Cognition 240 (2023) 105505

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

An information-theoretic approach to the typology of spatial demonstratives

Sihan Chen [a,*], Richard Futrell [b], Kyle Mahowald [c]

[a] Department of Brain and Cognitive Sciences, MIT, United States of America
[b] Department of Language Science, University of California, Irvine, United States of America
[c] Department of Linguistics, The University of Texas at Austin, United States of America

ARTICLE INFO

Keywords:
Information theory
Linguistic efficiency
Linguistic typology
Deixis
Spatial cognition

ABSTRACT

We explore systems of spatial deictic words (such as 'here' and 'there') from the perspective of communicative efficiency using typological data from over 200 languages Nintemann et al. (2020). We argue from an information-theoretic perspective that spatial deictic systems balance informativity and complexity in the sense of the Information Bottleneck (Zaslavsky et al., (2018). We find that under an appropriate choice of cost function and need probability over meanings, among all the 21,146 theoretically possible spatial deictic systems, those adopted by real languages lie near an efficient frontier of informativity and complexity. Moreover, we find that the conditions that the need probability and the cost function need to satisfy for this result are consistent with the cognitive science literature on spatial cognition, especially regarding the source–goal asymmetry. We further show that the typological data are better explained by introducing a notion of consistency into the Information Bottleneck framework, which is jointly optimized along with informativity and complexity.

1. Introduction

When Shakespeare's Hotspur says "Whither I go, thither shall you go too", he's using *whither* as an interrogative meaning "to where" and *thither* as a spatial demonstrative meaning "to there". When Edgar in *King Lear* says "Men must endure/Their going hence, even as their coming hither", *hence* means "from here" and hither "to here". This suite of spatial demonstrative (*here*, *hither*, *hence*, *there*, *thither*, *thence*, *where*, *whither*, *whence*) is largely lost in modern English—aside from *here* and *there*, some specialized use cases, and some fossilized expressions like "hither and yon" and "henceforth". English speakers now use *there* both to refer to a static location and for the "to" directional meaning. That is, one says "I was going there", not "I was going to there". In effect, the old meaning of *thither* has been entirely subsumed under the auspices of the word *there*.

Spatial demonstrative systems are a source of cross-linguistic variation (e.g., Levinson, 1996; Maldonado & Culbertson, 2020a; Nintemann, Robbers, & Hober, 2020; Stolz, Levkovych, Urdze, Nintemann, & Robbers, 2017) and have been studied as part of a broader body of work exploring how cross-linguistic variation affects, and is affected by, spatial cognition across cultures (e.g., Cadierno, 2004; Danziger, 2010; Gennari, Sloman, Malt, & Fitch, 2002; Jarvis & Pavlenko, 2008; Levinson, Kita, Haun, & Rasch, 2002; Levinson & Wilkins, 2006; Pederson et al., 1998). Some languages have more complex spatial demonstrative systems than others. Whereas English has only two levels to reflect distance in spatial demonstrative (*here* and *there*), Spanish has *aquí* (here), *ahí* (there, close by), *allí* (there, medium distance), and *allá* (there, far away). Other languages, like Malagasy, have systems that draw distinctions between spatial locations in which an item is visible or invisible. For instance, the suffix *-èto* is used to mean "here" when the object is visible but *-àto* when it is invisible. Moreover, some languages define spatial words in terms of landmarks (mountains, rivers, etc.), whereas others define space in terms of people (e.g., *here* means near the speaker and *there* means near the hearer), whereas others define spatial words in reference to a hypothetical "deictic center" (Levinson, 1996), meaning what constitutes *here* and *there* varies based on the imagined center of the particular discourse.

Why should languages converge on similar solutions? At the same time: why should some languages have more complicated spatial word systems than others? We argue that this convergence is an instance of a general functional pressure for efficiency in language (Gibson et al., 2019; Hawkins, 1994). We argue that as in other semantic domains (e.g., Kemp, Gaby, & Regier, 2019; Kemp & Regier, 2012; Mollica, Bacon, Xu, Regier, & Kemp, 2020; Mollica et al., 2021; Regier, Kay, & Khetarpal, 2007; Zaslavsky, Kemp, Tishby, & Regier, 2019; Zaslavsky, Regier, Tishby, & Kemp, 2019), there is a tradeoff between complexity and informativity. This tradeoff can be quantified using the

* Corresponding author.
E-mail address: sihanc@mit.edu (S. Chen).

https://doi.org/10.1016/j.cognition.2023.105505
Received 30 June 2022; Received in revised form 26 May 2023; Accepted 28 May 2023
Available online 18 August 2023

information bottleneck (Strouse & Schwab, 2017; Tishby, Pereira, & Bialek, 2000; Tishby & Zaslavsky, 2015; Zaslavsky, Kemp, Tishby, & Regier, 2019). Drawing on a typological database of spatial demonstratives across languages (Nintemann et al., 2020), we undertake an information-theoretic analysis of the cross-linguistic variation of place demonstrative systems across 5 major world regions.

To study the communicative efficiency of spatial demonstratives, we have to make a number of assumptions about both the world and about language users. Specifically, we have to estimate quantities like: How often do people refer to items that are close as opposed to far away? How often do people talk about where things were, as opposed to where they are going? How costly is it to confuse a word like "here" with a word like "there", as opposed to confusing a word like "here" with a word like "hence" ("from here")?

Here, we ask whether spatial demonstrative systems are optimized relative to statistical baselines and, if so, what assumptions must hold for that to be true. To fully characterize spatial demonstratives of the world's languages, we show that we must account for something not previously dealt with in information bottleneck work: a preference for consistency in paradigms. Introducing a notion of consistency, we show that a number of plausible, communicatively efficient systems are ruled out because they lack the consistency that characterizes real-world systems. We also show that, in an efficiency-based theory, the strong source/goal asymmetry found in many natural languages does not appear to originate from usage frequency, but rather from a stronger communicative penalty for confusing source words with place words than for confusing goal words with place words, this pattern emerges.

2. Background

2.1. Background on spatial demonstrative

The elements of language that we are concerned with are the class of deictic expressions used for space: meanings like HERE and THERE—which are closely related to spatial interrogatives (e.g., where/whether/whence) and to demonstratives (e.g., this/that/these/those) (e.g. Bühler, 1934; Coventry, Griffiths, & Hamilton, 2014; Coventry, Valdés, Castillo, & Guijarro-Fuentes, 2008; Diessel, 2006, 2012, 2012, 2019; Diessel & Coventry, 2020; Dixon, 2003; Fillmore, 1997; Hanks, 1990, 1990, 2011; Levinson, 2018; Levinson & Levinson, 2003; Perkins, 1992). While there has been considerable work on and debate about what it means to be a deictic expression, it is generally agreed that deictic are sensitive to context and involve joint attention between the communicators (Levinson, 2018; Levinson & Levinson, 2003).

Although these words may be naively interpreted in terms of referring to configurations in actual physical space, a variety of evidence from discourse analysis and experiments (e.g. Coventry et al., 2014, 2008; Enfield, 2003) suggests that they rather refer to positions within a subjective space defined by a particular discourse and in particular the body of the speaker (but see Peeters & Özyürek, 2016, for arguments against the body-centered view), a position going back to Bühler (1934). An important part of this position is that attention in a physical scene (e.g., eye gaze, pointing) has a major effect on the deictic expressions used (Coventry et al., 2010; García, Ehlers, & Tylén, 2017). And, when physical expressions are less available, deictic language is affected (Bangerter, 2004; Cooperrider, 2016; García et al., 2017).

This body of work has also shown a particularly prominent boundary between the more proximal levels and all other distance levels, perhaps because the more proximal language refers in particular to that which is within reach (Coventry et al., 2014, 2008; Kemmerer, 1999; Rocca, Wallentin, Vesper, & Tylén, 2019).

For our purposes, we use a notion of distance to formalize the meanings underlying deictic words. But this distance need not be thought of as a purely physical distance, but can instead represent the subjective distance between distance levels.

Table 1
English (a) and Maltese (b) spatial demonstratives.

	Goal	Place	Source
D3	there	there	from there
D2	there	there	from there
D1	here	here	from here

(a) English

	Goal	Place	Source
D3	hemm	hemm	minn hemm
D2	hemm	hemm	minn hemm
D1	hawn	hawn	minn hawn

(b) Maltese

Following Stolz et al. (2017) and Nintemann et al. (2020), we define the notion of "spatial demonstratives" as expressions that encode relative spatial properties. Many of these expressions are adverbs (e.g. *here*, *there* in English; *odavde* in Serbo-Croatian), but can also be adpositional phrases (e.g. *from here* in English; *cóng zhè lǐ* in Mandarin Chinese). The spatial relation encoded in these expressions can be divided into 2 dimensions: **distance level** – the distance of the referent with respect to the speaker, listener, or both, and **orientation** – the relative movement of the referent with respect to that deictic level. The key orientations we consider include PLACE, in which the referent is at a given distance level (e.g., *here*, *there*), GOAL, in which the referent is moving towards a distance level (e.g., *hither* and *thither* in Early Modern English), and SOURCE (e.g., *hence* and *thence*), in which the referent is moving *from* a distance level. We will discuss each of these 2 dimensions in detail below.

In Table 1, we show examples for two languages (English and Maltese) which, despite having very different words, have similar deictic systems. Both Maltese and English share the same strategy to partition the 3-by-3 meaning space using these terms: they use one term to represent D1-PLACE and D1-GOAL (the word 'here' in English), one term for PLACE and GOAL in both D2 and D3 (the word 'there' in English), one term only for D1-SOURCE ('from here'), and one term for SOURCE in both D2 and D3 ('from there').

Variation of distance levels in demonstrative systems. Much work on typological deixis (e.g., Anderson & Keenan, 1985; Dixon, 2003; Nintemann et al., 2020; Stolz et al., 2017) draws a distinction between person-oriented and deictic-center-oriented systems. A deictic-oriented system posits a deictic center around which the conversation is centered and that which is proximal to that center. Other systems are based more clearly on the position of the listeners. For instance, in Tagalog, "dito" is used when the referent is at a position near the speaker; "diyan" is used if the referent is at a position near the listener; and "doon" is used when the referent is far from both the speaker and the listener. Levinson (2018) complicates this picture, showing that systems with more than 2 deictic levels can have more complicated relationships with the position of the speaker and listener.

Moreover, some languages (e.g., Khwarshi, a language spoken in the Caucasus) use a combination of modalities. Different words are used if the referent is close by in general, close to the speaker, close to the listener, far away in general, far away from the speaker, and far away from the listener, respectively. The word also takes different forms depending on the grammatical genders. There are various other ways to categorize distance levels based on spatial and geographic features, such as altitude, upstream/downstream of a river, and visibility. We leave it to future work to incorporate speaker/listener systems and more geographic-based systems into this modeling framework.

For simplicity, we focus our analysis here on deictic-centric systems as categorized by Nintemann et al. (2020), in which spatial demonstratives are used relative to an imagined deictic center. Note that, by shoehorning all such systems into the same modeling framework, we

are necessarily collapsing over meaningful variation that exists among languages which are "deictic-centered". In future work, we believe it will be possible to more richly incorporate speaker/listener systems into this framework, as well as variation among languages on these dimensions.

Orientation: Place, goal, and source. In these spatial demonstrative systems, the form most likely to be formally marked is the SOURCE form (e.g., "from here", "from there"). When there is syncretism, it is most likely to be between PLACE and GOAL, or between all three. This pattern is not limited to just spatial demonstratives: across domains, languages tend to be more likely to mark sources than goals (Haspelmath, 2003; Jackendoff, 1983; Lakusta & Landau, 2012; Nikitina, 2009; Stolz, Lestrade, & Stolz, 2014). Georgakopoulos and Karatsareas (2017) gives a diachronic overview of how this played out historically in Greek, where goal markers were lost earlier than source markers. This pattern may fall out of a more general asymmetry between movement from a source and movement towards a goal.

There is robust evidence of a relationship between linguistic spatial reference systems and cognitive ones (Haun, Rapold, Janzen, & Levinson, 2011; Jackendoff, 1996; Jackendoff & Landau, 2013; Langacker, 2013; Levinson, 1996; Levinson et al., 2002; Pederson et al., 1998; Ünal, Ji, & Papafragou, 2021; Ünal, Richards, Trueswell, & Papafragou, 2021). Thus, the linguistic distinction is likely related to the fact that humans have an overall cognitive bias towards goals, as opposed to sources. They describe goals with more fine-grained distinctions (Papafragou, 2010; Regier & Zheng, 2007), and, in experiments, are more likely to focus on goal-directed movement than source movement (Lakusta & Landau, 2005, 2012; Regier, 1996). Children in particular seem to display a strong goal bias (Srinivasan & Barner, 2013) and overextend goal-directed meanings (e.g., assuming "weed the garden" means putting weeds into the garden, even when that contradicts evidence from world knowledge) (Johanson, Selimis, & Papafragou, 2019). Children also seem to produce GOAL markers earlier than SOURCE markers (see, e.g., Johanson et al., 2019, for Greek and English, Pléh, Vinkler, & Kálmán, 1997, for Hungarian, Dromi, 1979, for Hebrew).

Nikitina (2009), drawing on cross-linguistic evidence, suggests that "the meaning of Goal seems to be 'closer' to the meaning of Place than to the meaning of Source". In our work, we operationalize this by placing PLACE, GOAL, and SOURCE on a line, with GOAL and SOURCE as endpoints and with PLACE in the middle; we find that this configuration gives the best fit to the typological data. But, as Nikitina (2009) notes, there is an asymmetry. Do, Papafragou, and Trueswell (2020) show that, although goals are conceptually privileged, the frequency of source mentions increases when source is not in the common ground. Thus, there are likely pragmatic/communicative factors that underlie the source/goal asymmetry, while Chen, Trueswell, and Papafragou (2022) points out that the asymmetry might be due to an online attention bias, because it seems both SOURCE and GOAL are encoded in memory.

Does this mean that there is, in general, a greater penalty for confusing source words with place/goal words than for confusing place and goal? Or does the pattern fall out of the empirical need probability of discussing source events (which are, overall, less likely than goal events)? This is a question we address using the Information Bottleneck approach: whether the observed distribution of spatial adverb paradigms across languages can be explained merely by the prior probability of the various categories or whether there is evidence for a cognitive cost that differentially penalizes confusing source words with place words.

2.2. Past work on the efficient structure of semantic spaces

A large body of work has explored the way that languages efficiently break up semantic categories. This work typically quantifies a trade-off between complexity and informativity—earlier by measuring complexity and informativity explicitly (Regier, Kemp, & Kay, 2015) and, more recently, through the information bottleneck (Strouse & Schwab, 2017; Tishby et al., 2000; Tishby & Zaslavsky, 2015; Zaslavsky, Kemp, Tishby, & Regier, 2019). There is strong evidence that languages efficiently navigate this tradeoff by being maximally informative, given the complexity of the system.

One rich domain for these explorations has been color words (Gibson et al., 2017; Regier et al., 2007; Zaslavsky, Kemp, Regier, & Tishby, 2018). Some languages have more color words than others, with more precise boundaries. As a result, these languages can more precisely pick out particular parts of the color space. But that precision comes at a cost: greater complexity. If languages are efficient, there should be no languages that have more complex color systems but have less precision. And, broadly, this seems to be the case.

Through typological comparison, it has been shown that there is efficient structure in the semantic spaces of kinship terms (Kemp & Regier, 2012), numerals (Xu, Liu, & Regier, 2020), names of animals (Zaslavsky, Regier, Tishby, & Kemp, 2019), and season words (Kemp et al., 2019). There has also been work showing evidence for efficient structure in the organization of various grammatical systems, including indefinite pronouns (Denić, Steinert-Threlkeld, & Szymanik, 2021), tense systems (Mollica et al., 2020, 2021), quantifiers (Steinert-Threlkeld, 2020), and person systems (Zaslavsky, Maldonado, & Culbertson, 2021).

