ON THE RELATION BETWEEN CONTEXT DEPENDENCY AND VOCABULARY IN HUMAN LANGUAGE AND BIRDSONG

Takashi Morita[*1,2], Ryosuke O. Tachibana[3], Kazuo Okanoya[3,4,5], and Hiroki Koda[2,3]

[*]Corresponding Author: tmorita@alum.mit.edu
[1]SANKEN, Osaka University, Ibaraki, JAPAN
[2]Primate Research Institute, Kyoto University, Inuyama, JAPAN
[3]Graduate School of Arts and Sciences, the University of Tokyo, Tokyo, JAPAN
[4]Advanced Comprehensive Research Organization, Teikyo University, Tokyo, JAPAN
[5]Behavior and Cognition Joint Research Laboratory, RIKEN Center for Brain Science, Wako, JAPAN

Human language production is often characterized by long dependency on previous output records, or *context*. Recent studies on deep learning-based models of language generation have shown that the dependent context can range over hundreds of words. In this study, we investigated the relation between context dependency and vocabulary type/size, by case-studying English. We found that the long context dependency vanished when words were replaced with their grammatical categories and purely syntactic dependency was considered, which suggests that the long dependency in human language is largely attributed to semantic factors. We also performed clustering in word-embedding spaces and showed that a larger number of clusters (i.e., larger vocabulary) led to longer context dependency. Moreover, a parallel analysis of birdsong (of Bengalese finch, *Lonchura striata* var. *domestica*) revealed the opposite relation between context dependency and vocabulary size (i.e., larger vocabularies shortened dependency length) and moderate vocabulary sizes amounted to comparable dependency lengths to English syntax. We also showed that English phoneme sequences have a much shorter dependency length, which casts doubt on the previous generalization in comparative studies that birdsong is more homologous to human language phonology than to syntax.

1. Introduction

Human language production is often characterized by long dependency on previous output records, or *context* (Larson, 2017). Recent studies on deep learning-based models of language generation have shown that the dependent context can range over hundreds of words (Khandelwal, He, Qi, & Jurafsky, 2018; Dai et al., 2019). In this study, we investigated the relation between context dependency and another characteristic property of human language: rich vocabulary. Specifically, we clustered English words into various types/numbers of categories and analyzed its effect on the length of context dependency.

The results of the English analysis will be discussed in comparison with context dependency in birdsong, reproducing our recent study on Bengalese finch (*Lonchura striata* var. *domestica*; Morita, Koda, Okanoya, & Tachibana, 2021).

As pointed out above, the rich vocabulary is one of the most distinguished properties of human language; no other animal seems to handle tens of thousands of vocal categories like human language words. Thus, it is of interest to see whether birdsong exhibits language-like long-distance dependency if it consists of a larger vocabulary, and our previous study addressed this question using the same analytical paradigm as in this study. Here, we complement the cross-specific investigation of the relation between vocabulary size and context dependency through an analysis of English data.

2. Materials & Methods

2.1. *Measuring Context Dependency by Deep Language Modeling*

We measured context dependency in human language and birdsong by using a deep neural network (Khandelwal et al., 2018; Dai et al., 2019; Morita et al., 2021). We first trained a neural network on the language modeling task: the network computed the predictive probability $P(x_t \mid x_1, \ldots, x_{t-1})$ of each token x_t (e.g., word, phoneme, or birdsong syllable) conditioned on the preceding tokens $x_1, \ldots, x_{t-1}$ (i.e., context). Then, the trained network was used to compute the predictive probability of test tokens conditioned on the full and truncated contexts (Fig. 1A). Intuitively, truncation of a context decreases the predictive probability when dependent tokens are excluded from the context. Thus, the effective context length (ECL) was defined by the minimum length of the truncated context where the difference in the predictive probabilities based on the two contexts faded away; in practice, we adopted the canonical threshold of 1% difference in perplexity following the previous studies (Khandelwal et al., 2018; Dai et al., 2019; Morita et al., 2021).

