unannotated corpora without any reference to the semantics of the communication, which limits the range of analyses that can be performed (e.g., measures of compositionality are precluded because since it fundamentally a relationship between surface forms and their semantics). In terms of compute resources, we estimate that on the order of 150 GPU-hours (NVIDIA A6000 or equivalents) on an institutional cluster were used in the development of ELCC.

6 Conclusion

In this paper, we have introduced ELCC, a collection of emergent language corpora annotated with taxonomic metadata and suite of descriptive metrics derived from free and open source implementations of emergent communication systems introduced in the literature. ELCC also provides code for running these implementations, in turn, making those implementations more reproducible. This collection is the first of its kind in providing easy access to a variety of emergent language corpora. Thus, it enables new kinds of research on emergent communication as it serves as a foundation for research focused on the analysis of emergent languages themselves.

References

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.703. URL `https://www.aclweb.org/anthology/2020.emnlp-main.703`.

Frederic Blum, Carlos Barrientos, Adriano Ingunza, Damián E Blasi, and Roberto Zariquiey. Grammars across time analyzed (GATA): a dataset of 52 languages. *Scientific Data*, 10(1):835, 2023.

Brendon Boldt and David Mortensen. Xferbench: Benchmark for emergent languages using deep transfer learning. `https://github.com/brendon-boldt/xferbench/`, 2024a.

Brendon Boldt and David R. Mortensen. Mathematically modeling the lexicon entropy of emergent language. *arXiv*, 2211.15783, 2022. URL `https://arxiv.org/abs/2211.15783`.

Brendon Boldt and David R Mortensen. A review of the applications of deep learning-based emergent communication. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=jesKcQxQ7j.

Diane Bouchacourt and Marco Baroni. How agents see things: On visual representations in an emergent language game. *arXiv*, arXiv:1808.10696, 2018.

Nicolo' Brandizzi, Davide Grossi, and Luca Iocchi. RLupus: Cooperation through the emergent communication in The Werewolf social deduction game. *Intelligenza Artificiale*, 15(2):55–70, 2022. URL https://content.iospress.com/articles/intelligenza-artificiale/ia210081.

Kalesha Bullard, Douwe Kiela, Franziska Meier, Joelle Pineau, and Jakob Foerster. Quasi-equivalence discovery for zero-shot emergent communication. *arXiv*, arXiv:2103.08067, 2021.

Boaz Carmeli, Ron Meir, and Yonatan Belinkov. Emergent quantized communication. *arXiv*, arXiv:2211.02412, 2022.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. Anti-efficient encoding in emergent communication. *arXiv*, arXiv:1905.12561, 2019.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. *arXiv*, arXiv:2004.09124, 2020.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. Emergent communication at scale. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AUGBfDIV9rL.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. Babyai: A platform to study the sample efficiency of grounded language learning. *arXiv preprint arXiv:1810.08272*, 2018.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *CoRR*, abs/2306.13831, 2023.

Aritra Chowdhury, Alberto Santamaria-Pang, James R. Kubricht, Jianwei Qiu, and Peter Tu. Symbolic semantic segmentation and interpretation of covid-19 lung infections in chest ct volumes based on emergent languages. *arXiv*, arXiv:2008.09866, 2020a.

Aritra Chowdhury, Alberto Santamaria-Pang, James R. Kubricht, and Peter Tu. Emergent symbolic language based deep medical image classification. *arXiv*, arXiv:2008.09860, 2020b.

Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. Co-evolution of language and agents in referential games. *arXiv*, arXiv:2001.03361, 2020.

Kevin Denamganaï and James Alfred Walker. On (emergent) systematic generalisation and compositionality in visual referential games with straight-through gumbel-softmax estimator. *arXiv*, arXiv:2012.10776, 2020.

Roberto Dessì, Diane Bouchacourt, Davide Crepaldi, and Marco Baroni. Focus on what's informative and ignore what's not: Communication strategies in a referential game. *arXiv*, arXiv:1911.01892, 2019.

Roberto Dessì, Eugene Kharitonov, and Marco Baroni. Interpretable agent communication from scratch (with a generic visual processor emerging on the side). *arXiv*, arXiv:2106.04258, 2021.

C. M. Downey, Xuhui Zhou, Leo Z. Liu, and Shane Steinert-Threlkeld. Learning to translate by learning to communicate. *arXiv*, arXiv:2207.07025, 2022.

Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/`.

Kevin Eloff, Okko Räsänen, Herman A. Engelbrecht, Arnu Pretorius, and Herman Kamper. Towards learning to speak and hear through multi-agent communication over a continuous acoustic channel. *arXiv*, 2111.02827, 2023.

Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, and Bevil R. Conway. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790, 2017. doi: 10.1073/pnas.1619666114. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1619666114`.

Shangmin Guo, Yi Ren, Serhii Havrylov, Stella Frank, Ivan Titov, and Kenny Smith. The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *arXiv*, arXiv:1910.05291, 2019.