2.3. Typological database

We use a database of spatial demonstratives that appears in Nintemann et al. (2020). Their work's methodology draws on Stolz et al. (2017), which explores spatial interrogatives (e.g., *where*, *whither*, *whence*) typologically. Nintemann et al. (2020) report on the spatial demonstrative systems of languages across 5 major world regions (Africa, Americas, Asia, Europe, Oceania), drawing on reference grammars for the majority of evidence. For each language, they report the wordforms for each relevant distance level and by place, goal, and source. They also annotate whether the system has simple distance levels (e.g., proximal, medial, distal) or if it further sub-divides the spatial system along other dimensions, such as visible/invisible.

Nintemann et al. (2020) were interested, in part, in comparing the spatial demonstrative system to the spatial interrogative system. Because we are primarily interested in the structure of the spatial demonstrative system, we do not consider the spatial interrogative system. Nintemann et al. (2020) test several hypotheses in their work, focusing in particular on syncretism in orientations: that is, whether and how different orientations are referred to by the same form.

Nintemann et al. (2020) use the following schema for identifying syncretism in the spatial demonstrative system:

1. $P \neq G \neq S$
2. $P = G \neq S$
3. $P \neq G = S$
4. $P = S \neq G$
5. $P = G = S$

In the list above, an equal sign indicates that the two orientations on both sides of the sign are referred to by the same demonstrative. For instance, $P = G \neq S$ means that Place (P) and Goal (G) are referred to by the same demonstrative, but not Source (S) and Goal (G). For example, in English, 'here' can refer to both 'at here' (Place) and 'to here' (Goal), but 'from here' only refers to Source.

Nintemann et al. (2020) hypothesize that languages employ the same syncretism pattern in spatial demonstratives across different distance levels. For instance, if a language has syncretism for Place and Goal at one distance level, it tends to also have syncretism at other distance levels. We will address this prediction and its relation to the framework in Section 6.1. They also hypothesize that most languages employ one of three syncretism patterns: using the same demonstrative

for all orientations (Pattern 5 above), using different demonstratives for all orientations (Pattern 1 above), or using the same demonstrative for Place and Goal but another for Source (Pattern 2 above). As a third hypothesis, they offer that there is a rise in length of forms from Place via Goal to Source.

We argue that certain of these hypotheses fall out of information-theoretic motivations and can be tested using our framework. The last of these, that construction length rises from Place to Goal to Source, falls straightforwardly out of the general relationship between frequency and length (Haspelmath, 2021; Zipf, 1949) and, because Nintemann et al. (2020) show this relationship robustly in their work, we do not focus on it here. But we note that this pattern is broadly consistent with communicatively efficient patterns. We focus on two key hypotheses: that certain paradigms are much more common across languages than others and that languages prefer consistency in their syncretism patterns within spatial demonstratives. This second point, which we refer to as **consistency**, requires an additional step beyond the pure information bottleneck approach that we develop in Section 6.1.

Below, we re-formulate these hypotheses into two key factors that we aim to explain using an information-theoretic approach.

Orientation syncretism patterns. As mentioned above, the third hypothesis states that Patterns 1, 2, and 5 are widely attested in world languages. It is indeed the case in Nintemann et al. (2020): it is common for languages to have a three-way split between Place, Goal, and Source (Pattern 1), common to have no split (Pattern 5), and common for syncretism between Place and Goal with Source marked separately (Pattern 2). Meanwhile, it is rare for there to be syncretism between source and goal (with place as an outlier) or between source and place (with goal as an outlier). In spite of being rare, they are indeed attested in Nintemann et al. (2020): Balese for Pattern 3, and Northern Saami for Pattern 4. We show that in the information bottleneck approach, if a language values informativity more compared with complexity, Pattern 1 emerges, whereas if a language values complexity more than informativity, Pattern 5 emerges. Pattern 2 is likely to stem from the cognitive phenomenon of source-goal asymmetry mentioned in previous sections and operationalized here as a higher penalty for confusing Place and Source than for confusing Place and Goal.

Consistency. The second hypothesis states that the same syncretism pattern is likely to occur in both near distals and far distals. That is, it is unlikely (but not impossible) to be the case that a language follows one pattern (e.g., $P = G = S$) in the near distals and another pattern (e.g., $P \neq G = S$) in the far distals. We show that the information bottleneck approach, as currently proposed, does not necessarily lead to this result. Thus, in our third experiment, we propose adding a consistency component to the system. By adding a preference for consistency, this pattern emerges.

3. Information-theoretic formulation

We aim to model spatial demonstratives using the Information Bottleneck (IB) framework which was introduced into linguistics by Zaslavsky et al. (2018). The IB model has its ultimate origins in the physics literature (Tishby et al., 2000) and is a special case of the general theory of lossy compression (Berger, 2003; Harremoës & Tishby, 2007; Shannon, 1959). In this section, we review the framework and how we will apply it to our domain. We also discuss a number of extensions to the Information Bottleneck that will prove necessary to get an adequate model of the typological data.

3.1. Basic information bottleneck

Applied to natural language, the Information Bottleneck is a model of a communicative system: a mapping from mental representations of meaning to discrete forms. These discrete forms may be distinguished from each other by syntactic, morphological, or lexical means. On its own, the Information Bottleneck fundamentally provides a model of *what distinctions are made* without specifying *how* they are made.

The Information Bottleneck formalizes the intuition that an 'optimal' communicative system balances informativity on one hand with complexity on the other. Schematically, an optimal system should minimize an **objective function** of the form:

$$\text{Complexity} - \beta \cdot \text{Informativity}, \tag{1}$$

where β is a scalar value that determines how much a unit of Complexity should be traded off with a unit of Informativity. The scalar β can be seen as a conversion factor that converts Informativity into a common currency with Complexity. The meaning of Eq. (1) is that languages minimize Complexity and maximize Informativity.

The Information Bottleneck allows us to give precise definitions for both Informativity and Complexity in terms of information theory. Thus, it allows us (1) to evaluate the comparative optimality of real systems, by plugging them into Eq. (1), and (2) to derive mathematically optimal systems to which real systems can be compared, by finding minima of Eq. (1). The key information-theoretic concept behind the IB framework is mutual information, defined below.

Mutual information. Both the Informativity and Complexity terms in the IB framework will be defined in terms of **mutual information**: a statistical quantity that gives the most general measure of dependence between two random variables (Cover & Thomas, 2006). Given two random variables X and Y, the mutual information between X and Y is an average log-likelihood ratio:

$$\mathrm{I}[X : Y] = \sum_x \sum_y p(x, y) \log \frac{p(y \mid x)}{p(y)}, \tag{2}$$

where the x and y are possible values of the random variables X and Y respectively.

Intuitively, Eq. (2) quantifies how much the uncertainty about Y decreases when you know the value of X. When X and Y are independent – such that knowing X gives no information at all about Y – their mutual information is zero. The more dependent they are on each other, the higher their mutual information. As we will see below, mutual information admits a number of different interpretations, allowing it to serve both as a measure of Informativity and Complexity when applied to different variables.

Mathematical setup. Following the convention of Zaslavsky et al. (2018), we define the information bottleneck using three random variables:

1. U, a random variable over **world states**. In the case of spatial demonstratives, a world state is a pair $\langle r, \theta \rangle$ of a distance level r, consisting of one of a set of D discrete distance levels, and orientation θ, consisting of one of PLACE, GOAL, and SOURCE. Thus, there are $3 \times D$ world states. In this work we set the number of distance levels to $D = 3$.[1] The world state may refer to an objective physical space, or to a shared subjective space within a discourse.
2. M, a random variable over **meanings**: mental representations of world states. Each meaning corresponds to a distribution on world states, parameterized as below in Section 3.2. We assume that the relationship between world states and meanings

[1] In order to model actual world states, the distance levels r should likely vary continuously. We use a discrete number of distance levels for mathematical tractability.

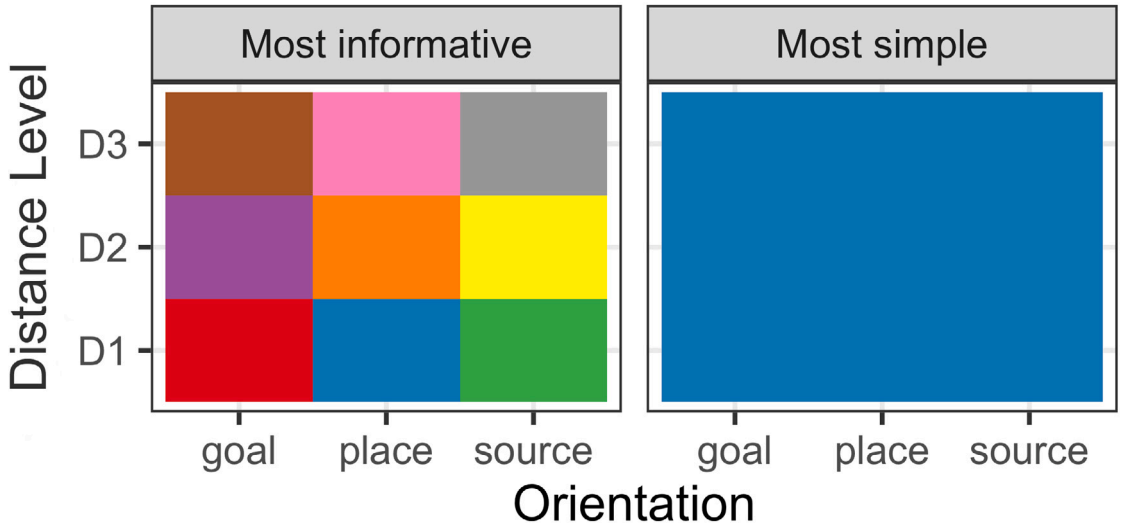

Fig. 1. An illustration of the complexity-informativity tradeoff. The horizontal axes indicate the orientation PLACE/GOAL/SOURCE distinctions. The vertical axes indicate the distance level. Each color represents a word. **Left**: a system with maximal informativity. In this case, each meaning has its own unique word. For a listener, there will be no confusion on what a speaker means when she utters a word. However, this system is also maximally complex, as it contains a total of 9 words. **Right**: a system with maximal simplicity (or minimal complexity). In this case, all 9 meanings are expressed by only 1 word. This system is the simplest but also the least informative.

is fixed and independent of language; presumably it is set by perceptual systems that map between world states and mental representations.

3. W, a random variable over discrete **words**. A **communicative system** q consists of a stochastic mapping from meanings to words:

   $$q : M \to W,$$

   or equivalently a conditional distribution $q(w|m)$ on words given meanings. The Information Bottleneck allows us to derive optimal systems q for a given M and U.

Following previous work (e.g. Zaslavsky et al., 2018), we assume that the 'need distribution' on meanings $p(m)$ and the conditional distribution on world states given meanings $p(u|m)$ – which reflects the perceptual relationship between cognitive meanings and world states – are fixed across languages. In reality, different cultural or geographic constraints may mean that the need distribution is not fixed across languages. Indeed, at least in the domain of color, studies such as Twomey, Roberts, Brainard, and Plotkin (2021) do suggest the need distribution varies across languages, and such variations seem to be related to geographic location and ecologic region. We leave it to future work to incorporate these differences into this approach.

Given this setting, the IB optimality of a system q with respect to meanings M and world states U is the difference of mutual information.

$$J_{\mathrm{IB}}[q] = \underbrace{\mathrm{I}[M : W]}_{\text{Complexity}} - \beta \cdot \underbrace{\mathrm{I}[W : U]}_{\text{Informativity}}. \tag{3}$$

In the equation, the mutual information between words and world states $\mathrm{I}[W : U]$ plays the role of Informativity, while the mutual information between *meanings* and words $\mathrm{I}[M : W]$ plays the role of Complexity. Below, we review the meanings and motivations for these terms.

Motivation: Informativity. The Informativity term uses the interpretation of mutual information as quantifying the amount of information contained in one variable about another. In this case, it gives the amount of information in the word W about the world state U. This interpretation is valid because mutual information quantifies the average reduction in uncertainty about the world state U that happens upon observation of a word W.

Motivation: Complexity. The complexity of a system, on the other hand, is defined using the mutual information of *meanings* and words. In its appearance here, mutual information represents the complexity of the *mapping* between meanings and words. It can be interpreted the number of distinctions about meaning encoded in the system.

Mutual information has been used in this sense in neuroscience as a general measure of complexity for action policies, that is, mappings from states to actions as implemented by agents (see Bhui, Lai, & Gershman, 2021; Lai & Gershman, 2021, for a review). It correlates with empirical measures of cognitive effort (Zénon, Solopchuk, & Pezzulo, 2019), and plays a role in models of the complexity of language production (Futrell, 2021). Under this measure, a communicative system has complexity 0 when all meanings are mapped to a single word—in that case, no computation at all is required to specify the word. A system has maximal complexity when each meaning is mapped to an individual unique word (see Fig. 1 for an illustration).

In the plain Information Bottleneck, complexity is quantified only by mutual information between meanings and words. However, it is possible that a more complete theory of communicative systems in natural language will require some more elaborate notion of complexity. Below, we will see that there is evidence that a full model of deictic words will require further constraints on *nondeterminism* and *consistency*, which can be implemented as an additional terms in an extended optimization objective.

Notion of optimality. We study the optimality of systems where optimality is defined using Eq. (3). This is a **multi-objective** optimization problem, meaning that multiple notions of optimality (informativity and complexity) are being optimized simultaneously. Furthermore, since informativity and complexity are related to each other mathematically, they trade off with each other. In particular, informativity is upper bounded by complexity:

$$\underbrace{\mathrm{I}[W : U]}_{\text{Informativity}} \leq \underbrace{\mathrm{I}[M : W]}_{\text{Complexity}},$$

so it is not possible to achieve arbitrarily high informativity with a system of low complexity. Essentially, by maximizing the mutual information of words and world states while minimizing the mutual information of meanings and words, we are finding a system which encodes *only* those distinctions of meaning which are relevant for distinguishing world states.

Table 2
The need distribution $p(m)$ based on spatial demonstrative frequency in Finnish, provided along with the actual spatial demonstratives.

	Goal	Place	Source
D3	tuonne 0.073	tuolla 0.030	tuolta 0.006
D2	sinne 0.185	siellä 0.072	sieltä 0.017
D1	tänne 0.393	täällä 0.160	täältä 0.065

When we plot a space defined by informativity and complexity, as we will do in Fig. 5, we see two regions: (1) an **achievable** region of possible systems below the black line, and (2) an **unachievable** region above the black line where there is no possible system q that simultaneously achieves the given value of informativity and complexity. The black line between these regions is the **efficient frontier**, defining a set of systems which are the best possible within the bounds of what is achievable in terms of IB optimality.

3.2. Parameterization

Summarizing the above, an Information Bottleneck model of a communicative system requires that we formulate two distributions: (1) a prior distribution on meanings $p(m)$ indicating how often a speaker needs to express a meaning, and (2) a conditional distribution on world states given meanings $p(u \mid m)$. We can then study the optimality of different systems $q(w \mid m)$.

Need distribution $p(m)$. We model the set of possible meanings using a three-way distinction of manner (place vs. goal vs. source) and a three-way distinction of distance (proximal, distal, and far-distal), giving $3 \times 3 = 9$ total possible meanings.

We set the prior probabilities on meanings $p(m)$ empirically, estimating the probability of each meaning $p(m)$ from the frequencies of the corresponding words in Finnish in Lexiteria.[2] We use these word frequencies because Finnish has a full unambiguous 3×3 distinction in its deictic words. In particular, it does not have syncretism of place and goal. The resulting probabilities are shown in Table 2. We use Finnish data in our main analyses for convenience, but it should not be assumed that the distribution of these terms will be the same across languages as they are in Finnish. The particular choice of our prior is not crucial to our findings, as discussed below, where we compare among priors.

Conditional distribution on world states $p(u \mid m)$. A world state is a tuple of a distance d and an orientation n. We model the distribution on world states conditional on meanings using **cost functions** which define a cost for confusing one distance d with another distance d', or one orientation n with another orientation n'.

