ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Invited reply

Cite this article: Morita T, Koda H. 2020 Difficulties in analysing animal song under formal language theory framework: comparison with metric-based model evaluation. *R. Soc. open sci.* **7**: 192069.
http://dx.doi.org/10.1098/rsos.192069

Received: 27 November 2019
Accepted: 9 January 2020

Subject Category:
Psychology and cognitive neuroscience

Subject Areas:
computational biology/cognition/behaviour

Authors for correspondence:
T. Morita
e-mail: tmorita@alum.mit.edu
H. Koda
e-mail: koda.hiroki.7a@kyoto-u.ac.jp

The accompanying comment can be viewed at http://dx.doi.org/10.1098/rsos.191772.

THE ROYAL SOCIETY PUBLISHING

Difficulties in analysing animal song under formal language theory framework: comparison with metric-based model evaluation

T. Morita and H. Koda

Primate Research Institute, Kyoto University, 41-2 Kanrin, Inuyama, Aichi 484-8506, Japan

TM, 0000-0002-3900-5410; HK, 0000-0002-0927-3473

1. Introduction

This is a reply to a comment from De Santo & Rawski (DS&R) [1] regarding our recent investigation of the gibbon song syntax [2]. The major objectives are to clarify the (i) difference between the proposed analytical method and formal language theory (FLT) advocated by DS&R and (ii) difficulties in applying the existing FLT-based analysis to animal studies.

2. Difficulties in FLT analysis

Figure 1*a* depicts typical FLT-based human language analysis. Herein, the researchers prepare a mathematically defined *language*—a set of strings of symbols—termed $\hat{L}$ in the figure. It is essential to identify the strings that belong to $\hat{L}$ and those outside $\hat{L}$, *including those not observed*: for example, unbounded centre-embeddings—though never observed—are often assumed to exist [3].

Once $\hat{L}$ is well defined, the researchers search for the *smallest* class of languages containing $\hat{L}$. This search begins with smaller classes; the researchers check whether each class contains $\hat{L}$ using mathematical theorems [3,4].

The *idealization*, i.e. the inference of $\hat{L}$, based on available information, is the most challenging aspect of FLT-based analysis. In the absence of systematic procedures, idealization may result in loss of reproducibility and even fabrication of data. However, linguists idealize languages unsystematically—without a constructive algorithm or evaluation metric. Moreover, idealization is not induced solely from collections of uttered sentences (corpora). Instead, factors like interpretation and grammatical judgement of sentences—seldom available from

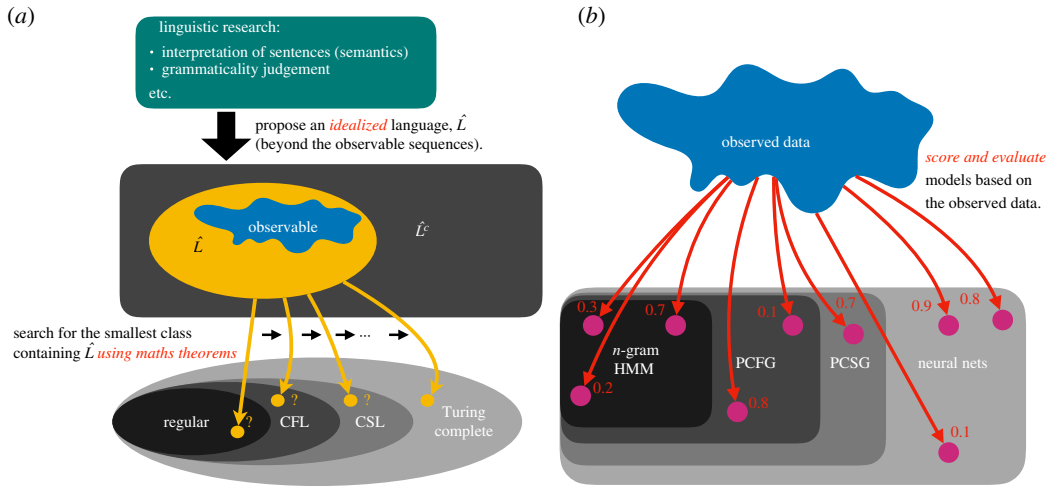

Figure 1. Schematic of research procedures in (*a*) FLT and (*b*) metric-based model comparison.

animal studies and often unreliable [5,6]—are used. Thus, without significant changes, idealization designed for human languages cannot be applied to animal data.