Shangmin Guo, Yi Ren, Agnieszka Słowik, and Kory Mathewson. Inductive bias and language expressivity in emergent communication. *arXiv*, arXiv:2012.02875, 2020.

Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *arXiv*, arXiv:1705.11192, 2017.

Bence Keresztury and Elia Bruni. Compositional properties of emergent languages in deep learning. *arXiv*, arXiv:2001.08618, 2020.

Eugene Kharitonov and Marco Baroni. Emergent language generalization and acquisition speed are not tied to compositionality. *arXiv*, arXiv:2004.03420, 2020.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. Entropy minimization in emergent languages. *arXiv*, arXiv:1905.13687, 2019.

Eugene Kharitonov, Roberto Dessì, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. EGG: a toolkit for research on Emergence of lanGuage in Games. `https://github.com/facebookresearch/EGG`, 2021.

Bohdan Khomtchouk and Shyam Sudhakaran. Modeling natural language emergence with integral transform theory and reinforcement learning. *arXiv*, arXiv:1812.01431, 2018.

Nur Geffen Lan, Emmanuel Chemla, and Shane Steinert-Threlkeld. On the spontaneous emergence of discrete and compositional signals. *arXiv*, arXiv:2005.00110, 2020.

Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *CoRR*, abs/2006.02419, 2020. URL `https://arxiv.org/abs/2006.02419`.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. *arXiv*, arXiv:1612.07182, 2016.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv*, arXiv:1804.03984, 2018.

David Kellogg Lewis. *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA, 1969.

Fushan Li and Michael Bowling. Ease-of-teaching and language structure from emergent communication. *arXiv*, arXiv:1906.02403, 2019.

Yaoyiran Li, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. Emergent communication pretraining for few-shot machine translation. *arXiv*, arXiv:2011.00890, 2020. doi: 10.18653/v1/2020.coling-main.416.

Ryan Lowe, Abhinav Gupta, Jakob Foerster, Douwe Kiela, and Joelle Pineau. On the interaction between supervision and self-play in emergent communication. *arXiv*, arXiv:2002.01093, 2020.

Diana Rodríguez Luna, Edoardo Maria Ponti, Dieuwke Hupkes, and Elia Bruni. Internal and external pressures on language emergence: least effort, object constancy and frequency. *arXiv*, arXiv:2004.03868, 2020.

Matéo Mahaut, Francesca Franzon, Roberto Dessì, and Marco Baroni. Referential communication in heterogeneous communities of pre-trained visual deep networks. *arXiv*, arXiv:2302.08913, 2023.

Daniela Mihai and Jonathon Hare. Avoiding hashing and encouraging visual semantics in referential emergent language games. *arXiv*, arXiv:1911.05546, 2019.

Daniela Mihai and Jonathon Hare. The emergence of visual semantics through communication games. *arXiv*, arXiv:2101.10253, 2021a.

Daniela Mihai and Jonathon Hare. Learning to draw: Emergent communication through sketching. *arXiv*, 2106.02067, 2021b.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Jesse Mu and Noah Goodman. Emergent communication of generalizations. *arXiv*, arXiv:2106.02668, 2021a.

Jesse Mu and Noah Goodman. Emergent communication of generalizations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17994–18007. Curran Associates, Inc., 2021b. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/9597353e41e6957b5e7aa79214fcb256-Paper.pdf`.

Xenia Ohmer, Michael Marino, Michael Franke, and Peter König. Mutual influence between language and perception in multi-agent communication games. *arXiv*, arXiv:2112.14518, 2021. doi: 10.1371/journal.pcbi.1010658.

Xenia Ohmer, Marko Duda, and Elia Bruni. Emergence of hierarchical reference systems in multi-agent communication. *arXiv*, arXiv:2203.13176, 2022.

Hugh Perkins. Neural networks can understand compositional functions that humans do not, in the context of emergent communication. *arXiv*, arXiv:2103.04180, 2021a.

Hugh Perkins. Texrel: a green family of datasets for emergent communications on relations. *arXiv preprint arXiv:2105.12804*, 2021b.

Eva Portelance, Michael C. Frank, Dan Jurafsky, Alessandro Sordoni, and Romain Laroche. The emergence of the shape bias results from communicative efficiency. *arXiv*, arXiv:2109.06232, 2021.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. *arXiv*, arXiv:2002.01365, 2020.

Mathieu Rita, Rahma Chaabouni, and Emmanuel Dupoux. "lazimpa": Lazy and impatient neural agents learn to communicate efficiently. *arXiv*, arXiv:2010.01878, 2020.

Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. On the role of population heterogeneity in emergent communication. *arXiv*, arXiv:2204.12982, 2022a.

Mathieu Rita, Corentin Tallec, Paul Michel, Jean-Bastien Grill, Olivier Pietquin, Emmanuel Dupoux, and Florian Strub. Emergent communication: Generalization and overfitting in lewis games. *arXiv*, arXiv:2209.15342, 2022b.