The cost for confusing distances d and d' is simply the absolute value of the difference between them:

$$C_{dd'} = |d - d'|. \tag{4}$$

The cost for confusing two orientations n and n' is given by three cost values C_{PG}, C_{PS}, and C_{GS} which are non-negative real numbers that define the cost for confusing place and goal, place and source, and goal and source respectively:

$$C_{nn'} = \begin{cases} 0 & \text{if } n = n' \\ C_{PG} & \text{if } n = \text{place and } n' = \text{goal or vice versa} \\ C_{PS} & \text{if } n = \text{place and } n' = \text{source or vice versa} \\ C_{GS} & \text{if } n = \text{goal and } n' = \text{source or vice versa.} \end{cases}$$

In addition, we have the constraint that the three cost values fall on a line (that is, the maximal cost of the three must be equal to the sum of the smaller two). We experiment with different values for the orientation costs in Experiment 2.

Finally, the costs are combined to form the probability distribution on world states given meanings using exponential decay with a decay rate parameter μ:

$$p(u \mid m) \propto \mu^{C_{dd'} + C_{nn'}}. \tag{5}$$

This means that a meaning m which corresponds to distance d and orientation n will give probability primarily to world states with matching d' and n', and also to other world states that are similar in distance and orientation. A sample distribution $p(u \mid m)$ under two different decay parameters is illustrated in Fig. 2.

Table 3
Free parameters of the information bottleneck and our model of the semantic domain of spatial demonstratives.

Parameter	Meaning
β	Tradeoff between informativity and complexity
μ	Decay in $p(u\|m)$
C_{PG}	Cost for confusing PLACE and GOAL
C_{PS}	Cost for confusing PLACE and SOURCE
C_{GS}	Cost for confusing GOAL and SOURCE

Summary of parameters. Table 3 shows the free parameters of the model. The need distribution $p(m)$ has additional parameters which are set empirically.

3.3. Generalizing the Information Bottleneck with further constraints

In order to model spatial demonstratives, the simple Complexity constraint of the basic Information Bottleneck will not be enough. This is for a combination of reasons involving the way that spatial demonstrative data is coded, as well as genuine linguistic phenomena that depart from the predictions of the basic Information Bottleneck.

Determinism. The Information Bottleneck generally predicts that optimal systems are **nondeterministic**: the mapping from a meaning to a word is a probabilistic function. This is an advantage in the case of semantic domains such as color words, where the mapping from perceptual space to color categories is indeed nondeterministic both across and within speakers: many speakers are unable to produce consistent color names for regions 'on the boundary' between color categories, and this variability is reflected straightforwardly in data sources such as the World Color Survey which give trial-level information on color labels provided by participants to color stimuli (Kay, Berlin, Maffi, Merrifield, & Cook, 2009).

In the case of spatial demonstratives, however, the available data sources such as Nintemann et al. (2020) provide only the mapping from distance levels and orientations to words. While in many cases this is a nondeterministic one-to-many mapping, we do not have data from which we could estimate the probability of using one particular word given one particular location described. The general IB framework does make fine-grained probabilistic predictions about how words will be used to describe meanings stochastically, but the available data do not allow these predictions to be tested. We believe that the full probabilistic structure of spatial demonstrative systems could only be investigated through quantitative experiments with a highly controlled meaning space.

In order to model the existing categorical descriptions of spatial demonstrative systems, we introduce a **determinism** constraint into the IB framework: we constrain systems to have a deterministic mapping from meanings to words. This Deterministic Information Bottleneck has been studied in the machine learning literature by Strouse and Schwab (2017). When considering only deterministic systems, the Information Bottleneck objective reduces to

$$J_{\text{DIB}} = \text{H}[W] - \beta \cdot \text{I}[W : U], \tag{6}$$

[2] https://lexiteria.com/word_frequency_list.html

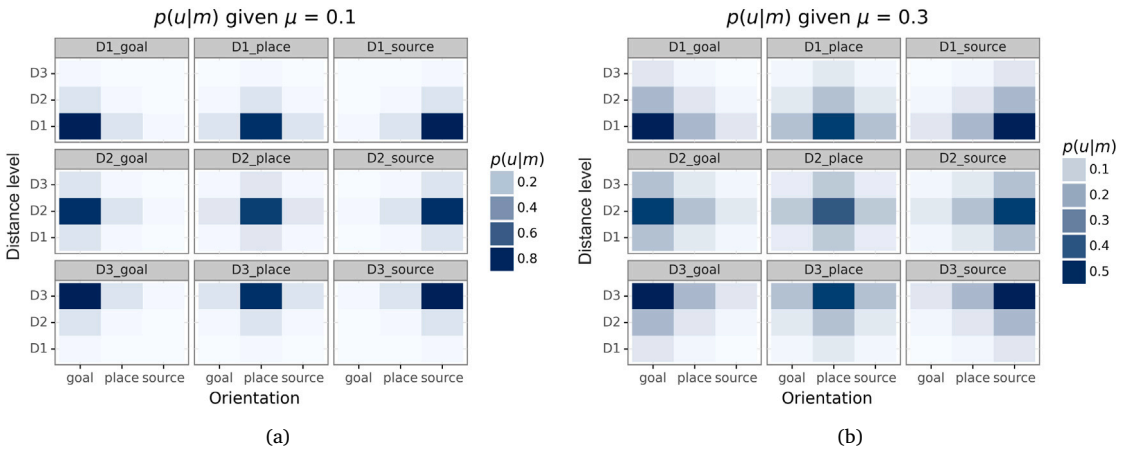

Fig. 2. Sample distributions of world states U (in x- and y-axes), conditioned on meaning M (in each facet), when the decay parameter $\mu = 0.1$ (left) and $\mu = 0.3$ (right). The color represents the conditional probability, from 0 (bright) to 1 (dark). Values in each facet sum up to 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where the Complexity term is replaced with the entropy over words produced, $\mathrm{H}[W]$.[3]

In order to find optimal deterministic systems, we can simply iterate through all the possible mappings from meanings to words and find the ones that score the highest on J_{DIB}. The number of all possible unique systems mapping m meanings to words can be calculated as the Stirling number of the second kind (Sequence A008277 in the Online Encyclopedia of Integer Sequences: Sloane et al., 2018).

For example, the number of possible systems given 9 meanings (3 distance levels and 3 orientations) is 21,146. We enumerate all these systems and compute their respective efficiency and informativity. Since these are all possible systems under the fixed number of meanings, the real systems will be a subset of them.

Consistency. Another factor that constrains spatial demonstrative systems, but which is not encompassed in the basic IB framework, is consistency or naturalness (Saldana, Herce, & Bickel, 2022). This notion captures the idea that not all paradigms are equally easy for humans to learn and use, even if they have similar complexity. In particular, it has been shown that paradigms with various kinds of similarity in their structure (what are often called "natural patterns", as in Baerman, 2004; Corbett, 2015; Noyer, 1992) are more common (Cysouw, 2009; Pertsova, 2007) and easier to learn (Johnson, Gao, Smith, Rabagliati, & Culbertson, 2021; Maldonado & Culbertson, 2020b; Nevins, 2015; Nevins, Rodrigues, & Tang, 2015; Noyer, 1992; Pertsova, 2011, 2012), which has been posited to drive typological patterns (Fedzechkina, Jaeger, & Newport, 2012; Hupp, Sloutsky, & Culicover, 2009; Johnson et al., 2021; Maldonado & Culbertson, 2020a, 2020b; Maldonado, Saldana, & Culbertson, 2020; Martin & Culbertson, 2020; Saldana et al., 2022). These constraints reflect a kind of 'system pressure' which may be distinct from communicative pressures (Haspelmath, 2014).

We operationalize these ideas as **consistency**. We say a deictic system is consistent when it has the same pattern of distinctions in each distance level and in each orientation. For example, in English, the word "here" is used to refer to both "place" and "goal" in the proximal level; similarly, the word "there" is used to also refer to both "place" and "goal" in the distance level. In addition, to indicate the orientation SOURCE, at both distance levels, the preposition "from" is used; similarly, to indicate orientations SOURCE or GOAL, no preposition is required. Therefore, English spatial demonstratives are consistent. In contrast, most of the enumerated deterministic systems are not consistent. An example is shown Fig. 3: here the consistent paradigm of English is juxtaposed with a random inconsistent paradigm with the same number of words.

It is not necessarily the case that optimization of the IB objective will produce consistent systems. This is because in the IB framework, complexity of a system is measured using mutual information, which does not explicitly penalize inconsistency. We will ultimately find that the attested linguistic systems are best modeled by an extended IB optimization that includes consistency as an additional constraint.

In order to quantify consistency of real and simulated systems, we introduce a **consistency score**, defined as the sum of the number of unique SOURCE/PLACE/GOAL patterns plus the number of unique distance level patterns in a given language, similar to the enumerative complexity in Ackerman and Malouf (2013). For instance, since English only has the pattern of "ABB" (using the same word to refer to both PLACE and GOAL) in all deictic levels and "CDD" in all orientations, English has a consistency score of 2, whereas the simulated system in Fig. 3 has a consistency score of 6. Hence, a lower consistency score indicates higher degree of consistency.

We use the consistency score for two purposes: (1) to quantify the actual consistency of attested system when compared to random simulated systems and to optimal simulated systems, and (2) to derive optimal systems where consistency is included as an additional constraint. The (deterministic) IB objective including consistency is

$$J_{\mathrm{Consistency}} = \mathrm{H}[W] - \beta \cdot \mathrm{I}[W : U] + \gamma \cdot \mathrm{S}[M : W], \tag{7}$$

where $\mathrm{S}[M : W]$ is the consistency score, and γ is a scalar parameter indicating how strongly inconsistency is penalized. When $\gamma = 0$, the consistency constraint has no effect. To preview the results below, we find a good fit to the attested deictic systems with $\gamma = 1$.

4. Experiment 1: Basic Information Bottleneck

In the first experiment, we compute the information plane for each of the real systems in our data set, as well as for simulated systems. The information plane plot reveals how close real deictic systems are to optimal systems as predicted by the basic Information Bottleneck, as described in Section 3.1.

We used the Appendices from Nintemann et al. (2020) in order to construct a machine-readable database of place demonstratives. We only analyzed languages where spatial demonstratives are used

[3] The entropy $\mathrm{H}[W] = -\sum_w p(w) \log p(w)$ represents the uncertainty in the random variable W. The IB objective reduces to Eq. (6) for deterministic systems because (1) the mutual information $\mathrm{I}[M : W] = \mathrm{H}[W] - \mathrm{H}[W \mid M]$ and (2) for deterministic systems, $\mathrm{H}[W \mid M] = 0$. The DIB objective can also be thought of as adding another term $\mathrm{H}[W \mid M]$ to the plain IB objective in Eq. (3), thus penalizing systems with nondeterminism (Strouse & Schwab, 2017).

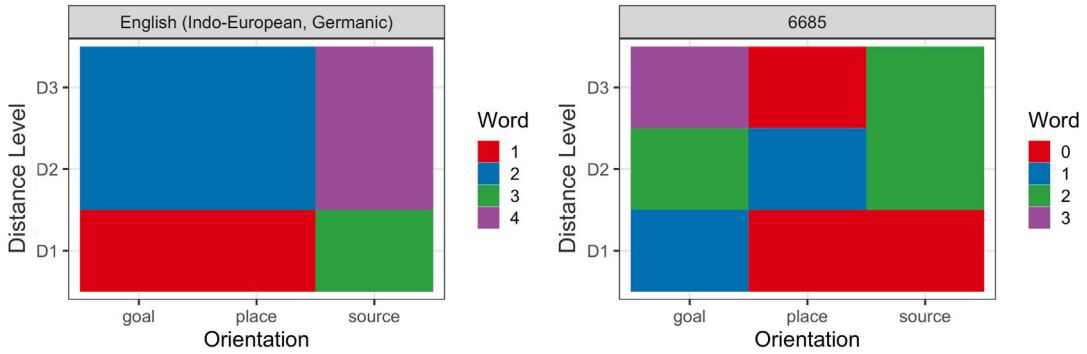

Fig. 3. Examples of consistent (left) and inconsistent (right) paradigms. The horizontal axis indicates the PLACE/SOURCE/GOAL distinctions. The vertical axis denotes the distance level. English (left) has syncretism for PLACE and SOURCE at all 3 distance levels, whereas a simulated paradigm (right) does not.

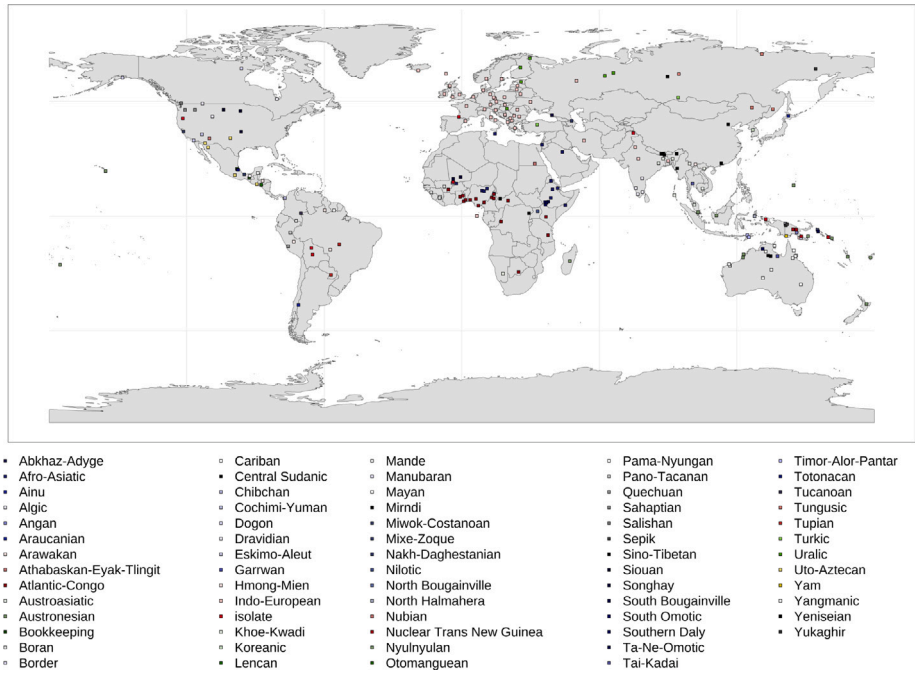

Fig. 4. Geographic distribution of the 220 languages investigated in this study, generated in R (R Core Team, 2014). Colors denote language families. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Source: The coordinates are taken from Glottolog (Hammarström & Forkel, 2022).

solely relative to an imaginary deictic center (Section 2.1) and left out languages that employ a speaker/listener-centric system or those that adapt a combination of deictic-centric and speaker/listener-centric approach. For instance, languages such as English and Maltese are included in our analysis, whereas languages such as Japanese are not. This left 220 languages for analysis (see Fig. 4 for the geographical distribution of the 220 languages).

4.1. Data processing

When there are multiple options for a particular cell, we concatenate those options together. For instance, Distal II in Tok Pisin is "longwe liklik" or "longwe" for each of place, goal, and source. When we concatenate these together, we see that there is syncretism between place, goal, and source. In Yuracaré, on the other hand, proximal Place is "ani" and Source is "ani" or "an=chi". Because the language has the option to distinguish place and source, we consider these cells as separate. Note that, for the analysis, all that matters is whether a particular cell is the same or different from another cell.

We highlight one coding choice which will have relevance for interpreting our figures: when a language has fewer than three distance levels, we assume as a convention that the second distance level encompasses both D2 and D3. This was motivated purely by the need to have a convention for the representation of two-level languages, and not motivated by any particular linguistic consideration.

Table 4(a) shows the deictic system for Tamil as presented by Nintemann et al. (2020), and Table 4(b) shows the same information re-coded, as it is presented to our model: the only thing that matters is whether a particular form is the same or different to another form in the paradigm. For our purposes, distinctions of form may be syntactic (involving multiple words) or morphological.

4.2. Choice of prior and parameters

We set the need distribution over meanings using the frequencies derived from Lexiteria for Finnish. We chose Finnish because it has canonical separation between three distance levels and between place/goal/source, with a single word associated with each. We consider a number of alternative need distributions below, and two alternative corpora besides Lexiteria in Appendix C.

