Languages optimize the trade-off between lexicon size and average utterance length: A case study of numeral systems

Milica Denić and Jakub Szymanik
University of Amsterdam

Abstract

Human languages vary in terms of which meanings they lexicalize, but there are important constraints on this variation. It has been argued that languages are under pressure to be simple (e.g., to learn or to cognitively represent) and to be informative (i.e., to allow for a precise communication), and that a good compromise between these two pressures determines which meanings get lexicalized (Kemp and Regier 2012 and much subsequent work). We argue that the way informativeness is operationalized in that line of work is problematic because it assumes a communication model where interlocutors communicate with monomorphemic expressions. In reality, however, morphosyntactically complex expressions greatly enrich the range of meanings we are able to express in many semantic domains. One such domain is number: many languages lexicalize few number meanings as monomorphemic expressions, but can precisely convey any number meaning using morphosyntactically complex numerals. We argue that a different notion of communicative efficiency plays a role in which meanings get lexicalized in semantic domains such as number: languages are trying to find a good compromise between the pressure to lexicalize as few meanings as possible (i.e, to minimize lexicon size) and the pressure to produce as morphosyntactically simple utterances as possible (i.e., to minimize average utterance length). This case study in conjunction with previous work on communicative efficiency suggests that, in order to explain which meanings get lexicalized across languages and across semantic domains, a more general approach may be that languages are finding a good compromise between not two but *three* pressures: be simple, be informative, and minimize average utterance length.

Keywords: numerals; number; simplicity; informativeness; average utterance length; trade-off

1 Introduction

Human languages vary in terms of which meanings they lexicalize as monomorphemic expressions. There are nonetheless important constraints on this variation: some meanings or meaning types are very frequently lexicalized, while others very rarely. Such constraints have been identified in both content and functional vocabulary. For instance, in the domain of content words, color terms have a well-established order of appearance in language evolution — certain semantic contrasts must be established before others (Berlin and Kay 1969). In addition, color terms label only convex regions of conceptual color space (Gärdenfors 2014, Jäger 2010). In the domain of function words, multiple semantic constraints have been identified for quantificational determiners (Horn 1972, Barwise and Cooper 1981, Keenan and Stavi 1986, Peters and Westerståhl 2006, Hackl 2009). For instance, the meaning akin to 'majority' is often lexicalized as a quantificational determiner across languages (e.g., English *most*), but the meaning akin to 'minority' never is (Hackl 2009). Where do these constraints come from? Three prominent answers to this questions include learnability (e.g., certain meanings are rarely or not lexicalized because they are harder to learn, cf. Hunter and Lidz 2013, Chemla, Buccola, and Dautriche 2019, Steinert-Threlkeld and Szymanik 2018, 2020, Maldonado and Culbertson 2022), syntax-semantics interface (e.g., certain meanings are not lexicalized because they would be degenerate due to how syntactic structures are semantically interpreted, cf. Fox 2002, Romoli 2015), and communicative efficiency (e.g., certain meanings are rarely or not lexicalized because lexicalizing them wouldn't significantly improve communication success between language users, cf. Kemp and Regier 2012 and much subsequent work).

In this paper, we won't discuss learnability and syntax-semantics interface explanations, and will focus instead on the communicative efficiency explanation. According to the most prominent version of the communicative efficiency explanation, languages are under two competing pressures: the pressure to be simple and the pressure to be informative, i.e., to allow for a precise communication. Natural languages have been argued to lexicalize the meanings which allow them to achieve a good compromise between these two pressures, i.e., *to optimize the simplicity/informativeness trade-off* (Kemp and Regier 2012, Regier, Kemp, and Kay 2015, Xu, Regier, and Malt 2016, Xu, Liu, and Regier 2020, Kemp, Xu, and Regier 2018, Zaslavsky, Kemp, Regier, and Tishby 2018, Steinert-Threlkeld 2019, 2021, Mollica, Bacon, Zaslavsky, Xu, Regier, and Kemp 2021, Denić, Steinert-Threlkeld, and Szymanik 2021, Denić, Steinert-Threlkeld, and Szymanik 2022, Zaslavsky, Maldonado, and Culbertson 2021, Uegaki 2022).

In this paper, we will refine the notion of communicative efficiency which plays a role in which meanings get lexicalized, making four main contributions.

First, we argue that the way informativeness is operationalized in the aforementioned work on communicative efficiency is problematic because it assumes a communication model where interlocutors communicate with monomorphemic expressions. While this simplification may be justified in certain semantic domains, in many semantic domains morphosyntactically complex expressions greatly enrich the range of meanings we are able to express. In other words, when morphosyntax is taken into account, languages may be able to reach very high levels of informativeness even in semantic domains in which they lexicalize very few meanings. Consider a semantic domain of natural numbers: many languages lexicalize few number meanings as monomorphemic expressions, but can precisely convey any natural number meaning using morphosyntactically complex numerals. For instance, even though English doesn't lexicalize the number meaning 61, English speakers are able to precisely convey it with a morphosyntactically complex numeral *sixty-one*. Morphosyntax thus needs to be taken into account when informativeness of a language is computed.

Second, if morphosyntax is taken into account when informativeness is computed, we demonstrate that the simplicity/informativeness trade-off optimization approach cannot explain which meanings get lexicalized, using number meanings and numeral systems as a case study.

Third, we propose another notion of communicative efficiency to explain which number meanings are lexicalized across languages. According to our proposal, languages are under the pressure to lexicalize as few meanings as possible (i.e, to minimize lexicon size) and the pressure to produce as morphosyntactically simple utterances as possible (i.e, to minimize average utterance length). Lexicon size and average utterance length are in competition: reducing the average morphosyntactic complexity of utterances will often require lexicalizing more meanings, and reducing the size of one's vocabulary will often result in needing expressions of greater morphosyntactic complexity to communicate. We demonstrate that natural languages numeral systems indeed optimize the trade-off between lexicon size and average utterance length.

Finally, once these results are in place, we abstract away from numeral systems and discuss which notion of communicative efficiency may explain what meanings get lexicalized across semantic categories more generally. We propose that, in light of the present case study on numeral systems *and* the previous work on simplicity/informativeness optimization approach, in order to explain which meanings get lexicalized across languages and across semantic domains, a more general approach may be that languages are finding a good compromise between not two but *three* pressures: be simple, be informative, and minimize average utterance length.

2 Background: Simplicity/informativeness trade-off

We start by introducing the notion of communicative efficiency as the simplicity/informativeness trade-off optimization.

Complexity One could distinguish between at least two sources of *complexity* (the opposite of *simplicity*) in language: (i) complexity of the lexicon of the language, and (ii) complexity of the morphosyntactic rules of the language. In most existing simplicity/informativeness trade-off analyses the focus is restricted to specific semantic categories and not to the entire language: this has motivated considering the lexicon as the sole source of complexity.[1]

Two approaches to operationalizing the complexity of the lexicon have been explored in the literature (Kemp and Regier 2012, Kemp et al. 2018, Steinert-Threlkeld 2021, Denić et al. 2022, Uegaki 2022, Xu et al. 2020). The first approach is the number of elements in the lexicon (i.e., morphemes). The second approach is rooted in assumptions about cognitive representations of morphemes in, for instance, the language of thought (Fodor 1975): a common assumption is that cognitive representations of morphemes are the shortest combination of primitives of the language of thought with an appropriate interpretation (cf. also van de Pol, Lodder, van Maanen, Steinert-Threlkeld, and Szymanik 2021, Katzir 2014, Feldman 2000, Goodman, Tenenbaum, Feldman, and Griffiths 2008, Piantadosi, Tenenbaum, and Goodman 2016, a.o.). According to that approach, the complexity of the lexicon may correspond to the sum of lengths of cognitive representations of morphemes in the lexicon.

