Modals in natural language optimize the simplicity/informativeness trade-off

Nathaniel Imel and Shane Steinert-Threlkeld

Introduction A language can be simple and uninformative (e.g. containing a single expression). A language can be complex and informative (e.g. containing unique expressions for each possible meaning). A language cannot be both simple and informative: these two pressures trade-off against each other. A recent line of work develops the idea that meanings cross-linguistically are optimized for efficient communication, i.e. they optimally balance these two competing pressures [1]. This approach successfully explains the semantic variation observed in domains both of content words (e.g. kinship [2], color [3]) and function words (e.g. quantifiers [4], indefinites [5], boolean connectives [6]). This paper shows that modals cross-linguistically [7, 8] can be seen as optimizing this trade-off.

Measures In modeling (efficient) communication with modals, we take the object of communication to be the correct transmission of a pair of a *force* and a *flavor*. At this level of modeling, the meaning of a modal is a set of such pairs, allowing us to capture variability in flavor (e.g. for English *may*) as well as variability in force, as recently argued to be present in Lilloet Salish [9], Nez Perce [10], Washo [11], and Old English [12].

We measure the complexity of a modal in terms of the shortest formula in a language of thought [13]. In particular, we use a basic propositional language with atoms for each possible force and each possible flavor. For a modal, we write a disjunctive normal form capturing all of the force-flavor pairs it can express, and then apply a minimization algorithm based on [14]. The complexity is the number of atoms in this shortest formula; the complexity of a language is the sum of the complexity of the modals therein.

We measure the informativeness of a modal system (following [4, 5]) in terms of the probability of successful communication between a speaker who wants to convey an intended force-flavor pair to a listener, who must guess which one is intended solely on the basis of hearing a modal expression from the speaker. More formally:

$$I(L) := \sum_{\mathbb{M}} P(\mathbb{M}) \sum_{m \in L} P(m|\mathbb{M}) \sum_{\mathbb{M}' \in m} P(\mathbb{M}'|m) \cdot u(\mathbb{M}', \mathbb{M})$$

$$\text{where } u(\mathbb{M}', \mathbb{M}) = 0.5 \cdot \mathbb{1}_{\text{force}(\mathbb{M})=\text{force}(\mathbb{M}')} + 0.5 \cdot \mathbb{1}_{\text{flavor}(\mathbb{M})=\text{flavor}(\mathbb{M}')}$$

Here, $P(\mathbb{M})$ is a prior probability (assumed to be uniform) over the pairs; $P(m|\mathbb{M})$ represents the speaker, where m is a modal, and $P(\mathbb{M}'|m)$ the listener. The utility function u gives partial credit: the listener gets half credit for correctly guessing each of the force and the flavor, and so full credit for correctly guessing the intended pair. Finally, *communicative cost* is inversely related to informativeness: $C(L) := 1 - I(L)$.

In the absence of a robust dataset of the modal systems of many languages, we proceed by generating a large number of artificial languages and using proposed semantic universals to measure how natural such languages are. In particular, Nauze [15] proposed what we may call the *Single Ambiguity Universal (SAU)*: a modal may be ambiguous in either force or flavor, but not both. For a given language, we measure its Nauze degree as the proportion of modals that satisfy the SAU. As a refinement, Vander Klok [16] suggested that within both the epistemic / root domains, the system as a whole may only exhibit one kind of ambiguity. See [8] for discussion. We record for each langauge whether or not it satisfies Vander Klok's refinement.

Results Figure 1 presents the main results. We experiment with a meaning space containing 2 forces and 3 flavors. Each point is a language; the x-axis is communicative cost,

and the y-axis is complexity. The black line is the Pareto frontier: the set of languages for which no other language is both simpler and more informative. Triangles are Vander Klok languages. The color of a language is its Nauze degree.

We catalog several particular results. All optimal languages (those on the frontier) satisfy Vander Klok's generalization, with the exception of a single language on the bottom-right, which corresponds to a language with a single, highly-ambiguous modal (à la the Washo verb -*e*ʔ [11]). In particular, the Vander Klok languages ($N = 2255$) have mean optimality of 0.957 compared to a mean optimality of 0.797 for the remaining languages ($N = 65023$). More generally: Nauze degree is highly correlated with optimality (Pearson $r = 0.55$). This shows that languages which have more modals satisfying Nauze's SAU tend to do better at optimizing the simplicity/informativeness trade-off.

Discussion To summarize: our experiments show (i) that modal systems optimized for efficient communication satisfy Vander Klok's generalization and (ii) that languages with more Nauze modals tend to be more efficient for communication. These results show that trading off very general pressures for simplicity and informativeness may shape the semantic variation in the modal systems of the world's languages.

