The forms and meanings of grammatical markers support efficient communication

Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, Charles Kemp

Abstract

Functionalist accounts of language suggest that forms are paired with meanings in ways that support efficient communication. Previous work on grammatical marking suggests that word forms have lengths that enable efficient production, and work on the semantic typology of the lexicon suggests that word meanings represent efficient partitions of semantic space. Here we establish a theoretical link between these two lines of work and present an information-theoretic analysis that captures how communicative pressures influence both form and meaning. We apply our approach to the grammatical features of number, tense, and evidentiality, and show that the approach explains both which systems of feature values are attested across languages and the relative lengths of the forms for those feature values. Our approach shows that general information-theoretic principles can capture variation in both form and meaning across languages.

1 Introduction

A primary goal of linguistic typology is to characterize and explain the diversity in extant linguistic systems compared to possible but unattested systems (Croft, 2002). Linguistic typology can be approached from a variety of perspectives (e.g., Newmeyer, 2005), but here we take a functional approach and build on a large body of work that has explored ways in which language supports efficient communication (von der Gabelentz, 1901; Zipf, 1949; J. Bybee, 2010; Hawkins, 2004, 2014). Recent work in this tradition has formalized communicative efficiency in terms of information theory and has used this formalization to demonstrate that linguistic forms and meanings support efficient communication (Gibson et al., 2019), but form and meaning are usually treated separately. On one hand, a substantial body of work has demonstrated that linguistic forms allow communication with a minimum of effort (Zipf, 1949; Greenberg, 1966; Piantadosi, Tily, & Gibson, 2011; Haspelmath, 2021), but this work typically does not explain the meanings associated with the forms in question. On the other hand, recent work in semantic typology has shown that word meanings within several semantic domains support efficient communication (Regier, Kemp, & Kay, 2015; Kemp, Xu, & Regier, 2018), but does not address the forms used to express these meanings. Here we show that an existing information-theoretic account of lexical semantics (Kemp et al., 2018, as formulated in Zaslavsky, Kemp, Regier, & Tishby, 2018) also accounts for classic ideas about coding efficiency from the literature on grammatical marking (Greenberg, 1966; Haspelmath, 2021). Connecting these lines of work illustrates how information theory provides a unified account of both the meanings encoded in natural language and the forms used to express them.

Our theoretical framework applies to both grammar and the lexicon, but we focus here on grammatical marking expressed by morphology, and in particular on the grammatical features of number, tense and evidentiality. We chose these features because they primarily convey semantic information and because each encodes a rich semantic dimension instead of a simple binary distinction. Number reflects the number of entities involved in one role of an event (e.g., four lion**s** chasing a giraffe). Tense refers to the location of an event in time (e.g., past, present, future). Evidentiality refers to the source of information (e.g., did I see it, or hear someone else describe it?). Grammatical features like these are core components of language, yet there is considerable variation in the size of grammatical feature inventories and the realization of grammatical features across languages. For example, the data analyzed in this paper include fifteen distinct morphological systems that languages use to mark grammatical number. Whereas English only distinguishes between singular and plural, Larike distinguishes between singular, dual, trial and plural (Laidig & Laidig,

1990). Accounting for the diversity of feature inventories and realizations across languages is therefore a significant challenge.

Our work builds on functionalist accounts of grammatical features from several areas of the literature. A longstanding line of work has used corpus analyses to show that the realizations of grammatical feature values are shaped by the principle of least effort (Zipf, 1949). Because speakers often need to convey the meanings associated with grammatical features, grammatical markers have short forms that are easy to produce, and the most frequent feature values may receive no overt marking (Greenberg, 1966; Haspelmath & Karjus, 2017; Haspelmath, 2021). A second line of work has used artificial language learning experiments and evolutionary models to demonstrate that learners restructure their input to produce systems that are simpler, easier to produce, and more informative (Fedzechkina, Jaeger, & Newport, 2012; Kanwal, Smith, Culbertson, & Kirby, 2017; Kurumada & Grimm, 2019; Fedzechkina & Jaeger, 2020), often in line with linguistic universals (e.g., Culbertson, Smolensky, & Legendre, 2012). Related work has also demonstrated that more easily acquired grammatical systems occur more frequently in the world's languages (Gentner & Bowerman, 2009; Saratsli, Bartell, & Papafragou, 2020). Our approach is broadly consistent with all of these research strands, but formally bridges them by providing an integrated account of both grammatical feature values and the forms used to express them.

The information-theoretic framework that we use formalizes the trade-off between informativeness and simplicity that languages must negotiate. Consider a speaker who wishes to convey some meaning (e.g., the number of empty coffee cups still on my desk) to a hearer (see Figure 1 top panel). A highly informative system allows the speaker to discriminate between many different meanings (e.g., many different numbers of cups), but this communicative precision can only be achieved if the system is far from simple. The trade-off between informativeness and simplicity has been discussed for many years in the literature on "competing motivations" (Haiman, 2010; von der Gabelentz, 1901; Du Bois, 1985) and several measures of morphological simplicity have been proposed (see Supplemental Information for a discussion). Here, we build on a recent account of lexical semantics (Zaslavsky et al., 2018; Zaslavsky, 2020) that is grounded in Rate–Distortion theory (Shannon, 1948, 1959, the branch of information theory characterizing efficient data compression), and that formalizes both informativeness and simplicity in information-theoretic terms. Within this framework, the simplicity dimension connects naturally with the notion of coding efficiency from the literature on grammatical marking (Hawkins, 2004; Haspelmath & Karjus, 2017). We will therefore argue that the trade-off between informativeness and simplicity helps to explain both which feature values are attested across languages and the relative lengths of the linguistic realizations of these feature values.

The next section introduces our theoretical framework and provides formal definitions of information loss (the inverse of informativeness) and complexity (the inverse of simplicity). We then provide an overview of number, tense and evidentiality across languages and introduce the typological data that we analyze. The first set of analyses focuses on meaning, and demonstrates that grammatical feature inventories achieve near-optimal trade-offs between informativeness and simplicity. The second set of analyses focuses on form, and demonstrates that the realizations of grammatical features enable concise communication.

2 Theoretical Framework

We build on the theoretical framework in Zaslavsky et al. (2018), which has been previously used to account for word meanings across languages, and show that the same framework can also be linked to aspects of linguistic form. The framework, illustrated in the top panel of Figure 1, assumes a speaker and a listener who wish to communicate about states of the world u drawn from the universe $\mathcal{U}$. The speaker is uncertain about the true state of the world, and their mental state is captured by a speaker distribution s over states in $\mathcal{U}$. To summarize this mental state the speaker generates a linguistic form f according to an encoder $q(f|s)$ which maps speaker distributions into forms. Upon receiving this form, the listener computes a distribution $\hat{s}$ that is intended to approximate the speaker distribution s. We assume that this distribution $\hat{s}$ is computed by carrying out Bayesian inference based on the encoder $q(f|s)$ and a prior $p(s)$ over speaker distributions, which gives the optimal $\hat{s}$ (Zaslavsky et al., 2018). The prior reflects communicative need, or the relative frequency with which speakers communicate about different states of the world (Gibson et al., 2019; Zaslavsky, Kemp, Tishby, & Regier, 2019; Karjus, Blythe, Kirby, & Smith, 2020).

An optimal encoder $q(f|s)$ should satisfy two criteria: it should allow the listener to accurately reconstruct

the speaker's mental state, and it should minimize production effort by ensuring that frequently-used forms are short. To formalize these criteria it will be convenient to represent a grammatical marker as a pair (m, f) that includes both a meaning (or feature value) m and a form (or realization) f. For number in English there are two such pairs: (SINGULAR, $\emptyset$) and (PLURAL, "-s"), where the empty set $\emptyset$ indicates that the singular is zero marked, or realized without an overt form. Given this representation we can decompose the encoder $q(f|s)$ into a meaning encoder $q_m(m|s)$ that maps speaker distributions into meanings and a form encoder $q_f(f|m)$ that maps meanings into forms. This two-stage encoding process is illustrated in Figure 1, and is used in the following sections to develop analyses that focus on efficiency of meaning and analyses that focus on efficiency of form.

The meaning encoder q_m is lossy, but for simplicity we assume that the form encoder q_f is lossless, which means that the listener is able to reconstruct the meaning m without error given the form f. In reality this assumption does not hold. Languages permit ambiguity in the linguistic signal, and ambiguity (including ambiguity arising from reanalysis) may have implications for the historical emergence of grammatical forms (Traugott, 2011). Assuming that q_f is lossless, however, is a natural starting point given the cross-linguistic data available to us.

2.1 Efficiency of meaning

An efficient encoder q_m achieves an optimal tradeoff between complexity and information loss (the inverse of informativeness). Following Zaslavsky et al. (2018), the formal definitions of complexity and information loss are grounded in the Information Bottleneck (IB) principle (Tishby, Pereira, & Bialek, 1999), which is a special type of a Rate–Distortion tradeoff. The complexity of an encoder measures how much information about the speaker's mental state is preserved in the meaning of a grammatical marker, and is defined as the mutual information between meanings and speaker distributions:

$$I(M; S) = H(M) - H(M|S), \tag{1}$$

which, as shown, can be formulated as the difference of two terms: the entropy over meanings $H(M)$ and the conditional entropy $H(M|S)$.

