On the Relationship between Zipf's Law of Abbreviation and Interfering Noise in Emergent Languages

Ryo Ueda
The University of Tokyo
ryoryoueda@is.s.u-tokyo.ac.jp

Koki Washio
The University of Tokyo
kwashio@is.s.u-tokyo.ac.jp

Abstract

This paper studies whether emergent languages in a signaling game follow Zipf's law of abbreviation (ZLA), especially when the communication ability of agents is limited because of interfering noises. ZLA is a well-known tendency in human languages where the more frequently a word is used, the shorter it will be. Surprisingly, previous work demonstrated that emergent languages do not obey ZLA at all when neural agents play a signaling game. It also reported that a ZLA-like tendency appeared by adding an explicit penalty on word lengths, which can be considered some external factors in reality such as articulatory effort. We hypothesize, on the other hand, that there might be not only such external factors but also some internal factors related to cognitive abilities. We assume that it could be simulated by modeling the effect of noises on the agents' environment. In our experimental setup, the hidden states of the LSTM-based speaker and listener were added with Gaussian noise, while the channel was subject to discrete random replacement. Our results suggest that noise on a speaker is one of the factors for ZLA or at least causes emergent languages to approach ZLA, while noise on a listener and a channel is not.

1 Introduction

There has recently been a growing interest in simulating languages spontaneously emerging among artificial agents, by training them to solve some tasks requiring communications. A primary motivation in this area is to pursue the development of artificial intelligence that can interact or communicate with human beings (e.g., Havrylov and Titov, 2017; Lazaridou et al., 2017, 2018; Lee et al., 2018). In addition to this line of research, some studies have investigated the characteristics of emergent languages, mainly concerned with to what extent they are similar to human languages or what kind of factor forms language-like protocols (e.g., Kottur et al., 2017; Harding Graesser et al., 2019; Chaabouni et al., 2020; Kharitonov et al., 2020).

Chaabouni et al. (2019), for example, studied the relationship between emergent languages and Zipf's law of abbreviation (ZLA), which is a universal tendency in human languages, where frequent words tend to be shorter (Zipf, 1935; Kanwal et al., 2017). To see whether emergent languages follow ZLA, they performed experiments in which agents played a signaling game. Their results suggested that emergent languages have an opposite tendency against ZLA. In other words, more frequent inputs are encoded into longer messages. They also reported that by giving an additional penalty on message lengths (Eq. 6), the emergence of a ZLA-like tendency was observed.

Zipf (1935) hypothesized that ZLA comes about between two conflicting pressures: one for accuracy and the other for efficiency. In a paradigm with human subjects using a simple artificial language, Kanwal et al. (2017), for instance, introduced some external factors for simulating the competing pressures, namely, money reward for precise and quick communications. In emergent-language simulations, the explicit penalty on message lengths (Eq. 6) of Chaabouni et al. (2019) can also be considered an external factor for ZLA.

However, we speculate that there might be not only such external factors but also internal factors (or implicit penalties) related to the cognitive abilities of human beings such as memory. Inspired by some concepts in psychology, we hypothesize at first in the following way:

Hypothesis 1. *ZLA appears due to some internal factors from the cognitive abilities of human beings, as well as external factors. In other words, human*

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 60–70
August 5–6, 2021. ©2021 Association for Computational Linguistics

beings assign shorter codes to frequent words so that they can avoid difficulty in their internal processes as much as possible.

Some studies in psychology suggested that in human beings, there is an output buffer of some sort that temporarily reserves some words to be spoken (Baddeley et al., 1975; Baddeley, 2003; Meyer et al., 2003; Damian et al., 2010; Baddeley and Hitch, 2019). The output buffer might decay over time, be overwhelmed by incoming inputs one after another, or be exposed to other disturbances. Such pressures, we thought, could be factors to shorten frequent words.

But how should they be modeled in the simulations of language emergence? Since artificial agents in simulations are not humans but often (recurrent) neural networks, it is not trivial to define equivalent pressures for them. To adopt such pressures into a signaling game, we propose modeling them into *noise* that interferes with the states of agents. Although the potential factors described above might be the matter of a speaker in a signaling game, we also propose adding noise to a listener for comprehensive research. The listener's short-term memory might also be limited due to similar reasons as the speaker. Besides, we try adding noise to a channel that spans the speaker and the listener, referring to a noisy-channel model (Shannon, 1948). Although a noisy channel is not probably pressure for efficiency but for accuracy, the assumption that redundancy contributes to accuracy seems to think implicitly of a listener as capable enough of correcting errors while maintaining necessary information, which is not trivial for neural agents. Therefore it is worth a try.

By the modeling and for the comprehensiveness, hypothesis 1 is revised as follows:

Hypothesis 2. *ZLA appears due to some of the three types of noises: noise on a speaker, noise on a listener, and noise on a channel.*

In our experimental setup, speaker and listener agents are exposed to Gaussian noise since they have continuous vectors as their states. On the other hand, the channel is exposed to discrete random replacements, as messages passing through it have discrete variables.

Our experiments suggest that noise on a speaker is one factor for ZLA or at least causes emergent languages to be closer to ZLA, whereas noise on a listener and a channel is not in our signaling game. Rather, the noise on a channel strengthened redundancy.

Our analysis reveals the following things. First, when noise interferes with a speaker agent, noise accumulation can make it difficult to generate long consistent messages. Second, when noise interferes with a listener agent, on the other hand, noise accumulation does not affect the overall tendency crucially: even if the listener agent "forgets" the prefix of a message, the suffix is sufficient for communications. Third, noise on a channel can be thought of as a pressure for accuracy rather than efficiency, which is consistent with an information-theoretic point of view and Zipf's hypothesis.

2 Background

Chaabouni et al. (2019) studied whether emergent languages follow ZLA when neural agents play a signaling game. As we largely refer to, we review their setups, methods, and results in this section.