Shane Steinert-Threlkeld. Paying attention to function words. *arXiv*, arXiv:1909.11060, 2019.

Agnieszka Słowik, Abhinav Gupta, William L. Hamilton, Mateja Jamnik, Sean B. Holden, and Christopher Pal. Structural inductive biases in emergent communication. *arXiv*, arXiv:2002.01335, 2020.

Mycal Tucker, Huao Li, Siddharth Agrawal, Dana Hughes, Katia Sycara, Michael Lewis, and Julie Shah. Emergent discrete communication in semantic spaces. *arXiv*, arXiv:2108.01828, 2021a.

Mycal Tucker, Huao Li, Siddharth Agrawal, Dana Hughes, Katia Sycara, Michael Lewis, and Julie A Shah. Emergent discrete communication in semantic spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10574–10586. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5812f92450ccaf17275500841c70924a-Paper.pdf.

Ryo Ueda, Taiga Ishii, and Yusuke Miyao. On the word boundaries of emergent languages based on harris's articulation scheme. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=b4t9_XASt6G.

Thomas A. Unger and Elia Bruni. Generalizing emergent communication. *arXiv: Artificial Intelligence*, 2020. URL https://arxiv.org/abs/2001.01772.

Oskar van der Wal, Silvan de Boer, Elia Bruni, and Dieuwke Hupkes. The grammar of emergent languages. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3339–3359, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.270. URL https://aclanthology.org/2020.emnlp-main.270.

Shunyu Yao, Mo Yu, Yang Zhang, Karthik Narasimhan, Joshua Tenenbaum, and Chuang Gan. Linking emergent and natural languages via corpus transfer. In *International Conference on Learning Representations (ICLR)*, 2022a.

Shunyu Yao, Mo Yu, Yang Zhang, Karthik R Narasimhan, Joshua B. Tenenbaum, and Chuang Gan. Linking emergent and natural languages via corpus transfer. *arXiv*, arXiv:2203.13344, 2022b.

Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier. Color naming reflects both perceptual structure and communicative need. *Topics in Cognitive Science*, 11(1):207–219, 2019. doi: https://doi.org/10.1111/tops.12395. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12395.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Łukasz Kuciński, Tomasz Korbak, Paweł Kołodziej, and Piotr Miłoś. Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. *arXiv*, arXiv:2111.06464, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work deals solely with emergent communication (i.e., synthetic data) which has been gathered from code bases with open source licenses. Research on emergent communication is still in "basic" stage with few, if any, immediate impacts outside the research community.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.
   (b) Did you mention the license of the assets? [Yes] See Section 4.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 1.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The data produced is synthetic and has little semantic content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A ECS-Level Metadata Example

See Figure 4.

```yaml
origin:
  source: https://github.com/brendon-boldt/emergent_communication_at_scale
  upstream_source:
    https://github.com/google-deepmind/emergent_communication_at_scale
  paper: https://openreview.net/forum?id=AUGBfDIV9rL
system:
  game_type: signalling
  data_source: natural
  game_subtype: discrimination
  observation_type: image
  observation_continuous: true
  seeding_available: true
  multi_step: false
  more_than_2_agents: true
  multi_utterance: false
  symmetric_agents: false
  variants:
    imagenet-1x10:
      n_receivers: 10
      n_senders: 1
    imagenet-10x10:
      n_receivers: 10
      n_senders: 10
    imagenet-5x5:
      n_receivers: 5
      n_senders: 5
    imagenet-1x1:
      n_receivers: 1
      n_senders: 1
    imagenet-10x1:
      n_receivers: 1
      n_senders: 10
```


Figure 4: Example of an ECS metadata file in the YAML format.

B Papers based on the signalling game

Mu and Goodman [2021a], Ohmer et al. [2022], Yao et al. [2022b], Rita et al. [2022a], Ohmer et al. [2021], Łukasz Kuciński et al. [2021], Portelance et al. [2021], Tucker et al. [2021a], Dessì et al. [2021], Bullard et al. [2021], Perkins [2021a], Mihai and Hare [2021a], Denamganaï and Walker [2020], Guo et al. [2020], Li et al. [2020], Rita et al. [2020], Chowdhury et al. [2020a,b], Lan et al. [2020], Chaabouni et al. [2020], Luna et al. [2020], Kharitonov and Baroni [2020], Ren et al. [2020], Słowik et al. [2020], Lowe et al. [2020], Keresztury and Bruni [2020], Dagan et al. [2020], Mihai and Hare [2019], Dessì et al. [2019], Guo et al. [2019], Steinert-Threlkeld [2019], Li and Bowling [2019], Kharitonov et al. [2019], Chaabouni et al. [2019], Khomtchouk and Sudhakaran [2018], Bouchacourt and Baroni [2018], Lazaridou et al. [2018], Havrylov and Titov [2017], Lazaridou et al. [2016], Mahaut et al. [2023], Carmeli et al. [2022], Rita et al. [2022b], Downey et al. [2022]