There is however a choice to be made in determining what exactly the morphemes in the lexicon are. This is due to the fact that, in many languages, multiple phonetic forms are associated with the same meaning. For instance, *-ty* in numerals *twen-ty, thir-ty, for-ty,...* is phonetically different from *-teen* in numerals *thir-teen, four-teen,...*, as well as from *ten*, albeit they all semantically denote the number 10 in English. As another

[1] Adding the complexity of morphosyntactic rules can be motivated for semantic categories where learners arguably acquire not only the lexicon, but also the category-specific morphosyntax.

example, consider $-s$ and $-en$ which contribute plural meaning in *cats* and *oxen*. Does this mean that English has multiple lexicon entries for 10, or does it have one lexicon entry for 10 which has different phonetic realizations (similarly for plural)? This question is still debated (cf. Haspelmath (2020) for a recent discussion), but a prominent approach is to take the latter route, that is, treat *ten, -teen, and -ty*, and similarly, *-s and -en* as different phonetic realization of a single abstract morpheme (cf. Halle and Marantz 1993 and much subsequent work). In general, the work in simplicity/informativeness trade-off tradition (explicitly or implicitly) adopts such an approach, measuring the complexity of the lexicon as the number of abstract morphemes (i.e., the number of lexicalized meanings). This choice needs to be kept in mind when the results of such studies are interpreted.

Informativeness The *informativeness* of a language measures how precisely its expressions allow its users to communicate the intended meanings. One way to formalize this notion is as the probability that the users of the language, i.e., a speaker and a listener, will communicate successfully, defined in (1) (Steinert-Threlkeld 2021, Denić et al. 2022, Uegaki 2022; cf. however Kemp and Regier 2012 for an information-theoretic notion of informativeness). This probability depends on the speaker distribution P_S (how likely is the speaker to use a word w from the lexicon of a language L given that they want to convey a meaning m from the set of meanings M?), the listener distribution P_L (how likely is the listener to guess a meaning m from the set of meanings M given that they have heard a word w from the lexicon of a language L?), and the need probability distribution P (how likely are the speaker and a listener to need to communicate about a meaning m from the set of meanings M?). This definition of informativeness ensures that if a language has a non-ambiguous word in the lexicon for every possible meaning, the informativeness of the language will be 1; on the other hand, semantic gaps (not having any word in the lexicon to express a meaning) and ambiguity (not having a non-ambiguous word in the lexicon to express a meaning) in general decrease informativeness of a language: by how much depends on P_L, P_S, and P.

(1) **Informativeness of a language:**

$$I(L) = \sum_{m \in M} \sum_{w \in L} P(m) P_L(m|w) P_S(w|m)$$

Simplicity/informativeness trade-off A language which has a non-ambiguous lexical item for every possible meaning would allow for a maximally precise communication (i.e., it would be maximally informative), but it would be complex. On the other hand, a language which only has one expression that can be used for any meaning would be simple, but completely uninformative. Simplicity and informativeness are in a tension: languages cannot both be minimally complex and maximally informative. This tension is known as the *simplicity/informativeness trade-off problem*. There can be many *optimal solutions* to this problem: the set of optimal solutions is called the *Pareto frontier*. A language is (Pareto) optimal if there is no language better at one of the simplicity and informativeness dimensions without being worse on the other. Remarkably, computational modeling of cross-linguistic semantic data has demonstrated that natural languages are at or very near the Pareto frontier — the sets of meanings that natural languages lexicalize are (in the proximity of) optimal solutions to the trade-off problem (Denić, Steinert-Threlkeld, and Szymanik 2021, Denić et al. 2022, Kemp and Regier 2012, Kemp et al. 2018, Steinert-Threlkeld 2019, 2021, Uegaki 2022, Xu et al. 2020, Zaslavsky et al. 2018, 2021). Importantly, that natural languages optimize the simplicity/informativeness trade-off is not inconsistent with cross-linguistic diversity found in many semantic domains: different Pareto-optimal languages can have very different properties and different natural languages may thus be (approaching) different optimal solutions to the trade-off problem.

3 Proposal: Revising the notion of communicative efficiency

The operationalization of informativeness as in definition (1) in Section 2 is problematic because it doesn't take into account that we can express many meanings which are not lexicalized in our language by means of morphosyntactically complex expressions. For instance, the English word *grandmother* is ambiguous between paternal and maternal grandmother. However, we are able to eliminate ambiguity and thus the possibility of miscommunication if we use syntactically complex *paternal grandmother* or *maternal grandmother*. If the computation of informativeness of a language relies only on the elements of the lexicon as in Section 2, how precisely language users are able to communicate with that language will be significantly underestimated.

One way to correct this would be to consider L in the definition (1) to be not the lexicon of the language, but all expressions of the language, that is, both the lexicon and morphosyntactically complex expressions.

When such a correction is made, does simplicity/informativeness trade-off optimization remain a viable explanation for which meanings get lexicalized in different semantic domains? In the present work, we will show that it doesn't. In order to do so, we will conduct a case study in a semantic domain in which many languages allow for a maximally informative communication: that is, any meaning in this semantic domain can be unambiguously conveyed with a (morphosyntactically complex) expression of that language. The semantic domain in question is the set of natural numbers in languages such as English, French and Serbian, among many others. Languages in which one can unambiguously refer to any natural number are called languages with a *recursive numeral system*. We will ask what determines which number meanings get lexicalized in recursive numeral systems.

We will contrast two hypotheses. The first hypothesis, summarized in (2), is the simplicity/informativeness trade-off optimization hypothesis adapted to semantic domains where languages reach maximal informativeness.

(2) **Hypothesis 1: Simplicity/informativeness trade-off**
In a semantic domain in which maximal informativeness is reached, languages lexicalize the smallest number of meanings which allows them to be maximally informative in view of the language's morphosyntax.

The second hypothesis, summarized in (3), is that languages balance between two different pressures in semantic domains where they reach maximal informativeness: (i) the pressure to minimize the complexity of the language and (ii) the pressure to minimize the average morphosyntactic complexity of utterances. We consider that the relevant dimension of complexity for both (i) and (ii) is the number of linguistic elements, rather than the length of their cognitive representations in, for instance, a language of thought (although one may explore the latter avenue in future work). More specifically, we consider that the complexity of a language is its lexicon size. We take the morphosyntactic complexity of an utterance to be the number of morphemes in it, which we will refer to as *utterance length (UL)*. Average utterance length ($\overline{UL}(S, L)$) for semantic domain S in language L is defined in (4). That pressures (i) and (ii) are in competition is due to the fact that reducing average utterance length will often require lexicalizing more meanings, and reducing the size of one's vocabulary will often result in needing expressions of greater length to communicate precisely.

(3) **Hypothesis 2: Lexicon size/utterance length trade-off optimization**
In a semantic domain in which maximal informativeness is reached, languages lexicalize those meanings which allow them to optimize the trade-off between lexicon size and average utterance length.

(4) Average utterance length $\overline{UL}$ for semantic domain S in language L, where $P(m)$ is the probability that the meaning m needs to be communicated and $UL(m, L)$ is the length of the utterance conveying meaning m in language L:

$$\overline{UL}(S, L) = \sum_{m \in S} P(m) UL(m, L)$$

In what follows, we will show that Hypothesis 1 cannot explain which number meanings are lexicalized in recursive numeral systems, while Hypothesis 2 can shed light on this question. In other words, when languages achieve the maximal degree of informativeness in a semantic domain, the pressures to minimize lexicon size and the pressure to minimize average utterance length are key to understanding which meanings are lexicalized across languages.