2.1 Signaling Game with a Power-law distribution

They extended a signaling game (Lewis, 1969) by making inputs be sampled from a power-law distribution. In the power-law distribution, the n-th most frequent input is sampled from a finite input space I at the probability $\propto 1/n$. Thus, if agents learned to assign frequent inputs to shorter messages, their communication protocol could be said to obey ZLA.

Let S and L be a speaker and a listener. Formally, the game procedure is as follows:

1. An input $i \in I$ is sampled from a power-law distribution. Let i_r be the r-th most frequent input. Then i_r is sampled at the probability $\propto r^{-1}$.

2. Given i, the speaker S generates a message m, i.e., $m = S(i)$. $m = x_1 \dots x_{|m|}$ is a string over an alphabet $A = \{a_1, \dots, a_{|A|-1}, \texttt{eos}\}$ s.t. $x_i \neq \texttt{eos}$ $(1 \leq i < |m|)$, $x_{|m|} = \texttt{eos}$, and $0 < |m| \leq \texttt{max\_len}$, where $|m|$ is the length of m and $\texttt{max\_len}$ is a hyperparameter. Note that $\texttt{eos} \in A$ stands for "end-of-sentence," and it is guaranteed to be attached to the end of each message[1].

3. Given m, the listener L generates an output, i.e., $o = L(m)$.

[1] One might think that eom (end-of-message) is better, but we follow the convention in the literature of neural language modeling.

4. The procedure is successful if $i = o$.

2.2 Training Method

Since players in a signaling game are neural networks, each input $i \in I$ is represented as a $|I|$-dimensional one-hot vector $\boldsymbol{i}$. Likewise, an output o is represented as a $|I|$-dimensional vector $\boldsymbol{o}$ s.t. $(\boldsymbol{o})_k > 0$ $(k = 1, \ldots, |I|)$ and $\sum_{k=1}^{|I|}(\boldsymbol{o})_k = 1$. Let $\mathcal{L}(\boldsymbol{i}, \boldsymbol{o}) = \mathcal{L}(\boldsymbol{i}, L(S(\boldsymbol{i}))$ be the cross-entropy error between $\boldsymbol{i}$ and $\boldsymbol{o} = L(S(\boldsymbol{i}))$:

$$\mathcal{L}(\boldsymbol{i}, \boldsymbol{o}) = -\sum_{k=1}^{|I|}(\boldsymbol{i})_k \log(\boldsymbol{o})_k, \tag{1}$$

where S is a speaker and L is a listener. Our purpose is to minimize its expectation $\mathbb{E}[\mathcal{L}]$, but the simple backpropagation algorithm is not applicable due to discrete messages $m = x_1 \ldots x_{|m|}$ sampled from a speaker. Chaabouni et al. (2019) used the following surrogate function, the gradient of which is an unbiased gradient estimator, with an auxiliary loss *entropy regularizer ER*:

$$\mathbb{E}[\mathcal{L}_S + \mathcal{L}_L + \mathrm{ER}] \tag{2}$$

$$\mathcal{L}_S = \mathrm{SG}(\mathcal{L}(\boldsymbol{i}, \boldsymbol{o}) - b) \sum_{t=1}^{|m|} \log P_{S,t}(x_t) \tag{3}$$

$$\mathcal{L}_L = \mathcal{L}(\boldsymbol{i}, \boldsymbol{o}) \tag{4}$$

$$\mathrm{ER} = -\frac{\lambda_{\mathcal{H}}}{N} \sum_{t=1}^{N} \mathcal{H}(P_{S,t}), \tag{5}$$

where b is a mean baseline added to reduce the estimate variance, $\mathrm{SG}(\cdot)$ denotes the stop-gradient operation[2], $P_{S,t}$ is the speaker's output layer at time step t defining a categorical distribution over an alphabet A, $P_{S,t}(x_t)$ is the probability of $x_t \in A$ being sampled at time step t, and $\mathcal{H}(\cdot)$ is the entropy function. Eq. 3 and Eq. 4 are derived by the approach of Schulman et al. (2015), which can be seen as the combination of REINFORCE-like method (Williams, 1992) and standard backpropagation. ER (Eq. 5) is added to encourage the exploration during training (Williams and Peng, 1991).

2.3 Anti-ZLA Emergent Languages

Chaabouni et al. (2019) reported, somewhat surprisingly, that the communication protocols had a clear anti-ZLA tendency when agents play a signaling game described in section 2.1. They also reported that a ZLA-like tendency appeared when they additionally imposed an *artificial length pressure* on messages:

$$\mathcal{L}'(\boldsymbol{i}, L(m), m) = \mathcal{L}(\boldsymbol{i}, L(m)) + \alpha \times |m|, \tag{6}$$

where m is a message, $|\cdot|$ denotes length, and $\alpha \geq 0$ is a hyperparameter.

Rita et al. (2020) took a quite similar approach and observed the emergence of ZLA. As well as imposing a length pressure on a speaker agent, they re-designed the architecture of a listener agent so that the listener would be *impatient* to recover $\boldsymbol{i}$ as soon as possible.

Note that both the length pressure (Eq. 6) and the architecture re-design in Rita et al. (2020) can be regarded as somewhat explicit losses, whereas we try to impose an implicit pressure on agents.

3 Setup

3.1 Game with Noise

For a game, we take almost the same design as Chaabouni et al. (2019), which was introduced in section 2.1. We additionally introduce a *channel* C over which messages move from speaker to listener: A listener L obtains a message $\tilde{m} = C(m)$ through a channel C, instead of receiving directly $m = S(i)$ from a speaker. Also, there are several differences in hyperparameter settings.