3. Non-FLT advantage of probabilistic context-free grammar

Based on difficulties due to idealization, it is more promising and practical to evaluate generative models of animal voice sequences from observed data, as depicted in figure 1*b*. This is noted to yield the best model among all hypothesized models; unlike FLT, it does not conclude that none of the models work. Thus, a broader range of models must be included for comparison, given an appropriate search algorithm. In [2], the hypothesis space for animal voice sequences was expanded to probabilistic context-free grammars (PCFGs) from previous regular domain. This difference between the FLT and proposed methods regarding search procedures has 'pushed the scientific community towards misguidance' with respect to idealization and *optimum-among-regulars* [7].

The *necessity* of models in a particular class is assessed using the proposed model comparison paradigm by defining the metric by its fit to data—*likelihood* under probabilistic settings. That is, a good model predicts the behaviour of real data with high probability and produces its generative simulation as realistically as possible. Natural language processing (NLP) adopts this metric; for example, the neural network parameters are optimized via likelihood maximization, which yields the current state-of-the-art language model [8–10]. Although the neural language modelling of animal voice sequences has not been studied extensively, it could become the empirical evidence for the necessity of superregular analysis if neural network models outperform the classical regular models [11–14]. The neural language model would also serve as the best animal voice sequence simulator currently available.

The likelihood metric is not suitable for non-neural, rule-based superregular models such as PCFG, as no remarkable advantages are exhibited over regular models for both human language [15] and animal voice sequences [2] (§4.2). Thus, NLP did not benefit much from superregularity prior to the deep learning era; the previous state-of-the-art architecture was smoothed *n*-gram models, which can only generate a subclass of regular languages (termed *strictly locally testable languages*) [16,17]. The results might appear to be counteractive, as the FLT proves that data with centre-embeddings can only be explained using superregular models. However, centre-embeddings are bounded and rare in real data [18]; therefore, they do not have a significant effect on statistical evaluation.

To study the advantages of superregular models, the *simplicity* of the models should be measured as well as the fit to data. The two submetrics can be balanced using the Bayesian posterior inference. A (non-regular) PCFG had greater posterior for human language data than regular grammars, which grow in size—decreasing the prior probability—to achieve the same degree of likelihood as the PCFG [15]. In [2], we showed that PCFG had the same advantage of compactness for analysis of gibbon song data (§4.3), to the extent that it outperformed regular grammars under the Bayesian metric (§4.1).

The compactness of PCFG probably arises from structural representation of frequent substrings. These statistical patterns are prominent even in child-directed speech [15] and animal voice sequences [19,20], where centre-embedding may not be observed. Improved versions of PCFG, such as *adaptor*

grammars, have been designed to better capture frequent substructures of sentences, rather than the centre-embeddings [21,22].

4. Execution costs

The processing of sentences using PCFG is generally based on a $O(n^3)$-time algorithm with respect to length of the sentence n [23,24]; this method is polynomial but highly expensive for practical applications. The algorithm also requires a working memory of size $O(n^2)$, which eventually exceeds the capacity of human and animals. Contrarily, the finite-state automata run in linear time and the memory size remains constant. As noted by DS&R, differences in the costs of execution are not incorporated in the Bayesian analysis, which only focuses on Marr's computational level of inquiry and refrains from discussing the algorithmic or implementation levels [25].

It may be noted that biological organisms may not compute the exact probabilities as defined by PCFG. A reliable approximation with fading memory is acceptable in practice. For example, recurrent neural networks—including biologically plausible variants—act as universal approximators and run in real time [11,13]. Actual algorithms and implementation used by humans and animals are considered efficient but difficult to interpret, as in the case of neural networks. Hence, studies at the computational level help us understand human and animal cognitive systems upon investigation of their interpretable representations.

5. What is FLT expected to do for animal cognition studies?

Herein, the difficulties in applying FLT-based analysis to studies on animals have been identified. However, this does not mean that FLT is futile. For example, the discovery of more efficient algorithms is always valuable. Perhaps what is expected from FLT-oriented *linguists* is the proposal of a systematic idealization procedure that runs on real animal data. Various important achievements in FLT cannot be exploited unless this fundamental technology is developed.

Data accessibility. This article has no additional data.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by the JSPS/MEXT KAKENHI (no. 4903(Evolinguistic), JP17H06380) and JST CREST no. 17941861 (no. JPMJCR17A4).