In light of these results, we will discuss what can be concluded more generally about pressures at play in determining which meanings get lexicalized — what do these results imply for semantic domains in which languages do not allow for the maximal degree of informativeness, which is arguably a more common scenario? Ultimately, the balance between three pressures — minimize lexicon size, minimize average utterance length and maximize informativeness — may be a promising path to explaining which meanings get lexicalized across semantic domains and languages.

4 Experiment

In this section, we present an experiment[2] investigating how close natural languages' recursive numeral systems are to trading off optimally between lexicon size and average utterance length (Hypothesis 2). We will evaluate this hypothesis by comparing natural languages to artificially generated languages. The artificial languages represent the space of possibilities in terms of lexicon size/average utterance length trade-off in recursive numeral systems, and will reveal what the (approximately) optimal solutions to this problem are. To preview the findings, the results speak in favor of natural languages' recursive numeral systems trading off optimally between lexicon size and average utterance length (Hypothesis 2), and against them trading off optimally between simplicity and informativeness (Hypothesis 1).

[2] All data and scripts used for the analysis are available at https://drive.google.com/file/d/1HnC0axDP24I4FKRujMSSZU5hXD74aQEh/view?usp=sharing.

4.1 Natural languages

We assume that numerals across languages semantically denote numbers (e.g., the numeral *two* denotes the number 2), noting that this is a simplification (see Bylinina and Nouwen (2020), Spector (2013)). We collected cross-linguistic data on number-denoting morphemes and how these are morphosyntactically combined to construct numerals denoting numbers 1-99 in the sample of languages of the *Numeral bases* chapter in *The World Atlas of Language Structures (WALS)* (Comrie 2013).[3] Out of 172 recursive numeral systems in Comrie (2013), 44 were excluded due to challenges with data collection or data interpretation.[4] 128 languages were thus included in the analysis.[5] *WALS* language samples are compiled with an aim to maximize genealogical and areal diversity of languages in them (Comrie, Dryer, Gil, and Haspelmath 2013) — we can thus have some confidence that we are analyzing a representative sample of world's languages' recursive numeral systems.

For each of the 128 studied numeral systems, for each numeral denoting a number in the range 1-99, its morphosyntactic components and their denotations were identified (cf. Table 1 for a few examples of numerals in Georgian); more details on how this was done will be provided in Section 4.2. The studied numeral systems differ in terms of which number meanings they lexicalize as monomorphemic expressions. Some recurring options are listed in (5):

(5) Lexicalized number concepts:

a. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (75 languages)
b. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20 (19 languages)
c. 1, 2, 3, 4, 5, 10 (9 languages)

Even when they lexicalize the same number meanings, languages sometimes make different choices in constructing morphosyntactically complex numerals. For instance, Kunama and Fulfulde lexicalize number meanings in (5c); Kunama constructs the numeral for number 9 as $10 - 1$, while Fulfulde does it as $5 + 4$. As another example, Greek and Georgian lexicalize number meanings in (5b); Greek constructs the numeral for number 45 as $4 \cdot 10 + 5$, while Georgian does it as $2 \cdot 20 + 5$. Morphosyntactically complex numerals for numbers 1-99 across languages reveal that addition, multiplication and subtraction are productively involved in their construction[6] (cf. also the discussion

[3] Two main sources were used to collect the cross-linguistic data. The primary source were descriptive grammars of individual languages, in most cases those referenced in Comrie (2013). When no descriptive grammar of a language was accessible to us, we used as a secondary source the data from the website https://lingweb.eva.mpg.de/channumerals/, maintained by Eugene Chen. This website is a collective effort of language scholars to document world's language's numeral systems. The sources used for each language can be found in Appendix at: https://drive.google.com/file/d/1HnC0axDP24I4FKRujMSSZU5hXD74aQEh/view?usp=sharing.

[4] For some of the languages from the sample in Comrie (2013), no complete description of the numeral system was accessible to us. Furthermore, a small number of languages were excluded due to difficulties with data interpretation. In most cases, this happened when morphosyntax of certain numerals was not aligned with their interpretation (e.g. in Zoque, the numeral for number 9 is morphologically 6+4; this is dubbed 'correct misinterpretation' in Hurford (1975)). Additionally, Danish was excluded for reasons explained in footnote 6 and Ainu was excluded for reasons explained in footnote 7.

[5] The list of analyzed languages and detailed descriptions of their numeral systems can be found in Appendix at: https://drive.google.com/file/d/1HnC0axDP24I4FKRujMSSZU5hXD74aQEh/view?usp=sharing.

[6] In the Danish numeral system, a morpheme denoting the fraction *half* is used to construct certain

Table 1: Georgian numerals for numbers 6, 30 and 62

Denoted number (numeral)	**Morphosyntactic make-up**
6 (*ekvsi*)	6 (*ekvsi*)
30 (*otsdaati*)	20 (*ots-*) + (*-da-*) 10 (*-ati*)
62 (*samotsdaori*)	3 (*sam-*) · 20 (*-ots-*) + (*-da-*) 2 (*-ori*)

of Hurford (1975, 2007) below) — these are sometimes, but not always, morphosyntactically overt. Furthermore, certain number-denoting morphemes play a special role in the construction of morphosyntactically complex numerals (the so-called *bases*). For instance, in 'base-10 languages', morphosyntactically complex numerals for numbers up to 99 are in general constructed according to the morphosyntactic pattern $x \cdot \mathit{10} + n$ (e.g., Greek). On the other hand, in 'base-20 languages' morphosyntactically complex numerals for numbers up to 99 are in general constructed according to the morphosyntactic pattern $x \cdot \mathit{20} + n$ (e.g., Georgian). Many languages behave as 'base-5' languages when it comes to the composition of numerals for numbers 6-9, which they construct as $5 + n$ (e.g., Fulfulde).

Hurford (1975, 2007) in his seminal work proposes that number-denoting morphemes are of the syntactic category Digits (D) or Multipliers (M). Roughly, morphemes denoting numbers which are 'bases' as 10 or 20 in 'base-10' or 'base-20' languages are of the syntactic category M; other morphemes are of the syntactic category D. He proposes that, across languages, numerals are constructed according to the morphosyntactic rules in (6). For instance, consider a language whose $D = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $M = \{10\}$ — in this language, $5 \cdot 10 + 6, 5 \cdot 10 - 6, 10 + 4, 10 - 4$ are examples of morphosyntactically well-formed expressions, while $5 \cdot 2, 5 + 3$ are examples of expressions which are not well-formed. All numerals in all natural languages in this study can be constructed using the grammar in (6).[7]

(6) NUMBER ⟶ D | PHRASE | PHRASE +/- NUMBER
PHRASE ⟶ M | NUMBER · M

4.2 Computing lexicon size and the average utterance length

Our measure of lexicon size is the number of lexicalized meanings. As we pointed out in Section 2, this choice influences how the results are interpreted: we will come back to this in the discussion.

complex numerals; for instance the numeral fifty in Danish contains morphemes three, half and twenty ((three - half) · twenty). This suggests that there is a limited use of division too involved in the composition of numerals. As, to our knowledge, the extent to which division can be used productively in numeral systems is not well understood (e.g., Hurford's morphosyntactic rules for numerals do not incorporate it), we assume for simplicity that division is not available, and exclude Danish from the corpus of natural languages.

[7] One language in *WALS* corpus, Ainu, has numerals which cannot be derived by this grammar. Hurford (1975) is aware of this and discusses the case of Ainu at length; he considers different extensions of the grammar in (6) to accommodate Ainu but provides arguments against each of them, thus leaving the problem of Ainu unresolved. Because we will rely on the grammar in (6) to generate artificial languages and to estimate the Pareto frontier (cf. Section 4.3), we excluded Ainu from the corpus of natural languages.