Following Morita et al. (2021), we adopted a Transformer with six layers, eight heads, hidden dimensionality of 512 (for both the self-attention and feed-forward modules), and relative position encoding (Dai et al., 2019) for language modeling. See Morita et al. (2021, S2 Text) for more details including the training procedure.

2.2. *Human Language Data*

We studied the English portion of Wiki40B dataset for word-level analysis of context dependency owing to syntax and/or semantics (Guo, Dai, Vrandečić, & Al-Rfou, 2020). The raw text data were tokenized into word(-like) sequences using the Stanza package of Python (Qi, Zhang, Zhang, Bolton, & Manning, 2020). This resulted in 1,542,787,693 training and 85,462,957 test tokens, which were chunked into 19,873,689 and 1,104,835 paragraphs respectively based on the tags in Wiki40B. Due to limitations in computational resources and time, we performed the analysis of context dependency on only 10,000 test paragraphs (but still amounting to 84,664,199 tokens) that were randomly selected among those

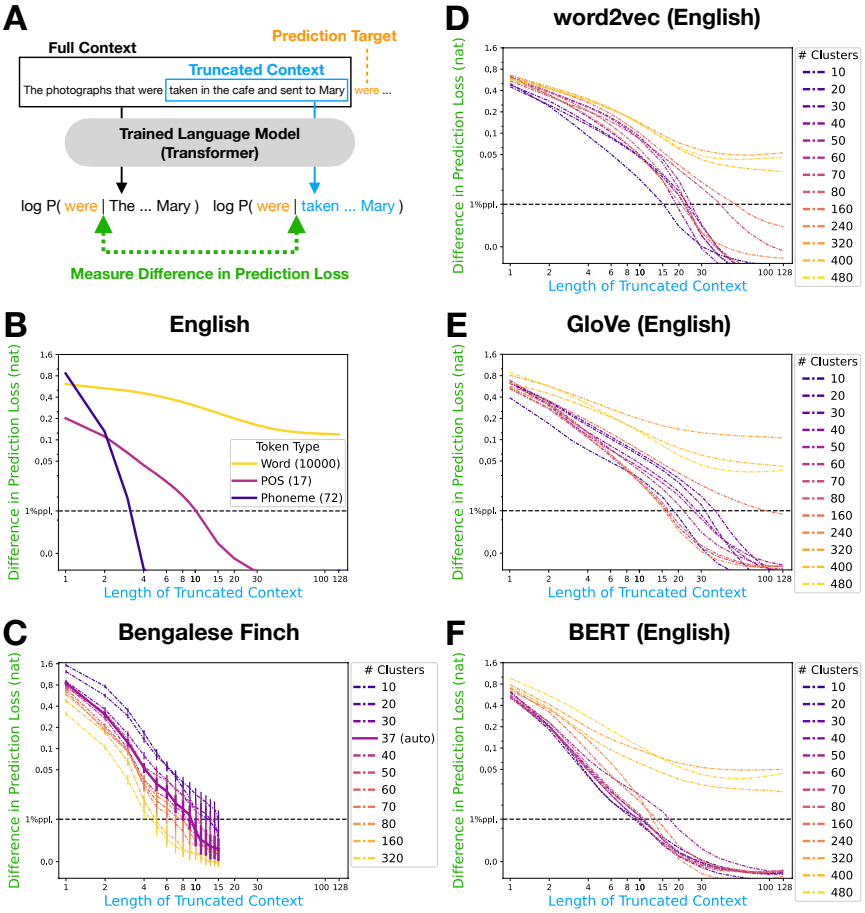

Figure 1. (A) Metric of context dependency based on deep language modeling. Log predictive probabilities of test tokens were computed conditioned on the full context (framed with a black line) and truncated context (framed with a light-blue line). The effective context length (ECL) is the minimum length of the truncated context where the prediction difference went below a threshold (1% in perplexity, or ppl.). (B–F) The differences in the mean loss (negative log probability) between the truncated- and full-context predictions of Wiki40B English words (B), part-of-speech (POS) tags (B), and k-means cluster labels in word2vec (D), GloVe (E), and BERT (F) word-embedding spaces; phoneme transcription of LibreSpeech (B); and Bengalese finch songs, consisting of syllable sequences (C). The x-axis corresponds to the length of the truncated context. Both of the axes are in log scale. The horizontal dashed line indicates the canonical threshold at 1% ppl. The error bars in (C) represent the 90% confidence intervals estimated from 10,000 bootstrapped samples, so that the loss difference is statistically significant ($p < 0.05$) if the lower side of the intervals is above the threshold.

containing at least 128 tokens.

Three major types of word representation were investigated:

- 10,000 most frequent word tokens, replacing the rest with a special label `<unk>`.
- Part-of-speech (POS) tags in the universal dependency format (Nivre et al., 2016), induced by a pretrained model provided in Stanza (Qi et al., 2020).
- L2-normalized k-means clustering of three pretrained word embeddings ($k := 10, 20, \ldots, 80, 160, 320, 400, 480$):
  - Word2vec embeddings (Mikolov, Chen, Corrado, & Dean, 2013).
  - GloVe embeddings (Pennington, Socher, & Manning, 2014)
  - Word embeddings of BERT (Devlin, Chang, Lee, & Toutanova, 2019).