3.2 Architectures

As speaker and listener agents have continuous vectors as their states, they are added with continuous noise. For simplicity, we choose a Gaussian noise sampled at each time step with replacement. Channels, on the other hand, are exposed to discrete noise, since they convey discrete symbols. We take a random replacement operation for the channel noise.

3.2.1 Speaker and Listener

The architectures of speaker and listener agents are based on a single-layer LSTM, following Chaabouni et al. (2019).

At training time, we add Gaussian noise to the cell states of the LSTM of each agent[3]. Formally,

[2] When we write $SG(x)$ instead of bare x, we regard x as a constant with respect to any parameters.

[3] We also tried simply shrinking the size of the agents' hidden layers to restrict their capacity, but it made it difficult to train the agents successfully. We leave it for future work

for $t > 0$,

$$(\boldsymbol{h}_{t+1}, \boldsymbol{c}_{t+1}) = \mathrm{LSTM}(\boldsymbol{x}_{t+1}, (\boldsymbol{h}_t, \hat{\boldsymbol{c}}_t)) \quad (7)$$

$$\hat{\boldsymbol{c}}_t = \boldsymbol{c}_t + \boldsymbol{\epsilon}_t \quad (8)$$

$$\boldsymbol{\epsilon}_t \sim \mathcal{N}(\cdot \mid \mathbf{0}, \sigma^2 E) \quad (9)$$

where $\sigma > 0$ is a standard deviation (SD), E is the identity matrix, $\mathcal{N}(\cdot \mid \mathbf{0}, \sigma^2 E)$ is a Gaussian distribution with a mean vector $\mathbf{0}$ and a variance-covariance matrix $\sigma^2 E$, and $\boldsymbol{\epsilon}_t$ is a sampled value from $\mathcal{N}(\cdot \mid \mathbf{0}, \sigma^2 E)$ at time step t. We denote by σ_S, σ_L the SDs for the speaker and listener architecture respectively.

At test time, we do not add noise for deterministic evaluation.

3.2.2 Channel

At training time, we think of a channel as being exposed to some noise so that the messages can be degraded during transportation. Such degradation is modeled as replacement: each symbol in a message is probabilistically replaced with another one. Note that each message is attached with `eos`, which is exceptionally protected from the replacement, since the effect of the insertion or deletion of `eos` is too strong for our purpose.

Formally, let A be an alphabet, $m = a_1 \dots a_n$ be an original message generated by the speaker, and $\tilde{m} = \tilde{a}_1 \dots \tilde{a}_n$ be transformed one. Then the probability distribution over $\tilde{a}_i \neq \texttt{eos}$ given $a_i \neq \texttt{eos}$ $(i = 1, \dots, n-1)$ is as follows:

$$p(\tilde{a}_i \mid a_i) = \begin{cases} 1 - \pi_C & (a_i = \tilde{a}_i) \\ \dfrac{\pi_C}{|A \backslash \{a_i, \texttt{eos}\}|} & (a_i \neq \tilde{a}_i) \end{cases}, \quad (10)$$

where π_C is a hyperparameter s.t. $0 \leq \pi_C \leq 1$. Let us call π_C a *channel replacement probability*.

At test time, the channel is free from noise so that we can perform deterministic examinations.

3.3 Optimization

3.3.1 Design and Estimation of Loss Function

We use almost the same loss function as Eq. 2. We modify ER (Eq. 5) into *Decayed Entropy Regularizer (DER)* and we define an additional auxiliary loss *Soft Max Length (SML)* in the following sections. Both DER and SML are introduced to prevent messages from being unnaturally long. Note that they themselves are not factors for ZLA in our assumption.

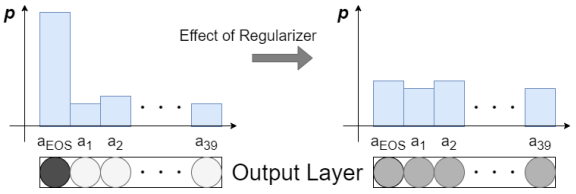

Figure 1: Illustration of the effect of the entropy regularizer

3.3.2 Decayed Entropy Regularizer

Chaabouni et al. (2019) used ER (Eq. 5) to encourage the exploration. However, ER might have an unexpected side-effect: They could lead messages to be unnecessarily long. We give an intuitive explanation as shown in Figure 1. Suppose that a speaker agent has learned a message pattern $m = x_1 \dots x_{|m|}$ for an input i. By the definition of the message, $x_{|m|} = \texttt{eos}$, indicating that the probability that `eos` is sampled is relatively higher at time step $|m|$. Then, the speaker's output layer $P_{S,|m|}$ at time step $|m|$ is updated so that the entropy $\mathcal{H}(P_{|m|})$ will be larger. It means that the probability of `eos` being sampled becomes lower, which might lead the message to be longer. Such an effect can cause an undesirable bias in emergent languages. Thus, we modify ER into *Decayed Entropy Regularizer (DER)* as:

$$\mathrm{DER} = -\frac{\lambda_{\mathcal{H}}}{Z} \sum_{t=1}^{N} \mathcal{H}(P_t) \times \rho_{\mathcal{H}}^{t-1}, \quad (11)$$

$$Z = \sum_{t=1}^{N} \rho_{\mathcal{H}}^{t-1}, \quad (12)$$

where $\rho_{\mathcal{H}}$ is a hyperparameter s.t. $0 < \rho_{\mathcal{H}} \leq 1$. DER is a weighted mean that puts a higher priority on the entropy at earlier time steps but lower on those at later. Therefore, it is expected to cancel the unnecessary effect of hindering `eos` emission at later time steps.