In our main analysis in Section 4.3, we first set our maximum number of distance levels to be 3, decay parameter μ to be 0.2, orientation

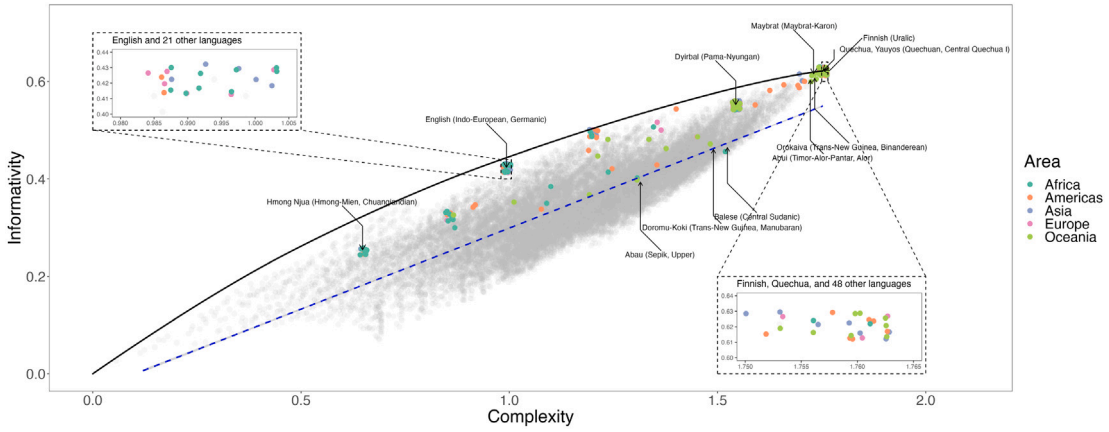

Fig. 5. Each colored point on the information plane represents an attested deictic system, and the gray points represent all the 21,146 hypothetically possible deterministic systems. The efficient frontier is marked by the black line. The points are jittered to avoid overlap. The blue dashed line represents the expected informativity among simulated systems given a complexity. Some languages close to the frontier or far from the frontier are labeled as an illustration. Since the points are heavily clustered, a zoomed-in view of two of the clusters are presented in the two panels: spatial deictic systems sharing similar complexity and informativity of English (left) and Finnish/Quechua (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Tamil place demonstratives (a) and coded demonstratives (b).

	Goal	Place	Source
distal	angee	angee	angeyrundu
proximal	ingee	ingee	ingeyrundu

(a) Tamil place demonstratives

	Goal	Place	Source
distal	C	C	D
proximal	A	A	B

(b) Tamil place demonstratives, coded.

confusion cost C_{PG} and C_{PS} to be 0.8 and 1.32, respectively. These parameters are selected because they broadly reflect the properties of attested spatial deictic systems. The maximum number of distance levels are set as 3 because most of our attested systems have fewer than 3 distinct distance levels. The confusion cost C_{PS} is set to be larger than C_{PG} because of the rich literature on the asymmetry between SOURCE and GOAL, as discussed in Section 2.1. The specific numbers of C_{PS} and C_{PG} are fitted to the data, that is, they are the ones that place the attested systems as close to the optimal frontier as possible. The choice of decay parameter μ does not affect the results for $\mu < 0.7$, as demonstrated in Appendix A. Later in Section 5, we will explore the effect of each parameter separately and see how they affect the efficiency of attested deictic systems.

4.3. Results: Efficient frontier

In Fig. 5, we plot the information plane for the set of 220 real systems, along with 21,146 random simulated ones. The black line indicates the efficient frontier if we allow the systems to be non-deterministic. As also pointed out in Zaslavsky et al. (2018), allowing non-determinism extends the efficient frontier by offering additional degrees of freedom that can be used to generate more efficient systems than are available under a deterministic approach. We also calculated the expected Informativity among simulated systems, under a given Complexity, which is plotted as the blue dashed line in the figure.

The real systems typically fall close to the optimal frontier, although there are some exceptions. Of the 220 systems in the sample, only 4 of them have below-expected Informativity relative to their Complexity ($I[M:W]$). These languages are Balese, Doromu-Koki, Bunoge Dogon, and Abau. Balese, the language that falls farthest from optimal on the plot, is noteworthy for being the only language in Nintemann et al. (2020)'s sample in which goal/source syncretism is required. Doromu-Koki, Bunoge Dogon, and Abau all have varying degrees of goal/source syncretism, as well. As we will see in Experiment 2, the reason that these languages appear sub-optimal is that we assume higher cost for confusing GOAL with SOURCE than for any other orientation confusion, and so a language with goal/source syncretism is dispreferred.

Relative to the majority of simulated systems, what causes the apparent optimality of real systems? One major factor is that natural language deictic systems rarely have jumps in distance levels. That is, a language is unlikely to use the same word for "here" as for "far over there" but a different word for "there". From the perspective of our information-theoretic framework, that makes sense since the cost of confusing meanings which are spatially distant is high. But the cost of confusing two nearby distance levels is relatively lower. This analysis further raises the question: which of the parameters in our model are driving these relationships? In Experiment 2, we undertake a series of explorations to uncover which choice of parameters would make the real systems appear most optimal.

In addition to asking what makes deictic systems closer to the efficient frontier than the random systems, we can ask why real deictic systems are often not exactly on the efficient frontier. We will take up this question in Section 6.1 where we discuss an additional constraint that appears to be operative in natural language—the constraint of consistency.

5. Experiment 2: Exploring factors that affect optimality

Experiment 1 found that natural deictic systems are very often closer to the efficient frontier than random systems for a particular setting of the IB model parameters. But what drives these results? Does the apparent efficiency of real systems emerge mostly from assumptions about the need distribution of the words? Or does it depend more on the cost parameters that penalize confusing, e.g., place with goal or goal with source? We take up these questions here.

In order to better characterize how different parameters affect the apparent optimality of communicative systems, we perform here comprehensive searches through parameter space, and study how the apparent optimality of systems changes under different parameter settings. Our goal here is to determine the minimal requirements on the parameters such that natural languages are well-modeled by the IB optimization.

In order to compare different model configurations, we need a metric for the optimality of a set of systems under those configurations.

With such a metric, we can compare the score across different sets of parameters. We adopt the metric of **a generalized version of Normalized Information Distance** (gNID) proposed in Zaslavsky et al. (2018). gNID, bounded by 0 and 1, reflects the distance between an attested system and the closest optimal system.

As shown in Zaslavsky et al. (2018), we can match each attested spatial deictic system $q_\ell(w|m)$ representing language ℓ with an optimal, non-deterministic spatial deictic system $q_{\beta_\ell}(w|m)$ by finding the trade-off parameter β_ℓ that minimizes the difference in the efficiency score between the attested system and the optimal system (Eq. (8)):

$$\beta_\ell = \underset{\beta}{\operatorname{argmin}}[(\mathrm{I}_{q_\beta}[W;M]-\beta\cdot\mathrm{I}_{q_\beta}[U;M])-(\mathrm{I}_{q_\ell}[W;M]-\beta\cdot\mathrm{I}_{q_\ell}[U;M])]. \quad (8)$$

Then, the gNID between $q_\ell(w|m)$ and $q_{\beta_\ell}(w|m)$ can be calculated by Eq. (9), where W'_ℓ and W'_{β_ℓ} denote random variables of other possible spatial demonstratives that could be produced given a meaning m:

$$\mathrm{gNID}(W_l, W_{\beta_\ell}) = 1 - \frac{\mathrm{I}[W_l, W_{\beta_\ell}]}{\max\{\mathrm{I}[W_l, W'_l], \mathrm{I}[W_{\beta_\ell}, W'_{\beta_\ell}]\}} \quad (9)$$

The lower gNID an attested system has, the closer it is to the optimal frontier. For example, Kodiak Alutiiq (see Fig. 5) has a gNID of 10^{-7}, corresponding to the fact that it lies nearly on the optimal frontier, occupying the point of maximal informativity and maximal complexity. On the other hand, Balese (see Fig. 5) has a gNID of 0.446, corresponding to the fact that it is far away from the optimal frontier.

Using the methods described above, for each set of parameters, we compute the average gNID over all attested spatial deictic systems and then compare gNID across different sets of parameters. Below we report how varying each parameter will affect the average gNID, or the information distance to the optimal frontier, of attested spatial demonstrative systems.

Factor 1: PLACE/GOAL/SOURCE *costs.* First, we can ask about the cost functions for confusing PLACE, GOAL, and SOURCE. Based on the prior literature, there are two major claims to consider. First, there is reason to believe that PLACE is intermediate between GOAL and SOURCE, meaning the penalty for confusing SOURCE and GOAL should be high (Nikitina, 2009). Second, we predict that it is less costly to confuse GOAL with PLACE than to confuse SOURCE with PLACE. This prediction falls out of the cognitive science literature (Papafragou, 2006, 2010; Regier, 1996) suggesting that source-directed movement tends to be more marked.

Factor 2: Choice of need distribution. In Experiment 1, the need distribution $p(m)$ is based on Finnish word frequency data. The resulting distribution exhibits two important properties: (1) the frequency decreases as the distance level increases, indicating that we tend to talk about things happening further away less, and (2) the frequency of PLACE is greater than the frequency of GOAL, which is itself greater than the frequency of SOURCE. To investigate the role of the need distribution in our model of deictic systems, we measure the average gNID among all attested systems under different permutations of the marginal PLACE/GOAL/SOURCE distribution, while keeping the marginal distance level distribution constant. For instance, suppose the marginal distribution used in Experiment 1 is p(PLACE) = a, p(GOAL) = b, and p(SOURCE) = c, where $a > b > c$ (henceforth denoted as PLACE > GOAL > SOURCE). One of the permutations could be SOURCE > PLACE > GOAL, where p(SOURCE) = a, p(PLACE) = b, and p(GOAL) = c.

5.1. Methods

In this analysis, we keep all other parameters identical to those in Experiment 1 and grid search different combinations of PLACE/GOAL/SOURCE confusion costs and need distribution permutations.

For the confusion costs, instead of varying C_{PG} and C_{PS}, which implicitly assumes the first claim, we represent PLACE, GOAL, SOURCE as coordinates on a 1D line. Without loss of generality, we assign the coordinate of PLACE (C_P) as 0. We vary the coordinates of SOURCE (C_S) and GOAL (C_G) in the interval $[-5, 5]$ and calculate the confusion cost C_{ij} as $C_{ij} = |C_i - C_j|$, where $i, j \in \{P, G, S\}$. Regarding the first claim, in our formulation, if C_S and C_G have opposite signs, PLACE lies in the middle between SOURCE and GOAL ("place-centric" henceforth); otherwise, PLACE is said to lie outside SOURCE and GOAL ("place-marginal" henceforth). For the second claim, if $C_{PS} > C_{PG}$, the cost setting is favoring GOAL over SOURCE, and otherwise, it is favoring SOURCE over GOAL.

Meanwhile, under each permutation of {PLACE/GOAL/SOURCE}, we compute the average gNID among attested systems. In addition, we also repeat the same analysis with two additional types of need distributions: a uniform distribution, where each meaning has an equal need probability (coded as "uniform prior"), and a distribution where we keep the decay with the distance levels but even out the need distribution within each distance level (coded as "PLACE = PLACE = PLACE").

5.2. Results

The results for the major qualitative categories of parameter configurations are shown in Fig. 6. Here, the black dots represent the average gNID among attested deictic systems of a potential arrangement of (C_G, C_S), categorized by which orientation is favored and whether it is place-centric or place-marginal. The horizontal axis shows the relative coordinates of PLACE (gray), GOAL (blue), and SOURCE (red) on a number line. Overall, the results show that the IB model best fits the typological data if (1) PLACE is in the middle of SOURCE and GOAL, and (2) the cost of confusing PLACE and GOAL is set to be lower than that of confusing PLACE and SOURCE—both of which patterns have been independently observed and motivated in the typological and psycholinguistic literature.

The results of permuting the need distribution are shown in Fig. 7. Interestingly, the need distribution we encounter in reality, PLACE > GOAL > SOURCE, is the second worst in terms of optimality and only better than a uniform distribution.

5.3. Which matters more: need distribution or PLACE /GOAL/SOURCE costs?

To summarize, we found that (1) when the cost of confusing PLACE/GOAL/SOURCE with one another is $C_{GS} > C_{PS} > C_{PG}$, which is the configuration that reflects the cognitive bias towards GOAL found in the literature, spatial demonstrative systems attested in real languages are closest to the optimal frontier; and (2) when the need distribution of PLACE/GOAL/SOURCE is PLACE > GOAL > SOURCE, which is the configuration demonstrated in text corpora, spatial demonstrative systems attested in real languages are merely second to last closest to the optimal frontier, compared with other permutations of PLACE/GOAL/SOURCE need distributions. This leaves one question open: which factor affects the average distance to the optimal frontier more? To answer this question, we take the results from the grid search and perform a Bayesian random-effects regression, using the `brms` package (Bürkner, 2017) and the formula below:

$$\mathrm{gNID} \sim (1 \mid \text{PLACE/GOAL/SOURCE cost}) + (1 \mid \text{need distribution permutation}). \quad (10)$$

We use a random-effects regression so that separate variances are fit to describe the effects of need distribution vs. cost configuration.[4] If the absolute value of random intercept for one factor is greater than that for another, we can say the former matters more compared with the latter, since the gNID is affected by the former factor to a greater extent.

[4] In the model, we set the priors as weakly informative ones, 4 chains, and 2000 iterations per chain, including 1000 iterations for warming up. To this end, for each combination of the confusion costs and the need distribution permutation, we obtain 4000 sets of fitted random intercepts for each factor.

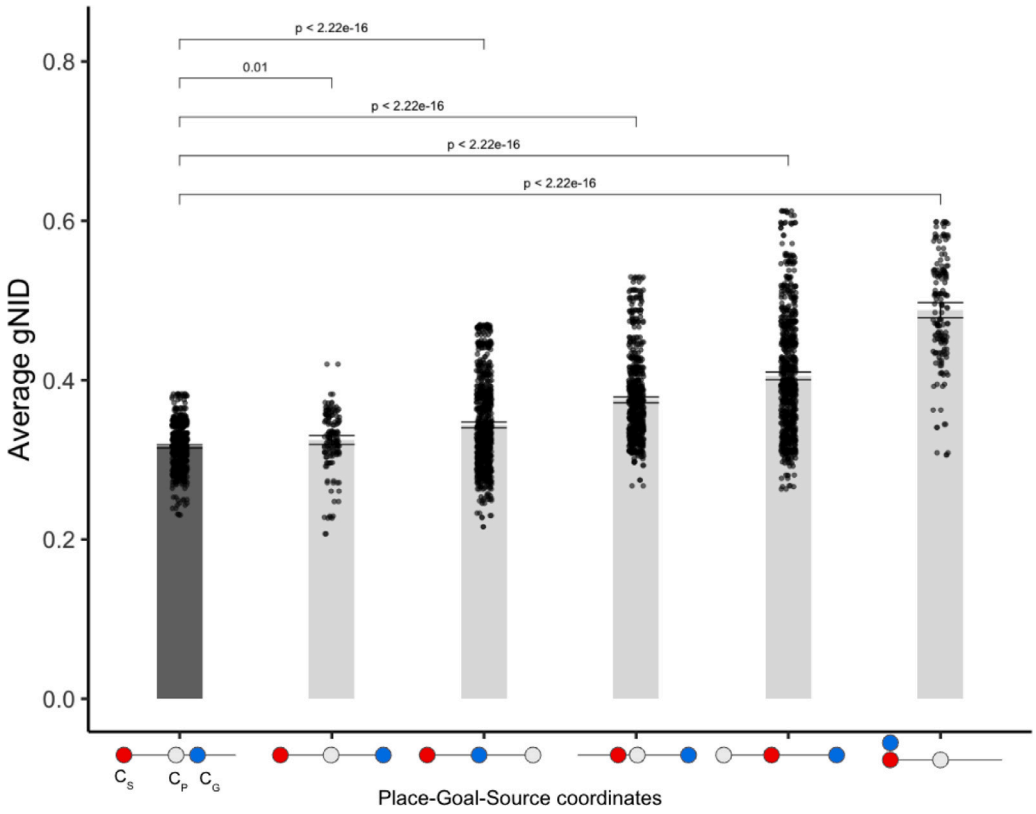

Fig. 6. Each black point represents the average gNID of attested spatial deictic systems, in an increasing order, under a (C_G, C_S) combination, categorized by the relative position of C_G (blue dot), C_S (red dot), and C_P (gray dot) on a number line. The error bar represents 95% bootstrapped confidence interval. First column (shaded dark gray, actually observed in human languages): GOAL favoring, PLACE centric ($C_{GS} > C_{PS} > C_{PG}$); Second column: no favoring, PLACE centric ($C_{GS} > C_{PS} = C_{PG}$); Third column: GOAL favoring, PLACE marginal ($C_{PS} > C_{PG}, C_{PS} > C_{GS}$); Fourth column: SOURCE favoring, PLACE centric ($C_{GS} > C_{PG} > C_{GS}$); Fifth column: SOURCE favoring, PLACE marginal ($C_{PG} > C_{PS}, C_{PG} > C_{GS}$); Sixth column: no favoring, PLACE marginal ($C_{PG} = C_{PS} > C_{GS} = 0$). The results show that if PLACE is in the middle of SOURCE and GOAL, and the cost of confusing PLACE and GOAL is set to be lower than that of confusing PLACE and SOURCE (i.e. $C_{GS} > C_{PS} = C_{PG}$), attested systems tend to be closer to the optimal frontier. The number shows the p-value in a t-test comparing the mean for the configuration observed in human languages (dark gray) with others. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The results are shown in Fig. 15. In each facet, a green dot represents the case where the random intercept of confusion cost is larger than the random intercept of need distribution orientation, under a combination of the two factors. There are a total of 192,000 data points in the figure, 154,403 of which are green, indicating that in most cases, the confusion cost affects the optimality of attested spatial demonstrative systems more than the need distribution. The relatively small effect of the need distribution mirrors the finding of Zaslavsky et al. (2018, §S7) that color naming systems appear optimal under several choices of need distribution.