  Only alphanumeric words included in the BERT's vocabulary (the smallest among the three) were clustered and the rest were labeled with `<unk>`.

The POS representation removed semantic distinctions among words (e.g., apple, dog $\rightarrow$ NOUN; give, run $\rightarrow$ VERB; a, the $\rightarrow$ DET) and, thus, allowed us to benchmark a purely syntactic context dependency. Word embeddings are real-valued vector representations that are derived ultimately from co-occurrence statistics of words. Word embeddings are known to reflect syntactic and semantic properties of words in numerical ways. Therefore, clustering in the embedded space allowed us to manipulate the vocabulary size and inquire into its relation with context dependency.

In addition to the word-level analyses above, we also assessed the context dependency in phoneme sequences (i.e., phonotactics). We used the phoneme transcription of the LibriSpeech corpus (Panayotov, Chen, Povey, & Khudanpur, 2015) that was induced by the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017).[1] The Transformer language model was trained on the `train-clean-360` portion of the corpus (consisting of 13,611,485 tokens extracted from 104,008 wav files), and `test-clean` portion (203,760 tokens from 2,620 wav files) was used for estimating the dependency length. We adopted the original wav files of the corpus as the unit of sequence.

2.3. *Birdsong Data*

We reproduced the estimation of context dependency in Bengalese finch song, originally reported in (Morita et al., 2021). In that study, we performed Transformer language modeling on syllable sequences produced by eighteen birds. Syllables were classified into discrete categories by an unsupervised clustering method with speaker normalization (see also Morita & Koda, 2020) and the resulting symbolic representation was used for the language modeling and analysis of context dependency. The number of syllable categories was automatically esti-

[1] Precomputed transcription is provided by Lugosch, Ravanelli, Ignoto, Tomar, and Bengio (2019) in `https://doi.org/10.5281/zenodo.2619474`.

mated by the clustering method as 37. We also reported the context dependency based on finer- and more coarse-grained classifications of the syllables (into 10, 20, ... 80, 160, 320 categories) by L2-normalized k-means clustering.[2] Here, we reproduced the analysis of context dependency for each of those syllable clusterings. The dataset consisted of 457,992 and 6,557 syllables (segmented into 7,779 and 100 sequences by recording sessions) for training and test respectively.

3. Results

While sequences of raw English words exhibited the context dependency beyond 128 tokens, the effective dependency length dropped to eleven when the words were replaced with their POS tag (Fig. 1B). The dependency in phoneme sequences was even shorter—only on four tokens—despite their larger vocabulary size than the POS categories.