3.3.3 Soft Max Length

Each message m is generated by sampling a symbol x_t at each time step t and concatenating them until either `eos` is sampled (self-termination) or the time step reaches `max_len` $- 1$ (forced termination). In the forced termination case, `eos` is attached to the end of the sequence. However, this generating procedure may cause a speaker agent to fail to learn to emit `eos` for some inputs, since message lengths are bounded regardless of the `eos` emission. To handle this problem, we introduce an

additional auxiliary loss *Soft Max Length (SML)* defined as:

$$\text{SML} = \lambda_{sml} \max(0, |m| - \texttt{eff\_max\_len}), \tag{13}$$

where m is a message, $|\cdot|$ denotes length, λ_{sml} is the coefficient of this term, and `eff_max_len` is a hyperparameter s.t. $0 \leq$ `eff_max_len` $\leq$ `max_len`.

3.3.4 Training and Implementation

We follow Chaabouni et al. (2019) on the rest of the training method: Agents are trained for 2500 episodes, each of which contains 100 mini-batches. Each mini-batches are made of 5120 inputs sampled from the power-law distribution with replacement. When the accuracy at test time reaches 0.99 or more, the training stops early. Note that we do not add any noise at test time.

The game and the training are implemented using the EGG toolkit (Kharitonov et al., 2019)[4].

3.4 Evaluating Communicative Effectiveness

As Lowe et al. (2019) pointed out, emergent communications have to be carefully examined in terms of effectiveness: even if something like communication emerges, agents might act without referring to signals from others. Since message lengths can vary in our signaling game, it is doubtful that every single symbol in a message conveys essential information. For example, it is not trivial whether `eos` is really end-of-sentence, since agents can use other symbols as "punctuations" or meaningless "blanks." The effective position of beginning-of-sentence is not trivial, either. Thus, apparent message lengths may differ from actual ones.

To evaluate effectiveness, we introduce *position-wise symbol effectiveness* and then *head/intermediate/tail effectiveness* to cover a weak point in the former.

Position-wise Symbol Effectiveness

First, to evaluate how informative symbols are distributed across positions, we introduce *position-wise symbol effectiveness*, which is a quite similar notion to *positional encoding* in Rita et al. (2020). Suppose a symbol x_k in a message $m = x_1 \dots x_k \dots x_{|m|}$ is informative enough. Then, a listener L is expected to fail to recover an input i correctly if x_k is replaced with another symbol y, i.e., $i \neq L(x_1 \dots y \dots x_{|m|})$. Based on this intuition, the symbol effectiveness $e(m, k)$ at position $k \in \{1, \dots, \texttt{max\_len}\}$ in a message $m = x_1 \dots x_{|m|}$ is defined as follows:

$$e(m,k) = \begin{cases} \dfrac{1}{|A'|} \displaystyle\sum_{a \in A'} \mathbb{1}_{i \neq L(m[x_k := a])} & (k < |m|) \\ 0 & (k \geq |m|) \end{cases} \tag{14}$$

$$A' = A \backslash \{x_k, \texttt{eos}\}, \tag{15}$$

where A is an alphabet, $m[x_k := a]$ denotes $x_1 \dots x_{k-1} a x_{k+1} \dots x_{|m|}$, and $\mathbb{1}_\phi$ is defined as

$$\mathbb{1}_\phi = \begin{cases} 1 & (\phi \text{ is true.}) \\ 0 & (\phi \text{ is false.}) \end{cases}. \tag{16}$$

By definition, $0 \leq e(m,k) \leq 1$. Low $e(m,k)$ means that symbol x_k is redundant, since the listener L can recover i from most of $m[x_k := a]$ ($a \in A'$). Otherwise, x_k is considered necessary for successful communications. Note that $\texttt{eos} = x_{|m|}$ is prevented from being replaced.

The value of $e(m,k)$ (Eq. 14) may vary depending on messages and speaker agents. That would make it difficult to perform straightforward evaluations for position-wise symbol effectiveness. To handle this problem, we also define $\overline{e}_k$, mean $e(m,k)$ across messages and across speaker agents. Formally, let $\mathcal{S} = \{S_1, \dots, S_{|\mathcal{S}|}\}$ be a set of $|\mathcal{S}|$ speaker agents trained with different random seeds. Then $\overline{e}_k$ is defined as:

$$\overline{e}_k = \frac{1}{|\mathcal{S}||I|} \sum_{S \in \mathcal{S}} \sum_{i \in I} e(S(i), k). \tag{17}$$

Head, Intermediate, and Tail Effectiveness

One may be interested in detecting whether the effectiveness is concentrated in the prefixes, infixes, or suffixes of messages. However, $\overline{e}_k$ (Eq. 17) do not seem good for this purpose: Since message lengths can vary, the effectiveness of infixes and suffixes can scatter across $\overline{e}_k$. Thus, we additionally introduce *head effectiveness* $\overline{e}_{head}$, *intermediate effectiveness* $\overline{e}_{med}$, and *tail effectiveness* $\overline{e}_{tail}$. Intuitively, $\overline{e}_{head}$ is mean effectiveness across the heads of messages (i.e., x_1 in $m = x_1 \dots x_{|m|}$) and across speaker agents. Similarly, $\overline{e}_{med}$ (resp. $\overline{e}_{tail}$) is mean effectiveness across the intermediate positions (resp. tails) of messages and across speaker

[4]The code for the EGG toolkit is found at https://github.com/facebookresearch/EGG. Our code is available at https://github.com/wedddy0707/noisyEGG.git.

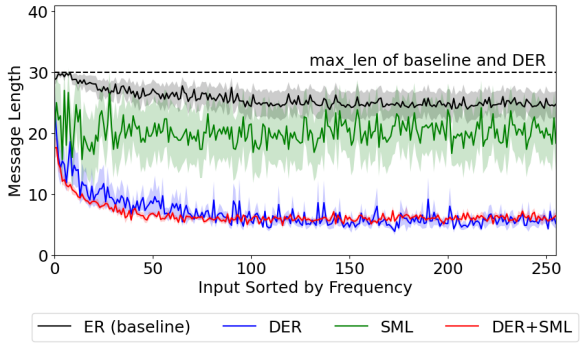

Figure 2: Mean message lengths across successful runs as a function of inputs sorted by frequency, when ER, DER, SML, and DER+SML are used respectively. The shaded areas represent one stanrd error of mean (SEM).