C Per system analysis

See Tables 3 to 6.

name	Token Count	Line Count	Tokens per Line	Tokens per Line SD
babyai-sr/GoToObj	130648	6116	21.361674	12.470737
babyai-sr/GoToObjLocked	272712	5629	48.447682	15.939260
babyai-sr/GoToObjLocked_ambiguous	229504	5507	41.674959	17.414703
babyai-sr/GoToObjLocked_ambiguous-freq_1	2605112	5179	503.014482	45.179870
babyai-sr/GoToObjLocked_ambiguous-freq_2	1061520	5396	196.723499	70.458489
babyai-sr/GoToObjLocked_ambiguous-freq_32	67248	5496	12.235808	3.993043
babyai-sr/GoToObjLocked_ambiguous-freq_4	402248	5728	70.224860	30.849604
babyai-sr/GoToObjLocked_ambiguous-msg_16	511840	5514	92.825535	33.765546
babyai-sr/GoToObjLocked_ambiguous-msg_32	855744	5508	155.363834	58.659355
babyai-sr/GoToObjLocked_ambiguous-msg_4	103228	5730	18.015358	6.603910
babyai-sr/GoToObjUnlocked	118752	6077	19.541221	7.060342
babyai-sr/GoToObjUnlocked-freq_1	1666456	6006	277.465201	205.399768
babyai-sr/GoToObjUnlocked-freq_2	333552	5777	57.737926	28.293252
babyai-sr/GoToObjUnlocked-freq_32	48616	6001	8.101316	0.894576
babyai-sr/GoToObjUnlocked-freq_4	193176	5762	33.525859	13.813609
babyai-sr/GoToObjUnlocked-msg_16	273008	6038	45.214972	18.173718
babyai-sr/GoToObjUnlocked-msg_32	469440	5765	81.429315	33.131090
babyai-sr/GoToObjUnlocked-msg_4	58588	5759	10.173294	4.351285
corpus-transfer-yao-et-al/cc	42977805	2865187	15.000000	0.000000
corpus-transfer-yao-et-al/coco_2014	1241745	82783	15.000000	0.000000
ec-at-scale/imagenet-10x1	2500000	250000	10.000000	0.000000
ec-at-scale/imagenet-10x10	2500000	250000	10.000000	0.000000
ec-at-scale/imagenet-1x1	2500000	250000	10.000000	0.000000
ec-at-scale/imagenet-1x10	2500000	250000	10.000000	0.000000
ec-at-scale/imagenet-5x5	2500000	250000	10.000000	0.000000
egg-discrimination/4-attr_4-val_3-dist_0-seed	110000	10000	11.000000	0.000000
egg-discrimination/4-attr_4-val_3-dist_1-seed	110000	10000	11.000000	0.000000
egg-discrimination/4-attr_4-val_3-dist_2-seed	110000	10000	11.000000	0.000000
egg-discrimination/6-attr_6-val_3-dist_0-seed	110000	10000	11.000000	0.000000
egg-discrimination/6-attr_6-val_3-dist_1-seed	110000	10000	11.000000	0.000000
egg-discrimination/6-attr_6-val_3-dist_2-seed	110000	10000	11.000000	0.000000
egg-discrimination/6-attr_6-val_9-dist_0-seed	110000	10000	11.000000	0.000000
egg-discrimination/6-attr_6-val_9-dist_1-seed	110000	10000	11.000000	0.000000
egg-discrimination/6-attr_6-val_9-dist_2-seed	110000	10000	11.000000	0.000000
egg-discrimination/8-attr_8-val_3-dist_0-seed	110000	10000	11.000000	0.000000