References

1. De Santo A, Rawski J. In press. What can formal language theory do for animal cognition studies? *R. Soc. open sci.*
2. Morita T, Koda H. 2019 Superregular grammars do not provide additional explanatory power but allow for a compact analysis of animal song. *R. Soc. open sci.* **6**, 190139. (doi:10.1098/rsos.190139)
3. Bar-Hillel Y, Perles M, Shamir E. 1961 On formal properties of simple phrase structure grammars. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* **14**, 143–172. (Reprinted by Addison-Wesley 1964.)
4. Huybregts R. 1984 The weak inadequacy of context-free phrase structure grammars. In *Van Periferie Naar Kern* (eds G de Haan, M Trommelen, W Zonneveld), pp. 81–99. Dordrecht, The Netherlands: Foris Publications.
5. Gibson E, Fedorenko E. 2010 Weak quantitative standards in linguistics research. *Trends Cogn. Sci.* **14**, 233–234. (https://doi.org/10.1016/j.tics.2010.03.005)
6. Linzen T, Oseki Y. 2018 The reliability of acceptability judgments across languages. *Glossa J. Gen. Linguist.* **3**, 100. (doi:10.5334/gjgl.528)
7. Berwick RC, Okanoya K, Beckers GJ, Bolhuis JJ. 2011 Songs to syntax: the linguistics of birdsong. *Trends Cogn. Sci.* **15**, 113–121. (doi:10.1016/j.tics.2011.01.002)
8. Mikolov T, Karafiát M, Burget L, Černocký J, Khudanpur S. 2010 Recurrent Neural Network Based Language Model. In *INTERSPEECH-2010, 11th Annual Conf. of the International Speech Communication Association, Makuhari, Chiba, Japan, 26–30 September*, pp. 1045–1048. ISCA.
9. Graves A. 2013 Generating sequences with recurrent neural networks. (https://arxiv.org/abs/1308.0850).
10. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. 2019 Transformer-XL: attentive language models beyond a fixed-length context. (https://arxiv.org/abs/1901.02860)
11. Jin L, Gupta MM, Nikiforuk PN. 1995 Universal approximation using dynamic recurrent neural networks: discrete-time version. In *Proc. of ICNN'95 - Int. Conf. on Neural Networks, Perth, Australia, 27 November–1 December*, vol. 1, pp. 403–408. IEEE.
12. Siegelmann HT, Sontag ED. 1995 On the computational power of neural nets. *J. Comput. Syst. Sci.* **50**, 132–150. (doi:10.1006/jcss.1995.1013)
13. Maass W, Natschläger T, Markram H. 2002 Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* **14**, 2531–2560. (doi:10.1162/089976602760407955)
14. Weiss G, Goldberg Y, Yahav E. 2018 On the practical computational power of finite precision RNNs for language recognition. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, July (Volume 2: Short Papers)*, pp. 740–745. Association for Computational Linguistics.
15. Perfors A, Tenenbaum JB, Regier T. 2011 The learnability of abstract syntactic principles. *Cognition* **118**, 306–338. (doi:10.1016/j.cognition.2010.11.001)
16. Katz SM. 1987 Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoust.*

Speech Signal Process. **35**, 400–401. (doi:10.1109/TASSP.1987.1165125)

17. Kneser R, Ney H. 1995 Improved backing-off for N-gram language modeling. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Detroit, MI, 9–12 May*, vol. 1, pp. 181–184. IEEE.
18. Karlsson F. 2007 Constraints on multiple center-embedding of clauses. *J. Linguist.* **43**, 365–392. (doi:10.1017/S002222670 7004616)
19. Payne RS, McVay S. 1971 Songs of humpback whales. *Science* **173**, 585–597. (doi:10.1126/science.173.3997.585)
20. Okanoya K. 2004 Song syntax in Bengalese finches: proximate and ultimate analyses. *Adv. Stud. Behav.* **34**, 297–345. (doi:10.1016/S0065-3454(04)34008-8)
21. Johnson M, Griffiths TL, Goldwater S. 2007 Adaptor grammars: a framework for specifying compositional nonparametric Bayesian Models. In *Advances in Neural Information Processing Systems 19* (eds B Schölkopf, JC Platt, T Hoffman), pp. 641–648. Cambridge, MA: MIT Press.
22. O'Donnell TJ. 2015 *Productivity and reuse in language: a theory of linguistic computation and storage*. Cambridge, MA: MIT Press.
23. Baker JK. 1979 Trainable grammars for speech recognition. In *Speech communication: papers presented at the 97th meeting of the Acoustical Society of America, Cambridge, MA, 11–15 June*, pp. 547–550.
24. Stolcke A. 1995 An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Comput. Linguist.* **21**, 165–201.
25. Marr D. 1982/2010 *Vision: a computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.