5.4. Discussion

Varying the parameters of the IB model, we found that the attested spatial demonstrative systems behave more similarly to optimal systems predicted by the model when the cost of confusing SOURCE and GOAL is greater than that of confusing PLACE and SOURCE, which then is bigger than that of confusing PLACE and GOAL, a realistic constraint reported in the cognitive science literature. Although another realistic constraint on need distribution that PLACE is more frequently used than GOAL, which is more frequently used than SOURCE does not make the attested spatial demonstrative systems behave more similarly to predicted optimal systems, we demonstrate that the second constraint plays a minor role in affecting the optimality of attested systems, compared with the first constraint.

Our results also shed light on which model components are more important in terms of explaining deictic patterns. In particular, the results are relatively invariant to changes in the need probabilities (with the uniform prior performing better than the place > goal > source prior), whereas the model fit is dramatically worse when the orientation confusion costs are misspecified. This finding suggests that the reason for the goal–source asymmetry in typology arises from asymmetries in the cost function and not from need probability.

6. Experiment 3: Information bottleneck with consistency

In all the simulations above, there often emerge optimal systems that are unlike real systems. These are not consistent: e.g., they have different PLACE/GOAL/SOURCE paradigms for one distance level as compared to another. Examples are shown in Fig. 8.

There are no parameters in our model, nor in prior work on the information bottleneck, that can account for the fact that these inconsistent paradigms are dispreferred. In this section, we propose a new framework to account for this preference for consistency within the IB framework, through the addition of a new constraint.

6.1. Consistency

In this section, we examine whether the systems considered optimal under the information theory framework are the ones actually attested in real languages.

Although the demonstratives in different languages vary, how they are used to encode different meanings shows some commonality. For example, recall Table 1, which lists the spatial demonstratives in English and Maltese, respectively. They partition the space in the same way: one word for proximal PLACE and GOAL, one word for distal PLACE

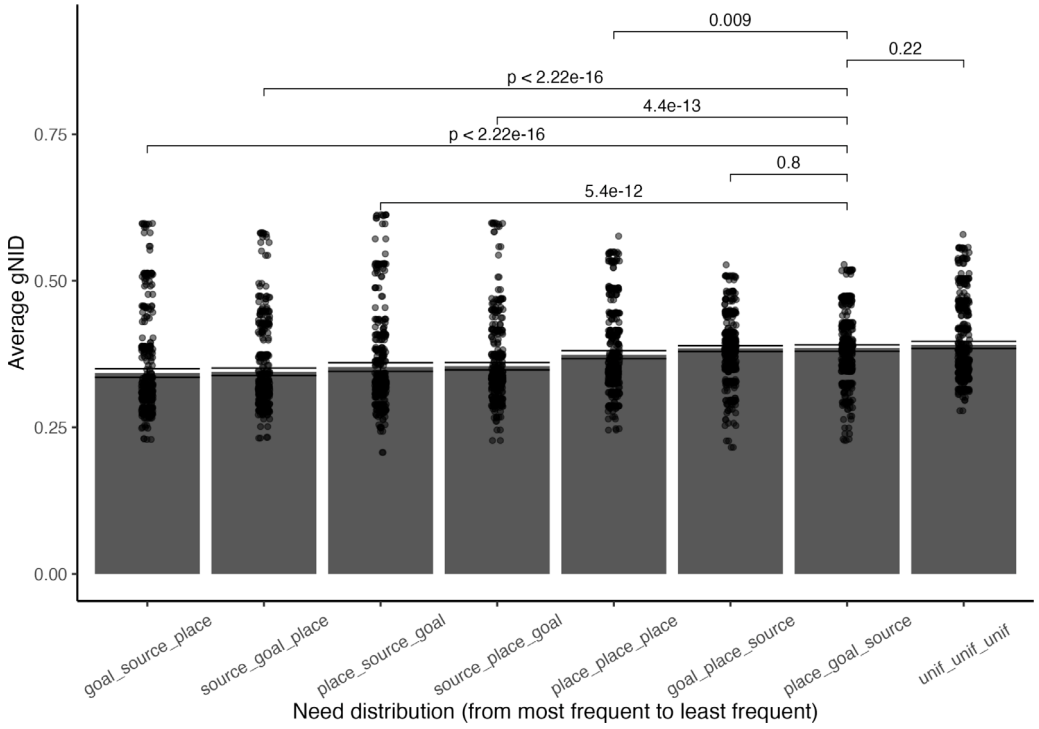

Fig. 7. Each black point represents the average gNID of attested spatial deictic systems, in an increasing order, under each need distribution. The error bar represents 95% bootstrapped confidence interval. First column: GOAL > SOURCE > PLACE; Second column: SOURCE > GOAL > PLACE; Third column: PLACE > SOURCE > GOAL; Fourth column: SOURCE > PLACE > GOAL; Fifth column: PLACE = PLACE = PLACE; Sixth column: GOAL > PLACE > SOURCE; Seventh column: PLACE > GOAL > SOURCE (the actual need distribution); Eighth column: uniform need distribution. The numbers represent the *p*-value of a *t*-test comparing the mean of PLACE > GOAL > SOURCE with those in other configurations.

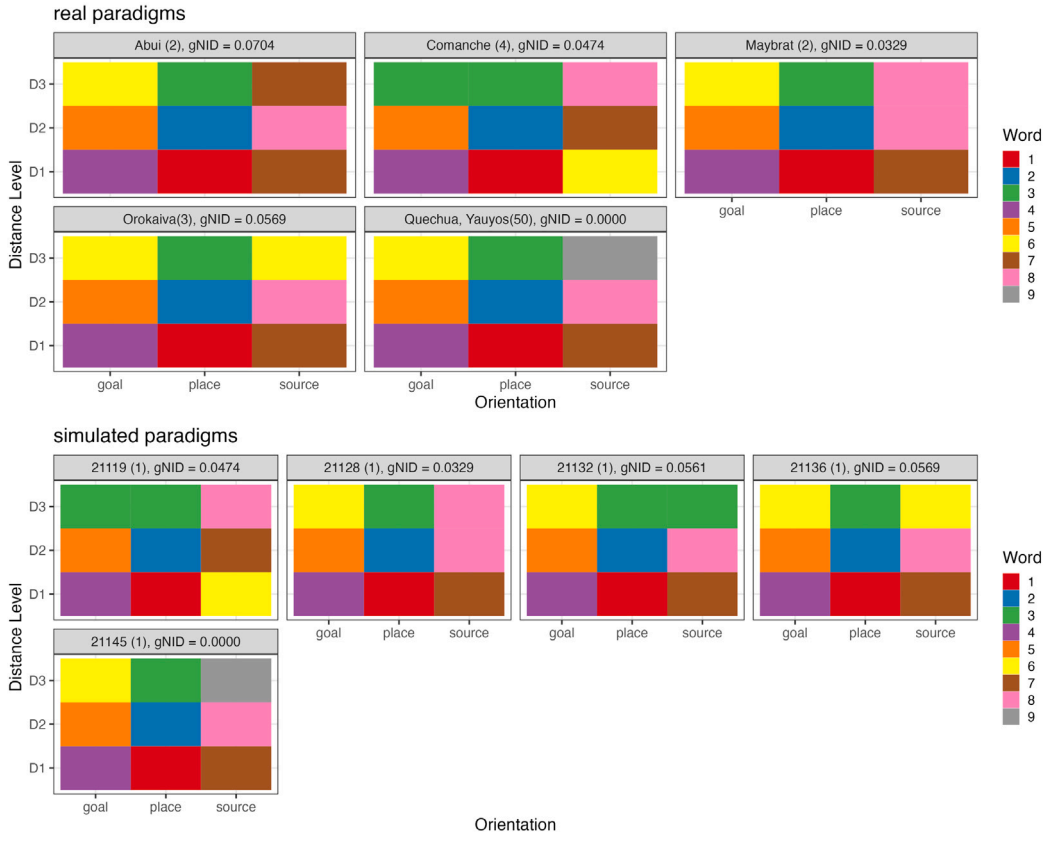

Fig. 8. The 5 most efficient real paradigms (top) and the optimal simulated paradigms (bottom) from Experiment 1. The horizontal axis indicates the PLACE/SOURCE/GOAL distinctions, whereas the vertical axis denotes the distance level. The real paradigms are exemplified by their languages. The number in the parenthesis indicates the number of languages in a given paradigm. gNID denotes the distance to the frontier for each paradigm.

and GOAL, and a separate set of words for conveying SOURCE information. We now call a particular strategy to partition the meaning space a **paradigm**. Each simulated system has only one paradigm, whereas many real systems might share the same paradigm. In fact, the 220 real systems in the database only utilize 34 out of the 21,146 possible paradigms. As we will demonstrate below, these real paradigms vary in their frequency.

The bottom panel Fig. 8 shows all the simulated paradigms located on the optimal frontier. The top panel shows all 5 real paradigms closest to the optimal frontier. Among the 5 simulated paradigms on the optimal frontier, only three (21,119, 21,128 and 21,145) are attested in our database. Simulated paradigm 21,119 merges PLACE and GOAL in distance level D3 while keeping other meanings distinct, similar to Bengali and 3 other languages. Simulated paradigm 21,145 gives every meaning its own word, which is the paradigm Quechua and 49 other languages adopt. Simulated paradigm 21,128 gives every meaning its own word except merging SOURCE in D2 and D3. The other 2 simulated paradigms are not attested. This is to say, only a few of the possible optimal paradigms are actually adopted by real languages.

Meanwhile, if real languages are optimized in communicative efficiency, we would expect the paradigms closest to the optimal frontier to all be adopted by many languages. However, we see a clear disparity in terms of the number of languages adopting each paradigm. For instance, in the top panel of Fig. 8, Yauyos Quechua shares the same paradigm with 49 other languages, whereas Maybrat, Comanche, and other 4 languages are the only languages utilizing their respective paradigm, despite being very close to the optimal frontier. This indicates that information theory alone fails to predict why some paradigms are favored but not others, suggesting that some other factors might be affecting such preference.

One of the additional factors appears to be consistency. Consistent paradigms that are close to the optimal frontier tend to be adopted by more languages, compared with inconsistent paradigms. In this section we adopt the consistency score defined in Section 3.3.

6.2. Basic analysis

We calculate the consistency score of the real systems and the random systems generated in Experiment 1, using the same set of parameters. In Fig. 9, we plot the consistency as well as the distance to the optimal frontier for each real paradigm and the simulated paradigm, faceted by the number of words used. The size indicates the number of languages in that paradigm. The most frequent real paradigms (shown as bigger-sized circles), utilized by a majority of the real languages in our database, fall into the bottom left corner in each facet, indicating that they tend to be consistent in addition to being efficient in balancing informativity and complexity. In contrast, infrequent paradigms (shown as smaller-sized circles) are generally located away from the lower-left corner in each facet, meaning that they are either not consistent or efficient. Meanwhile, simulated paradigms are scattered across the plot. The graph suggests that consistency, combined with the distance to the information-theoretic optimal frontier, is a good predictor of the typological frequency of real language demonstrative paradigms.

6.3. Constructing optimal paradigms

Now we examine whether the optimal paradigms considered by both information-theoretic and consistency constraints would be utilized by most languages. To do so, we extend Eq. (6) by adding a consistency term. The new optimality score is shown in Eq. (11):

$$J_{DIB}[q] = \underbrace{\mathrm{H}[W]}_{\text{Complexity}} - \beta \cdot \underbrace{\mathrm{I}[W:U]}_{\text{Informativity}} + \gamma \cdot \underbrace{\mathrm{S}[W:M]}_{\text{Consistency}}, \tag{11}$$

where $S[W:M]$ is the consistency score from Section 3.3.

Then, we search for the most efficient simulated system under each pair of (β, γ) values, where we let $(\beta, \gamma) \in [1, 10] \times [1, 10]$. The paradigms located on the new optimal frontier are shown in Fig. 10. The optimal paradigms now resemble real systems much more closely. For example, the optimal paradigm under $(\beta, \gamma) = (3.090, 1)$, where GOAL and PLACE are merged in each distance level and the by-distance level distinction is kept, is shared by Irish and 15 other languages.

One difference between real systems and the optimal systems under Eq. (11) is that the optimal systems with fewer words show a clear preference in merging distance levels D1 and D2, instead of merging D2 and D3 as natural languages do. We believe this disparity stems from the way our data is coded: in particular we assumed that, if a language distinguishes fewer than 3 distance levels, the last distance level extends out to D3 (see Section 4.1). This can possibly be resolved in reformulating the world state, a possible direction for future research.

In Fig. 10, we label optimal systems that differ from attested ones only by merging D1 and D2 instead of D2 and D3 with a suffix *-like*, and a total of 34 languages, including English, belong to this category. Interestingly, the paradigm under $(\beta, \gamma) = (1.072, 1)$ is not attested in any language, probably because this paradigm does not make distance level distinctions at all, while all the languages in Nintemann et al. (2020) have at least 2 distance levels. This suggests that in real languages, distinguishing distance levels might be more prioritized than distinguishing orientations. Meanwhile, the most popular paradigm shared by 71 languages, where all 3 orientations are distinguished and D2/D3 are merged, is not among the optimal paradigms, since distinguishing all orientations adds to the paradigm's complexity.

7. Discussion

Spatial demonstratives, a class of adverbs or adpositional phrases that encode spatial relations, vary both in forms and meanings they encode across languages in the world (Levinson, 1996). But there exist common patterns in how meanings are expressed by spatial demonstratives (Nintemann et al., 2020). Why are certain paradigms preferred over others? In this work, from an information-theoretic perspective (Strouse & Schwab, 2017; Tishby et al., 2000; Tishby & Zaslavsky, 2015; Zaslavsky et al., 2018), we have argued that spatial demonstrative systems in natural languages are communicatively optimized, relative to random statistical baselines. We have demonstrated that a deviation from appropriate parameter choices results in a less good fit to real languages, as measured by the residual complexity; the optimal choice of parameters reflects findings from past studies (e.g. Chen et al., 2022; Do et al., 2020; Lakusta & Landau, 2012; Nikitina, 2009; Papafragou, 2010; Regier & Zheng, 2007; Srinivasan & Barner, 2013) that humans exhibit a strong bias towards GOAL compared with SOURCE. Then, we show that in addition to informativity and complexity within the information-theoretic framework, consistency is another constraint real languages tend to satisfy, in that real languages tend to be consistent, in addition to balancing between informativity and complexity.