The average utterance length is computed according to the formula in (4), adapted in (7) for the the semantic domains of natural numbers in the range 1-99. In (7), $UL(n, L)$ is the length of the numeral (i.e., the number of morphemes in it) of the language L denoting the number n and $P(n)$ is the probability that the number n needs to be communicated. We assume that the probabilities that different numbers need to be communicated follow a power-law distribution as in (8) (cf. Dehaene and Mehler 1992, Piantadosi 2016, Xu et al. 2020). Qualitatively, this probability distribution captures that the larger the number n, the lower the need to talk about it. Finding the number of morphemes in a numeral required studying the morphosyntactic patterns according to which numerals for numbers 1-99 are constructed in each of the 128 languages. In most languages, the relevant morphosyntacic patterns were described in the descriptive grammars we consulted, or were otherwise easy to detect. There were however more difficult cases, even in languages we are closely familiar with (for instance, is there a phonetic variant of the morpheme for number 2 in English *twelve*?). In such cases, we applied the following decision rule: if a numeral is an exception to an established morphosyntactic pattern in a language, but shows (phonetic or orthographic, depending on the available data) elements of morpheme(s) which should have been there if the morphosyntactic pattern was respected, we assume that the numeral follows the morphosyntactic pattern but with phonetic variants not seen elsewhere; otherwise we assume that the numeral is monomorphemic. For instance, English is a base-10 language and because of this we may expect that the numeral for 12 will be built from morphemes for 2 and 10; as the numeral *twelve* has elements of other phonetic realizations of the morpheme for the number 2 (e.g., *two*, *twe(n?)-* from *twenty*), we assume that English *twelve* incorporates the morpheme for number 2 (*twe*-?) and 10 (-*lve*?). Of course, one could argue for the application of an alternative decision rule (unless the morphosyntactic pattern is transparent, assume that the numeral is monomorphemic). It turns out that in practice applying this alternative decision wouldn't qualitatively alter the results; we will elaborate on this point in the discussion section, once the results are in place.

(7)
$$\overline{UL}([1,99], L) = \sum_{n \in [1,99]} P(n) UL(n, L)$$

(8) **Prior over numbers:**
$$P(n) \propto n^{-2}$$

4.3 Estimating Pareto frontier

In order to evaluate whether natural languages optimize the trade-off between lexicon size and average utterance length, we need to establish how close they are to the Pareto frontier, i.e., how close they are to the optimal solutions to the lexicon size/average utterance length trade-off problem.

In order to do that, we use an evolutionary algorithm to estimate the Pareto frontier (cf. Steinert-Threlkeld 2019, 2021, Denić et al. 2021, Denić et al. 2022), which involves generations of many artificial languages (i.e., artificial numeral systems). Before we explain how the evolutionary algorithm works, we explain what an artificial language is.

For our purposes, an artificial language consists of (a) a lexicon of number-denoting morphemes, which are of category D or M (cf. the grammar in (6)), and (b) a set of numerals for numbers 1-99 generated according to the morphosyntactic rules in (6) from the lexicon, where each numeral is the shortest expression (or one of them, in case of a tie) for a given number. Due to computational constraints, two restrictions are imposed on the search for the shortest expression for a given number in a language. (1) Expressions of depth x (i.e., expressions with at most x number-denoting morphemes and $x - 1$ arithmetic operators) are incrementally constructed from expressions of lower depths (e.g., expressions of depth 2 are constructed by combining expressions of depth 1). However, at all depths, the meaning of expressions is restricted to be a natural number in $[1, 200]$. (2) If no expression for a number meaning is found with a depth of at most 7, the search is abandoned, and the language is discarded.

The evolutionary algorithm works as follows. First, the generation 0 is created, which consists of 2000 artificial languages. The lexicons of these artificial languages are generated by drawing two random samples of numbers between 1-99; these stand for morphemes of category D and M respectively. As natural languages tend to have very few morphemes of category M (often only 1 or 2), we restrict the size of the random sample for the category M to at most 5 (no such restriction is imposed on the category D). The numerals for numbers 1-99 of these languages are generated as explained above. The dominant languages of a generation (those for which there is no language which is better on one dimension of lexicon size and average utterance length without being worse on the other) each give rise to an equal number of offspring languages, which are obtained via a small number of mutations (between 1 and 3; these mutations included removing a number morpheme from D or M, adding a number morpheme to D or M, and interchanging a number morpheme in D or M, making sure that M is no larger than 5) from dominant languages. The dominant languages from generation 0 together with their offspring languages constitute generation 1, whose size is limited to 2000 languages. This process is repeated for 100 generations. Finally, the dominant languages are selected from the union of the last generation and the natural languages. Each of these dominant languages is a point in a two-dimensional (lexicon size and average utterance length) space; we do spline interpolation of these points to form a Pareto frontier.

4.4 Results

The natural languages and the artificial languages generated through the 100 generations of the evolutionary algorithm are plotted in Figure 1. The estimated Pareto frontier is plotted as the black curve in Figure 1.

We first note that natural languages all lie along or very close to the Pareto frontier in Figure 1. This speaks in favor of natural languages' recursive numeral systems optimizing the trade-off between lexicon size and average utterance length (Hypothesis 2).

What about Hypothesis 1 — do natural languages' recursive numeral systems optimize the trade-off between simplicity and informativeness? The results from Figure 1 speak against this. To see why, recall what it would mean for recursive numeral systems to optimize the trade-off between simplicity and informativeness. As these numeral systems are maximally informative, optimizing the simplicity/informativeness trade-off boils down to lexicalizing the smallest number of meanings allowing for maximal informativeness. That there is quite some variation among natural languages in terms of their lexicon size (cf. Figure 1) speaks against the hypothesis that these systems optimize the simplicity/informativeness trade-off. In fact, if they were optimizing simplicity/informativeness trade-off, one would expect many languages to only lexicalize the number 1, as this would suffice to construct all numerals using the morphosyntactic rules in (6)!

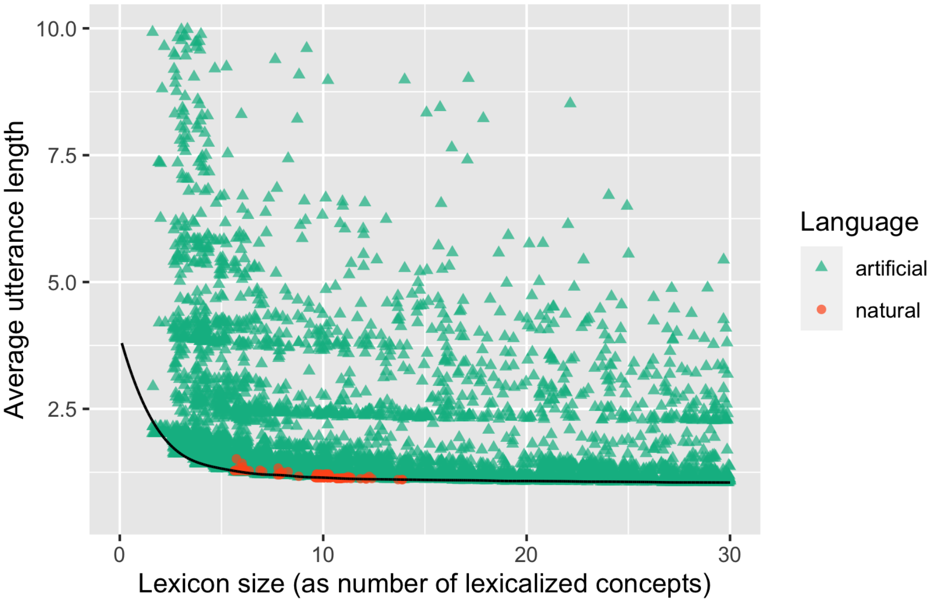

Figure 1: Experiment: Lexicon size and average utterance length of natural languages compared to artificial languages generated via an evolutionary algorithm. Natural languages lie at or very close to the Pareto frontier (black curve).