The informativeness of an encoder is negatively related to the expected information loss associated with each communicative interaction. Following Regier et al. (2015) and Zaslavsky et al. (2018) we define this information loss as the expected Kullback-Leibler divergence $\mathrm{KL}(s||\hat{s})$ between the speaker distribution s and the listener's reconstruction $\hat{s}$ of that distribution:

$$E[\mathrm{KL}[S||\hat{S}]] = \sum_{s,m} p(s)q_m(m|s)\mathrm{KL}[s||\hat{s}_m], \tag{2}$$

where $\hat{s}_m$ is the reconstructed distribution for occasions on which the speaker chooses meaning m.

Every possible mapping from speaker distributions to meanings corresponds to a point in a two-dimensional space where the dimensions represent complexity and information loss. Some points in this space cannot be achieved by any possible language—for example, in any realistic setting it is impossible for an encoder to achieve both zero complexity and zero information loss. The boundary separating achievable points from unachievable points is a special case of a Pareto frontier known as the IB theoretical limit, and encoders along this continuous frontier achieve optimal trade-offs between complexity and information loss. These encoders are optimal in the sense that complexity cannot be reduced without increasing information loss, and information loss cannot be reduced without increasing complexity.

Given this theoretical framework, we can ask whether attested grammatical feature inventories achieve near-optimal trade-offs between complexity and information loss. For any given feature, applying the framework requires three components to be specified: the universe of world states $\mathcal{U}$, the speaker distributions for each objective world state s_u, and the prior on speaker distributions $p(s)$. Given these components we can compute the complexity and information loss of both attested and hypothetical systems, and trace out the IB Pareto frontier of systems that achieve optimal trade-offs between complexity and information loss (Tishby et al., 1999; Zaslavsky et al., 2018).

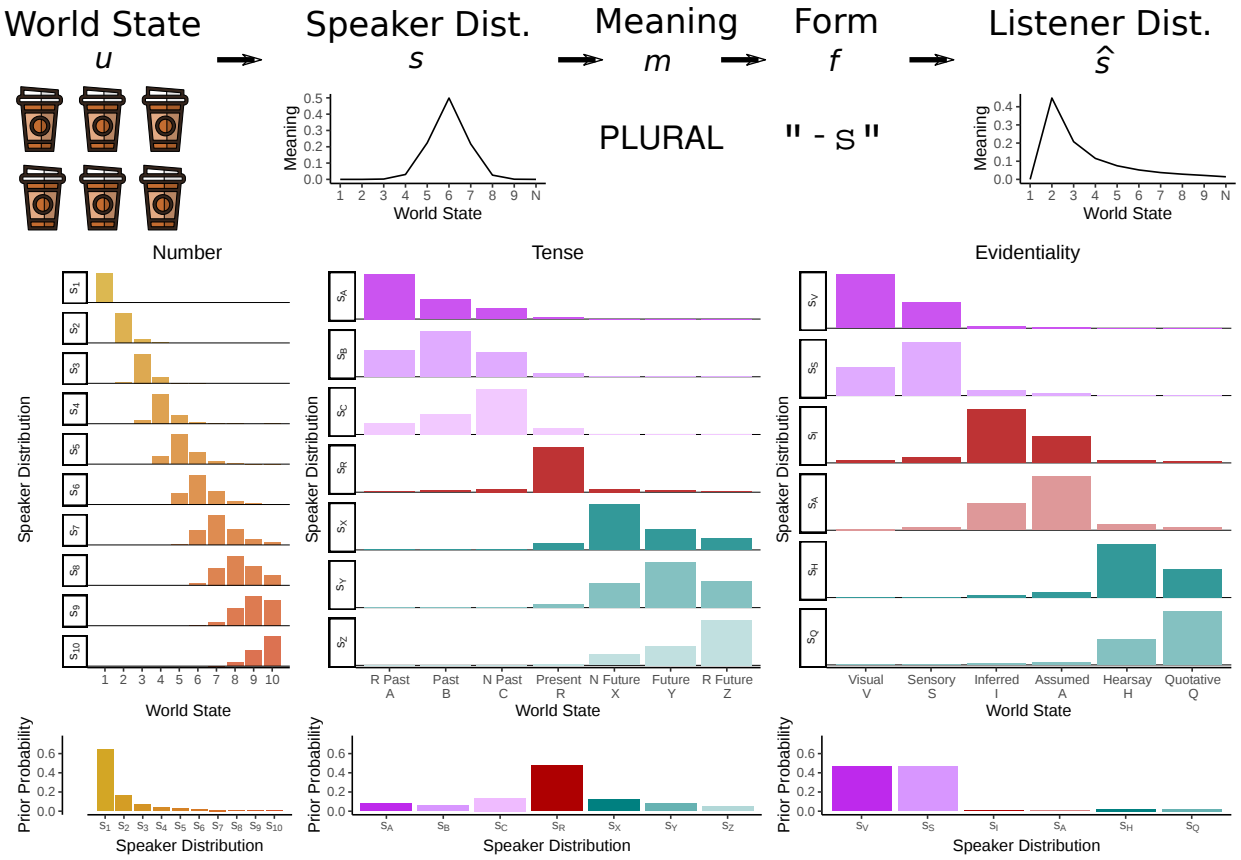

Figure 1: Communicative scenario along with speaker distributions and priors for number, tense and evidentiality. Top panels: Communicative scenario illustrating how a speaker generates a form which is then used by a listener to reconstruct the speaker distribution s over world states. In reality the form would not be uttered in isolation but rather combined with the noun "cup" to generate the utterance "cups." Center panels: Speaker distributions s_u for number, tense and evidentiality. Bottom panels: Priors $p(s)$ on the three sets of speaker distributions.

	Language	System	Complexity	Info. Loss	Front. Dist.	Form Corr.
Number	Pirahã	(GENERAL, ∅)	0.00	1.44	0.00	NA
	Russian	(SG, ∅)(PL, "ы")	0.94	0.55	0.01	1.00
	Larike	(SG, "mane")(DU_O, "matua")(TR_O, "matidu")(PL, "mati")	1.13	0.40	0.03	1.00
	Murrinh-Patha	(SG, "nukunu")(DU, "penintha")(PAUC, "peneme")(PL, "pigunu")	1.43	0.16	0.01	0.16
	Sursurunga	(SG, "i")(DU, "diar")(PAUC, "ditul")(GPAUC, "dihat")(PL, "di")	1.47	0.14	0.02	0.44
Tense	West Greenlandic	(ABCR, ∅)(XYZ, "ssa")	0.81	0.50	0.04	1.0
	Japanese	(ABC, "た")(RXYZ, ∅)	0.85	0.47	0.04	1.00
	Wolof	(ABC, "naa")(R, "nge")(XYZ, "dinaa")	1.52	0.08	0.00	1.00
	Hixkaryana	(A, "ye")(B, "yako")(C, "no")(RXYZ, "yaha")	1.26	0.42	0.21	0.32
	Zulu	(A, "a")(BC, "ile")(R, ∅)(X, "za")(YZ, "yaku")	2.01	0.02	0.00	0.88
Evidentiality	Sissala	(VSIA, ∅)(HQ, "ɛ")	0.28	0.15	0.00	1.0
	Abkhaz	(VS, ∅)(IAHQ, "заарԥн")	0.36	0.12	0.01	1.0
	Quechua	(VS, "mi")(IA, "chi")(HQ, "shi")	0.42	0.08	0.00	0.99
	Turkish	(V, ∅)(SIAHQ, "mış")	1.00	0.26	0.23	1.0
	Barasano	(V, "ka")(S, "ruyu")(IA, "ra")(HQ, "yu")	1.35	0.01	0.00	-0.09

Table 1: Example inventories for grammatical number, tense and evidentiality. Each system includes a single representative form for each meaning, and orthographic forms are shown instead of phonemic forms. The meanings for number are: GENERAL-"the noun can be expressed without reference to number" (Corbett, 2000, pg. 10); SG-Singular; PL-plural; DU-dual; TR-trial; PAUC-paucal (a few); GPAUC-greater paucal (a bunch); GPL-greater plural; and optional values are shown using the subscript $_O$. For tense, A, B and C denote distant, near, and immediate past, R denotes present, and X, Y and Z denote immediate, near and remote future. For evidentiality, V and S denote visual and sensory, I and A denote inferred and assumed, and H and Q denote hearsay and quotative. Frontier distance shows the Euclidean distance between a system and the corresponding Pareto frontier in Figure 2, and small values indicate efficiency of meaning. Form correlations show correlations between optimal and observed form lengths (see Figure 4) and large values indicate efficiency of form.