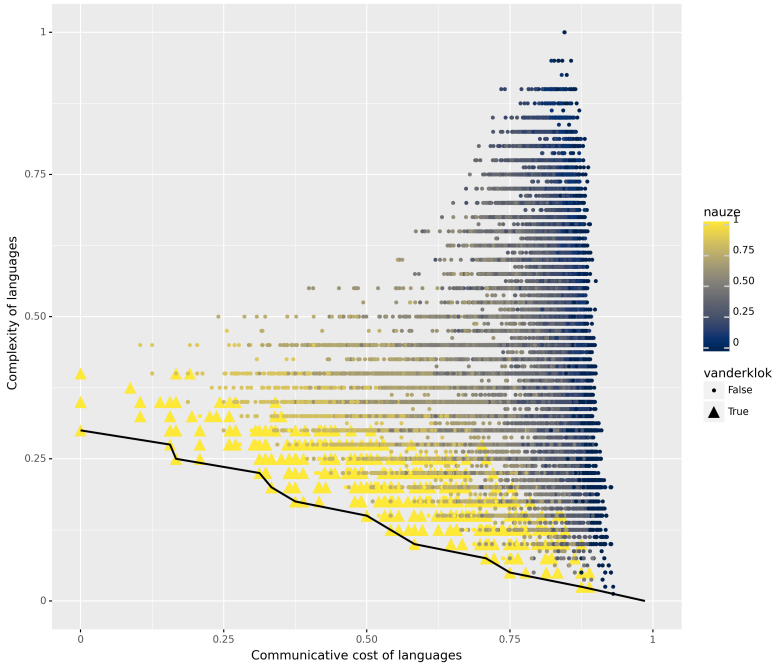

Figure 1: The modal systems sampled, plotted with communicative cost on the x-axis and complexity on the y-axis. Black: the Pareto frontier of optimal languages. Triangles satisfy Vander Klok's generalization. Color corresponds to Nauze degree.

1. Kemp, C. *et al.* Semantic Typology and Efficient Communication. *Annual Review of Linguistics,* 1–23 (2018).
2. Kemp, C. *et al.* Kinship Categories across Languages Reflect General Communicative Principles. *Science* **336,** 1049–1054 (2012).
3. Zaslavsky, N. *et al.* Efficient Compression in Color Naming and Its Evolution. *Proceedings of the National Academy of Sciences* **115,** 7937–7942 (2018).
4. Steinert-Threlkeld, S. Quantifiers in Natural Language: Efficient Communication and Degrees of Semantic Universals. *Entropy* **23,** 1335 (2021).
5. Denić, M. *et al. Complexity/Informativeness Trade-off in the Domain of Indefinite Pronouns* in *Proceedings of Semantics and Linguistic Theory (SALT 30)* **30** (2020), 166–184.
6. Uegaki, W. The Informativeness / Complexity Trade-off in the Domain of Boolean Connectives. *Linguistic Inquiry* (2021).
7. Kratzer, A. *The Notional Category of Modality* in *Words, Worlds, and Context* (eds Eikmeyer, H.-J. *et al.*) 38–74 (Walter de Gruyter, 1981).
8. Matthewson, L. *Modality* in *The Cambridge Handbook of Formal Semantics* (eds Aloni, M. *et al.*) 525–559 (Cambridge University Press, Cambridge, 2019).
9. Rullmann, H. *et al.* Modals as Distributive Indefinites. *Natural Language Semantics* **16,** 317–357 (2008).
10. Deal, A. R. Modals Without Scales. *Language* **87,** 559–585 (2011).
11. Bochnak, M. R. *Variable Force Modality in Washo* in *Proceedings of North-East Linguistic Society (NELS) 45* (eds Bui, T. *et al.*) (2015), 105–114.
12. Yanovich, I. Old English *motan, Variable-Force Modality, and the Presupposition of Inevitable Actualization. *Language* **92,** 489–521 (2016).
13. Piantadosi, S. T. *et al.* The Logical Primitives of Thought: Empirical Foundations for Compositional Cognitive Models. *Psychological Review* **123,** 392–424 (2016).
14. Feldman, J. Minimization of Boolean Complexity in Human Concept Learning. *Nature* **407,** 630–633 (2000).
15. Nauze, F. D. *Modality in Typological Perspective* (Universiteit van Amsterdam, 2008).
16. Vander Klok, J. *Restrictions on Semantic Variation: A Case Study on Modal System Types* in *Workshop on Semantic Variation* (2013).