Datasheet for the Emergent Language Corpus Collection (ELCC)

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

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

DOI	https://doi.org/10.57967/hf/2533
Repository	https://huggingface.co/datasets/bboldt/elcc
Maintainer	Brendon Boldt
License	CC BY 4.0, MIT
Availability	Public
Host	Hugging Face Datasets
n corpora	73
Size on disk	700 MiB
Croissant URL	https://huggingface.co/datasets/bboldt/elcc/raw/main/croissant.json

Table 1: Quick Reference

1 Introduction

This document is the datasheet[1] for ELCC which is under submission at the Datasets and Benchmarks Track at NeurIPS 2024. Throughout the document, the AUTHORS refers to Brendon Boldt and David Mortensen as specified in the title block. The MAINTAINER refers to Brendon Boldt. All data in ELCC was derived either directly from the AUTHORS's work or from free and open source, publicly available codebases.

2 Motivation

For what purpose was the dataset created? ELCC was created for studying the linguistic and statistical properties of corpora generated by emergent communication systems (ECSs). ECSs are simulations of language evolution from scratch using neural-network based agents. ELCC is the first collection of emergent language corpora derived from a variety of ECSs.

Who created the dataset and on behalf of which entity? The AUTHORS created the dataset. Brendon Boldt is a PhD student at Carnegie Mellon University's Language Technologies Institute. David Mortensen is a research professor at the same department and university.

Who funded the creation of the dataset? The AUTHORS are funded by the department. There is no grant associated with ELCC.

3 Composition

What do the instances that comprise the dataset represent? Instances can be taken in two senses with respect to ELCC. First, each instance is a corpus generated by a particular ECS. Second, each instance within a corpus is an array of tokens derived from the utterances of agents with in that particular ECS.

How many instances are there total? There are 73 corpora generated from 8 distinct ECSs. The corpora altogether have 5.6×10^6 lines and 1.0×10^8 tokens. The median corpus size is 1.0×10^4 lines and 1.1×10^5 tokens.

Does the dataset contain all possible instances or is it a sample of instances from a larger set? No. There are theoretically infinitely many ECSs; in practice, not all existing ECSs in the literature have been included. Although ELCC is intended to be representative, there was no rigorous procedure in place to ensure that the particular corpora generated were truly representative of some particular "universal" set. Each ECS could, in theory, produce an infinitely number of utterances for a corpus, so a corpus is only a random sampling of utterances from the ECS.

What data does each instance consist of? Each corpus consists of an array of integer arrays. Each corpus additionally has metadata associated with describing basic statistical properties about the corpus as well as taxonomic features of the ECS which generated it.

Is there a label or target associated with each instance? No.

Is any information missing from individual instances? No.

[1] Gebru et al., 2021 (https://arxiv.org/abs/1803.09010)

Are relationships between individual instances made explicit? Some corpora are generated from different hyperparameter settings of the same ECS. This is information is present in the metadata associated with each corpus.

Are there recommended data splits? No.

Are there any errors, sources of noise, or redundancies in the dataset? It is possible that there are errors in the metadata about the corpora due to curation errors. Similarly, the corpora could have errors in their generation process due to bugs in the associated software.

Is the dataset self-contained, or does it link to or otherwise rely on external resources? The data itself is self contained. The code which generates the data references external repositories which are publicly available with free software licenses on websites such as GitHub. These repositories are either maintained by the MAINTAINER or will be made available otherwise should the original source become inaccessible.

Does the dataset contain data that might be considered confidential? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

4 Collection Process

How was the data associated with each instance acquired? Each corpus was generated by a free and open source implementation of the particular ECS. The individual ECSs come from various authors across the body of literature on emergent communication. Each corpus was generated by the MAINTAINER directly running the ECS.

What mechanisms or procedures were used to collect the data? See previous response.

If the dataset is a sample from a larger set, what was the sampling strategy? The AUTHORS aimed for a representative sample of ECSs from literature, although this was constrained by the availability of free and open source implementations as well as total time available for curation. The particular hyperparameter settings within each ECS were selected because they were either already present in the ECS (e.g., used in the original paper) or demonstrated a variation of some particular hyperparameter (e.g., vocabulary size).

Who was involved in the data collection process? The data was collected directly by the MAINTAINER.

Over what timeframe was the data collected? The original dataset was created from March 2024 to June 2024.

Were any ethical review processes conducted? No.

5 Preprocessing, Cleaning, and Labelling

Was any preprocessing, cleaning, or labelling of the data done? At the ECS level, labelling the taxonomic features of the ECSs (i.e., in the metadata) was done by the MAINTAINER. At the corpus level, multiple utterances from the same or different agents from a given episode were concatenated together to form a single line in the dataset. Otherwise the data in the corpora are taken directly from the ECSs.

Was the "raw" data saved in addition to the preprocessed, cleaned, and/or labelled data? Yes. In the aforementioned cases of concatenating multiple utterances, the unconcatenated (or "structured") versions are also included in the dataset.

Is the software that was used to preprocessed, clean, and/or label the data available? Yes, it is part of the same repository where the data is available.

6 Uses

Has the dataset been used for any tasks already? No.

Is there a repository that links to any or all papers or systems that use the dataset? No.

What (other) tasks could the dataset be used for? This data can be used for linguistics and machine learning research on emergent communication and/or synthetic data. The potential uses of this data coincide, more or less, with those of synthetic data mimicking natural language.

Is there anything about the composition of the dataset or the way it was collected and preprocessed, cleaned, and/or labelled that might impact future uses? The data in ELCC does not contain the full extent of emergent languages or ECSs. Any conclusions drawn from the dataset must be done so acknowledging that fact.

Are there tasks for which the dataset should not be used? No.

7 Distribution

Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created? The dataset is public.

How will the dataset be distributed? The dataset is publicly available as a HuggingFace dataset at `https://huggingface.co/datasets/bboldt/elcc`.

When will the dataset be distributed? The dataset was made publicly available June 5, 2024.

Will the dataset be distributed under a copyright or other intellectual property license, and/or under applicable terms of use? The code contributed by the MAINTAINER is licensed under the MIT license, and the data is licensed under the CC BY 4.0.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

8 Maintenance

Who will be supporting, hosting, maintaining the dataset? The MAINTAINER.

How can the owner, curator, and/or manager of the dataset be contacted? See the repository at `https://huggingface.co/datasets/bboldt/elcc`. Discussion can be opened on that repository; contact information for the MAINTAINER will be kept up-to-date in the repository.

Is there an erratum? No.

Will the dataset be updated? The dataset will updated to address any significant bugs. New ECSs and/or hyperparameters may be added in the future.

Will older versions of the dataset continue to be supported, hosted, and/or maintained? Yes. The Hugging Face repository will keep old versions available.

If others want to extend, augment, build on, and/or contribute to the dataset, is there a mechanism for them to do so? Yes. Others are free to fork the repo to make their own changes (as is permitted by the free and open source license). People working with ELCC are encouraged to open pull requests with any contributions they may produce in their research (e.g., bug fixes, new ECSs).