egg-discrimination/8-attr_8-val_3-dist_1-seed	110000	10000	11.000000	0.000000
egg-discrimination/8-attr_8-val_3-dist_2-seed	110000	10000	11.000000	0.000000
egg-reconstruction/4-attr_4-val_10-vocab_10-len	110000	10000	11.000000	0.000000
egg-reconstruction/6-attr_6-val_10-vocab_10-len	110000	10000	11.000000	0.000000
egg-reconstruction/8-attr_8-val_10-vocab_10-len	110000	10000	11.000000	0.000000
generalizations-mu-goodman/cub-concept	1333330	133333	10.000000	0.000000
generalizations-mu-goodman/cub-reference	1333330	133333	10.000000	0.000000
generalizations-mu-goodman/cub-set_reference	1333330	133333	10.000000	0.000000
generalizations-mu-goodman/shapeworld-concept	1164800	166400	7.000000	0.000000
generalizations-mu-goodman/shapeworld-reference	1164800	166400	7.000000	0.000000
generalizations-mu-goodman/shapeworld-set_reference	1164800	166400	7.000000	0.000000
nav-to-center/lexicon_size_11	65528	10000	6.552800	2.521193
nav-to-center/lexicon_size_118	58664	10000	5.866400	2.167199
nav-to-center/lexicon_size_17	59936	10000	5.993600	2.282533
nav-to-center/lexicon_size_174	63129	10000	6.312900	2.434747
nav-to-center/lexicon_size_25	61659	10000	6.165900	2.323010
nav-to-center/lexicon_size_255	60054	10000	6.005400	2.263000
nav-to-center/lexicon_size_37	62753	10000	6.275300	2.396604
nav-to-center/lexicon_size_54	58778	10000	5.877800	2.197195
nav-to-center/lexicon_size_7	61295	10000	6.129500	2.315368
nav-to-center/lexicon_size_80	60250	10000	6.025000	2.256097
nav-to-center/temperature_0.1	74939	10000	7.493900	3.180623
nav-to-center/temperature_0.167	72255	10000	7.225500	2.995171
nav-to-center/temperature_0.278	75732	10000	7.573200	3.106580
nav-to-center/temperature_0.464	79810	10000	7.981000	3.564778
nav-to-center/temperature_0.774	65665	10000	6.566500	2.527326
nav-to-center/temperature_1.29	62566	10000	6.256600	2.364647
nav-to-center/temperature_10	63105	10000	6.310500	2.397059
nav-to-center/temperature_2.15	62019	10000	6.201900	2.314160
nav-to-center/temperature_3.59	58786	10000	5.878600	2.187661
nav-to-center/temperature_5.99	61106	10000	6.110600	2.289491
rlupus/21-player.run-0	7131411	1001	7124.286713	445.837385
rlupus/21-player.run-1	7196469	999	7203.672673	396.806665
rlupus/21-player.run-2	7212723	1000	7212.723000	404.660045
rlupus/9-player.run-0	565164	1003	563.473579	14.708875
rlupus/9-player.run-1	417924	1010	413.786139	124.676603
rlupus/9-player.run-2	414612	1000	414.612000	124.794461
rlupus/9-player.run-3	416538	1003	415.292124	124.887337