Finally, it is interesting to notice another property of the data in Figure 1. Natural languages lie along the region of the Pareto frontier where both average utterance length and lexicon size are relatively low. In other words, it would seem that there are levels of average utterance length and lexicon size above which natural languages are, in a sense, not willing to go as far as numeral systems are concerned, even if it is still possible to achieve Pareto-optimality with such higher levels of average utterance length and lexicon size. This would in turn suggest that the lexicon size/average utterance length trade-off optimization is not the only pressure shaping numeral systems: there seem to exist additional pressures, namely to keep lexicon size and average utterance length below a certain level, pushing towards certain regions of the Pareto frontier.

5 Discussion

The results of the experiment reported in Section 4 speak in favor of natural languages' recursive numeral systems trading off optimally between lexicon size and average utterance length (Hypothesis 2), and against them trading off optimally between simplicity and informativeness (Hypothesis 1). We will first discuss in greater detail the interpretation of these findings (Section 5.1). We will then consider these findings from three perspectives: (1) related work (Section 5.2); (2) the specifics of our analysis (Section 5.3); and (3) a more general question of which notion of communicative efficiency may explain which meanings get lexicalized across languages (Section 5.4).

5.1 What number meanings are lexicalized across languages?

Numeral systems differ in terms of which number meanings they lexicalize as monomorphemic expressions (cf. (5) for some examples). Languages seem to lexicalize the following numbers from the range 1-99: (1) the first n number meanings, with n varying across languages, often the first five or the first 10, and (2) a couple of additional number meanings such as 10 and/or 20 as in (5b) or (5c) — these are often morphemes of category M (cf. (6) for the morphosyntactic roles of morphemes of categories D and M). Outside of the 1-99 range, languages lexicalize number meanings such as 100, 1000, 1000000 as morphemes of category M. Why don't languages lexicalize only the first n numbers, which have the highest prior probability according to the probability distribution in (8)? According to the results of the experiment reported in Section 4, natural languages' recursive numeral systems trade off optimally between lexicon size and average utterance length. This suggests that the reason why languages lexicalize number meanings such as 10, 20, 100, 1000, 1000000 is because these allow to construct numerals denoting large(r) numbers using very few morphemes, which reduces the average length of numerals.

One may wonder however to what extent this reduction is significant given the probability distribution in (8), according to which large numbers rarely need to be communicated, and whether the result reported in Section 4 is solely a consequence of (1) above, namely, of the fact that languages often lexicalize the first few number meanings. In other words, does lexicalizing the first n number meanings suffice for a nearly optimal lexicon size/average utterance length trade-off?

To investigate this, we generated a space of artificial languages with the following properties: (a) they lexicalize the first n numbers, with n between 2 and 10; (b) in addition to these, they lexicalize at most two other numbers smaller than 100; (c) at most 2 of number meanings they lexicalize are lexicalized as morphemes of category M. We sampled 50000 languages from this space[8] and analyzed their lexicon size/average utterance length trade-off. We plot them in Figure 2 (*artificial_first_n* languages) against the natural and artificial languages from the main experiment. *Artificial_first_n* languages are not all clustered close to the Pareto frontier (black curve) in Figure 2, which demonstrates that lexicalizing the first n numbers does not guarantee a nearly optimal lexicon size/average utterance length trade-off.

This analysis shows that lexicalizing number meanings such as 10 and/or 20 in addition to the first n number meanings are good choices in terms of the optimality of the trade-off between lexicon size and average utterance length. Of course, these might not be the *only* good choices: what is the space of optimal choices for lexicalization of number meanings and whether additional pressures are at play pushing natural languages towards a subset of that space of optimal choices is an interesting avenue for future work.

5.2 Related work

We discuss three lines of related work. We first discuss how to resolve the tension between our results, and those of Xu et al. (2020) who argue that numeral systems (recursive numeral systems included) optimize the simplicity/informativeness trade-off. We then discuss the connection between the present proposal to three studies highlighting the role of (notions related to) utterance length in language structure: Zipf (1949), Piantadosi, Tily, and Gibson (2011), Haspelmath (2021), Mollica et al. (2021) and Carcassi and Sbardolini (2021) (Section 5.2.2). Finally, we discuss how our results relate to the so-called *packing strategy* proposal (Hurford 1975, 2007) for how morphosyntactically complex numerals are constructed.

[8] Due to computational constraints, it was not possible to analyze the entire space of such languages. Numerals for numbers 1-99 in the sampled languages are generated in the same way as for artificial languages in the main experiment (cf. Section 4.3).

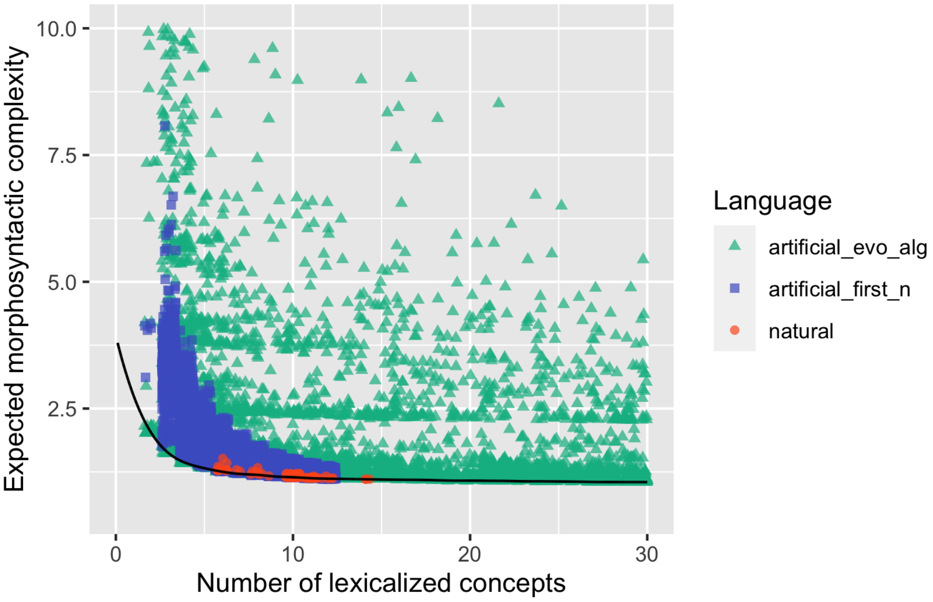

Figure 2: Lexicon size and average utterance length of natural languages compared to artificial languages generated via an evolutionary algorithm (*artificial_evo_lang*) as in Figure 1, with 50 000 artificial languages which lexicalize the first n numbers and at most 2 additional number meanings (*artificial_first_n*). *Artificial_first_n* languages are not all clustered close to the Pareto frontier (black curve), which demonstrates that lexicalizing the first n numbers does not guarantee a nearly optimal lexicon size/average utterance length trade-off.