2.2 Efficiency of form

We now consider the mapping q_f from meanings to forms, or strings of phonemes. Because this mapping is assumed to be lossless, efficiency is purely a matter of minimizing expected form length. The entropy $H(M)$ gives a lower bound on expected form length (Shannon, 1948), and an efficient mapping q_f is expected to assign a form to meaning m that has length close to

$$h(m) = -\log p(m) = \log \sum_s q_m(m|s)p(s). \tag{3}$$

In reality, natural language mappings q_f are likely to yield expected codelengths that do not come especially close to the entropy $H(M)$ (Futrell, 2017; Pate, 2017; Pimentel, Nikkarinen, Mahowald, Cotterell, & Blasi, 2021). These mappings, however, may nevertheless reflect a pressure towards brevity. Equation 3 suggests that shorter forms should be used for more frequent meanings, and we will examine whether this inverse relationship between form length and frequency holds in our data.

Previous work on grammatical marking (Haspelmath, 2021; Fenk-Oczlon, 2001) and the lexicon (Zipf, 1949; Piantadosi et al., 2011; Bentz & Ferrer Cancho, 2016) has emphasized the notion of coding efficiency, and has demonstrated that forms tend to be paired with meanings in ways that allow utterances to be relatively concise. Similar results have emerged from studies of phonetic realization (Aylett & Turk, 2004), online word choice (Mahowald, Fedorenko, Piantadosi, & Gibson, 2013), and word ordering (Hahn, Jurafsky, & Futrell, 2020). Our theoretical approach is consistent with all of these results, but goes beyond them by considering efficiency of form within a framework that also captures efficiency of meaning.

Considering efficiency of form and meaning within a single framework is important because the two are connected via the entropy $H(M)$. Minimizing the expected codelength for an efficient code (i.e. minimizing $H(M)$) can only be achieved if the encoder q_m generates the same meaning for every speaker distribution. A system of this kind is maximally simple but also maximally uninformative (i.e. the information loss in Equation 2 is maximized). The pressure towards minimizing codelengths must therefore trade off against a pressure towards informative communication (Ferrer i Cancho & Solé, 2003). This tradeoff is especially clear for deterministic encoders q_m, for which the complexity measure in Equation 1 is equivalent to the entropy $H(M)$. Most (but not all) of the grammatical feature systems that we analyze are deterministic to a good first approximation: for example, in the English number system the singular is consistently used for individual items and the plural is consistently used for multiple items.

3 Framework instantiations

Now that we have introduced our general theoretical framework, we show how it can be applied to the grammatical features of number, tense and evidentiality. To evaluate our theory we will make several simplifying assumptions (Cooper & Guest, 2014; Guest & Martin, 2020). First, while number, tense and evidentiality can reflect multiple semantic dimensions (e.g., numerosity vs individuality, absolute vs relative time), we focus on a single grammaticalized semantic dimension for each. Each of these semantic dimensions can be expressed using a variety of strategies (e.g., numerals, adverbs, and verbal constructions), but for tractability we focus on systems that use obligatory morphological markers. As a result of these simplifications, the set of unique feature inventories in our sample is relatively small, and many languages are coded using a single feature value that spans the entire dimension. This kind of inventory is maximally simple (and therefore technically optimal), but also maximally uninformative. The languages coded in this way may not make use of obligatory grammatical markers, but typically rely on other linguistic constructions or contextual information for conveying information about the semantic dimensions in question. Evaluating the efficiency of these alternative communicative strategies is an important challenge for future work, and we return to this issue in the Discussion.

A second assumption is that the prior on speaker distributions $P(s)$ and the speaker distributions themselves are invariant across cultures. Previous studies have made similar assumptions (Kemp & Regier, 2012; Zaslavsky et al., 2018), and in all cases these assumptions should be viewed as rough first approximations that can be subsequently relaxed using data from studies that directly estimate culture-specific priors (Rácz, Passmore, Sheard, & Jordan, 2019). Third, our operationalization of production effort is relatively coarse, and we treat this quantity as a binary variable (marker present vs marker absent) or define it as the length of a marker's orthographic representation. Considering phonetic structure would allow for a more satisfying operationalization, but is not possible given the data available to us. Finally, the grammatical systems considered in our analyses (including the examples in Table 1) are idealizations that are best treated as high-level summaries of a more complex reality. Within any individual language, there may be departures from our idealizations and these differences may be irregular (e.g., the English plural is not marked for some nouns like *deer*) or context-dependent (e.g., in Hunzib evidentiality is marked only in the past tense).

The following sections introduce additional assumptions made when analyzing each of the three grammatical features and describe our samples of attested languages. These samples are drawn from a diverse set of language families and geographic regions, but are convenience samples that aim for breadth of coverage rather than tight control over genealogical or geographic relationships. Given the nature of our analyses, controlling for historical descent and geographic region does not seem essential, and the more important question is the extent to which our samples cover the space of attested feature inventories. Some extant inventories are almost certainly missing from our samples, and it will be valuable to revisit our analyses if and when larger data sets become available.

3.1 Number

Although the underlying semantic dimension for number is probably the natural numbers,[1] we consider only natural numbers less than or equal to ten for simplicity. The universe $\mathcal{U}$ therefore includes 10 world states, one for each number considered. Number marking in English distinguishes between singular (1) and plural (> 1), but some languages have more precise systems. For example, Murrinh-Patha distinguishes between singular (1), dual (2), paucal (3-6) and plural (> 5) (Corbett, 2000). While English and Murrinh-Patha require a speaker to always use the most specific marker, some languages allow speakers a choice between specific and less specific markers. For example, Larike distinguishes between singular (1), optional dual (2), optional trial (3) and plural (> 2) which means that the plural is always an alternative to the dual and to the trial (Laidig & Laidig, 1990).

When communicating about a state including n items, the speaker distribution s_n is intended to capture the speaker's uncertainty about the precise number of items present. Speaker distributions associated with the ten possible world states are shown in the left center panel of Figure 1, and these distributions are based on data from a timed, high-contrast estimation task (Cheyette & Piantadosi, 2020). The prior distribution $p(s)$

[1] We only consider numerical amounts in our analysis, leaving the dimension of individualization (Grimm, 2018) for future work.

captures the relative frequencies with which speakers attempt to convey the ten different meanings. Usage frequencies for number have been extensively studied using cross-linguistic corpora (Dehaene & Mehler, 1992; Piantadosi, 2016), and these studies suggest that $p(s)$ can be roughly approximated as an inverse square law (bottom left panel of Figure 1).

For our analysis, we compiled and coded the number marking inventories from 37 languages in Corbett (2000), representing 15 language families and 15 unique encoding systems. Five of these inventories are shown in Table 1. We adopt the standard linguistic conventions for number distinctions as labels (glossed in the caption). The main challenge in coding number systems is that indeterminate meanings (paucal: 3-6, plural: > 2, greater paucal: 6-8, greater plural: > 9) vary slightly across languages. For example, we code plural in Murrinh-Patha as greater than 5 because there are non-optional meanings for all lower numbers; whereas, plural in English is greater than two because there is only one other meaning. When a distinction is optional, we assume that the speaker chooses the two possible meanings equally often (see Supplemental Information for further details).

To study how meanings are realized as forms, we compiled number forms for a subset of 33 languages. Number is marked in a variety of ways across languages, and we considered only nominal and pronominal marking of grammatical number. For brevity, the systems in Table 1 show a single form for each meaning, but our data set and analyses allow for multiple forms per meaning. For example, the Russian data include number forms for different combinations of case and gender.

3.2 Tense

Tense is analogous to number but the underlying dimension is time rather than quantity. Tense marking in English distinguishes between past, present and future, but some languages have more elaborate tense inventories that specify not only whether an event is in the past or future, but also how far in the past or future it is. For example, Hixkaryana distinguishes between events in the immediate past (same day or previous night), near past (past few months) and remote past (Derbyshire, 1979). Researchers in formal semantics and artificial intelligence have developed precise representations of tense that could potentially be used in frameworks like ours (Rescher & Urquhart, 1971; McCarthy & Hayes, 1981; Allen, 1983; Reichenbach, 1947), but we take a simpler approach that can be readily applied across languages and is similar to that of Velupillai (2016b). We formulate $\mathcal{U}$ as a set of seven temporal intervals: remote past (A), near past (B) and immediate past (C), present (R), immediate future (X), near future (Y) and remote future (Z). These intervals are not sufficient to capture the tense inventory of every language in full: for example, Comrie (1985) reports that Kiksht, a language of the US Pacific Northwest, distinguishes between six or seven past tense categories. Our seven-interval timeline is therefore a pragmatic choice that allows us to represent the tense inventories of many but not all of the languages of the world.[2]

As in our number analysis, we pair each element of $\mathcal{U}$ with a speaker distribution s, and the seven meaning distributions are shown in the center panel of Figure 1. These distributions are intended to capture the uncertainty that speakers maintain over the exact time of an event: for example, s_a captures uncertainty about an event that actually took place in the remote past. To formulate these distributions we postulate major boundaries between past, present, and future, and minor boundaries between the three pasts (remote, near and immediate) and between the three futures. The distributions are defined in terms of two parameters κ and λ that specify how sharply probability mass decreases across minor and major boundaries. We set $\kappa = 0.5$ and $\lambda = 0.1$, which means that distributions drop by factors of 2 and 10 across minor and major boundaries respectively. Our results are qualitatively robust to variation in the speaker distributions as long as there is an appreciable decrease across minor boundaries ($\kappa \leq 0.75$) and a reasonable distinction between major and minor boundaries (κ and λ are not equivalent or near equivalent).