Manipulation of the vocabulary size by word-embedding clustering showed that larger vocabularies led to longer context dependency (Fig. 1D–F). By contrast, the context dependency of Bengalese finch song decreased as finer-grained syllable classifications were adopted (Fig. 1C; due to the small size of test data, the dependency length was estimated considering the statistical significance of the difference between full- and truncated-context predictions against the threshold with $p < 0.05$, following Morita et al., 2021). When coded by the automatically detected vocabulary consisting of 37 syllable categories, the birdsong exhibited effective dependency on eight tokens. Smaller vocabularies with 10–30 categories increased the dependency length to ten to fifteen tokens. Conversely, larger vocabularies decreased the dependency length, up to five when 160/320 syllable categories were assumed.

4. Discussion

We found that the word-level context dependency of English became drastically shorter when tokens were replaced with their POS tag. This result suggests that the long context dependency in human language is mostly attributed to semantic factors and purely syntactic dependencies can be handled relatively locally, by referring to eleven recent tokens in the production history. This view is also consistent with the positive correlation between the number of word-embedding clusters and the dependency length; finer-grained clusterings of embeddings recovered more semantic information in the original continuous space, which led to the longer dependency length in turn.

By contrast, finer-grained classification of Bengalese finch syllables decreased the dependency length (as originally reported in Morita et al., 2021). This find-

[2] All the clustering results are publicly available in `https://doi.org/10.17605/OSF.IO/R6PAQ`.

ing indicates that expanding vocabulary size does not lead to language-like long-distance dependency by itself. Specifically, minor acoustic variations in Bengalese finch syllables—which are encoded in fine-grained classifications but ignored in coarse-grained classifications—are unlikely to carry semantics-like information (Okanoya, 2007; Berwick, Okanoya, Beckers, & Bolhuis, 2011; Miyagawa, Berwick, & Okanoya, 2013). Instead, acoustic properties of the birdsong syllables are known to be affected by surrounding syllables (Wohlgemuth, Sober, & Brainard, 2010) and, thus, paying more attention to local context improved the prediction of fine-grained syllable categories (e.g., five previous tokens when 160 and 320 categories are assumed). Such local interactions are more akin to human language phonotactics; as demonstrated by our analysis of phoneme sequences in English, phonotactic dependencies are much shorter (four tokens; although longer and potentially unbounded phonotactic dependencies have been suggested for limited patterns in several languages; Sapir & Hoijer, 1967) than syntactic dependencies encoded by POS sequences (eleven tokens).

Meanwhile, it should be noted that Bengalese finch song—under the assumption of a moderate number of syllable categories (10–40)—exhibited longer dependency on eight to fifteen tokens than English phonotactics (four tokens). The estimated dependency length of the birdsong instead amounted to that of English syntax (eleven tokens), which casts doubt on the previous generalization in comparative studies that birdsong is more homologous to human language phonology than to syntax (Berwick et al., 2011). The non-local context dependency comparable to human language syntax also implies that the birdsong cannot be modeled efficiently by traditional n-gram grammars (Hosino & Okanoya, 2000) because exponentially more transitional rules are needed as the dependency length increases. The long dependencies would be captured more succinctly by latent structures as in hidden Markov models (Rabiner, 1989; Katahira, Suzuki, Okanoya, & Okada, 2011) and hierarchical grammars (Berwick, 2015; Morita & Koda, 2019), or by distributed representation as in biological/artificial neural networks (Nishikawa, Okada, & Okanoya, 2008; Dai et al., 2019).