	# successful runs
ER (baseline)	16
DER	7
SML	6
DER+SML	11

Table 1: The number of successful runs out of 16.

agents. Formally, let $\mathcal{S} = \{S_1, \ldots, S_{|\mathcal{S}|}\}$ be as above. Then $\overline{e}_{head}$, $\overline{e}_{med}$, and $\overline{e}_{tail}$ are defined as follows:

$$\overline{e}_{head} = \frac{1}{|\mathcal{S}||I|} \sum_{S \in \mathcal{S}} \sum_{i \in I} e(S(i), 1) = \overline{e}_1 \quad (18)$$

$$\overline{e}_{med} = \frac{1}{|\mathcal{S}||I|} \sum_{S \in \mathcal{S}} \sum_{i \in I} e\left(S(i), \left\lfloor \frac{|S(i)|}{2} \right\rfloor\right) \quad (19)$$

$$\overline{e}_{tail} = \frac{1}{|\mathcal{S}||I|} \sum_{S \in \mathcal{S}} \sum_{i \in I} e(S(i), |S(i)| - 1), \quad (20)$$

where $\lfloor \cdot \rfloor$ is a floor function.

4 Experiments

4.1 Hyperparameter Setting

In all our experiments, the size $|I|$ of an input space was set to 256, the size $|A|$ of an alphabet was 40, the size of hidden layers was 100 for both agents, and the entropy regularizer coefficient $\lambda_{\mathcal{H}}$ was 1. The hyperparameters σ_S, σ_L, and π_C for noise varied through sections.

We define a training run ending with an accuracy higher than 0.99 as a *successful* run.

4.2 Effects of DER and SML

Before conducting the main experiments, we show the effect of DER (Eq. 11) and SML (Eq. 13). For a baseline model, we used the existing entropy regularizer ER (Eq. 5), setting $\lambda_{\mathcal{H}} = 1$ and $\texttt{max\_len} = 30$. For a model with DER, $(\lambda_{\mathcal{H}}, \rho_{\mathcal{H}}) = (1, 1/2)$. For a model with SML (and ER), $\lambda_{\mathcal{H}} = 1$ and $(\texttt{max\_len}, \texttt{eff\_max\_len}) = (40, 30)$. For a model with DER+SML, $(\lambda_{\mathcal{H}}, \rho_{\mathcal{H}}) = (1, 1/2)$ and $(\texttt{max\_len}, \texttt{eff\_max\_len}) = (40, 30)$.

setting	Spearman ρ
no noise	0.327 $(p = 5.9 \times 10^{-71})$
noise $\sigma_S = 1/4$	0.113 $(p = 1.5 \times 10^{-6})$
noise $\sigma_S = 1/2$	0.109 $(p = 6.9 \times 10^{-7})$
noise $\sigma_S = 1$	0.008 $(p = 7.7 \times 10^{-1})$
noise $\sigma_L = 1/4$	0.273 $(p = 6.6 \times 10^{-32})$
noise $\sigma_L = 1/2$	0.280 $(p = 5.9 \times 10^{-20})$
noise $\sigma_L = 1$	0.268 $(p = 1.4 \times 10^{-22})$
noise $\pi_C = 0.01$	0.261 $(p = 3.3 \times 10^{-37})$
noise $\pi_C = 0.05$	0.236 $(p = 6.3 \times 10^{-21})$
noise $\pi_C = 0.1$	0.249 $(p = 8.6 \times 10^{-27})$

Table 2: Spearman correlations between input frequency ranks and message length ranks in successful runs in various noise conditions.

To see the overall tendency, we show the mean message lengths across successful runs for each model in Figure 2. The mean lengths are longer when ER is used. In particular, the ones of the baseline model are near $\texttt{max\_len} = 30$. On the other hand, the mean lengths are shorter when DER is used. That suggests that DER prevents messages from being unnecessarily longer.

To check the effects on learning, in addition, Table 1 shows the number of successful runs out of 16 for each model. Although apparent tendencies in Figure 2 are similar between the DER and DER+SML model, Table 1 suggests that it is easier to learn with the DER+SML model which has 5 more successful runs than the SML model.

4.3 Effects of Noise

In this section, we show the influence of noise on a speaker, listener, and channel. We used the DER+SML model with the same hyperparameters as in the previous section. We examined the effect of each noise by varying σ_S, σ_L, and π_C. Note that σ_S is the standard deviation of noise on a speaker, σ_L is the one on a speaker, and π_C is the channel replacement probability.

4.3.1 Noise on a Speaker

To examine the effect of noise on a speaker, $(\sigma_S, \sigma_L, \pi_C)$ was set to $(1/4, 0, 0, 0)$, $(1/2, 0, 0)$,

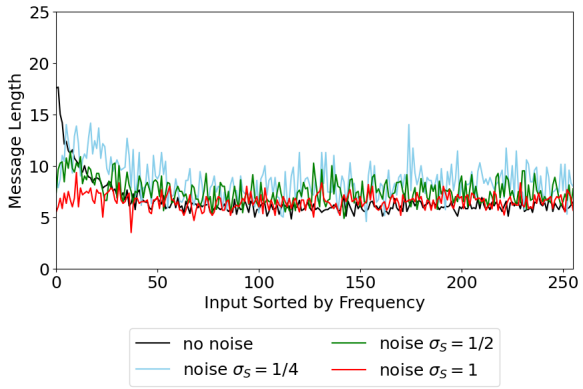

Figure 3: Mean message lengths across successful runs as a function of inputs sorted by frequency, when $(\sigma_S, \sigma_L, \pi_C) = (0, 0, 0)$, $(1/4, 0, 0)$, $(1/2, 0, 0)$, and $(1, 0, 0)$ respectively.