We estimated the prior $p(s)$ using a two-step process. In the first step we used estimates of past, present and future from an analysis of social media (Park et al., 2017). The resulting counts yield a distribution of $[0.274, 0.475, 0.251]$ over the coarse categories of past, present and future. Second, we used frequencies of temporal adverbs such as *yesterday*, *last week* and *last month* (see Table S1 for the complete list) to distribute probability mass among the three levels of remoteness within both past and future categories. All frequencies were derived from the Google N-gram English corpus (Michel et al., 2011) for 1985, the year of

[2]We focus in this paper on absolute tense, leaving relative tense for future work.

publication for the source of many of our tense systems (Dahl, 1985). The prior distribution resulting from the two-step process is shown in the bottom center panel of Figure 1.

For our analysis, we compiled tense inventories for 157 languages, representing 73 language families and 16 unique inventories. Our sample was largely taken from Dahl (1985). To study how meanings are realized as forms, we compiled forms for a subset of 33 languages. Languages were selected with a bias toward languages with more linguistic forms and toward grammars that made the relevant information especially clear.

The major challenge encountered in assembling the data is that tense is often hard to separate from aspect and modality (J. L. Bybee, Perkins, & Pagliuca, 1994). For example, in some languages the primary distinction is between perfective and imperfective (roughly whether an action is complete or incomplete) rather than between past and future. Some languages include markers for categories (e.g., past perfective) that combine tense and aspect. When consulting our primary sources, we made our best judgment about whether a language could be represented in our coding scheme without distorting it too greatly, and excluded two languages (Hawaiian and Ewe) for which our scheme seemed especially inadequate. A second and less fundamental challenge is that languages which make remoteness distinctions do not include categories that are precisely equivalent. Our coding scheme distinguishes between remote (more than 7 days distant), near (between 1 and 7 days) and immediate (on the same day), and we fitted each language into this scheme as best we could.

3.3 Evidentiality

Evidentiality is a grammatical feature that conveys the source of a piece of information: for example, whether the speaker saw an event or heard it described by another person. There is no standard characterization of the space of possible sources, and we therefore formulate $\mathcal{U}$ as the set of six sources distinguished by Aikhenvald in her typology of evidential systems (Aikhenvald, 2004). In principle our framework allows these sources to be located within a multidimensional space, but for simplicity we order them along a single dimension that is consistent with Willett (1988)'s hierarchy of evidentiality values and roughly captures distance from the speaker. The first source is visual perception, and the second includes all senses other than vision. Next comes inference from visual evidence (e.g., learning there was a fire by seeing smoke), followed by assumption. The penultimate source combines general world knowledge (e.g., "it is known") and hearsay, and the final source is quoted speech. Languages group these six sources in different ways. For example, Quechua (Table 1) has markers for direct evidence (visual and sensory perception), indirect evidence (inference and assumption), and reported evidence (hearsay and quotation). In contrast, Turkish makes a simple partition between firsthand (visual sources) and non-firsthand (all other information sources).

Psychological evidence from Western populations suggests that speakers are often uncertain about the source of information retrieved from memory (Johnson, Hashtroudi, & Lindsay, 1993), but there are no detailed characterizations of how this uncertainty is distributed across different kinds of sources. We therefore specify the speaker distributions using the same hierarchical approach described for tense. Following Willett (1988)'s hierarchy, we assume major boundaries between perception (visual and sensory), reasoning (inference and assumption) and external report (hearsay and quotation speech), and minor boundaries within each of these three pairs. As for our tense analysis we use parameters κ and λ that specify how sharply probability mass decreases across minor and major boundaries. Our results are again qualitatively robust to variation in these parameters, and as before we set $\kappa = 0.5$ and $\lambda = 0.1$. The resulting speaker distributions are shown on the right center panel of Figure 1.

Although evidentiality occurs in around a quarter of the world's languages, few corpora are available for languages with fine-grained evidentiality inventories. We therefore estimate the prior $p(s)$ using a single corpus of Cuzco Quechua text (Rios, Göhring, & Volk, 2008). Quechua groups the six sources in $\mathcal{U}$ into three pairs, and we therefore divide corpus frequencies evenly within these pairs to produce the prior shown in the bottom right panel of Figure 1. For evidentiality in particular, the data available for grounding assumptions about the prior and speaker distributions are relatively limited, and our results should be viewed as tentative conclusions only.

We conducted our analysis on a set of 184 extant languages, representing 61 language families and 16 unique inventories. Descriptions of all languages were taken from Aikhenvald (2004), and five are represented in Table 1. To study how meanings are realized as forms, we compiled forms for a subset of 31 languages.

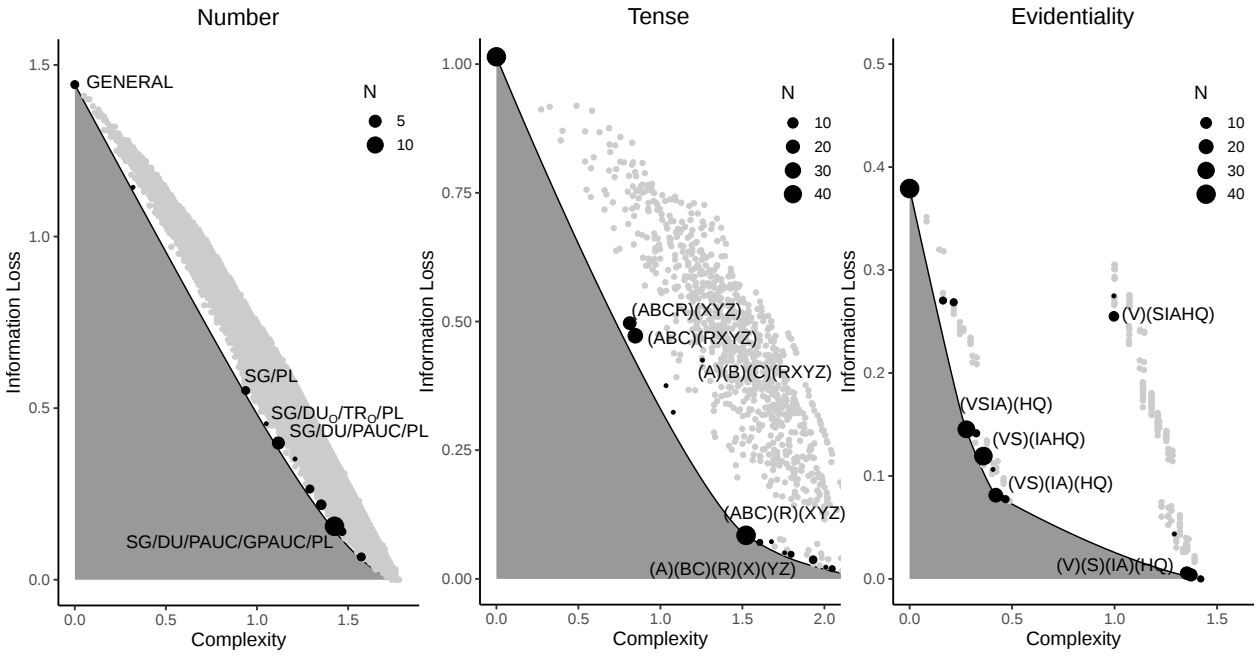

Figure 2: Analysis of the meaning of grammatical markers. (a)-(c) Trade-offs between information loss and complexity for number, tense and evidentiality. Attested inventories (black points) and unattested systems (grey points) are plotted in the space of all possible grammatical systems. Systems that achieve optimal trade-offs lie along the Pareto frontier (solid line), and the shaded region below the line shows trade-offs that are impossible to achieve.

Similar to tense, there was some difficulty separating evidentiality from other grammatical features including mood, mirativity (grammaticalized surprise), aspect and tense. For example, in Mansi, Svan and Turkish, evidentiality correlates with mirativity, and the non-firsthand marker can be used when the speaker has perfect visual evidence but witnesses something so surprising that they do not believe it. Evidentiality can also interact with genre, register and person systems. For example, in Meithei, the source of information is not always with respect to the speaker (first person) but can be calculated with respect to the listener (second person). As with tense, we tried our best to encode each system as faithfully as possible, but acknowledge that our encoding of evidentiality represents a starting point only and that future work will be required (see Supplemental Information for further discussion).

4 Analysis of meaning

We first analyze the feature values or meanings captured by each language in our data set, and the next section analyzes the forms that realize these meanings. For each of the three grammatical features, the space of possible encoders q_m is shown in Figure 2. Encoders that achieve optimal trade-offs between information loss and complexity lie along the Pareto frontier, shown here as a solid line, and the dark grey region below the curve shows trade-offs that are impossible to achieve. Attested inventories are shown as black points, and the light gray points include all possible inventories that partition $\mathcal{U}$ into non-overlapping feature values. Attested inventories (black points) are generally closer to the Pareto frontier than are unattested inventories (light gray points), suggesting that attested inventories for number, tense and evidentiality are near-optimal. The Supplemental Information includes a quantitative analysis that supports this conclusion strongly for number and tense and less strongly for evidentiality. It also shows that our model accounts better for attested inventories than an alternative approach previously applied to tense marking (Bacon, 2020) that defines the complexity of an inventory as the number of markers that it includes.