7.1. Why consistency?

As we have demonstrated in Section 6.1, most real languages have some degree of consistency in their paradigms, which our information-theoretic approach alone does not predict. In fact, languages could be more efficient by, e.g., having access to the additional degrees of freedom that come with being able to use different syncretism patterns at different distance levels.

There are several possible explanations for the consistency preference. From a learning perspective, consistent spatial deictic paradigms tend to have low Kolmogorov complexity: in other words, fewer words need to be used to describe the pattern in the paradigm (Li, Vitányi, et al., 2008), leading to relative ease and readiness of acquisition of paradigms in natural languages (Ehret, 2014) and those in novel, artificial languages (Johnson et al., 2021; Maldonado & Culbertson, 2020a, 2020b). Broadly speaking, the tendency towards minimizing

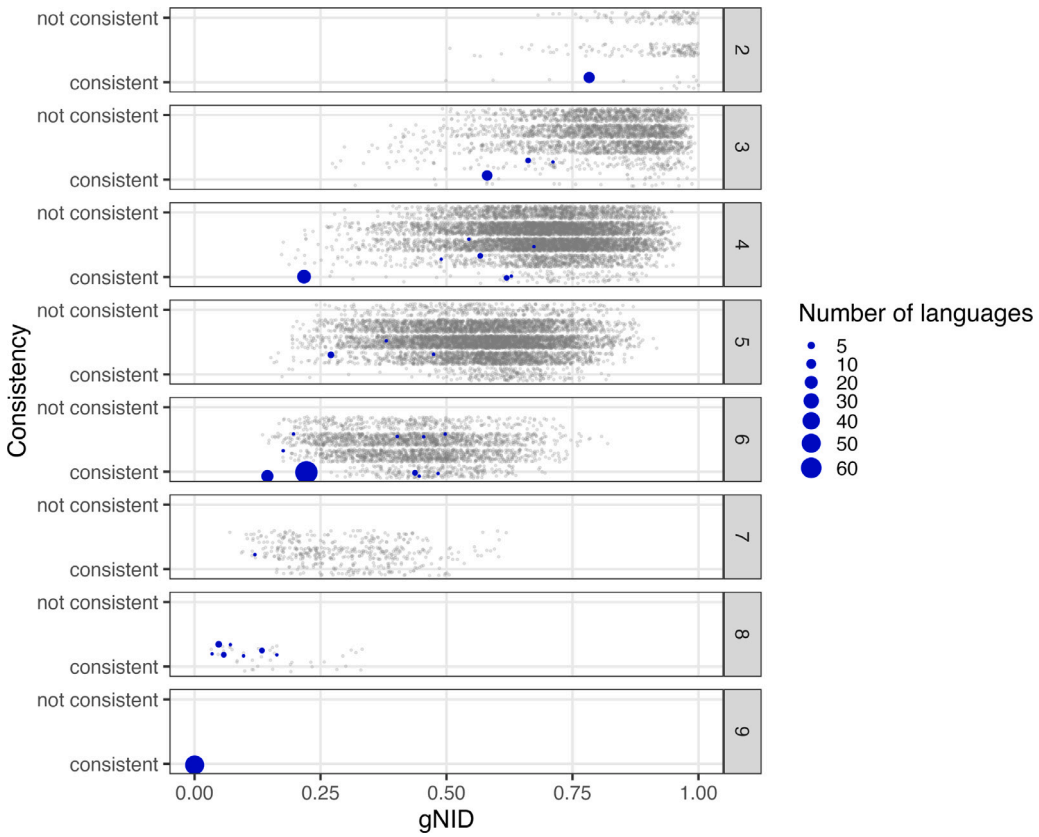

Fig. 9. Paradigms utilized by most real languages tend to be both consistent and close to the optimal frontier. Each blue dot represents one of the 34 unique paradigms attested in the real languages, whereas each gray dot represents all the possible paradigms. The horizontal axis is its distance to the optimal frontier, and the vertical axis is its consistency (high to low). The size indicates the number of languages in each paradigm. The plot is faceted by the number of words used in the paradigm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the description length of a paradigm is often reflected in historical linguistics, where, for example, in Indo-European historical linguistics, analogy seems to frequently play a role in shaping how ancient or archaic languages evolve into its modern descendants, in that irregular inflection/declension patterns are often replaced by regular ones (Fortson, 2011). Such tendencies have also shown in the lab: that structures gradually emerge as languages are passed down between generations (Kirby, Cornish, & Smith, 2008).

7.2. Limitations of our consistency formulation

The consistency metric used in this work is rudimentary in that it only broadly classifies different paradigms in terms of the number of distinct distance level patterns and the number of distinct orientation patterns. However, this metric still lacks granularity, in that it fails to take the need distribution into account. For example, if a particular meaning is rarely used in everyday conversation, a lack of syncretism in this meaning should be considered more consistent than a lack of syncretism in a meaning that is frequently used. Hence, in future studies, a more frequency-based metric, preferably an information-theoretic one, should be developed to operationalize consistency.

7.3. What influences the evolution of spatial demonstratives

Past studies have been focused on the evolution of communicative systems in semantic domains such as colors. Since it is impossible to conduct experiments on speakers from hundreds or even thousands of years ago, Zaslavsky, Garvin, Kemp, Tishby, and Regier (2022) instead investigate a rapidly-evolving language: Nafaara, spoken in Ghana and Côte d'Ivoire. They show that the language has acquired several new color terms while keeping the trade-off between informativity and complexity optimal. Spatial demonstratives in languages today, albeit merely speculative, might have been through the same process of traversing along the optimal frontier. For instance, as mentioned in the beginning of this paper, English used to have 6 distinct spatial deictic words, distinguishing between 3 different orientations and 2 different distance levels. However, the GOAL demonstratives *hither* and *thither* merged with *here* and *there*, respectively, whereas the SOURCE demonstratives *hence* and *thence* were replaced with prepositional phrases *from here* and *from there*. Meanwhile, all the 3-way orientation distinctions in other Germanic languages (such as Dutch, German, Danish, and Icelandic) are very much preserved. This is possibly because English has been extensively acquired as a second language, and the vast presence of L2 speakers leads to a simplification in the paradigm (McWhorter, 2007). Both paradigms, as shown in Fig. 5, are very close to the optimal frontier (the archaic English paradigm is the same as Dyirbal).

7.4. The continuous nature of distance levels

One limitation of our approach is that we assume that the space of distance levels is fixed and discrete, and that the mapping from meanings to words is deterministic. In reality, the space is likely continuous and word usage is likely to be stochastic. For instance, while objects that are extremely close to the deictic center are likely to be referred to by a D1 word (and never D2 or D3 words), it is likely objects that are somewhat farther away may be referred to using D1 or D2 in a way that is random or depends on the situation.

We believe that the Information Bottleneck approach could be used to make quantitative predictions about where boundaries between words would be found in this continuous, stochastic setting. Such a study would require data on the usage characteristics of these spatial words in a known spatial layout and situation, which is not currently available to the best of our knowledge.

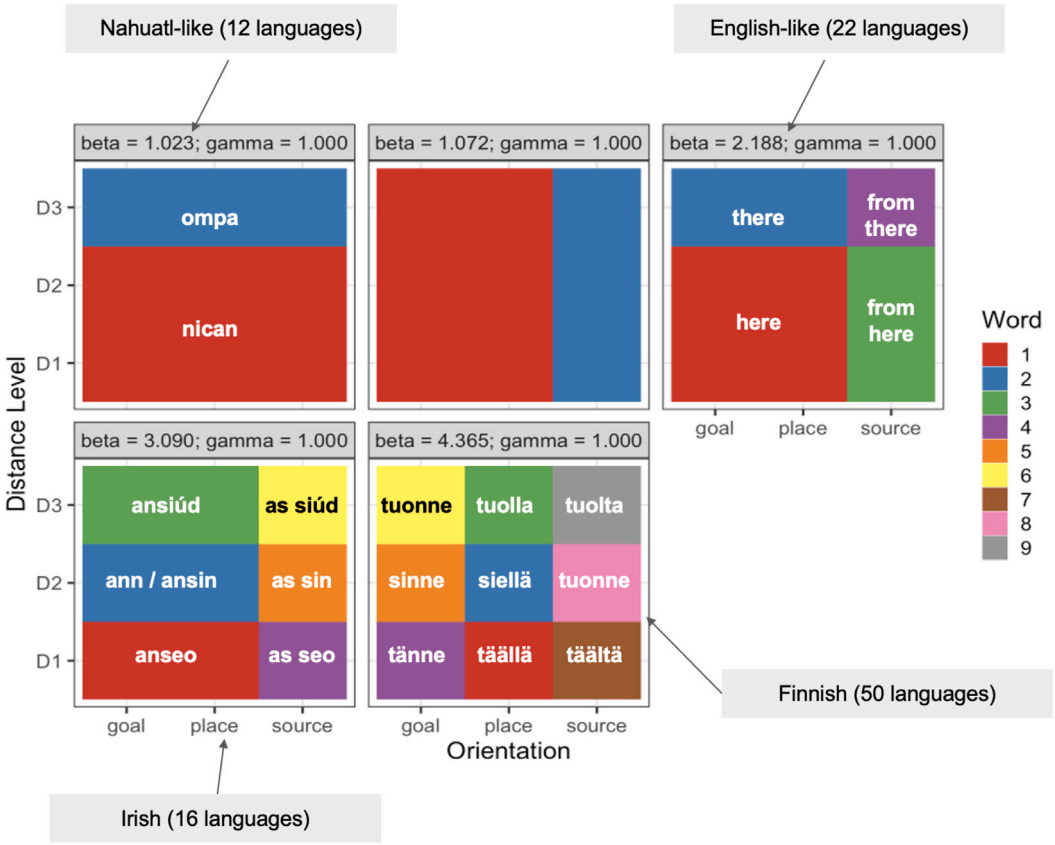

Fig. 10. The most optimal and consistent simulated paradigms. The β and γ values on each facet indicate the smallest (β, γ) pair corresponding with that system. Gray boxes indicate an example language and the number of languages utilizing that paradigm. A "like" suffix is attached when the real paradigm only differs from the simulated paradigm by merging D1 and D2 instead of merging D2 and D3. About half of real languages fall into one of these patterns.

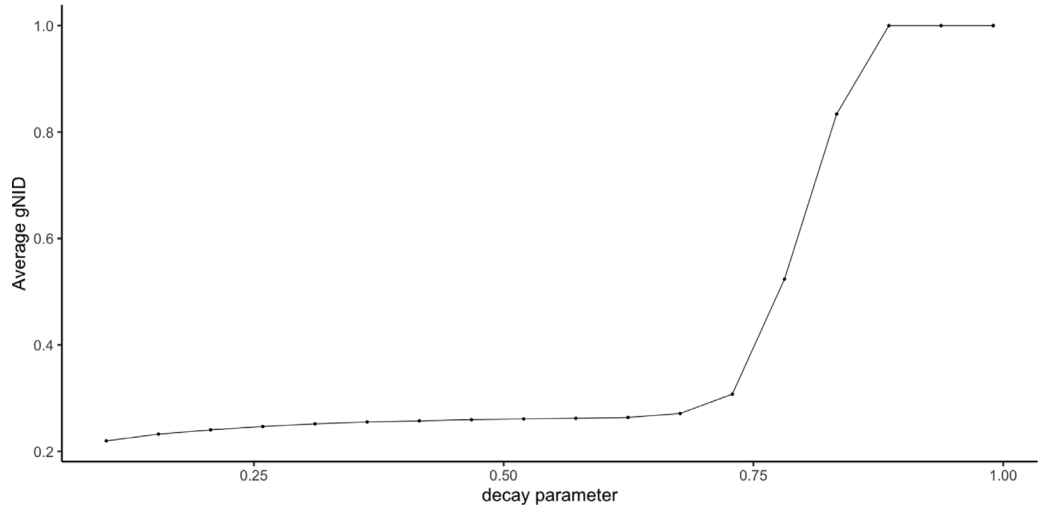

Fig. 11. The average gNID for different values of μ.

8. Conclusion

Overall, we have shown that the information bottleneck can explain major patterns in the typology of spatial demonstratives. Supplementing the information bottleneck with a preference for consistent paradigms, a preference motivated by the human preference for regularity in learning and memorizing, improves the explanatory performance of the model.

Our analyses also show that there is value not just in asking whether observed features of languages can be explained by a drive towards

efficiency, but by examining how model assumptions affect the ability of an efficiency-based model to fit the linguistic data. We found that using a cost function motivated by the observed source/goal asymmetry in cognition more generally led to real languages that looked more efficient. This match between cognitive-plausible model assumptions and fit to linguistic data not only provides further validation for the hypothesis that communicative efficiency drives linguistic behavior,it also suggests that efficiency-based approaches to linguistics can provide novel insight on underlying cognitive processes and can, as with our cost functions, provide evidence convergent with evidence from cognitive behavioral experiments.

Data availability

All the materials can be found at https://github.com/cshnican/spatial_demonstratives.

CRediT authorship contribution statement

Sihan Chen: Conceptualization, Methodology, Software, Investigation, Validation, Formal analysis, Data curation, Writing, Writing – original draft, Writing – review & editing, Visualization. **Richard Futrell:** Conceptualization, Methodology, Software, Investigation, Validation, Formal analysis, Data curation, Writing, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Kyle Mahowald:** Conceptualization, Methodology, Software, Investigation, Validation, Formal analysis, Data curation, Writing, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Data availability

All our data and analyses are available online on Github, with the link provided in the manuscript.

Acknowledgments

We would like to express our gratitute to Julia Nintemann, Maja Robbers, and Nicole Hober for the typological work that made our study possible, sharing materials, and providing feedback. We wish to thank Ted Gibson for his discussions and feedback. We would also like to thank the audience at Cog Lunch at MIT Department of Brain and Cognitive Sciences, at the 35th Annual Conference on Human Sentence Processing, and at the 2022 SIGTYP Workshop, from whom we received feedback.

Appendix A. The decay parameter μ

Here we test the effect of changing the decay parameter on the conditional distribution of world states given a meaning $p(u|m)$. From Eq. (5), a low μ indicates $p(u|m)$ decreases very quickly as the number of mismatches between world state u and meaning m increases. Therefore, in this situation, the cost for a person to confuse different distance levels and orientation will be very high, since such a cost is proportional to $-\ln\mu$. On the other hand, a high μ suggests a relatively low cost for confusing distance levels and orientation. In this analysis, we kept other parameters constant, let $\mu \in [0.05, 0.99]$, and compute the average gNID among attested spatial deictic systems.

The results in Fig. 11 show that as long as $\mu < 0.75$, the decay parameter has a minor effect on the optimality of attested spatial deictic systems. In other words, unless we assign a very low cost to confusing between distance levels and orientations, attested spatial deictic systems tend to stay close to the optimal frontier.

Table 5
English (a) and Fake English (b) spatial demonstratives.

	Goal	Place	Source
D3	there	there	from there
D2	there	there	from there
D1	here	here	from here

(a) English

	Goal	Place	Source
D3	there	there	from there
D2	here	here	from there
D1	here	here	from here

(b) Fake English

Appendix B. Alternative complexity measures

In this study, similar to Zaslavsky et al. (2018), we defined complexity as the mutual information between word W and meaning M (which was reduced to $H[W]$ due to determinism). Meanwhile, a reviewer pointed out that our formulations of complexity and consistency are closely related, and they suggested two alternative ways to operationalize complexity: first, to use the log number of words $\log|W|$ instead of the entropy $H[W]$; and second, to combine complexity and consistency into one single metric, namely, the minimal description length (MDL) of a spatial demonstrative paradigm. Here we discuss these two alternative metrics and their implementation.