5.2.1 Numeral systems and the simplicity/informativeness trade-off approach: Xu et al. (2020)

Xu et al. (2020) analyze the simplicity/informativeness trade-off in 24 *restricted* and 6 *recursive* numeral systems. Restricted numeral systems don't have numerals for all numbers: most of them have numerals for only the first few numbers, and use a quantifier such as *many* for any higher number. For instance, the language Krenak only has numerals for numbers 1-3 (Hammarström 2010), and the language Rama only has numerals for numbers 1-5 (Grinevald 1990). Furthermore, the few numerals in restricted numeral systems are often monomorphemic. Recall that recursive numeral systems too have few monomorphemic numerals — in fact, not that many more than in restricted numeral systems — but a numeral for any number can be morphosyntactically constructed. Recursive numeral systems are considered by Xu et al. (2020) to be maximally informative when it comes to communicating about number meanings, while restricted numeral systems have lower degrees of informativeness. On the other hand, according to Xu et al.'s (2020) approach to measuring complexity, the six studied recursive numeral systems are more complex than most of the 24 studied restricted numeral systems. Simplifying somewhat, this is because they assume that the complexity measure of a language should incorporate both the complexity of the lexicon and the complexity of morphosyntactic rules (cf. discussion in 2). As recursive numeral systems always have morphosyntactic rules for building numerals but restricted numeral systems often don't have any, recursive numeral systems tend to have a greater measure of complexity than restricted numeral systems. Recursive numeral systems are thus more complex and more informative than restricted numeral systems in Xu et al.'s (2020) study. Xu et al. (2020) further argue that natural

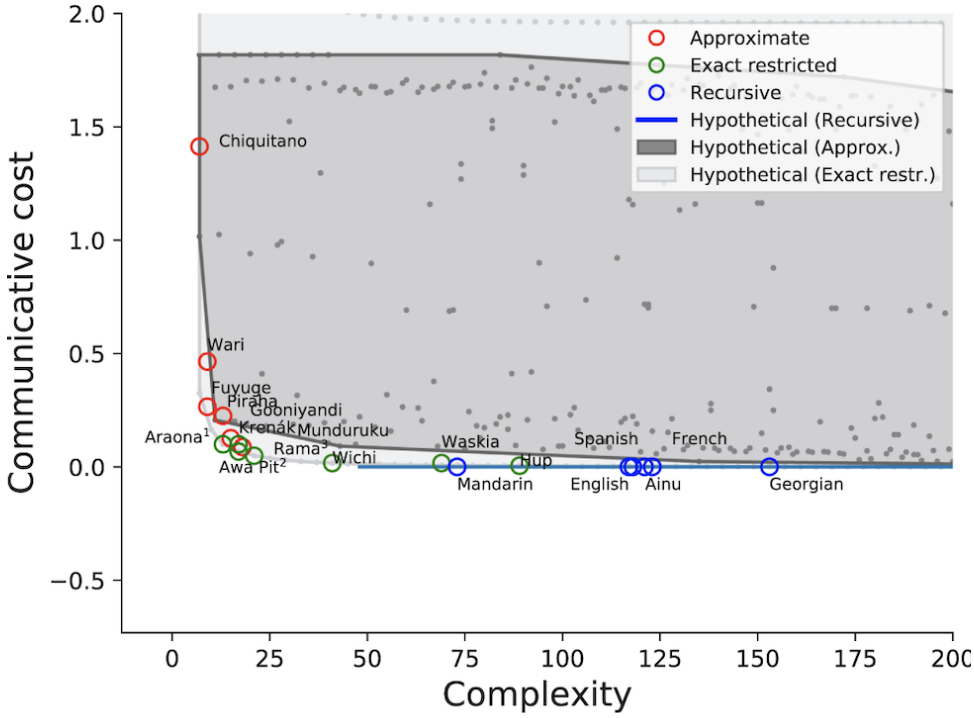

Figure 3: Reproduction of Figure 4b from Xu et al. (2020) plotting complexity and communicative cost (the opposite of simplicity and informativeness) measures of natural languages' numeral systems (red, green and blue circles) and artificial (hypothetical) numeral systems. Blue circles represent the 6 natural languages' recursive numeral systems, and the blue line are hypothetical recursive numeral systems. The Pareto-optimal recursive numeral system is the left-most point of the blue line, and natural languages' recursive numeral systems are not clustered close to it.

languages' numeral systems optimize the simplicity/informativeness trade-off, without making a distinction between restricted and recursive languages.

Their conclusion may seem to be in tension with our findings. However, a careful examination of Xu et al.'s (2020) results reveals that, while restricted numeral systems are indeed close to being Pareto-optimal in trading off simplicity and informativeness, recursive numeral systems are not. Given that recursive numeral systems are maximally informative, if they were optimizing the simplicity/informativeness trade-off, we would expect to find them in the proximity of the minimally complex numeral system which is maximally informative (cf. Hypothesis 1 in Section 3.) According to Xu et al.'s (2020) results (cf. Figure 4b in Xu et al. 2020, reproduced here as Figure 3), the maximally informative system with minimal complexity has complexity ≈ 50. Strikingly, the six recursive numeral systems examined by Xu et al. (2020) have much higher complexity than that, ranging between ≈ 75 and ≈ 150. The results of Xu et al. (2020) thus also speak against recursive numeral systems optimizing the simplicity/informativeness trade-off — in other words, there is no tension between the results of Xu et al. (2020) and our findings. Of course, Xu et al.'s (2020) conclusions still hold for restricted numeral systems which are much closer to being Pareto-optimal than recursive numeral systems (cf. Figure 3. We will come back to this point in Section 5.4.

5.2.2 Utterance length

Zipf (1949) and related work We have argued that languages are under pressure to minimize lexicon size, i.e., the number of lexicalized meanings, as well as to minimize average utterance length, i.e., the average morphosyntactic complexity of produced expressions. The latter pressure is an instance of Zipf's principle of least effort (Zipf 1949), according to which humans are prone to spending the least amount of effort to accomplish a task. In the case of utterance length minimization, humans would be prone to spending the least amount of morphosyntactic structure-building effort to accomplish a communicative task.

Zipf's principle of least effort has mostly been discussed in connection to morpheme or word length: meanings that need to be conveyed more frequently tend to be associated to shorter linguistic forms, where length is operationalized as the number of phonemes in the linguistic form (for data and refinements, see Piantadosi et al. 2011, Haspelmath 2021, Mollica et al. 2021).

It would thus seem that the principle of least effort is operating on two levels, shaping which meanings get lexicalized and how long the linguistic forms associated to them should be: (i) given the number of meanings languages decide to lexicalize, they lexicalize specifically those meanings which allow them to construct as morphosyntactically simple utterances as possible; and (ii) among the lexicalized meanings, the more frequent ones are associated with shorter linguistic forms.

Carcassi and Sbardolini (2021) There is another recent proposal that gives a central role to utterance length in explaining typological patterns, put forward by Carcassi and Sbardolini (2021) for the semantic domain of Boolean connectives. They aim to explain the so-called *nand*-puzzle: across languages, connectives *and* and *or* are often lexicalized; in addition to them, the negated disjunction *nor* is sometimes lexicalized, but a negated conjunction **nand* is never lexicalized (Horn 1972). Carcassi and Sbardolini (2021) propose that the systems of Boolean connectives are optimal solutions to two pressures. The first pressure is the desire to minimize production effort when expressing observations (what we have called *average utterance length*). For instance, if languages don't lexicalize *nand*, they have to use syntactically more complex *not (A and B)* instead of *A nand B* to express the observation that at least one of A and B is false: lexicalizing *nand* thus reduces the average utterance length. The second pressure, labeled *conceptual complexity* in Carcassi and Sbardolini (2021), relates to how sentences with connectives are used to update the information shared by conversational participants (= *context*), which is modeled as the set of possible worlds (Stalnaker 1978, 1999). According to their proposal, upon accepting a sentence *A nor B*, the conversational participants update the context C as follows: first, they restrict C to the set of worlds where A is false, and then they further restrict C to the set of worlds where B is false. On the other hand, upon accepting a sentence *A nand B*, the conversational participants update the context C as follows: first they consider a hypothetical context C', obtained by restricting the original context C to the set of worlds where A is true, followed by a further restriction of C' to the set of worlds where B is true. They then remove the hypothetical context C' from the original context C. Carcassi and Sbardolini (2021) assume that how context is updated by *A nand B* is more conceptually complex that how context is updated by *A nor B* because of the creation of the hypothetical context C' in the former but not in the latter case. They propose that languages are under pressure to minimize the total conceptual complexity of contextual updates of the connectives they lexicalize. They argue that the *nand*-puzzle can be explained on the assumption that languages optimize the trade-off between average utterance length and the conceptual complexity of contextual update procedures of lexicalized connectives.