Although most attested inventories lie close to the Pareto frontier, there are a handful of notable exceptions. There are no clear outliers for number, and for tense, the single outlier is Hixkaryana. The Hixkaryana tense inventory (see Table 1) is unusual because it includes a relatively large number of categories but does

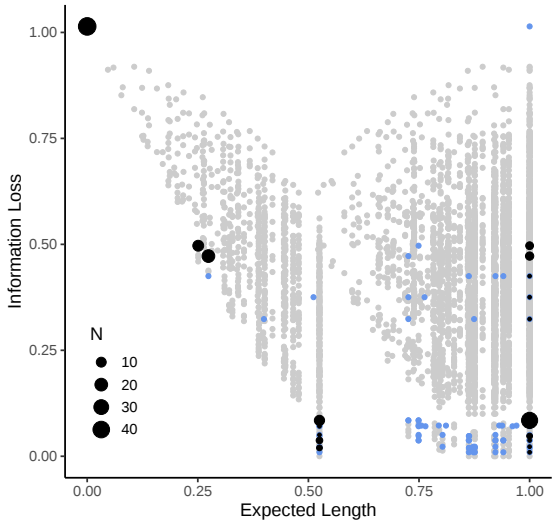

Figure 3: Zero marking analysis of tense systems. Trade-off between information loss and expected length when zero-marking is allowed for all tense systems ($N = 157$). Black dots show attested systems (size denotes frequency), blue dots show all ways to zero-mark at most one feature value in an attested system, and grey dots show possible but unattested systems.

not distinguish between present and future. For evidentiality, there are two notable groups of outliers driven by the same principle: our model predicts that distinctions at the level of Willett (1988)'s clusters should be made before distinctions within these clusters. For a few two term systems, including Turkish, Mansi and Meithei, and one three term system, Siona, a distinction between visual and sensory information is made before distinctions are made between all of Willett (1988)'s levels. For a more detailed discussion of individual languages and outliers see the Supplemental Information. The general conclusion from this discussion is that most attested systems are qualitatively similar to optimal systems.

For each of the plots in Figure 2, traversing the Pareto frontier from top left to bottom right generates a hypothetical evolutionary trajectory that makes predictions about the order in which distinctions are introduced if a system grows more complex over time (Zaslavsky et al., 2018; Zaslavsky, 2020). The Supplemental Information includes a detailed analysis of these trajectories and shows that they recapitulate some patterns previously identified by work in linguistic typology. Following Greenberg (1963), these patterns are often formulated as universal constraints on possible systems: for example, one such universal states that if a number system has a trial, then it also has a dual. Our theory broadly captures this and other known patterns, but is most compatible with the view that they are strong regularities that emerge from soft functional constraints instead of strict universals that hold without exception (M. S. Dryer, 1998; N. Evans & Levinson, 2009).

5 Analysis of form

We now analyze form length for number, tense and evidentiality. Before comparing form lengths across languages, we normalize lengths within each system to allow for the fact that lengths may be systematically longer in some languages (e.g., those with relatively small phoneme inventories) than others. Our first analysis asks whether feature values that are "zero-marked" (i.e. not overtly expressed, as for the nominal singular marker in English) tend to be more frequent than other feature values belonging to the same system (Greenberg, 1966; Haspelmath, 2021). To address this question we use a coarse form of normalization that assigns a length of 0 to any feature value that is zero-marked and lengths of 1 to all other feature values. Figure 3 plots the information loss from our analysis of meaning against expected length for tense, and therefore shows how informativeness of meaning trades off against brevity of form. The Supplemental Information contains an analogous plot for evidentiality but not number, because our sample of number markers includes

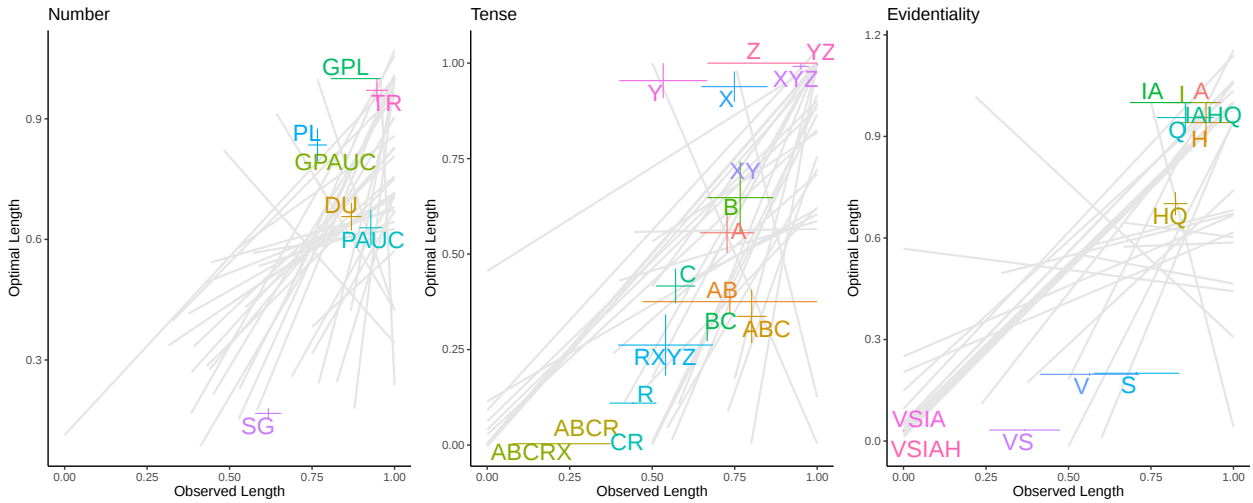

Figure 4: Analyses of the form of grammatical markers. Relationship between optimal and observed codelengths for a subsample of number, tense and evidentiality systems. Within each language, forms were unit normalized and the lengths of multiple forms for the same feature value within a language were averaged. Gray lines show trend lines for each language, and each colored data point shows an average across all languages that include a given feature value. Error bars show standard error of the mean, and the vertical error bars occur due to normalization and because the optimal length for a marker (e.g., past, or abc) depends on whether or not it is optional. Colors are arbitrary but help to distinguish overlapping error bars.

pronominal forms for which zero-marking does not occur. The black dots represent attested systems, and the light blue dots include all permutations of systems that use zero marking for at most one category in an attested system. The small grey dots show all ways to apply zero-marking to unattested systems. Attested systems with zero marking overwhelmingly tend to zero-mark the most frequent feature value and therefore lie along the Pareto frontier. The remaining attested systems explicitly mark all grammatical feature values and appear as a column of black dots with expected length equal to one. The Supplemental Information includes a statistical analysis suggesting that whether or not a tense system uses zero-marking can be partially predicted by the information loss of the meaning encoding system. When information loss is high, zero-marking provides relatively large reductions in expected length and is relatively likely to be used. In contrast, systems with low information loss have little to gain by zero-marking and are relatively likely to be explicit.

We now ask more generally whether the frequency of a grammatical marker is inversely related to the length of its form. Form length should ideally be measured in phonemes, but we do not have phonemic transcriptions for all languages in our samples and therefore use orthographic length as a rough proxy for phonemic length. Within each system, form lengths are normalized so that the longest form has length 1. We compare these "observed lengths" to predicted or "optimal lengths," where the optimal length for a marker with probability $p(m)$ is the surprisal $-\log(p(m))$. Frequent markers have short optimal lengths, and the optimal length of each marker can be interpreted as the number of bits used to represent the marker given an optimal code. Figure 4 shows that observed and optimal lengths are positively correlated across our samples of number, tense and evidentiality systems, and correlations for selected languages are shown in the final column of Table 1. The labeled data points in Figure 4 are based on averages across all systems that share a given feature value, and the gray lines are regression lines based on lengths from individual languages. Some individual languages (gray lines) represent exceptions to the general trend — for example, Tamil and Seneca have slightly shorter forms for future tense than for past (ABC) or past/present (ABCR) tenses. In general, however, languages tend to assign relatively short forms to markers that are high in frequency.

The results in Figure 4 are highly compatible with previous discussions of coding efficiency and grammatical marking. Most relevant to our approach is the work of Haspelmath (2021), who uses a broad range of grammatical patterns to demonstrate that more frequent grammatical feature values tend to have relatively short forms, and explains this result using the same functional-adaptive principles invoked by our theory.

Relative to this prior work our main contribution is to suggest that coding efficiency should not be considered in isolation, but rather trades off against a pressure for informative communication.

6 Discussion

We presented an account of grammatical marking which suggests that number, tense and evidentiality systems across languages achieve efficient tradeoffs between informativeness and simplicity. Our results align with related results previously reported for domains including color naming (Zaslavsky et al., 2018), kin naming (Kemp & Regier, 2012), quantifiers (Steinert-Threlkeld, 2020), numeral systems (Xu, Liu, & Regier, 2020), and indefinite pronouns (Denić, Steinert-Threlkeld, & Szymanik, 2020), and with a broader literature that characterizes ways in which language supports efficient communication (Gibson et al., 2019). Within this literature there are studies that focus on meaning (e.g., Kemp et al., 2018) and studies that focus on form (e.g., Piantadosi et al., 2011), but few that address both meaning and form (e.g., Dautriche, Mahowald, Gibson, & Piantadosi, 2017; Tamariz, 2008). Our work suggests how form and meaning can be brought together in an integrated information-theoretic framework.