Table 3

name	Unique Tokens	Unique Lines	EoS Token Present	EoS Padding
babyai-sr/GoToObj	5	653	False	False
babyai-sr/GoToObjLocked	6	788	False	False
babyai-sr/GoToObjLocked_ambiguous	6	1253	False	False
babyai-sr/GoToObjLocked_ambiguous-freq_1	5	5171	False	False
babyai-sr/GoToObjLocked_ambiguous-freq_2	4	3078	False	False
babyai-sr/GoToObjLocked_ambiguous-freq_32	3	18	False	False
babyai-sr/GoToObjLocked_ambiguous-freq_4	6	3241	False	False
babyai-sr/GoToObjLocked_ambiguous-msg_16	9	4428	False	False
babyai-sr/GoToObjLocked_ambiguous-msg_32	3	1887	False	False
babyai-sr/GoToObjLocked_ambiguous-msg_4	2	1362	False	False
babyai-sr/GoToObjUnlocked	7	521	False	False
babyai-sr/GoToObjUnlocked-freq_1	7	4614	False	False
babyai-sr/GoToObjUnlocked-freq_2	8	3820	False	False
babyai-sr/GoToObjUnlocked-freq_32	4	41	False	False
babyai-sr/GoToObjUnlocked-freq_4	7	2766	False	False
babyai-sr/GoToObjUnlocked-msg_16	13	1740	False	False
babyai-sr/GoToObjUnlocked-msg_32	15	1430	False	False
babyai-sr/GoToObjUnlocked-msg_4	3	400	False	False
corpus-transfer-yao-et-al/cc	391	309405	True	True
corpus-transfer-yao-et-al/coco_2014	902	82783	True	False
ec-at-scale/imagenet-10x1	20	161235	False	False
ec-at-scale/imagenet-10x10	20	126775	False	False
ec-at-scale/imagenet-1x1	20	145834	False	False
ec-at-scale/imagenet-1x10	20	120182	False	False
ec-at-scale/imagenet-5x5	20	169505	False	False
egg-discrimination/4-attr_4-val_3-dist_0-seed	10	240	True	True
egg-discrimination/4-attr_4-val_3-dist_1-seed	10	220	True	True
egg-discrimination/4-attr_4-val_3-dist_2-seed	9	187	True	True
egg-discrimination/6-attr_6-val_3-dist_0-seed	8	2326	True	True
egg-discrimination/6-attr_6-val_3-dist_1-seed	10	3279	True	True
egg-discrimination/6-attr_6-val_3-dist_2-seed	9	1976	True	True
egg-discrimination/6-attr_6-val_9-dist_0-seed	9	2883	True	True
egg-discrimination/6-attr_6-val_9-dist_1-seed	9	1015	False	False
egg-discrimination/6-attr_6-val_9-dist_2-seed	10	2499	True	True
egg-discrimination/8-attr_8-val_3-dist_0-seed	10	2610	True	True
egg-discrimination/8-attr_8-val_3-dist_1-seed	10	2789	True	True
egg-discrimination/8-attr_8-val_3-dist_2-seed	9	2656	True	True
egg-reconstruction/4-attr_4-val_10-vocab_10-len	7	228	True	True
egg-reconstruction/6-attr_6-val_10-vocab_10-len	8	1373	True	True
egg-reconstruction/8-attr_8-val_10-vocab_10-len	8	1464	False	False
generalizations-mu-goodman/cub-concept	23	27163	False	False
generalizations-mu-goodman/cub-reference	23	39457	False	False
generalizations-mu-goodman/cub-set_reference	23	35042	False	False
generalizations-mu-goodman/shapeworld-concept	17	12481	False	False
generalizations-mu-goodman/shapeworld-reference	17	7683	False	False
generalizations-mu-goodman/shapeworld-set_reference	17	28061	True	False
nav-to-center/lexicon_size_11	8	2317	False	False
nav-to-center/lexicon_size_118	61	4392	False	False
nav-to-center/lexicon_size_17	15	3124	False	False
nav-to-center/lexicon_size_174	40	3226	False	False
nav-to-center/lexicon_size_25	12	1961	False	False
nav-to-center/lexicon_size_255	37	3706	False	False
nav-to-center/lexicon_size_37	22	2440	False	False
nav-to-center/lexicon_size_54	43	4911	False	False
nav-to-center/lexicon_size_7	7	1937	False	False
nav-to-center/lexicon_size_80	35	3486	False	False
nav-to-center/temperature_0.1	4	1437	False	False
nav-to-center/temperature_0.167	4	1313	False	False
nav-to-center/temperature_0.278	10	1308	False	False
nav-to-center/temperature_0.464	4	1498	False	False
nav-to-center/temperature_0.774	7	1639	False	False
nav-to-center/temperature_1.29	9	2100	False	False
nav-to-center/temperature_10	64	8793	False	False
nav-to-center/temperature_2.15	64	8643	False	False
nav-to-center/temperature_3.59	64	9044	False	False
nav-to-center/temperature_5.99	64	9263	False	False
rlupus/21-player.run-0	21	1001	False	False
rlupus/21-player.run-1	21	999	False	False
rlupus/21-player.run-2	21	1000	False	False
rlupus/9-player.run-0	9	1003	False	False
rlupus/9-player.run-1	9	1010	False	False
rlupus/9-player.run-2	9	1000	False	False
rlupus/9-player.run-3	9	1003	False	False

