COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 49 (2025) e70145
© 2025 Cognitive Science Society LLC.
ISSN: 1551-6709 online
DOI: 10.1111/cogs.70145

Assessing Pressures Shaping Natural Language Lexica

Jeanne Bruneau--Bongard,[a,b] Emmanuel Chemla,[a,c]
Thomas Brochhagen[b]

[a] *Département d'Études Cognitives, Laboratoire de Sciences Cognitives et Psycholinguistique, ENS, PSL University*
[b] *Department of Translation and Language Sciences, Universitat Pompeu Fabra*
[c] *Earth Species Project, Berkeley, CA*

Received 26 February 2025; received in revised form 20 October 2025; accepted 2 November 2025

Abstract

Human languages balance communicative informativity with complexity, conveying as much as needed through the simplest means required to do so. Yet, these concepts—informativity and complexity—have been operationalized in various ways, and it remains unclear which definitions best capture empirical linguistic patterns. A particularly successful operationalization is that offered by the Information Bottleneck framework, which suggests a balance between complexity and informativity across domains like color, kinship, and number. However, we show that the notion of complexity employed by this framework has some counterintuitive consequences. Focusing on color terms, we then study to what extent this and other notions of complexity play a role in explaining cross-linguistic regularity. We propose a method to assess their explanatory contributions; and to probe whether they enter in a joint optimization or in a trade-off competition. This offers a more general framework to study language change and the forces that shape it, where instead of showing that a given model is compatible with existing data, the data is used to adjudicate between candidate measures.

Keywords: Information theory; Language complexity; Semantic typology; Computational linguistics; Language change

1. Current models of language change

Languages across the world carve up reality in different ways. For instance, some languages distinguish between colors that other languages amalgamate into a single word. Such

Correspondence should be sent to Jeanne Bruneau--Bongard, Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Études Cognitives, ENS, PSL University, EHESS, CNRS, 29 rue d'Ulm, 75005 Paris, France. E-mail: jeanne.bruneau--bongard@ens.psl.eu lexical discrepancies have been studied in the domain of colors (Zaslavsky et al. 2018), as well as in many others, such as person (Zaslavsky et al. 2021), number (Denić and Szymanik 2023), tense (Mollica et al. 2021), and pronominal systems (Saldana and Maldonado 2024). Beyond such variation, languages also follow similar abstract organizational patterns in their vocabularies (Kemp and Regier 2012; Zaslavsky et al. 2018; Kemp et al. 2018; Xu et al. 2020; Brochhagen and Boleda 2022; Jackson et al. 2019; Carlsson et al. 2023). These patterns have been argued to reflect the result of an interplay between common but nondeterministic desiderata for these languages (Kemp and Regier 2012; Zaslavsky et al. 2018). Under this view, recurring universal patterns and their variations can be explained as the outcome of conflicting pressures for languages to help convey as much information as possible, while being as simple as possible.

The Information Bottleneck (IB) framework (Tishby et al. 1999) has been used as an operationalization of this trade-off. Rooted in information theory, this framework provides a characterization of compression of *meanings* into *words* as an optimization of a trade-off between informativity and a notion of complexity which derives from analytic principles. This approach has gained popularity in recent years, notably because it has been shown to fit the organization of meaning across languages in numerous semantic domains remarkably well (Zaslavsky et al. 2019; 2021; 2022; Tucker et al. 2022; Carlsson et al. 2023; Mollica et al. 2020).

Here, we first identify aspects of the framework that could be different. For instance, the IB complexity measure that is typically used comes directly from abstract theories of information, and it has not been put in competition with alternatives that could be relevant for direct linguistic processes, such as the borrowing of new words across languages or the existence of synonyms. Similarly, the proposed form of a trade-off is only one of several ways that different constraints may influence language change. Having identified these sources of alternative models, we thus propose a method to *compare* between various potential models of semantic evolution. Very recently, Tucker et al. (2025) have demonstrated how incorporating pressures for utility in addition to complexity and informativity provides a better fit to the World Color Survey data than the classic IB model. Our study provides new empirical and systematic methods to explore the space of theoretical options of pressures that shape language and the way they interact, including but also beyond the IB framework. Our results in the domain of color suggest a nuanced and multifaceted picture of language change, in which the cross-linguistic organization of lexica is explained by a more diverse notion of complexity, and in which interactions between two or more evolutionary pressures may be involved.

Below, we first introduce the IB framework, its associated notions of informativity and complexity, and the way these are measured in general. In Experiment 1, we explore more practically how the IB complexity measure behaves in situations at stake in language change. In Experiment 2, we present other plausible complexity measures, and their combinations, to eventually test which may be involved in the process of language change. Lastly, in Experiment 3, we relax the usual assumption of a trade-off between informativity and complexity and ask whether a linear combination of pressures can better explain real-world data.