Replacing $H[W]$ *with number of words.* One reviewer was suggesting using the number of words as the complexity metric, in order to get around the issue that complexity and consistency are closely related. For instance, if the number of words is 9, the complexity is maximized, and in this case, the consistency in our definition can only take the value of 2. We do not see much improvement other than making both complexity and consistency count-based, since the number of words and consistency are also closely related (see Fig. 12).

Incorporating complexity and consistency into one single metric by taking an MDL approach. Simple operationalizations of MDL are not sufficient to create a unified measure of complexity and consistency. We adopted a similar approach to Denić et al. (2021): counting the minimal number of distance levels and orientations needed to describe a demonstrative, along with logical operations such as AND and OR. For instance, consider the spatial demonstrative system for English (Table 5(a)):

For each demonstrative, let us consider the minimal number of distance levels and orientations needed to fully describe it:

- *here*: $(G \cup P) \cup D_1 \rightarrow 3$ features (This means the full definition of the demonstrative "here" is a one that describes GOAL and PLACE at distance level D_1)
- *there*: $(G \cup P) \cup (D_2 \cup D_3) \rightarrow 4$ features
- *from here*: $S \cup D_1 \rightarrow 2$ features
- *from there*: $S \cup (D_2 \cup D_3) \rightarrow 3$ features

Therefore, when we sum them up, under this formulation, English has a complexity of 12. Similarly, Finnish has a complexity of 18 since every word will have 2 features.

However, let us consider the case of Fake English (Table 5(b)), where we simply replace the demonstratives at (D_2, PLACE) and (D_2, GOAL) from "there" to "here". Fake English is an example of inconsistent system, as it does not show syncretism of distance levels at different orientations. It is also not attested in the database in Nintemann et al. (2020). However, let us consider the MDL formulation:

- *here*: $(G \cup P) \cup (D_1 \cup D_2) \rightarrow 4$ features (This means the full definition of the demonstrative "here" is a one that describes GOAL and PLACE at distance level D_1)

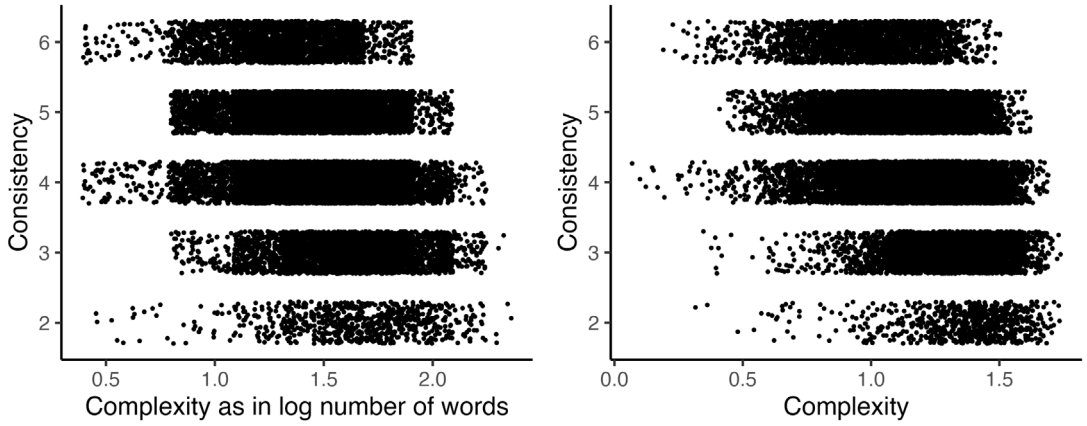

Fig. 12. Plotting consistency against different formulations of complexity. Left: consistency vs. complexity defined as the log number of demonstratives in a system; Right: consistency vs. complexity defined as the mutual information between words and meanings, as in the main text. It can be seen that both metrics are related to consistency, which is also supported by statistics ($R^2 \approx 0.18$ in both cases).

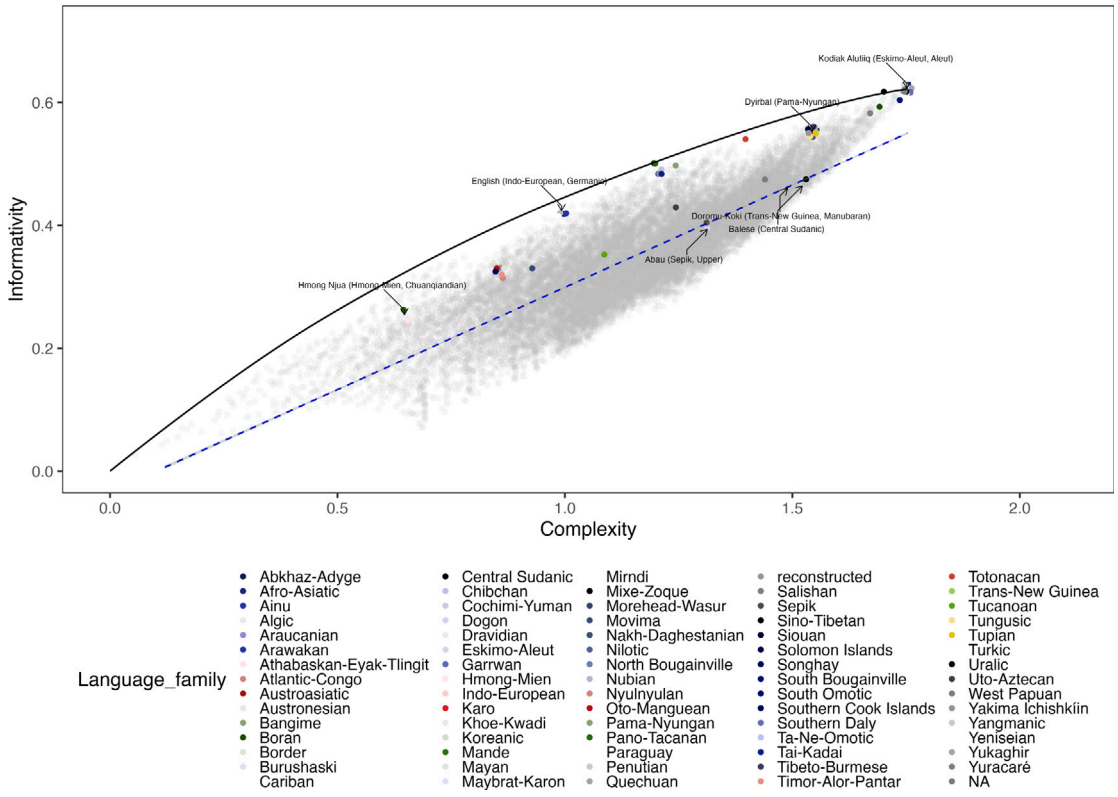

Fig. 13. Similar plot as Fig. 5, except we only sampled and presented 1 language per language family.

- *there*: $(G \cup P) \cup D_3 \rightarrow 3$ features
- *from here*: $S \cup D_1 \rightarrow 2$ features
- *from there*: $S \cup (D_2 \cup D_3) \rightarrow 3$ features

The MDL complexity is still 12. In other words, although Fake English is clearly less consistent than English, their MDL complexity is the same. Hence, MDL is probably not a metric that can incorporate both complexity and consistency. As a result, in the main text, we kept our current metrics for complexity and consistency.

Although the MDL theory above does not capture systematicity in our sense, more sophisticated description languages may do so. Another challenge for linking MDL with the IB sense of complexity is the fact that mutual information as complexity metric depends on the quantitative real-valued probabilities of words and meanings, while MDL approaches typically model discrete phenomena.

Appendix C. Alternative need distribution sources

In the main text (see Section 5), we approximated the need distribution of different meanings $p(m)$ by the word frequency distribution of Finnish spatial demonstratives, since in Finnish, each meaning has its own, unique spatial demonstrative. We drew the Finnish word frequency data from Lexiteria, which contains the word frequency data on the internet between 2009 and 2011. However, a disadvantage

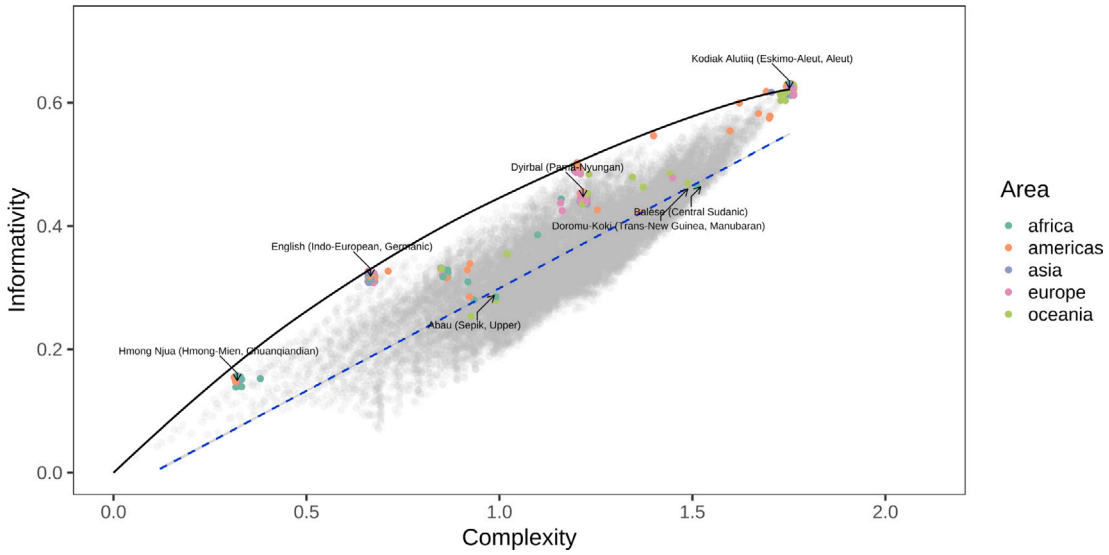

Fig. 14. Similar plot as Fig. 5, except we assume the first distance level encompasses D1 and D2, instead of just D1.

Table 6
Finnish spatial demonstrative frequency from three different corpora: Lexiteria, WorldLex (Gimenes & New, 2016), and OpenSubtitles (Lison & Tiedemann, 2016), as a proxy for the need distribution $p(m)$, and the average gNID among real spatial demonstrative systems under each frequency distribution.

	Lexiteria	WorldLex	Opensubtitles
D1, place, tänne	232,946	10,913	297,313
D1, goal, täällä	94,887	3,931	121,392
D1, source, täältä	38,402	1,926	45,204
D2, place, sinne	109,576	11,724	139,150
D2, goal, siellä	42,923	6,431	54,970
D2, source, sieltä	10,006	4,233	111,80
D3, place, tuonne	43,016	2,245	49,141
D3, goal, tuolla	17,587	448	21,193
D3, source, tuolta	3,850	480	4,928
Average gNID for attested systems	**0.239**	**0.271**	**0.237**

of Lexiteria is that this database is not open-source. Therefore, here we present analysis from two other databases: WorldLex (Gimenes & New, 2016), a database of Twitter and blog word frequencies, and OpenSubtitles (Lison & Tiedemann, 2016), a database of movie and TV subtitles. For each spatial demonstrative in the Worldlex corpus, we add its Twitter frequency and blog frequency together.

The results are shown in Table 6: the last row shows the average generalized normalized information distance (gNID, Zaslavsky et al., 2018) among real deictic systems, under each corpus. As also explained in Section 5, the lower the average gNID is, the closer the real deictic systems are to the optimal frontier. The average gNID from Lexiteria and OpenSubtitles are very close to each other (0.239 vs. 0.237, respectively), while they are relatively far from the gNID calculated from WorldLex (0.271), probably because the frequency distribution in WorldLex does not decay with respect to distance level, like those in Lexiteria and Opensubtitles. However, the discrepancy is still relatively small compared to the variation in average gNID under different PLACE/GOAL/SOURCE costs (see Section 5).

Appendix D. Results by language family

The dataset used in this study (Nintemann et al., 2020) sampled approximately 50 languages from each of the 5 geographic regions around the globe. However, they did not control for language relatedness, as many languages from the same region are closely related. For example, 15 out of the 50 languages in Europe are Slavic. As a result, our finding in Fig. 5 did not rule out a possibility that the efficiency of spatial deictic systems are mainly pushed by some large language families. For example, it could be possible that the spatial deictic systems in Indo-European languages are efficient due to chance, and since Indo-European languages constitute a large portion of the database, they can easily inflate the results shown in Fig. 5.

To address this potential confound of language relatedness, in Fig. 13, we present the information plane in the same way as Fig. 5, but instead of plotting all 220 languages that we investigate, we randomly sample 1 language per language family. All attested spatial deictic systems shown in the plot as colored points do not share a common ancestor with each other. Since a large portion of them are still located very close to the optimal frontier, we can say that language relatedness is an unlikely confound in our study.

Appendix E. Results by merging distance levels D1 and D2

In Section 4.1, we treat every attested language as having three distance levels. However, there exist languages (e.g. English) that only distinguish two distance levels. Hence, in that section, we adopt a convention that if the spatial deictic system in a language only distinguishes two distance levels, we assume the second distance level extends out, encompassing both D2 and D3. As one anonymous reviewer points out, there is an arbitrary decision made by us, with no theoretical motivation. In this analysis, we repeat Experiment 1 by assuming if a language has only two distance levels, the first distance level includes both D1 and D2, instead of just D1. The results are shown in Fig. 14, suggesting that there are no qualitative differences in our results that depend on our choice in distance level merging.

Appendix F. Effects of need distribution permutation and place/goal/source cost on optimality

See Fig. 15 for the analysis results.

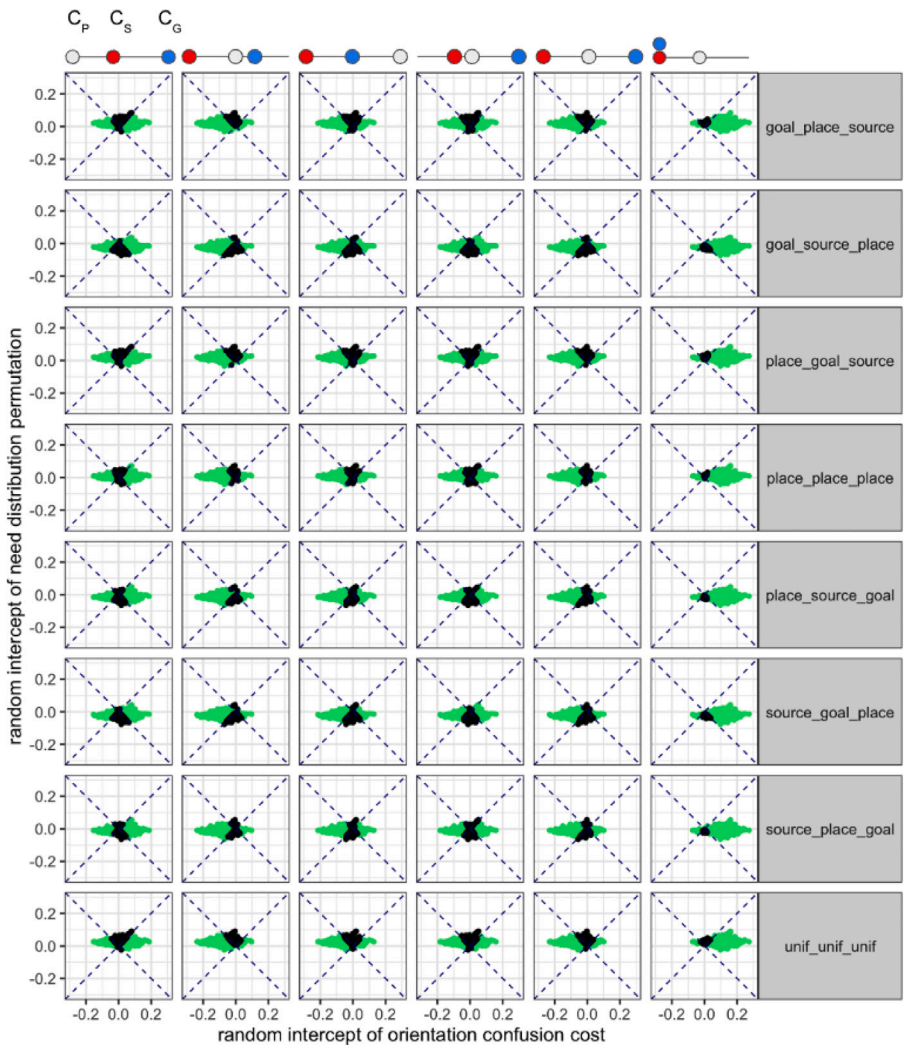

Fig. 15. The fitted random intercept per sample for PLACE/GOAL/SOURCE confusion costs (x-axis) and for need distribution permutations (y-axis), under each combination of these two factors (shown in each individual facet), in the regression of Eq. (10). Color code: green—the random intercept for the confusion cost has a higher absolute value than that for the need distribution permutation; black—the random intercept for the confusion cost has a lower absolute value than that for the need distribution permutation. Most of the samples (154,403 out of 192,000) are colored green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*(3), 429–464.