Table 4

name	1-gram Entropy	1-gram Normalized Entropy	Entropy per Line
babyai-sr/GoToObj	1.237631	0.533019	26.437867
babyai-sr/GoToObjLocked	0.986990	0.381820	47.817369
babyai-sr/GoToObjLocked_ambiguous	1.724020	0.666942	71.848479
babyai-sr/GoToObjLocked_ambiguous-freq_1	1.463654	0.630362	736.239281
babyai-sr/GoToObjLocked_ambiguous-freq_2	1.385921	0.692961	272.643237
babyai-sr/GoToObjLocked_ambiguous-freq_32	0.358125	0.225952	4.381954
babyai-sr/GoToObjLocked_ambiguous-freq_4	1.996955	0.772528	140.235868
babyai-sr/GoToObjLocked_ambiguous-msg_16	2.555153	0.806061	237.183454
babyai-sr/GoToObjLocked_ambiguous-msg_32	1.350560	0.852108	209.828108
babyai-sr/GoToObjLocked_ambiguous-msg_4	0.922138	0.922138	16.612643
babyai-sr/GoToObjUnlocked	1.993155	0.709976	38.948688
babyai-sr/GoToObjUnlocked-freq_1	0.896426	0.319313	248.726924
babyai-sr/GoToObjUnlocked-freq_2	2.116083	0.705361	122.178237
babyai-sr/GoToObjUnlocked-freq_32	1.643569	0.821785	13.315074
babyai-sr/GoToObjUnlocked-freq_4	2.165184	0.771254	72.589650
babyai-sr/GoToObjUnlocked-msg_16	2.608207	0.704837	117.930014
babyai-sr/GoToObjUnlocked-msg_32	2.940985	0.752769	239.482412
babyai-sr/GoToObjUnlocked-msg_4	1.453225	0.916883	14.784087
corpus-transfer-yao-et-al/cc	1.398306	0.162386	20.974592
corpus-transfer-yao-et-al/coco_2014	6.599321	0.672235	98.989817
ec-at-scale/imagenet-10x1	3.980879	0.921089	39.808790
ec-at-scale/imagenet-10x10	3.908713	0.904391	39.087127
ec-at-scale/imagenet-1x1	4.121796	0.953694	41.217964
ec-at-scale/imagenet-1x10	3.975498	0.919844	39.754975
ec-at-scale/imagenet-5x5	4.213196	0.974842	42.131963
egg-discrimination/4-attr_4-val_3-dist_0-seed	2.996740	0.902109	32.964144
egg-discrimination/4-attr_4-val_3-dist_1-seed	2.494699	0.750979	27.441685
egg-discrimination/4-attr_4-val_3-dist_2-seed	2.564778	0.809097	28.212561
egg-discrimination/6-attr_6-val_3-dist_0-seed	2.581470	0.860490	28.396171
egg-discrimination/6-attr_6-val_3-dist_1-seed	2.887394	0.869192	31.761330
egg-discrimination/6-attr_6-val_3-dist_2-seed	2.573849	0.811959	28.312341
egg-discrimination/6-attr_6-val_9-dist_0-seed	2.861929	0.902838	31.481224
egg-discrimination/6-attr_6-val_9-dist_1-seed	2.462500	0.776832	27.087504
egg-discrimination/6-attr_6-val_9-dist_2-seed	2.750845	0.828087	30.259294
egg-discrimination/8-attr_8-val_3-dist_0-seed	2.426752	0.730525	26.694277
egg-discrimination/8-attr_8-val_3-dist_1-seed	2.556315	0.769528	28.119469
egg-discrimination/8-attr_8-val_3-dist_2-seed	2.802140	0.883977	30.823535
egg-reconstruction/4-attr_4-val_10-vocab_10-len	2.296329	0.817969	25.259614
egg-reconstruction/6-attr_6-val_10-vocab_10-len	2.573243	0.857748	28.305674
egg-reconstruction/8-attr_8-val_10-vocab_10-len	2.295767	0.765256	25.253441
generalizations-mu-goodman/cub-concept	3.752944	0.829644	37.529443
generalizations-mu-goodman/cub-reference	3.103881	0.686159	31.038812
generalizations-mu-goodman/cub-set_reference	3.213538	0.710400	32.135376
generalizations-mu-goodman/shapeworld-concept	3.226724	0.789420	22.587066
generalizations-mu-goodman/shapeworld-reference	3.224439	0.788861	22.571074
generalizations-mu-goodman/shapeworld-set_reference	3.365556	0.823385	23.558893
nav-to-center/lexicon_size_11	2.805418	0.935139	18.383341
nav-to-center/lexicon_size_118	3.767532	0.635255	22.101847
nav-to-center/lexicon_size_17	3.186153	0.815521	19.096524
nav-to-center/lexicon_size_174	3.245330	0.609803	20.487443
nav-to-center/lexicon_size_25	2.804201	0.782212	17.290421
nav-to-center/lexicon_size_255	3.534679	0.678513	21.227163
nav-to-center/lexicon_size_37	3.028477	0.679117	19.004602
nav-to-center/lexicon_size_54	3.754792	0.691966	22.069917
nav-to-center/lexicon_size_7	2.758577	0.982625	16.908697
nav-to-center/lexicon_size_80	3.457586	0.674088	20.831957
nav-to-center/temperature_0.1	1.994309	0.997155	14.945156
nav-to-center/temperature_0.167	1.981753	0.990876	14.319155
nav-to-center/temperature_0.278	1.986637	0.598037	15.045198
nav-to-center/temperature_0.464	1.982692	0.991346	15.823868
nav-to-center/temperature_0.774	2.311150	0.823248	15.176164
nav-to-center/temperature_1.29	2.754878	0.869067	17.236170
nav-to-center/temperature_10	4.905167	0.817528	30.954056
nav-to-center/temperature_2.15	4.966695	0.827782	30.802945
nav-to-center/temperature_3.59	5.340638	0.890106	31.395475
nav-to-center/temperature_5.99	5.266347	0.877724	32.180539
rlupus/21-player.run-0	4.062520	0.924915	28942.558214
rlupus/21-player.run-1	4.196960	0.955523	30233.522552
rlupus/21-player.run-2	3.997152	0.910033	28830.352466
rlupus/9-player.run-0	3.079577	0.971498	1735.260160
rlupus/9-player.run-1	3.119583	0.984119	1290.840311
rlupus/9-player.run-2	3.090164	0.974838	1281.218984
rlupus/9-player.run-3	3.111235	0.981485	1292.071343

