Thanks for the very constructive feedback. Due to lack of space, we only address here the major issues that were raised. We will however incorporate all feedback in our paper revision.

Power-law distribution of input referents (R1/R3). We agree with the reviewers that our assumption that words in natural language are power law-distributed because their referents in the world are is unwarranted. A more careful characterization for our setup is that the inputs to the Speaker represent abstract word types (which are definitely power law-distributed in languages); the task of the Speaker agent is to map these abstract types to phonological/orthographic forms and vice versa for the Listener agent. This brings our setup closer to the case of natural language; we will rephrase this in the introduction and discussion accordingly.

Uniform input distribution (R1/R3). Agents' messages are very long also when the input distribution is uniform, see Fig 1 (to be included in Supplementary with more settings, that follow the same pattern). Their average length is significantly larger than MT messages with uniform inputs (t-test, $p < 10^{-9}$).

Quantitative support for "anti-efficiency" claim (R1). Instead of running correlations which make assumptions about the underlying distribution, we have run the randomization test of Ferrer-i-Cancho et al. (CogSciJ 2013). We note $E = \sum_{i=1}^{1000} p_i \times l_i$ the mean length of messages, where p_i is the probability of the type i and l_i is the length of the corresponding message. A language that respects ZLA is characterized by a small E (optimal coding, OC, is associated with $min(E)$). Under H_0, the mean length of the encoding coincides with the mean length of a random permutation of messages across types. Also, we adopt Ferrer-i-Cancho et al. (CogSciJ 2013) definition of "left p-value" and "right p-value". If left p-value$\leq$ 0.005, the studied encoding is *significantly small* (characterized by significantly smaller E than random permutations), if right p-value$\leq$ 0.005, it is *significantly large*, corresponding to our notion of anti-efficiency. We observe in Table 1 (to be included in Supplementary with more settings, that confirm the same pattern) that H_0 is not rejected only for MT, which, as we mentioned in the paper, approaches a random length distribution for large a. OC, natural languages, and emergent language *with* Speaker-length regularization are significantly more efficient than chance. Importantly, the Emergent language results confirm LSTMs' natural preference for long messages (E approaching `max_len`) and *significant* anti-efficiency (right p-value$\approx$ 0).

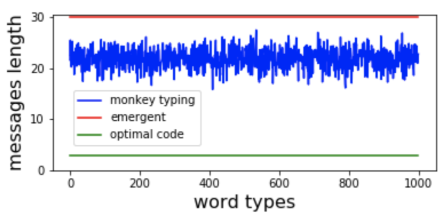

Figure 1: Mean message length per word type across successful runs, `max_len`=30, a=40. Word types are uniformly distributed.

code	E	left p-value	right p-value
OC	2.29	< 0.005	1
MT	21.30	0.81	0.18
Emergent	29.40	1	< 0.005
Regularized (α=0.5)	7.22	< 0.005	1
English	3.68	< 0.005	1
Arabic	3.14	< 0.005	1

Table 1: Randomization test results for `max_len`=30, a=40. OC: Optimal Coding, MT: Monkey Typing. To be comparable with previous studies, we use the same parameters as in Ferrer-i-Cancho et al. (CogSciJ 2013).

Specific points

***R1**: There are a couple of cases where numbers get averaged [...], and I'm unclear about what's being averaged.*

Figure 2: average length of all rank-i messages across successful runs. Figure 3: average pairwise distance across all considered non-trained Listeners. We will clarify accordingly in the paper.

***R2**: It is interesting to speculate whether this is caused by a peculiarity in LSTM dynamics, and whether encoders with alternative architectures (such as hierarchical tree-based encoders) distinguish different features.*

Very interesting idea; we have indeed preliminary results suggesting that a Transformer listener may be less anti-efficient than LSTM. To be further explored in future work.

***R2**: The authors do not state whether the length penalty affects communication success.*

Convergence is slower with smaller number of successful runs (depending on the coefficient α) in this case. We will report this in the paper.

***R3**: Somewhat unsurprisingly, the developed protocols implement "anti- efficient" encoding.*

We were actually surprised by this. Ours is the first successful protocol ever to display a significant anti-efficient effect (compare to natural languages and animal communication systems in Ferrer-i-Cancho et al CogSciJ 2013).

***R3**: The authors mentioned they use top 1000 most frequent words from natural languages. Do they have the same degree (exponent) of a power-law distribution as in synthetic referents experiment?*

The natural languages corpora follow a power-law distribution with an exponent between -0.81 and -0.92 (we used -1 in the artificial language).