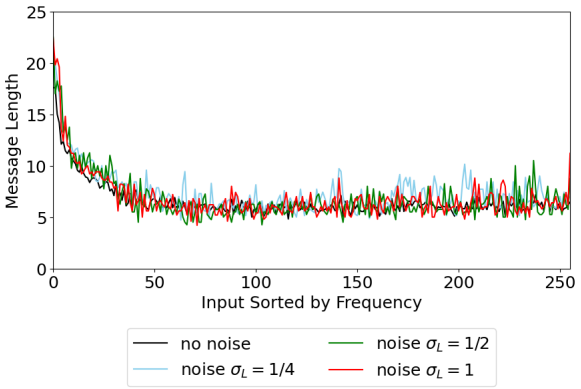

Figure 5: Mean message lengths across successful runs as a function of inputs sorted by frequency, when $(\sigma_S, \sigma_L, \pi_C) = (0, 0, 0)$, $(0, 1/4, 0)$, $(0, 1/2, 0)$, and $(0, 1, 0)$ respectively.

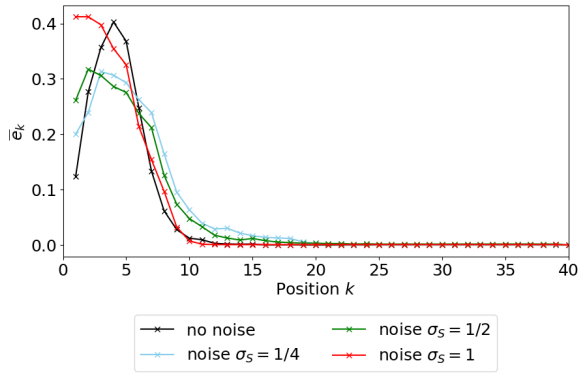

Figure 4: $\overline{e}_k$ in successful runs, when $(\sigma_S, \sigma_L, \pi_C) = (0, 0, 0)$, $(1/4, 0, 0)$, $(1/2, 0, 0)$, and $(1, 0, 0)$ respectively.

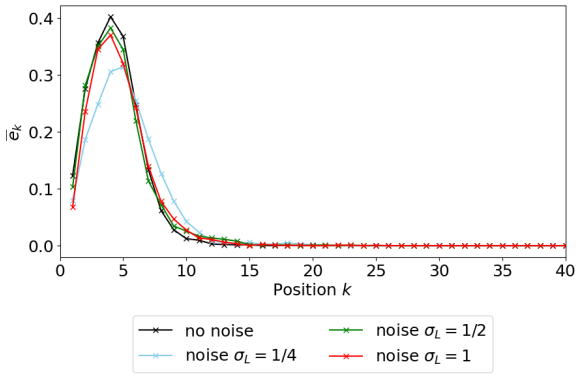

Figure 6: $\overline{e}_k$ in successful runs, when $(\sigma_S, \sigma_L, \pi_C) = (0, 0, 0)$, $(0, 1/4, 0)$, $(0, 1/2, 0)$, and $(0, 1, 0)$ respectively.

and $(1, 0, 0)$. 7 out of 16, 8 out of 16, and 6 out of 32 runs were successful for each setting.

To see the overall tendency, we show mean message lengths for each model in Figure 3[5]. The tendency shifts from anti-ZLA to the one between ZLA and anti-ZLA as σ_S gets bigger.

In addition, we show Spearman correlations between input frequency ranks and message length ranks in Table 2. Intuitively, $\rho < 0$ implies ZLA and $\rho > 0$ implies anti-ZLA. According to Table 2, ρ gets smaller as σ_S gets bigger, which is consistent with the observation in Figure 3.

To check the symbol effectiveness, we show $\overline{e}_k$ (Eq. 17) in Figure 4. Judging from Figure 4, the effectiveness at an earlier position becomes higher as σ_S gets bigger. We also show $\overline{e}_{head}$, $\overline{e}_{med}$, and $\overline{e}_{tail}$ (Eq. 18, Eq. 19, and Eq. 20) in Figure 9. In Figure 9, the bigger σ_S is, the higher $\overline{e}_{head}$ and $\overline{e}_{med}$ are, indicating that the former halves of messages become more informative by the effect of noise on a speaker.

These results suggest that noise on a speaker is a factor for ZLA, or at least causes message lengths to be closer to ZLA. One possible reason is that noise accumulation over time made it difficult for a speaker agent to generate long consistent messages.

4.3.2 Noise on a Listener

Next, to investigate the effect of noise on a listener, $(\sigma_S, \sigma_L, \pi_C)$ was set to $(0, 1/4, 0)$, $(0, 1/2, 0)$, and $(0, 1, 0)$. 7 out of 16, 4 out of 32, and 5 out of 16 runs were successful for each setting.

To see the overall tendency, mean message lengths are shown in Figure 5. The apparent tendencies are quite similar among all the settings

[5]There are some messages of length `max_len` =40 while other messages are much shorter. We excluded the former in Figure 3 because otherwise the mean lines would have unnatural peaks and impair readability. As a result, 4 out of 1792, 30 out of 2048, and 7 out of 1526 data points were removed for $\sigma_S = 1/4$, $1/2$, and 1 respectively.