Anderson, S. R., & Keenan, E. L. (1985). Deixis. *Language Typology and Syntactic Description*, *3*(4), 259–308.

Baerman, M. (2004). Directionality and (un) natural classes in syncretism. *Language*, *80*(4), 807–827.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, *15*(6), 415–419.

Berger, T. (2003). Rate-distortion theory. In *Wiley encyclopedia of telecommunications*. Wiley Online Library.

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21.

Bühler, K. (1934). *Sprachtheorie*. Fischer.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Cadierno, T. (2004). Expressing motion events in a second language: A cognitive typological perspective. *Cognitive Linguistics, Second Language Acquisition, and Foreign Language Teaching*, 13–49.

Chen, Y., Trueswell, J., & Papafragou, A. (2022). Source-goal asymmetry in motion events: Sources are robustly encoded in memory but overlooked at test. 35th Annual Conference on Human Sentence Processing.

Cooperrider, K. (2016). The co-organization of demonstratives and pointing gestures. *Discourse Processes*, *53*(8), 632–656.

Corbett, G. G. (2015). Morphosyntactic complexity: A typology of lexical splits. *Language*, 145–193.

Coventry, K. R., Griffiths, D., & Hamilton, C. J. (2014). Spatial demonstratives and perceptual space: Describing and remembering object location. *Cognitive Psychology*, *69*, 46–70, Publisher: Elsevier.

Coventry, K. R., Lynott, D., Cangelosi, A., Monrouxe, L., Joyce, D., & Richardson, D. C. (2010). Spatial language, visual attention, and perceptual simulation. *Brain and Language*, *112*(3), 202–213, Publisher: Elsevier.

Coventry, K. R., Valdés, B., Castillo, A., & Guijarro-Fuentes, P. (2008). Language within your reach: Near–far perceptual space and spatial demonstratives. *Cognition*, *108*(3), 889–895.

Cover, T., & Thomas, J. (2006). *Elements of information theory*. Hoboken, NJ: John Wiley and sons.

Cysouw, M. (2009). *The paradigmatic structure of person marking*. OUP Oxford.

Danziger, E. (2010). Deixis, gesture, and cognition in spatial frame of reference typology. *Studies in Language. International Journal Sponsored By the Foundation "Foundations of Language"*, *34*(1), 167–185.

Denić, M., Steinert-Threlkeld, S., & Szymanik, J. (2021). Complexity/informativeness trade-off in the domain of indefinite pronouns. In *Semantics and linguistic theory, vol. 30* (pp. 166–184).

Diessel, H. (2006). Demonstratives, joint attention, and the emergence of grammar. *Cognitive Linguistics*.

Diessel, H. (2012). Bühler's two-field theory of pointing and naming and the deictic origins of grammatical morphemes. *Grammaticalization and Language Change: New Reflections*, 37–50, Publisher: John Benjamins Amsterdam.

Diessel, H. (2019). 13 Deixis and demonstratives. *Semantics-Interfaces*, 463, Publisher: Walter de Gruyter GmbH & Co KG.

Diessel, H., & Coventry, K. R. (2020). Demonstratives in spatial language and social interaction: An interdisciplinary review. *Frontiers in Psychology*, *11*, Article 555265, Publisher: Frontiers Media SA.

Dixon, R. M. (2003). Demonstratives: A cross-linguistic typology. *Studies in Language. International Journal Sponsored By the Foundation "Foundations of Language"*, *27*(1), 61–112.

Do, M. L., Papafragou, A., & Trueswell, J. (2020). Cognitive and pragmatic factors in language production: Evidence from source-goal motion events. *Cognition*, *205*, Article 104447.

Dromi, E. (1979). More on the acquisition of locative prepositions: An analysis of Hebrew data. *Journal of Child Language*, *6*(3), 547–562.

Ehret, K. (2014). Kolmogorov complexity of morphs and constructions in English. *Linguistic Issues in Language Technology*, *11*, 43–71.

Enfield, N. J. (2003). The definition of what-d'you-call-it: semantics and pragmatics of recognitional deixis. *Journal of Pragmatics*, *35*(1), 101–117, Publisher: Elsevier.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902.

Fillmore, C. J. (1997). *Lectures on deixis*. CSLI publications.

Fortson, B. W. (2011). *Indo-European language and culture: An introduction*. John Wiley & Sons.

Futrell, R. (2021). An information-theoretic account of semantic inference in word production. *Frontiers in Psychology*, *12*, Article 672408.

García, J. O. P., Ehlers, K. R., & Tylén, K. (2017). Bodily constraints contributing to multimodal referentiality in humans: the contribution of a de-pigmented sclera to proto-declaratives. *Language & Communication*, *54*, 73–81.

Gennari, S. P., Sloman, S. A., Malt, B. C., & Fitch, W. T. (2002). Motion events in language and cognition. *Cognition*, *83*(1), 49–79.

Georgakopoulos, T., & Karatsareas, P. (2017). A diachronic take on the Source–Goal asymmetry. *Space in Diachrony*, *188*.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., et al. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785–10790.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., et al. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*, 389–407.

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, *48*(3), 963–972.

Hammarström, H., & Forkel, R. (2022). Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web Journal*, *13*(6), 917–924.

Hanks, W. F. (1990). *Referential practice: Language and lived space among the Maya*. University of Chicago Press.

Hanks, W. F. (2011). 11. Deixis and indexicality. *Foundations of Pragmatics*, *1*, 315, Publisher: Walter de Gruyter.

Harremoës, P., & Tishby, N. (2007). The information bottleneck revisited or how to choose a good distortion measure. In *Information theory, 2007. ISIT 2007. IEEE international symposium on* (pp. 566–570). IEEE.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In *The new psychology of language* (pp. 217–248). Psychology Press.

Haspelmath, M. (2014). On system pressure competing with economic motivation. In B. MacWhinney, A. Malchukov, & E. Moravcsik (Eds.), *Competing motivations in grammar and usage* (pp. 197–208). Oxford University Press.

Haspelmath, M. (2021). Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics*, *57*(3), 605–633.

Haun, D. B., Rapold, C. J., Janzen, G., & Levinson, S. C. (2011). Plasticity of human spatial cognition: Spatial language and cognition covary across cultures. *Cognition*, *119*(1), 70–80.

Hawkins, J. (1994). *A performance theory of order and constituency, vol. 73*. Cambridge, UK: Cambridge University Press.

Hupp, J. M., Sloutsky, V. M., & Culicover, P. W. (2009). Evidence for a domain-general mechanism underlying the suffixation preference in language. *Language and Cognitive Processes*, *24*(6), 876–909.

Jackendoff, R. (1983). *Semantics and cognition, vol. 8*. MIT Press.

Jackendoff, R. (1996). The architecture of the linguistic-spatial interface. *Language and Space*, *1*, 30.

Jackendoff, R., & Landau, B. (2013). Spatial language and spatial cognition. In *Bridges between psychology and linguistics* (pp. 157–182). Psychology Press.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Johanson, M., Selimis, S., & Papafragou, A. (2019). The source-goal asymmetry in spatial language: language-general vs. language-specific aspects. *Language, Cognition and Neuroscience*, *34*(7), 826–840.

Johnson, T., Gao, K., Smith, K., Rabagliati, H., & Culbertson, J. (2021). Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems. *Journal of Language Modelling*, *9*(1), 97–150.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. CSLI Publications Stanford, CA.

Kemmerer, D. (1999). Near and far in language and perception. *Cognition*, *73*(1), 35–63, Publisher: Elsevier.

Kemp, C., Gaby, A., & Regier, T. (2019). Season naming and the local environment. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 539–545).

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In *Psychology of learning and motivation, vol. 74* (pp. 195–232). Elsevier.

Lakusta, L., & Landau, B. (2005). Starting at the end: The importance of goals in spatial language. *Cognition*, *96*(1), 1–33.

Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal paths. *Cognitive Science*, *36*(3), 517–544.

Langacker, R. W. (2013). Reference-point constructions. In *Mouton classics* (pp. 413–450). De Gruyter Mouton.

Levinson, S. C. (1996). Language and space. *Annual Review of Anthropology*, *25*(1), 353–382.

Levinson, S. C. (2018). Introduction: demonstratives: patterns in diversity. In *Demonstratives in cross-linguistic perspective* (pp. 1–42). Cambridge University Press.

Levinson, S. C., Kita, S., Haun, D. B., & Rasch, B. H. (2002). Returning the tables: Language affects spatial reasoning. *Cognition*, *84*(2), 155–188.

Levinson, S. C., & Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity, no. 5*. Cambridge University Press.

Levinson, S. C., & Wilkins, D. P. (2006). *Grammars of space: Explorations in cognitive diversity*. Cambridge University Press.

Li, M., Vitányi, P., et al. (2008). An introduction to kolmogorov complexity and its applications. Springer.

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the tenth international conference on language resources and evaluation*.

Maldonado, M., & Culbertson, J. (2020a). Here, there and everywhere: An experimental investigation of the semantic features of indexicals. https://ling.auf.net/lingbuzz/005628.

Maldonado, M., & Culbertson, J. (2020b). Person of interest: Experimental investigations into the learnability of person systems. *Linguistic Inquiry*, 1–42.

Maldonado, M., Saldana, C., & Culbertson, J. (2020). Learning biases in person-number linearization. (pp. 163–176). PsyArXiv Preprints, https://psyarxiv.com/5s2r8/; DOI:10.31234/osf.io/5s2r8.

Martin, A., & Culbertson, J. (2020). Revisiting the suffixing preference: Native-language affixation patterns influence perception of sequences. *Psychological Science*, *31*(9), 1107–1116.

McWhorter, J. (2007). *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford University Press on Demand.

Mollica, F., Bacon, G., Xu, Y., Regier, T., & Kemp, C. (2020). Grammatical marking and the tradeoff between code length and informativeness. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., & Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, *118*(49).

Nevins, A. (2015). Productivity and portuguese morphology. *Romance Languages and Linguistic Theory 2013: Selected Papers from'Going Romance'Amsterdam 2013*, *8*, 175.

Nevins, A., Rodrigues, C., & Tang, K. (2015). The rise and fall of the L-shaped morphome: diachronic and experimental studies. *Probus*, *27*(1), 101–155.

Nikitina, T. (2009). *Subcategorization pattern and lexical meaning of motion verbs: a study of the source/goal ambiguity*. Walter de Gruyter GmbH & Co. KG.

Nintemann, J., Robbers, M., & Hober, N. (2020). *Here–Hither–Hence and related categories: A cross-linguistic study*. Walter de Gruyter.

Noyer, R. R. (1992). *Features, positions and affixes in autonomous morphological structure* (Ph.D. thesis), Massachusetts Institute of Technology.

Papafragou, A. (2006). Spatial representations in language and thought. In *ITRW on experimental linguistics*.

Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science*, *34*(6), 1064–1092.

Pederson, E., Danziger, E., Wilkins, D., Levinson, S., Kita, S., & Senft, G. (1998). Semantic typology and spatial conceptualization. *Language*, *74*(3), 557–589.

Peeters, D., & Özyürek, A. (2016). This and that revisited: A social and multimodal approach to spatial demonstratives. *Frontiers in Psychology*, *7*.

Perkins, R. D. (1992). *Deixis, grammar, and culture* (pp. 1–255). John Benjamins Publishing Company.

Pertsova, K. (2007). *Learning form-meaning mappings in presence of homonymy: A linguistically motivated model of learning inflection*. Los Angeles: University of California.

Pertsova, K. (2011). Grounding systematic syncretism in learning. *Linguistic Inquiry*, *42*(2), 225–266.

Pertsova, K. (2012). Logical complexity in morphological learning: effects of structure and null/overt affixation on learning paradigms. In *Annual meeting of the Berkeley linguistics society, vol. 38* (pp. 401–413).

Pléh, C., Vinkler, Z., & Kálmán, L. (1997). Early morphology of spatial expressions in Hungarian children: A childes study. *Acta Linguistica Hungarica*, *44*(1/2), 249–260.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. MIT Press.
Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441.
Regier, T., Kemp, C., & Kay, P. (2015). 11 Word meanings across languages support efficient communication. In *The handbook of language emergence, vol. 87* (p. 237). Wiley Online Library.
Regier, T., & Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. *Cognitive Science*, *31*(4), 705–719.
Rocca, R., Wallentin, M., Vesper, C., & Tylén, K. (2019). This is for you: social modulations of proximal vs. distal space in collaborative interaction. *Scientific Reports*, *9*(1), 1–14.
Saldana, C., Herce, B., & Bickel, B. (2022). More or less unnatural: Semantic similarity shapes the learnability and cross-linguistic distribution of unnatural syncretism in morphological paradigms. *Open Mind*, *6*, 183–210.
Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. In *IRE national convention record, Pt. 4* (pp. 142–163).
Sloane, N. J., et al. (2018). The on-line encyclopedia of integer sequences. Published electronically at https://oeis.org.
Srinivasan, M., & Barner, D. (2013). The amelia bedelia effect: World knowledge and the goal bias in language acquisition. *Cognition*, *128*(3), 431–450.
Steinert-Threlkeld, S. (2020). Quantifiers in natural language optimize the simplicity/informativeness trade-off. In *Proceedings of the 22nd Amsterdam colloquium* (pp. 513–522).
Stolz, T., Lestrade, S., & Stolz, C. (2014). *The crosslinguistics of zero-marking of spatial relations*. De Gruyter (A).
Stolz, T., Levkovych, N., Urdze, A., Nintemann, J., & Robbers, M. (2017). *Spatial interrogatives in Europe and beyond*. De Gruyter Mouton.
Strouse, D., & Schwab, D. J. (2017). The deterministic information bottleneck. *Neural Computation*, *29*(6), 1611–1630.
Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. arXivpreprintphysics/0004057.
Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop* (pp. 1–5). IEEE.
Twomey, C. R., Roberts, G., Brainard, D. H., & Plotkin, J. B. (2021). What we talk about when we talk about colors. *Proceedings of the National Academy of Sciences*, *118*(39), Article e2109237118.
Ünal, E., Ji, Y., & Papafragou, A. (2021). From event representation to linguistic meaning. *Topics in Cognitive Science*, *13*(1), 224–242.
Ünal, E., Richards, C., Trueswell, J. C., & Papafragou, A. (2021). Representing agents, patients, goals and instruments in causative events: A cross-linguistic investigation of early language and cognition. *Developmental Science*, *24*.
Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication. *Open Mind*, *4*, 57–70.
Zaslavsky, N., Garvin, K., Kemp, C., Tishby, N., & Regier, T. (2022). The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*.
Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.
Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2019). Communicative need in colour naming. *Cognitive Neuropsychology*, *37*, 312–324.
Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.
Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. In *41st Annual meeting of the cognitive science society*.
Zénon, A., Solopchuk, O., & Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, *123*, 5–18.
Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.