Table 5

name	2-gram Entropy	2-gram Conditional Entropy
babyai-sr/GoToObj	1.544519	0.306888
babyai-sr/GoToObjLocked	1.147285	0.160295
babyai-sr/GoToObjLocked_ambiguous	1.978413	0.254393
babyai-sr/GoToObjLocked_ambiguous-freq_1	2.071162	0.607508
babyai-sr/GoToObjLocked_ambiguous-freq_2	1.538991	0.153070
babyai-sr/GoToObjLocked_ambiguous-freq_32	0.420175	0.062050
babyai-sr/GoToObjLocked_ambiguous-freq_4	2.406197	0.409242
babyai-sr/GoToObjLocked_ambiguous-msg_16	3.097571	0.542418
babyai-sr/GoToObjLocked_ambiguous-msg_32	1.463220	0.112660
babyai-sr/GoToObjLocked_ambiguous-msg_4	1.717505	0.795367
babyai-sr/GoToObjUnlocked	2.497606	0.504450
babyai-sr/GoToObjUnlocked-freq_1	1.092966	0.196541
babyai-sr/GoToObjUnlocked-freq_2	2.877898	0.761815
babyai-sr/GoToObjUnlocked-freq_32	1.731359	0.087790
babyai-sr/GoToObjUnlocked-freq_4	2.979210	0.814026
babyai-sr/GoToObjUnlocked-msg_16	3.043978	0.435771
babyai-sr/GoToObjUnlocked-msg_32	3.157215	0.216230
babyai-sr/GoToObjUnlocked-msg_4	2.307255	0.854029
corpus-transfer-yao-et-al/cc	2.059689	0.661383
corpus-transfer-yao-et-al/coco_2014	12.884451	6.285130
ec-at-scale/imagenet-10x1	6.811992	2.831113
ec-at-scale/imagenet-10x10	6.328754	2.420041
ec-at-scale/imagenet-1x1	6.882813	2.761016
ec-at-scale/imagenet-1x10	6.375876	2.400379
ec-at-scale/imagenet-5x5	7.137788	2.924591
egg-discrimination/4-attr_4-val_3-dist_0-seed	4.434835	1.438094
egg-discrimination/4-attr_4-val_3-dist_1-seed	3.550278	1.055580
egg-discrimination/4-attr_4-val_3-dist_2-seed	3.544613	0.979835
egg-discrimination/6-attr_6-val_3-dist_0-seed	3.917021	1.335551
egg-discrimination/6-attr_6-val_3-dist_1-seed	4.308021	1.420628
egg-discrimination/6-attr_6-val_3-dist_2-seed	3.738390	1.164541
egg-discrimination/6-attr_6-val_9-dist_0-seed	4.371053	1.509123
egg-discrimination/6-attr_6-val_9-dist_1-seed	3.578326	1.115826
egg-discrimination/6-attr_6-val_9-dist_2-seed	4.070906	1.320061
egg-discrimination/8-attr_8-val_3-dist_0-seed	3.504384	1.077631
egg-discrimination/8-attr_8-val_3-dist_1-seed	3.712531	1.156216
egg-discrimination/8-attr_8-val_3-dist_2-seed	4.006086	1.203946
egg-reconstruction/4-attr_4-val_10-vocab_10-len	3.212115	0.915787
egg-reconstruction/6-attr_6-val_10-vocab_10-len	3.750294	1.177051
egg-reconstruction/8-attr_8-val_10-vocab_10-len	3.515011	1.219244
generalizations-mu-goodman/cub-concept	5.686797	1.933852
generalizations-mu-goodman/cub-reference	5.641346	2.537465
generalizations-mu-goodman/cub-set_reference	5.509904	2.296366
generalizations-mu-goodman/shapeworld-concept	6.040857	2.814134
generalizations-mu-goodman/shapeworld-reference	5.908455	2.684016
generalizations-mu-goodman/shapeworld-set_reference	6.409305	3.043749
nav-to-center/lexicon_size_11	4.240224	1.434806
nav-to-center/lexicon_size_118	5.389004	1.621472
nav-to-center/lexicon_size_17	4.655472	1.469319
nav-to-center/lexicon_size_174	4.717891	1.472561
nav-to-center/lexicon_size_25	4.106729	1.302528
nav-to-center/lexicon_size_255	5.098629	1.563950
nav-to-center/lexicon_size_37	4.335838	1.307361
nav-to-center/lexicon_size_54	5.463441	1.708649
nav-to-center/lexicon_size_7	4.123176	1.364599
nav-to-center/lexicon_size_80	5.001934	1.544348
nav-to-center/temperature_0.1	3.405157	1.410848
nav-to-center/temperature_0.167	3.469046	1.487293
nav-to-center/temperature_0.278	3.396763	1.410126
nav-to-center/temperature_0.464	3.377160	1.394467
nav-to-center/temperature_0.774	3.777791	1.466642
nav-to-center/temperature_1.29	4.202502	1.447624
nav-to-center/temperature_10	8.121348	3.216181
nav-to-center/temperature_2.15	7.739814	2.773120
nav-to-center/temperature_3.59	8.433494	3.092856
nav-to-center/temperature_5.99	8.660965	3.394618
rlupus/21-player.run-0	6.956412	2.893892
rlupus/21-player.run-1	7.403071	3.206111
rlupus/21-player.run-2	7.039882	3.042730
rlupus/9-player.run-0	5.883233	2.803656
rlupus/9-player.run-1	5.925070	2.805487
rlupus/9-player.run-2	5.979073	2.888910
rlupus/9-player.run-3	5.865222	2.753987

Table 6