Our analysis assumed that the function of each grammatical feature is to convey information about a single underlying dimension. Our approach, however, can be directly applied to settings in which the conceptual universe $\mathcal{U}$ combines multiple semantic dimensions: for example, both person and number (Zaslavsky, Maldonado, & Culbertson, 2021). Applying the framework in this way provides a new perspective on grammatical paradigms, and may help to explain attested patterns of syncretism. A further possible extension is to allow for additional functions that grammatical features may serve: for example, some features (e.g., case; Mollica & Kemp, 2020) may convey information about structural dependencies via indexing, and others (e.g., grammatical gender; Dye, Milin, Futrell, & Ramscar, 2017) may convey information about what forms should be expected next. Frequent linguistic units such as grammatical markers are especially likely to have multiple functions (Haspelmath, 2003; Zipf, 1949), and capturing the full range of these functions is a major challenge for quantitative approaches.

Our analysis focused only on grammatical marking but other linguistic strategies are available for communicating about number, time, and information source, including the use of quantifiers, temporal adverbs, and modal verbs. The languages in our data sets that do not mark number, tense and or evidentiality rely on these other strategies. Studying grammatical marking and other individual strategies in isolation is a natural first step, but future work should aim to allow for multiple different strategies when evaluating communicative efficiency.

Our work suggests that systems of grammatical markers achieve efficient trade-offs between informativeness and simplicity, but does not capture the historical processes that led to this outcome. It is possible that efficient trade-offs could arise in the absence of communicative pressures (Caplan, Kodner, & Yang, 2020), but recent work on cultural evolution and language acquisition suggests that language learning and use impose pressures towards informativeness and simplicity (Kirby, Tamariz, Cornish, & Smith, 2015; Carstensen, Xu, Smith, & Regier, 2015; Carr, Smith, Culbertson, & Kirby, 2020). On this account, the pressure towards informativeness applies during cooperative language use (e.g., Fay, Garrod, Roberts, & Swoboda, 2010), and the pressure towards simplicity applies during language learning (e.g., Hudson Kam & Newport, 2005). There is now a sizable body of evidence in language acquisition showing that learners reshape their input to learn languages that are simpler, easier to produce and more informative than their input (Fedzechkina et al., 2012; Kanwal et al., 2017; Kurumada & Grimm, 2019; Fedzechkina & Jaeger, 2020; Culbertson & Smolensky, 2012).

Connecting our approach with a model of historical language change may help to address two additional questions left open by our results. Our approach helps to explain the range of grammatical systems observed across languages, but does not explain why some systems are more frequent than others, or why any particular language has the systems that it does. One possibility is that different cultures impose different functional constraints, but a second possibility is that variation across languages reflects a set of crystallized historical accidents. If grammatical systems were initialized randomly, selective pressures over time may lead them to converge on a relatively small set of attractors, and the relative frequencies of these attractors could be explained by the relative sizes of their basins of attraction. Phylogenetic analyses have provided insight into the evolution of both semantic and grammatical systems (Jordan, 2011; Haynie & Bowern, 2016; M. Dunn,

Greenhill, Levinson, & Gray, 2011), and a similar approach could be productively applied to our data.

Although we focused on number, tense and evidentiality marking, our general approach can be applied both to other grammatical features and to the lexicon. In all cases, the goal is to simultaneously explain both the meanings captured by a linguistic system and the relative lengths of the forms that express those meanings. Grammar and the lexicon have traditionally been explored somewhat separately, but information-theoretic analyses can help to characterize how both support efficient communication.

7 Materials & Methods

7.1 Treatment of Data

Languages were primarily sampled from monographs surveying the grammatical features of number (Corbett, 2000), tense (Dahl, 1985), and evidentiality (Aikhenvald, 2004) with the goal of including as many distinct attested systems as possible. All data and code used in the analyses are available in an OSF repository: `https://osf.io/s5b7h/`

7.2 Specifying encoder distributions

Because usage data are not available for many languages in the data set, encoders $q(m|s)$ were determined using a maximum entropy assumption. This assumption is only relevant for languages that have optional distinctions. The unique encoders in our analysis are shown in detail in the Supplemental Information.

7.3 Speaker Distributions

For number, the nth speaker distribution is given by the analytical form:

$$S(u|n) \propto p(u) \exp\left(-\frac{p(n)}{\lambda_n}(n-u)^2 \right), \tag{4}$$

where the λ_i are empirically estimated precision parameters. To avoid having speaker distributions with no uncertainty due to values smaller than numerical precision, we added 10^{-5} to each state u and renormalized. For tense and evidentiality, the speaker distribution associated with state u^* is given by:

$$S(u|u^*) \propto B^{\lambda} b^{\kappa}, \tag{5}$$

where B and b are the number of major and minor boundaries separating u and u^* and λ and κ are the discount rates across major and minor boundaries respectively. For our analyses, λ and κ were set to 0.1 and 0.5 respectively.

7.4 The IB Pareto-frontier

The trade-off between complexity and information loss is given by the IB objective function (Tishby et al., 1999):

$$\mathcal{F}_{\beta}[q(m|s)] = I(M;S) - \beta I(M;U). \tag{6}$$

Following Zaslavsky et al. (2018), the Pareto-frontier was computed using reverse deterministic annealing (Tishby et al., 1999). For number, the β schedule was 2^x for x from 4 to 0 by 0.001 increments. For tense, the β schedule was 2^x for x from 5 to 0 by 0.001 increments. For evidentiality, the β schedule was 2^x for x from 5 to 0 by 0.001 increments.

8 Acknowledgments

We thank Naomi Baes for her meticulous work in compiling the number forms. A preliminary version of this work was presented at the Annual Meeting of the Cognitive Science society in 2020. This work was supported in part by a BCS fellowship in computation (N.Z.) and by ARC FT190100200 (C.K.).

References

Aikhenvald, A. (2004). *Evidentiality*. Oxford University Press.

Akiyama, N., & Akiyama, C. (2012). *Japanese grammar*. Simon & Schuster.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, *26*(11), 832–843.

Andronov, M. S. (1980). *The brahui language*. Nauka.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Bacon, G. I. (2020). *Evaluating linguistic knowledge in neural networks* (Unpublished doctoral dissertation). University of California, Berkeley.

Baerman, M., Brown, D., & Corbett, G. G. (2015). *Understanding and measuring morphological complexity*. Oxford University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bentz, C., & Ferrer Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In *Proceedings of the leiden workshop on capturing phylogenetic algorithms for linguistics* (pp. 1–4).

Bentz, C., Soldatova, T., Koplenig, A., & Samardžić, T. (2016). A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the workshop on computational linguistics for linguistic complexity*.

Berger, H. (1998). *Die burushaski-sprache von hunza und nager*. Harrassowitz.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their university and evolution*. University of California Press.

Boas, F. (1927). Annotated version of grammar of the Kutenai language, by Pater Philippo Canestrelli; additional notes on the Kutenai language. *International Journal of American Linguistics*, *45*, 61–94.

Boas, F., & Deloria, E. C. (1941). *Dakota grammar* (Vol. 23). US Government Printing Office.

Bohnemeyer, J. (2002). *The grammar of time reference in yukatek maya*. Lincom.

Borgman, D. M. (1999). Sanuma. In R. M. W. Dixon & A. Y. Aikhenvald (Eds.), *The amazonian languages*. Cambridge University Press.

Bromley, H. M. (1981). *A grammar of lower grand valley dani*. The Australian National University.

Bruce, L. (1984). *The alamblak language of papua new guinea (east sepik)*. The Australian National University.

Brunner, J. (2010). Phonological length of number marking morphemes in the framework of typological markedness. In *Between the regular and the particular in speech and language* (pp. 5–28). Peter Lang.

Buechel, E. (1939). *A grammar of lakota: The language of the teton sioux indians*. Swift.

Bugenhagen, R. D. (1991). *A grammar of mangap-mbula: An austronesian language of papua new guinea*. The Australian National University.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Bybee, J. L., Perkins, R. D., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. University of Chicago Press.

Caplan, S., Kodner, J., & Yang, C. (2020). Miller's monkey updated: Communicative efficiency and the statistics of words in natural language. *Cognition*, *205*.

Carlson, R. (1994). *A grammar of supyire*. de Gruyter.

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, *202*.

Carstensen, A., Xu, J., Smith, C., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*.

Chafe, W. (2015). *A grammar of the seneca language*. University of California Press.

Chang, A. H.-c. (2006). *A reference grammar of paiwan* (Unpublished doctoral dissertation). The Australian National University.

Chelliah, S. L. (1997). *A grammar of meithei*. de Gruyter.

Chen, R. (2019). *Plural forms in the world's languages* (Unpublished doctoral dissertation). Universität Leipzig.