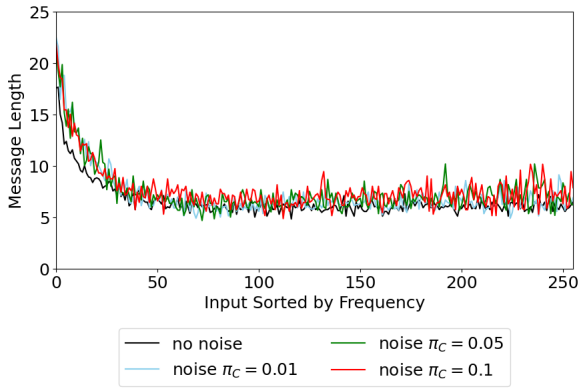

Figure 7: Mean message lengths across successful runs as a function of inputs sorted by frequency, when $(\sigma_S, \sigma_L, \pi_C) = (0, 0, 0)$, $(0, 0, 0.01)$, $(0, 0, 0.05)$, and $(0, 0, 0.1)$ respectively.

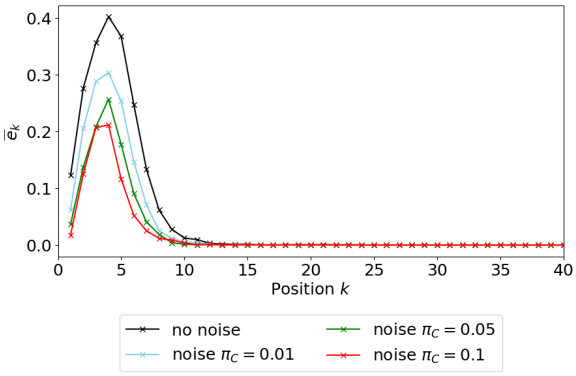

Figure 8: $meaneff_k$ in successful runs, when $(\sigma_S, \sigma_L, \pi_C) = (0, 0, 0)$, $(0, 0, 0.01)$, $(0, 0, 0.05)$, and $(0, 0, 0.1)$ respectively.

including 'no noise,' showing clear anti-ZLA tendencies. Spearman correlations in Table 2 also suggest anti-ZLA tendencies.

To check the symbol effectiveness, we show $\overline{e}_k$ (Eq. 17) in Figure 6. In Figure 6, $\overline{e}_k$ for $\sigma_L > 0$ shows similar tendencies to those for 'no noise,' although the peak of $\overline{e}_k$ for $\sigma_L = 1/2$ is lower than the other results.. $\overline{e}_{head}$, $\overline{e}_{med}$, and $\overline{e}_{tail}$ (Eq. 18, Eq. 19, and Eq. 20) are shown in Figure 9. According to Figure 9, $\overline{e}_{head}$ for $\sigma_L > 0$ tends to be smaller than the one for 'no noise,' but the overall tendencies seem similar (e.g., $\overline{e}_{head} < \overline{e}_{med} < \overline{e}_{tail}$).

These results suggest that noise on a listener is not a crucial factor for changing a tendency in emergent languages. The listener's short-term memory is thought to have been limited due to noise accumulation over time, as $\overline{e}_{head}$ got smaller. However, even if there was no noise, informative symbols tended to be located in the latter half of messages, i.e, $\overline{e}_{head} < \overline{e}_{med} < \overline{e}_{tail}$, which is one possible reason why noise on a listener did not crucially affect the overall tendency.

4.3.3 Noise on a Channel

Finally, to check the effect of noise on a channel, $(\sigma_S, \sigma_L, \pi_C)$ was set to $(0, 0, 0.01)$, $(0, 0, 0.05)$, and $(0, 0, 0.1)$. 9 out of 16, 6 out of 32, and 7 out of 32 runs were successful for each setting.

To see the overall tendency, mean message lengths are shown in Figure 7. The apparent results for $\pi_C > 0$ are similar to the one for 'no noise,' showing clear anti-ZLA tendencies. Spearman correlations in Table 2 also suggest anti-ZLA tendencies.

To check the symbol effectiveness, we show $\overline{e}_k$ (Eq. 17) in Figure 8. In Figure 8, $\overline{e}_k$ becomes lower entirely as π_C gets bigger. $\overline{e}_{head}$, $\overline{e}_{med}$, and $\overline{e}_{tail}$ (Eq. 18, Eq. 19, and Eq. 20) are shown in Figure 9. In Figure 9, $\overline{e}_{head}$, $\overline{e}_{med}$, and $\overline{e}_{tail}$ become lower as π_C gets bigger. Remember that low $e(m, k)$ (Eq. 14) means that the symbol at position k in m is redundant. Thus, lower $\overline{e}_k$, $\overline{e}_{head}$, $\overline{e}_{med}$, and $\overline{e}_{tail}$ indicate that symbols are redundant on the whole.

These results suggest that redundancy was facilitated due to the noise on a channel. It is consistent with Zipf's hypothesis and a noisy-channel model.

5 Discussion

Our experiments suggest that noise on a speaker is a factor for ZLA, while noise on a listener and a channel is not in our signaling game.

One possible reason for the noise on a speaker is that noise accumulation matters as time goes. At each trial, the speaker agent gets an input $\boldsymbol{i}$ and transforms it into an initial hidden state $\boldsymbol{h}_0$. The hidden states need to maintain the input $\boldsymbol{i}$ in some way for emitting consistent symbols. But noise accumulates over time and is harmful to their memory, which may cause frequent messages to be shorter. However, the result per se shows a neutral tendency between ZLA and anti-ZLA. Our implicit length pressure might not have been strong enough, or there might have been some problems with the agents' architectures.