Acknowledgements

This study was supported by JSPS Grant-in-aid for Scientific Research on Innovative Areas #4903 (Evolinguistic; JP17H06380) to HK and KO, JSPS Grant-in-Aid for Scientific Research (JP19KT0023, JP21H03781) to ROT, and for Early-Career Scientists (JP21K17805) to TM, the JST Core Research for Evolutional Science and Technology 17941861 (JPMJCR17A4) to HK and ACT-X 21454934 (JPMJAX21AN) to TM, and the Mitsubishi Foundation Research Grants in the Natural Sciences (202111014) to TM. We also gratefully acknowledge the support of the Academic Center for Computing and Media Studies, Kyoto University, regarding the use of their supercomputer system.

References

Berwick, R. C. (2015). Mind the gap. In Á. J. Gallego & D. Ott (Eds.), *50 years later: Reflections on Chomsky's aspects* (pp. 1–12). Cambridge, MA: MITWPL.

Berwick, R. C., Okanoya, K., Beckers, G. J., & Bolhuis, J. J. (2011). Songs to syntax: the linguistics of birdsong. *Trends in Cognitive Science*, *15*(3), 113–121.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2978–2988). Florence, Italy.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota.

Guo, M., Dai, Z., Vrandečić, D., & Al-Rfou, R. (2020). Wiki-40B: Multilingual language model dataset. In *Proceedings of the 12th language resources and evaluation conference* (pp. 2440–2452). Marseille, France.

Hosino, T., & Okanoya, K. (2000). Lesion of a higher-order song nucleus disrupts phrase level complexity in bengalese finches. *Neuroreport*, *11*(10), 2091–2095.

Katahira, K., Suzuki, K., Okanoya, K., & Okada, M. (2011). Complex sequencing rules of birdsong can be explained by simple hidden Markov processes. *PLOS ONE*, *6*(9), 1–9.

Khandelwal, U., He, H., Qi, P., & Jurafsky, D. (2018). Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 284–294). Melbourne, Australia.

Larson, B. (2017). Long distance dependencies. *Oxford Bibliographies*.

Lugosch, L., Ravanelli, M., Ignoto, P., Tomar, V. S., & Bengio, Y. (2019). Speech model pre-training for end-to-end spoken language understanding. In *Proceedings of INTERSPEECH* (pp. 814–818). Graz, Austria.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of INTERSPEECH* (pp. 498–502). Stockholm, Sweden.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 1st international conference on learning representations (ICLR)*. Scottsdale, Arizona.

Miyagawa, S., Berwick, R., & Okanoya, K. (2013). The emergence of hierarchical

structure in human language. *Frontiers in Psychology*, *4*, 71.

Morita, T., & Koda, H. (2019). Superregular grammars do not provide additional explanatory power but allow for a compact analysis of animal song. *Royal Society Open Science*, *6*(7), 190139.

Morita, T., & Koda, H. (2020). Exploring TTS without T using biologically/psychologically motivated neural network modules (ZeroSpeech 2020). In *Proceedings of interspeech 2020* (pp. 4856–4860). Shanghai, China.

Morita, T., Koda, H., Okanoya, K., & Tachibana, R. O. (2021). Measuring context dependency in birdsong using artificial neural networks. *PLOS Computational Biology*, *17*(12), e1009707.

Nishikawa, J., Okada, M., & Okanoya, K. (2008). Population coding of song element sequence in the Bengalese finch hvc. *European Journal of Neuroscience*, *27*(12), 3273–3283.

Nivre, J., Marneffe, M.-C. de, Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 1659–1666). Portorož, Slovenia.

Okanoya, K. (2007). Language evolution and an emergent property. *Current Opinion in Neurobiology*, *17*(2), 271–276.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5206–5210). South Brisbane, Queensland, Australia.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations* (pp. 101–108). Online.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Sapir, E., & Hoijer, H. (1967). *The phonology and morphology of the Navaho language.* Berkeley, CA: University of California Press.

Wohlgemuth, M. J., Sober, S. J., & Brainard, M. S. (2010). Linked control of syllable sequence and phonology in birdsong. *Journal of Neuroscience*, *30*(39), 12936–12949.