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified theory of numerosity perception. *Nature Human Behavior*, *4*. doi: https://doi.org/10.1038/s41562-020-00946-0
Cole, P. (1982). *Imbabura quechua*. North Holland.
Comrie, B. (1985). *Tense*. Cambridge University Press.
Conrad, R. J., & Wogiga, K. (1991). *An outline of bukiyip grammar*. The Australian National University.
Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research*, *27*, 42–49.
Corbett, G. G. (2000). *Number*. Cambridge University Press.
Croft, W. (1990). *Typology and universals*. Cambridge University Press.
Croft, W. (2002). *Typology and universals*. Cambridge University Press.
Culbertson, J., & Smolensky, P. (2012). A Bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive Science*, *36*(8), 1468–1498.
Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.
Curnow, T. J. (1997). *A grammar of awa pit* (Unpublished doctoral dissertation). The Australian National University.
Dahl, Ö. (1985). *Tense and aspect systems*. Basil Blackwell.
Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*(8), 2149–2169.
Davidson, M. (2002). *Studies in southern wakashan (nootkan) grammar* (Unpublished doctoral dissertation). State University of New York at Buffalo.
Dedrick, J. M., & Casad, E. H. (1999). *Sonora yaqui language structures*. University of Arizona Press.
Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.
Dench, A. C. (1995). *Martuthunira: A language of the pilbara region of western australia*. The Australian National University.
Denić, M., Steinert-Threlkeld, S., & Szymanik, J. (2020). Complexity/informativeness trade-off in the domain of indefinite pronouns. In *Semantics and linguistic theory 2020*.
Derbyshire, D. C. (1979). *Hixkaryana*. North-Holland.
Dez, J. (1980). *Structures de la langue malgache*. Publications orientalistes de France.
Dickens, P. (1991). Jul'hoan orthography in practice. *South African Journal of African languages*, *11*(3), 99–104.
Dixon, R. M. W. (1988). *A grammar of boumaa fijian*. University of Chicago Press.
Dol, P. H. (1999). *A grammar of maybrat* (Unpublished doctoral dissertation). Rijksuniversiteit Leiden.
Donaldson, T. (1980). *Ngiyambaa: The language of the wangaaybuwan*. Cambridge University Press.
Donohue, M. (1999). Syntactic categories in Tukang Besi. *Revue québécoise de linguistique*, *27*, 71–90.
Drapeau, L. (2014). *Grammaire de la langue innue*. Presses de l'Université du Québec.
Dryer, M., & Haspelmath, M. (2020). *The World Atlas of Language Structures Online*. Retrieved from `https://zenodo.org/record/3731125` doi: 10.5281/ZENODO.3731125
Dryer, M. S. (1998). Why statistical universals are better than absolute universals. In *Papers from the 33rd regional meeting of the Chicago linguistic society* (pp. 123–145).
Du Bois, J. W. (1985). Competing motivations. In *Iconicity in syntax* (pp. 343–365). Benjamins Amsterdam.
Du Feu, V. (1996). *Rapanui: A descriptive grammar*. Routledge.
Dum-Tragut, J. (2009). *Armenian: Modern eastern armenian*. Benjamins.
Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79–82.
Dunn, M. J. (1999). *A grammar of chukchi* (Unpublished doctoral dissertation). The Australian National University.
Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In *Perspectives on morphological organization* (pp. 212–239). Brill.
Eisele, J. C. (1999). *Arabic verbs in time: Tense and aspect in cairene arabic*. Harrassowitz.
Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448.
Evans, N. D. (1995). *A grammar of kayardild*. de Gruyter.

Everett, D. L. (1986). Pirahã. In *Handbook of amazonian languages.* de Gruyter.
Everett, D. L., & Kern, B. (1997). *Wari: The pacaas novos language of western brazil.* Routledge.
Facundes, S. d. S. (2000). *The language of the apurina people of brazil(maipure/arawak)* (Unpublished doctoral dissertation). State University of New York at Buffalo.
Fallou, N. (2003). *Wolof.* Lincom.
Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351–386.
Fedzechkina, M., & Jaeger, T. F. (2020). Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission. *Cognition*, *196*, 104115.
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.
Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In *Typological studies in language* (Vol. 45, pp. 431–448).
Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, *100*(3), 788–791.
Foley, W. A. (1986). *The papuan languages of new guinea.* Cambridge University Press.
Fortescue, M. (1984). *West greenlandic.* Croom Helm.
Franklin, K. J. (1971). *A grammar of kewa, new guinea.* The Australian National University.
Futrell, R. L. (2017). *Memory and locality in natural language* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
Gentner, D., & Bowerman, M. (2009). Why some spatial semantic categories are harder to learn than others: The typological prevalence hypothesis. In J. Guo, E. Lieven, N. Budwig, S. Ervin-Tripp, K. Nakamura, & S. Ozcaliskan (Eds.), *Crosslinguistic approaches to the psychology of language: Research in the tradition of dan isaac slobin* (pp. 465–480).
Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Science*, *23*(5), 389–407.
Gordon, L. (1986). The development of evidentials in Maricopa. In *Evidentiality: The linguistic coding of epistemology* (pp. 75–88). Abex.
Greenberg, J. H. (1963). *Universals of language.* MIT press.
Greenberg, J. H. (1966). *Language universals with special reference to feature hierarchies.* The Hague: Mouton.
Grimm, S. (2018). Grammatical number and the scale of individuation. *Language*, *94*(3), 527–574.
Grinevald, C. G. (1990). *A grammar of rama.* Universite de Lyon.
Grinevald Craig, C. (1977). *Jacaltec: The structure of jacaltec.* The University of Texas Press.
Grout, L. (1859). *The isizulu: a grammar of the zulu language.* Trübner & Company.
Grunwald, P., & Vitányi, P. (2004). Shannon information and Kolmogorov complexity. *arXiv preprint cs/0410002*, *1*.
Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789-802.
Hagman, R. S. (1977). *Nama hottentot grammar.* Indiana University.
Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, *117*(5), 2347–2353.
Haiman, J. (2010). Competing motivations. In J. J. Song (Ed.), *The oxford handbook of linguistic typology.*
Harbour, D. (2003). *Elements of number theory* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
Harbour, D. (2014). Paucity, abundance, and the theory of number. *Language*, *90*(1), 185–229.
Hardman, M. J. (1986). Data-source marking in the Jaqi languages. In W. L. Chafe & J. Nichols (Eds.), *Evidentiality: The linguistic coding of epistemology.*
Harriehausen, B. (1990). *Hmong njua.* Niemeyer.
Harrison, R., Harrison, M., & García, C. (1981). *Diccionario zoque de copainalá.* SIL.
Haspelmath, M. (1993). *A grammar of lezgian.* de Gruyter.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In *The new psychology of language* (Vol. 2, pp. 1–30). Psychology Press.

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

Haspelmath, M., & Karjus, A. (2017). Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency. *Linguistics*, *55*(6), 1213–1235.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.

Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford University Press.

Hayashi, M., & Spreng, B. (2005). Is Inuktitut tenseless? In *Proceedings of the 2005 annual meeting of the canadian linguistics association*.

Haynie, H. J., & Bowern, C. (2016). Phylogenetic approach to the evolution of color term systems. *Proceedings of the National Academy of Sciences*, *113*(48), 13666–13671.

Heath, J. (1999). *A grammar of koyraboro (koroboro) senni*. Köppe.

Hewson, J., & Bubenik, V. (1997). *Tense and aspect in indo-european languages: Theory, typology, diachrony*. Benjamins.

Horton, A. E. (1949). *A grammar of luvale*. Witwatersrand University Press.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Hvitfeldt, E. (2020). themis: Extra recipes steps for dealing with unbalanced data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=themis` (R package version 0.1.2)

Innes, G. (1966). *An introduction to grebo*. University of London (School of Oriental & African Studies).

Jacob, J. M. (1968). *Introduction to cambodian*. Oxford University Press.

Jeon, L., Li, J., Mauney, S., Navarro, A., & Wittke, J. (2015). A basic sketch grammar of Gĩkũyũ. *Rice Working Papers in Linguistics*, *6*.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3.

Jones, W., & Jones, P. (1991). *Barasano syntax*. SIL.

Jordan, F. M. (2011). A phylogenetic analysis of the evolution of Austronesian sibling terminologies. *Human Biology*, *83*(2), 297–321.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Karjus, A., Blythe, R. A., Kirby, S., & Smith, K. (2020). Communicative need modulates competition in language change. *arXiv preprint arXiv:2006.09277*, *1*.

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

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kimball, G. D. (1985). *A descriptive grammar of koasati (louisiana)* (Unpublished doctoral dissertation). Tulane University.

Kimball, G. D. (1991). *Koasati grammar*. University of Nebraska Press.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Krejnovich, E. (1968). O grammaticheskom vyrazhenii imennykh klassov v glagole ketskogo jazyka. In *Studia ketica: Linguistique* (pp. 139–195).

Kruspe, N. (2004). *A grammar of semelai*. Cambridge University Press.

Kuhn, M., & Wickham, H. (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. [Computer software manual]. Retrieved from `https://www.tidymodels.org`

Kunnap, A. (2002). On the Enets evidential suffixes. *Linguistica Uralica*, *38*(2), 145–154.

Kurumada, C., & Grimm, S. (2019). Predictability of meaning in grammatical encoding: Optional plural marking. *Cognition*, *191*, 103953.