Noise on a listener is not a crucial factor for ZLA in our setting. Judging from symbol effectiveness, the latter halves of messages tend to be more informative than the former when noise interferes with the listener. It means that the listener could "forget" the former halves of messages. In

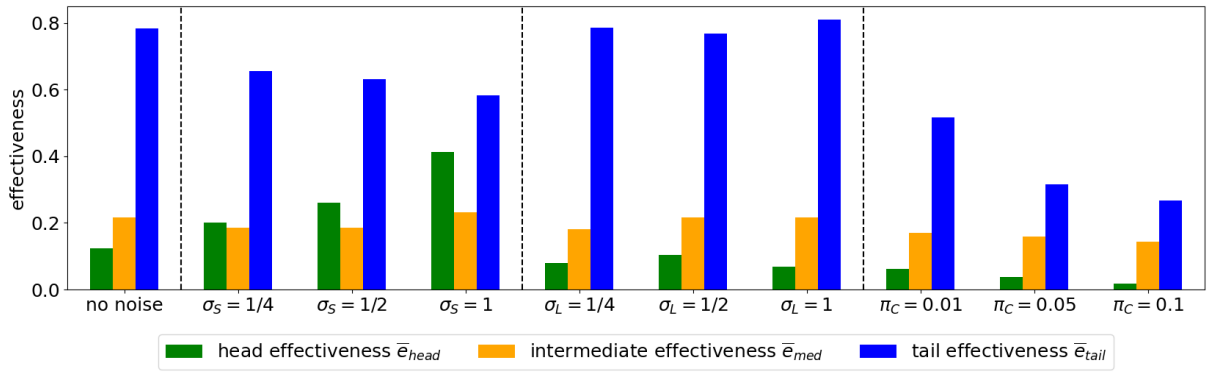

Figure 9: $\overline{e}_{head}$, $\overline{e}_{med}$, and $\overline{e}_{tail}$ in successful runs under various noise conditions.

the first place, however, the former halves are less informative even if there is no noise. That may be why noise on a listener did not affect the overall tendency. Noise on a channel seems to facilitate the redundancy of messages, which is consistent with Zipf's hypothesis and a noisy-channel model.

To help agents with learning, we used the two auxiliary loss DER (Eq. 11) and SML (Eq. 13) which are somewhat artificial. In particular, the usage of SML conflicts a bit with our original goal to give rise to ZLA by an implicit penalty, as SML is similar to an artificial length pressure (Eq. 6).

6 Conclusion

In this paper, we simulated the emergence of language and checked whether the emergent languages follow Zipf's law of abbreviation (ZLA). Inspired by some psychological concepts, we proposed exposing architectures to some noise during training. Our experiments were conducted under several noise conditions. The results suggested that noise on a speaker agent is one factor for ZLA, whereas neither noise on a listener nor noise on a channel is in our signaling game.

Our main contribution is to propose a potential factor for ZLA instead of an external length pressure and to demonstrate that noise imposing internal difficulty on a speaker agent may cause ZLA.

However, there are several problems and limitations in addition to what is discussed in section 5. First, we could not try the combination of noises. One might be interested in combining the noises on a speaker, listener, and channel, but we failed to train agents stably under such conditions. It is simply because it became much more difficult for agents to learn under several noises.

Second, our signaling game did not contain any *contexts*. As an input space was no more complex than having the order by frequency, emergent languages could only have a unigram-like structure. However, according to Piantadosi et al. (2011), word predictability considering contexts is a better predictor of word length than unigram probabilities. From a more realistic point of view, therefore, contexts should be considered in some ways. Moreover, if agents are forced to remember contexts, noise on a listener may also be a factor for ZLA, making the listener *impatient*.

We leave these issues for future work.

Acknowledgment

We would like to thank Professor Yusuke Miyao for supervising our research, Jason Naradowsky for fruitful discussions and proofreading, and the anonymous reviewers for helpful suggestions. The first author would also like to thank his colleagues Taiga Ishii and Hiroaki Mizuno as they have encouraged each other in their senior theses.

References

Alan D. Baddeley. 2003. Working memory and language: an overview. *Journal of Communication Disorders*, 36(3):189 – 208.

Alan D. Baddeley and Graham J. Hitch. 2019. The phonological loop as a buffer store: An update. *Cortex*, 112:91 – 106.

Alan D. Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14(6):575 – 589.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 4427–4442, Online. Association for Computational Linguistics.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. Anti-efficient encoding in emergent communication. In *Advances in Neural Information Processing Systems*, volume 32, pages 6293–6303. Curran Associates, Inc.

Markus Damian, Jeff Bowers, Hans Stadthagen-Gonzalez, and Katharina Spalek. 2010. Does word length affect speech onset latencies when producing single words? *Journal of experimental psychology. Learning, memory, and cognition*, 36:892–905.

Laura Harding Graesser, Kyunghyun Cho, and Douwe Kiela. 2019. Emergent linguistic phenomena in multi-agent communication games. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3700–3710, Hong Kong, China. Association for Computational Linguistics.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2149–2159.

Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 2017. Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45 – 52.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. EGG: a toolkit for research on emergence of lanGuage in games. In *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*, pages 55–60, Hong Kong, China. Association for Computational Linguistics.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2020. Entropy minimization in emergent languages. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5220–5230, Virtual. PMLR.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, Copenhagen, Denmark. Association for Computational Linguistics.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. 2018. Emergent translation in multi-agent communication. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

David K. Lewis. 1969. *Convention: A Philosophical Study*. Wiley-Blackwell.

Ryan Lowe, Jakob N. Foerster, Y-Lan Boureau, Joelle Pineau, and Yann N. Dauphin. 2019. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 693–701. International Foundation for Autonomous Agents and Multiagent Systems.

Antje S Meyer, Ardi Roelofs, and Willem J.M Levelt. 2003. Word length effects in object naming: The role of a response criterion. *Journal of Memory and Language*, 48(1):131 – 147.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Mathieu Rita, Rahma Chaabouni, and Emmanuel Dupoux. 2020. "LazImpa": Lazy and impatient neural agents learn to communicate efficiently. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 335–343, Online. Association for Computational Linguistics.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. 2015. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, volume 28, pages 3528–3536. Curran Associates, Inc.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423.

R. J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Ronald J. Williams and Jing Peng. 1991. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3:241–268.

George Kingsley Zipf. 1935. *The Psychobiology of Language*. Houghton-Mifflin, New York, NY, USA.