Our proposal shares with Carcassi and Sbardolini (2021) the idea that languages are under pressure to minimize average utterance length. Where the two proposals differ is what the competing pressure is: in our case, the pressure to minimize how many meanings are lexicalized, while in Carcassi and Sbardolini (2021) the competing pressure relates to another aspect of language use, namely, how conversational participants proceed to update contextual information with sentences with lexicalized connectives.

5.2.3 Packing strategy

Recall that Hurford (1975, 2007) proposes that morphosyntactically complex numerals are constructed according to the grammar in (6) (see Section 4.1). He also observes that this grammar overgenerates — for a given lexicon, it may predict that multiple numerals for the same number are available, which is usually not the case in natural languages. For instance, consider a language L whose $D = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $M = \{10, 20\}$. The grammar in (6) predicts that the numeral for 80 in L can be constructed as $4 \cdot 20$, as $8 \cdot 10$, as $5 \cdot 20 - 20$, as $10 \cdot 10 - 20$, etc. How does L 'decide' on one of these options? Our results suggests that utterance length (i.e., the number of morphemes in a numeral) is a major factor, otherwise natural languages wouldn't be nearly optimal in trading off between lexicon size and average utterance length (recall that numerals of artificial languages in our study are the shortest expressions with appropriate denotations constructed according to the grammar in (6)). This would for instance explain why natural languages with a lexicon as that of L may be more likely to construct the numeral for number 80 as $4 \cdot 20$ or as $8 \cdot 10$ rather than $5 \cdot 20 - 20$ or $10 \cdot 10 - 20$. However, how would L choose between $4 \cdot 20$ and $8 \cdot 10$, which have the same number of morphemes?

Hurford (1975, 2007) proposes a constraint called *packing strategy* to select between multiple expressions denoting the same number, according to which the sister constituent of Number category in the expression should denote the highest possible number (cf. the grammar in (6) for morphosyntactic configurations of Number category). As Hurford (1975) discusses in detail, packing strategy covers a lot of empirical ground. However, it also faces many empirical challenges. For instance, it predicts that languages with $D = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $M = \{10, 20\}$ should construct 80 as $4 \cdot 20$ and not as $8 \cdot 10$ because 10 and 20 would be the sister constituent to Number category in these constructions and $20 > 10$. However, in our corpus, out of 19 language with this lexicon, there doesn't seem to be a clear tendency for one or the other option: approximately the same number of languages opts for $4 \cdot 20$ as for $8 \cdot 10$.

To our knowledge, there haven't been attempts to improve on the packing strategy proposal and address its empirical challenges. It is thus an interesting avenue for future research: what considerations, in addition to utterance length, influence how languages construct their numerals when grammar allows for multiple options?

5.3 Choice points

Our analysis of lexicon size/average utterance length in numeral systems included (at least) three important choice points.

The first choice point relates to how the lexicon size is computed. As explained in Section 4.2, we consider the lexicon size to be the number of abstract morphemes (i.e., the number of lexicalized meanings), rather than the total number of different phonetic realizations of different meanings. This simply means that our result shows that languages optimize the trade-off between the number of lexicalized meanings and the average utterance length, and it doesn't show that they optimize the trade-off between the number of phonetic form-meaning pairs and the average utterance length. Of course, an important question that remains open is why languages have different phonetic realizations of a single abstract morpheme.

The second choice point relates to how the number of morphemes in a numeral is computed (cf. Section 4.2). Recall that, when the number of morphemes in a numeral was difficult to establish, we applied the following decision rule: if a numeral is an exception to an established morphosyntactic pattern in a language, but shows (phonetic or orthographic, depending on the available data) elements of morpheme(s) which should have been there if the morphosyntactic pattern was respected, we assume that the numeral follows the morphosyntactic pattern but with phonetic variants not seen elsewhere; otherwise we assume that the numeral is monomorphemic. The alternative would have been to treat all difficult cases as monomorphemic. Importantly, however, applying the alternative decision rule (treat all difficult cases as monomorphemic) wouldn't alter the results qualitatively. With such an alternative decision rule, natural languages would have slightly larger lexicon sizes and slightly lower average utterance lengths. In other words, they would be shifted slightly down-right-wards in Figure 1. It is easy to see that they would remain close to the Pareto frontier under such a shift.

The third choice point relates to the prior over numbers. As explained in Section 4.2, we assume that the probabilities that different numbers need to be communicated follow a power-law distribution as in (8) (cf. Dehaene and Mehler 1992, Piantadosi 2016, Xu et al. 2020). Importantly, we also assume that this prior is the same for all cultures/languages (at least those with a recursive numeral system). These assumptions may ultimately prove to be wrong. One could however test the robustness of our findings under different assumptions about priors.

5.4 Outlook — three pressures

We have argued that, in a semantic domain in which languages achieve the maximal degree of informativeness, they lexicalize those meanings which allow them to optimize the trade-off between lexicon size and average utterance length (Hypothesis 2). They do not lexicalize the smallest number of meanings which allows them to be maximally informative in view of the language's morphosyntax (Hypothesis 1).

However, semantic domains in which languages do not allow for the maximal degree of informativeness are arguably more common. Consider for example colors: morphosyntactically complex expressions can improve how precisely we communicate about colors compared to the communication with monomorphemic color terms only, but the improvement is arguably limited (*sea blue*, *sky blue*, etc.). Furthermore, there are cross-linguistic differences with respect to whether the maximal degree of informativeness is achieved even within a single semantic domain: we have seen that recursive numeral systems achieve the maximal degree of informativeness for the semantic domain of natural numbers, while restricted numeral systems do not. It is an interesting question why restricted numeral systems do not achieve the maximal degree of informativeness: do they lack morphemes for arithmetic operators, or are some other components of morphosyntactic rules in (6) lacking in these languages? Be that as it may, the simplicity/informativeness trade-off approach seems to provide a good explanation for which number meanings get lexicalized in restricted numeral systems (cf. Xu et al. 2020 and the discussion in Section 5.2.1). In other words, in semantic domains in which the maximal degree of informativeness cannot be reached, considerations related to maximizing informativeness seem to play a role in determining which meanings get lexicalized.

We thus have evidence that (at least) three pressures are shaping which meanings get lexicalized across languages: minimize complexity of the lexicon, minimize average utterance length and maximize informativeness. We thus propose the following hypothesis to be pursued in future work: *languages lexicalize those meanings which allow them to be optimal solutions to the three pressures.* In the extreme case where informativeness is maximal (as in recursive numeral systems), this reduces to finding optimal solutions to the complexity of the lexicon/average utterance length trade-off problem, for which we have seen evidence in the present study. In the other extreme case where interlocutors communicate about meanings from a semantic domain using single monomorphemic words only (possibly in restricted numeral systems), this reduces to finding optimal solutions to the simplicity of the lexicon/informativeness trade-off problem, as in Xu et al. (2020).

6 Conclusion

In this paper, we ask what explains which meanings get lexicalized across languages. We pursue the explanation according to which languages lexicalize meanings which allow them to support efficient communication. We however put forward a novel proposal for what it means for a language to support efficient communication.

In particular, we show that the standard approach to communicative efficiency — the simplicity/informativeness trade-off optimization approach — cannot explain which meanings get lexicalized in natural languages' recursive numeral systems. We put forward a different approach to communicative efficiency, according to which languages are under the pressure to lexicalize as few meanings as possible (i.e, to minimize lexicon size) and the pressure to produce as morphosyntactically simple utterances as possible (i.e, to minimize average utterance length). We show that natural languages' recursive numeral systems indeed optimize the lexicon size/average utterance length trade-off.