Laidig, W. D., & Laidig, C. J. (1990). Larike pronouns: Duals and trials in a central Moluccan language. *Oceanic Linguistics*, *29*(2), 87–109.

Lamberti, M., & Sottile, R. (1997). *The wolaytta language*. Köppe.
Lee, H. S. (1991). *Tense, aspect, and modality: A discourse-pragmatic analysis of verbal affixes in korean from a typological perspective* (Unpublished doctoral dissertation). University of California, Los Angeles.
Lindfors, A. (2003). *Tense and aspect in swahili* (Unpublished doctoral dissertation). Uppsala University.
Lohitare, L. D., Lohammarimoi, D. L., Peter, D. T., & Joseph, P. L. (2012). *Didinga grammar book*.
Lombard, D. P. (1985). *Introduction to the grammar of northern sotho*. van Schaik.
Macaulay, M. A. (1996). *A grammar of chalcatongo mixtec*. University of California Press.
Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.
Maslova, E. (2003). *A grammar of kolyma yukaghir*. de Gruyter.
McCarthy, J., & Hayes, P. J. (1981). Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence* (pp. 431–450). Elsevier.
McGregor, W. (1990). *A functional grammar of gooniyandi*. Benjamins.
Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122.
Merlan, F. (1982). *Mangarayi*. North Holland.
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.
Miller, W. R. (1965). *Acoma grammar and texts*. University of California Press.
Mollica, F., & Kemp, C. (2020). An efficient communication analysis of morpho-syntactic grammatical features. In *Proceedings of the 42th annual conference of the cognitive science society.*
Murane, E. (1974). *Daga grammar*. SIL.
Nedjalkov, I. (1997). *Evenki*. Routledge.
Newman, P. (2000). *The hausa language. an encyclopedic reference grammar*. Yale University Press.
Newmeyer, F. J. (2005). *Possible and probable languages: A generative perspective on linguistic typology*. Oxford University Press.
Nguyen, D.-H. (1997). *Vietnamese*. Benjamins.
Nichols, J. (2019). Why is gender so complex? Some typological considerations. In *Grammatical gender and linguistic complexity volume i: General issues and specific studies.* Language Sciences Press.
Nikolaeva, I. (2014). *A grammar of tundra nenets*. De Gruyter Mouton.
Noonan, M. P. (1992). *A grammar of lango*. de Gruyter.
Osborne, C. R. (1974). *The tiwi language*. Australian Institute of Aboriginal Studies.
Owens, J. (1985). *A grammar of harar oromo (northeastern ethiopia)*. Buske.
Panfilov, V. Z. (1962). *Grammatika nivxskogo jazyka*.
Park, G., Schwartz, H. A., Sap, M., Kern, M. L., Weingarten, E., Eichstaedt, J. C., ... Seligman, M. E. (2017). Living in the past, present, and future: Measuring temporal orientation with language. *Journal of Personality*, *85*(2), 270–280.
Parks, D. R. (1976). *A grammar of pawnee*. Garland.
Pate, J. (2017). Optimization of American English, Spanish, and Mandarin Chinese over time for efficient communication. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society.*
Payne, D. (1990). Yagua. In *Handbook ofamazonian languages* (Vol. 2). de Gruyter.
Penchoen, T. G. (1973). *Tamazight of the ayt ndhir*. Undena Publications.
Piantadosi, S. T. (2016). A rational analysis of the approximate number system. *Psychonomic Bulletin & Review*, *23*, 877–886.
Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.
Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., & Blasi, D. (2021). How (non-)optimal is the lexicon? In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4426–4438). Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.350
Pitman, D. (1980). *Bosquejo de la gramática araona*. Instituto Lingüístico de Verano.
Popjes, J., & Popjes, J. (1986). Canela-Krahô. In *Handbook of amazonian languages* (Vol. 1, pp. 128–199). de Gruyter.

Praulinš, D. (2012). *Latvian: An essential grammar*. Routledge.
R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`
Rácz, P., Passmore, S., Sheard, C., & Jordan, F. M. (2019). Usage frequency and lexical class determine the evolution of kinship terms in Indo-European. *Royal Society Open Science*, *6*.
Refsing, K. (1986). *The ainu language: The morphology and syntax of the shizunai dialect*. Aarhus University Press.
Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In *The handbook of language emergence* (pp. 237–263). Wiley.
Reichenbach, H. (1947). The tenses of verbs. In *The language of time: A reader* (pp. 71–79). Oxford University Press.
Rescher, N., & Urquhart, A. (1971). *Temporal logic*. Springer-Verlag/Wien.
Rios, A., Göhring, A., & Volk, M. (2008). A Quechua-Spanish parallel treebank. In *Lot occasional series* (Vol. 12, pp. 53–64). LOT, Netherlands Graduate School of Linguistics.
Roberts, J. R. (1987). *Amele*. Croom Helm.
Romero-Figueroa, A. (1997). *A reference grammar of warao*. Lincom.
Rood, D. S. (1976). *Wichita grammar*. Garland.
Rounds, C. (2009). *Hungarian: An essential grammar*. Routledge.
Saeed, J. (1999). *Somali*. Benjamins.
Salminen, T. (1997). *Tundra nenets inflection*. Suomalais-Ugrilainen Seura.
Sapir, J. D. (1965). *A grammar of diola fogny*. Cambridge University Press.
Saratsli, D., Bartell, S., & Papafragou, A. (2020). Cross-linguistic frequency and the learnability of semantics: Artificial language learning studies of evidentiality. *Cognition*, *197*, 104194.
Schachter, P., & Otanes, F. T. (1972). *Tagalog reference grammar*. University of California Press.
Schiffman, H. F., & Harold, F. (1999). *A reference grammar of spoken tamil*. Cambridge University Press.
Seiler, W. (1985). *Imonda, a papuan language*. The Australian National University.
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.
Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE Nat'l. Cony. Rec.*, *4*(142-163).
Smeets, C. J. (1989). *A mapuche grammar* (Unpublished doctoral dissertation). Rijksuniversiteit Leiden.
Sridhar, S. N. (1990). *Kannada*. Routledge.
Steinert-Threlkeld, S. (2020). Quantifiers in natural language optimize the simplicity/informativeness trade-off. In *Proceedings of the 22nd Amsterdam colloquium* (p. 513-522).
Stevenson, R. C. (1969). *Bagirmi grammar*. University of Khartoum.
Stump, G. (2017). The nature and dimensions of complexity in morphology. *Annual Review of Linguistics*, *3*, 65–83.
Svantesson, J.-O. (1991). Tense, mood and aspect in Mongolian. *Working Papers (Lund University, Department of Linguistics)*, *38*, 189–204.
Svenonius, P. (2002). Icelandic case and the structure of events. *The Journal of Comparative Germanic Linguistics*, *5*, 197–225.
Swanton, J. (1911). *Haida*. US Government Printing Office.
Szagun, G. (1978). On the frequency of use of tenses in English and German children's spontaneous speech. *Child Development*, *49*(5), 898–901.
Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, *3*(2), 259–278.
Terrill, A. (1999). *Lavukaleve: a papuan language of the solomon islands* (Unpublished doctoral dissertation). Australian National University, Canberra.
Thornell, C. (1997). *The sango language and its lexicon* (Unpublished doctoral dissertation). Lund University.
Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37th annual allerton conference on communication, control and computing.*
Topping, D. M., & Dungca, B. C. (1973). *Chamorro reference grammar*. University of Hawaii Press.

Traugott, E. C. (2011). Grammaticalization and mechanisms of change. In *The oxford handbook of grammaticalization.*

Underhill, R. (1976). *Turkish grammar* (Vol. 460). MIT Press.

Velupillai, V. (2016a). *Hawai'i creole english: A typological analysis of the tense-mood-aspect system.* Springer.

Velupillai, V. (2016b). Partitioning the timeline: A cross-linguistic survey of tense. *Studies in Language*, *40*(1), 93–136.

von der Gabelentz, G. (1901). *Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse.* Leipzig: Chr. Herm. Tauchnitz.

Voorhoeve, C. (1965). *The flamingo bay dialect of the asmat language.* The Hague.

Watkins, L. J. (1984). *A grammar of kiowa.* University of Nebraska Press.

Wdzenczny, D. (n.d.). Temporal expression in Wichí nominals. *Santa Barbara Papers in Linguistics*, *22*.

Willett, T. (1988). A cross-linguistic survey of the grammaticization of evidentiality. *Studies in Language.*, *12*(1), 51–97.

Wilson, D. (1974). *Suena grammar.* SIL.

Wolfart, H. C. (1996). Sketch of Cree, an Algonquian language. In *Handbook of north american indians* (Vol. 17, pp. 390–439). Smithsonian Institution.

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57–70.

Yeon, J., & Brown, L. (2013). *Korean: A comprehensive grammar.* Routledge.

Zaslavsky, N. (2020). *Information-theoretic principles in the evolution of semantic systems* (Unpublished doctoral dissertation). The Hebrew University of Jerusalem.

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 T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd annual meeting of the cognitive science society* (pp. 938–944).

Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. In *41st annual conference of the Cognitive Science Society.*

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press.