Importantly, recursive numeral systems are an extreme case in the sense that they allow for a maximally informative communication about natural numbers. Arguably, achieving the maximal degree of informativeness is not possible in many semantic domains, and there is evidence that the pressure to maximize informativeness plays a role in determining which meanings are lexicalized in such domains (cf. Section 5.4). We thus put forward a more general proposal how supporting efficient communication shapes which meanings get lexicalized across languages:

(9) **Simplicity/informativeness/average utterance length trade-off**
Languages lexicalize those meanings which allow them to solve the trade-off problem between the simplicity, informativeness, and average utterance length in a (nearly) optimal way.

Funding

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007- 2013) / ERC Grant Agreement n. STG 716230 CoSaQ.

References

Jon Barwise and Robin Cooper. Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence*, pages 241–301. Springer, 1981.

Brent Berlin and Paul Kay. *Basic color terms: Their universality and evolution*. Univ of California Press, 1969.

Lisa Bylinina and Rick Nouwen. Numeral semantics. *Language and Linguistics Compass*, 14(8):e12390, 2020.

Fausto Carcassi and Giorgio Sbardolini. Updates and boolean universals. 2021.

Emmanuel Chemla, Brian Buccola, and Isabelle Dautriche. Connecting content and logical words. *Journal of Semantics*, 36(3):531–547, 2019.

Bernard Comrie. Numeral bases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/chapter/131`.

Bernard Comrie, Matthew S. Dryer, David Gil, and Martin Haspelmath. Introduction. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/chapter/s1`.

Stanislas Dehaene and Jacques Mehler. Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1):1–29, 1992.

Milica Denić, Shane Steinert-Threlkeld, and Jakub Szymanik. Complexity/informativeness trade-off in the domain of indefinite pronouns. *Semantics and Linguistic Theory*, 30:166, mar 2021. doi: 10.3765/salt.v30i0.4811.

Milica Denić, Shane Steinert-Threlkeld, and Jakub Szymanik. Complexity/informativeness trade-off in the domain of indefinite pronouns. In *Semantics and linguistic theory*, volume 30, pages 166–184, 2021.

Milica Denić, Shane Steinert-Threlkeld, and Jakub Szymanik. Indefinite pronouns optimize the simplicity/informativeness trade-off. *Cognitive Science*, 2022.

Jacob Feldman. Minimization of Boolean complexity in human concept learning. *Nature*, 407(6804):630–633, 2000.

Jerry A. Fodor. *The language of thought*, volume 5. Harvard University Press, 1975.

Danny Fox. Antecedent-contained deletion and the copy theory of movement. *Linguistic inquiry*, 33(1):63–96, 2002.

Peter Gärdenfors. *The Geometry of Meaning*. The MIT Press, 2014.

Noah D Goodman, Joshua B Tenenbaum, Jacob Feldman, and Thomas L Griffiths. A rational analysis of rule-based concept learning. *Cognitive science*, 32(1):108–154, 2008.

Colette G Grinevald. A grammar of rama. 1990.

Martin Hackl. On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics*, 17(1):63–98, 2009.

Morris Halle and Alec Marantz. Distributed morphology and the pieces of inflection. In Kenneth Hale and Samuel Jay Keyser, editors, *The view from Building 20*, pages 111–176. MIT press, Cambridge, 1993.

Harald Hammarström. *Rarities in numeral systems.* De Gruyter Mouton, 2010.

Martin Haspelmath. The morph as a minimal linguistic form. *Morphology*, 30(2):117–134, 2020.

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

Laurence Robert Horn. *On the semantic properties of logical operators in English.* University of California, Los Angeles, 1972.

Tim Hunter and Jeffrey Lidz. Conservativity and learnability of determiners. *Journal of Semantics*, 30(3):315–334, 2013.

James R Hurford. *The linguistic theory of numerals*, volume 16. Cambridge University Press, 1975.

James R Hurford. A performed practice explains a linguistic universal: Counting gives the packing strategy. *Lingua*, 117(5):773–783, 2007.

Gerhard Jäger. Natural Color Categories Are Convex Sets. In Maria Aloni, Harald Bastiaanse, Tikitu de Jager, and Katrin Schulz, editors, *Logic, Language, and Meaning: Amsterdam Colloquium 2009*, pages 11–20. 2010.

Roni Katzir. A cognitively plausible model for grammar induction. *Journal of Language Modelling*, 2(2):213–248, 2014.

Edward L Keenan and Jonathan Stavi. A semantic characterization of natural language determiners. *Linguistics and philosophy*, pages 253–326, 1986.

Charles Kemp and Terry Regier. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054, 2012. doi: 10.1126/science.1218811.

Charles Kemp, Yang Xu, and Terry Regier. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4:109–128, 2018. doi: 10.1146/annurev-linguistics-011817-045406.

Mora Maldonado and Jennifer Culbertson. Person of interest: Experimental investigations into the learnability of person systems. *Linguistic Inquiry*, 53(2):295–336, 2022.

Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, 118(49), 2021.

Stanley Peters and Dag Westerståhl. *Quantifiers in language and logic.* Oxford University Press, 2006.

Steven T Piantadosi. A rational analysis of the approximate number system. *Psychonomic bulletin & review*, 23(3):877–886, 2016.

Steven T Piantadosi, Harry Tily, and Edward Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9): 3526–3529, 2011.

Steven T Piantadosi, Joshua B Tenenbaum, and Noah D Goodman. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123(4):392, 2016.

Terry Regier, Charles Kemp, and Paul Kay. Word meanings across languages support efficient communication. In Brian MacWhinney and William O'Grady, editors, *The Handbook of Language Emergence*, volume 87, pages 237–263. Wiley-Blackwell, Hoboken, NJ, 2015. doi: 10.1002/9781118346136.ch11.

Jacopo Romoli. A structural account of conservativity. *Semantics-Syntax Interface*, 2(1): 28–57, 2015.

Benjamin Spector. Bare numerals and scalar implicatures. *Language and Linguistics*

Compass, 7(5):273–294, 2013.

Robert C Stalnaker. Assertion. In *Pragmatics*, pages 315–332. Brill, 1978.

Robert C Stalnaker. *Context and content: Essays on intentionality in speech and thought*. Clarendon Press, 1999.

Shane Steinert-Threlkeld. Quantifiers in natural language optimize the simplicity/informativeness trade-off. In *Amsterdam Colloquium 2019*, 2019. URL http://events.illc.uva.nl/AC/AC2019/uploaded_files/inlineitem/Steinert-Threlkeld_Quantifiers_in_natural_language_.pdf.

Shane Steinert-Threlkeld. Quantifiers in natural language: Efficient communication and degrees of semantic universals. *Entropy*, 23(10):1335, 2021.

Shane Steinert-Threlkeld and Jakub Szymanik. Learnability and semantic universals. *Semantics and Pragmatics*, 2018.

Shane Steinert-Threlkeld and Jakub Szymanik. Ease of learning explains semantic universals. *Cognition*, 195:104076, 2020.

Wataru Uegaki. The informativeness/complexity trade-off in the domain of boolean connectives. *Linguistic Inquiry*, pages 1–39, 2022.

Iris van de Pol, Paul Lodder, Leendert van Maanen, Shane Steinert-Threlkeld, and Jakub Szymanik. Quantifiers satisfying semantic universals are simpler. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

Yang Xu, Terry Regier, and Barbara C. Malt. Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, 40(8):2081–2094, 2016. doi: 10.1111/cogs.12312.

Yang Xu, Emmy Liu, and Terry Regier. Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, 4: 57–70, 2020. doi: 10.1162/opmi_a_00034.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018.

Noga Zaslavsky, Mora Maldonado, and Jennifer Culbertson. Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 2021.

George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA, 1949.