Recursive numeral systems optimize the trade-off between lexicon size and average morphosyntactic complexity

Milica Denić[1] and Jakub Szymanik[2]

[1] *Tel Aviv University*
[2] *University of Trento*

Abstract

Human languages vary in terms of which meanings they lexicalize, but this variation is constrained. It has been argued that languages are under two competing pressures: the pressure to be simple (e.g., to have a small lexicon) and to allow for informative (i.e., precise) communication, and that which meanings get lexicalized may be explained by languages finding a good way to trade off between these two pressures (Kemp and Regier 2012 and much subsequent work). However, in certain semantic domains, languages can reach very high levels of informativeness even if they lexicalize very few meanings in that domain. This is due to productive morphosyntax and compositional semantics, which may allow for construction of meanings which are not lexicalized. Consider the semantic domain of natural numbers: many languages lexicalize few natural number meanings as monomorphemic expressions, but can precisely convey very many natural number meanings using morphosyntactically complex numerals. In such semantic domains, lexicon size is not in direct competition with informativeness. What explains which meanings are lexicalized in such semantic domains? We will propose that in such cases, languages need to solve a different kind of trade-off problem: the trade-off 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 morphosyntactic complexity of utterances). To support this claim, we will present a case study of 128 natural languages' numeral systems, and show computationally that they achieve a near-optimal trade-off between lexicon size and average morphosyntactic complexity of numerals. This study in conjunction with previous work on communicative efficiency suggests that languages' lexicons are shaped by a trade off between not two but *three* pressures: be simple, be informative, and minimize average morphosyntactic complexity of utterances.

Keywords: numerals; number; simplicity; informativeness; average morphosyntactic complexity; trade-off

1 Introduction

Human languages vary in terms of which meanings they lexicalize into simple morphemes.[1] 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, no language lexicalizes a quantificational determiner whose meaning is *not every* (Horn 1972). 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 improve communicative efficiency of a language, 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 the pressure to be simple (e.g., to have a small lexicon) while simultaneously being under the pressure to be informative, i.e., to allow for precise communication. These two pressures are in competition. For instance, if a language only has one word for colors, it will have a small lexicon, but it won't allow for very precise communication about colors. Adding more words for various colors would allow for more precise communication, but at the cost of having a larger lexicon. 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, Kemp, Xu, and Regier 2018, Zaslavsky, Kemp, Regier, and Tishby 2018, Xu, Liu, and Regier 2020, Steinert-Threlkeld 2019, 2021, Mollica, Bacon, Zaslavsky, Xu, Regier, and Kemp 2021, Denić, Steinert-Threlkeld, and Szymanik 2021, 2022, Zaslavsky, Maldonado, and Culbertson 2021, Uegaki 2022).

However, in certain semantic domains, it is possible to reach very high levels of informativeness even if very few meanings from that semantic domain are lexicalized. This is due to productive morphosyntax and compositional semantics, which may allow for construction of meanings which are not lexicalized. Consider a semantic domain of natural numbers in languages with a so-called *recursive numeral system*[2]: these are languages in

[1] We consider that a meaning is lexicalized if there is a simple morpheme carrying this meaning in a language. Whether this morpheme can or can't stand alone as a single word is not part of our criterion for lexicalization. For instance, English lexicalizes the number meaning *six* into a morpheme which can stand alone as a single word, and it lexicalizes plural meaning in a morpheme *-s*, among others, which cannot stand alone as a single word.

[2] We will use the the term 'number' for arithmetic values, and the term 'numeral' for linguistic expressions denoting numbers. For instance, in English, the numeral *one* denotes the number 1.

which (practically)[3] any natural number meaning can be expressed precisely (e.g., English). Many such languages lexicalize few natural number meanings, but can precisely convey (practically) any natural number meaning using morphosyntactically complex numerals (e.g., *sixty-one* in English). In such semantic domains, lexicon size is not in direct competition with informativeness. What explains which meanings are lexicalized in these domains?

We will propose that, in semantic domains in which productive morphosyntax enables precise communication even with very few lexicalized meanings, 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 morphosyntactic complexity of utterances). Lexicon size and average morphosyntactic complexity of utterances are in competition in such domains: reducing average morphosyntactic complexity of utterances will often require lexicalizing more meanings, and reducing the size of the lexicon will often result in needing utterances of greater morphosyntactic complexity to communicate. *We thus propose that, in such domains, languages lexicalize those meanings which allow them to optimize the trade-off between lexicon size and average morphosyntactic complexity of utterances.*

We will evaluate this proposal within a case study on lexicalized number meanings in languages with a recursive numeral system, and present evidence that recursive numeral systems indeed optimize the trade-off between lexicon size and average morphosyntactic complexity of numerals. This conclusion is in tension with Xu et al. (2020), who have instead argued that numeral systems —including recursive numeral systems — optimize the simplicity/informativeness trade-off. We review their results in Section 4; we will argue that their results show that other types of numeral systems optimize the simplicity/informativeness trade-off, but that they do not show that recursive numeral systems optimize the simplicity/informativeness trade-off.

Furthermore, we will discuss several other lines of work (Zipf 1949, Piantadosi, Tily, and Gibson 2011, Haspelmath 2021, Mollica et al. 2021, Carcassi and Sbardolini 2022) which share with the present work the idea that speakers attempt to minimize complexity of their utterances, and explain in what way they are different from the proposal we are pursuing.

Finally, once our results are in place, we will abstract away from numeral systems and discuss which notion of communicative efficiency may explain what meanings get lexicalized across semantic categories more generally. We will propose that, in light of the present case study on numeral systems *and* the previous work on simplicity/informativeness optimization (e.g., Kemp and Regier 2012), 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 morphosyntactic complexity of utterances.

[3]Whether recursive numeral systems truly allow to express *any* natural number, or (just) a very large range of natural numbers is debated. Greenberg (1990) argues that even recursive numeral systems have a ceiling, i.e., the largest number that can be generated by that system — e.g., for English, Greenberg (1990) suggests that this number is $10^{36} - 1$. Comrie (2020) however contests this claim, arguing that English allows reference to any natural number. For the purposes of our study, it won't matter much whether recursive numeral systems allow to refer to any natural number, or (just) to a very large range of natural numbers: as we will explain later on, we will investigate numeral expressions cross-linguistically for numbers up to 100, and all studied languages had all numerals in that range.

2 Experiment

We will present an experiment investigating how close natural languages' recursive numeral systems are to trading off optimally between lexicon size and average morphosyntactic complexity of numerals. We will evaluate this by comparing natural languages to artificially generated languages. The artificial languages represent the space of possibilities in terms of lexicon size/average morphosyntactic complexity of numerals trade-off in recursive numeral systems, and will reveal what the (approximately) optimal solutions to the trade-off problem are.

2.1 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 numerals denoting numbers 1-99 and their morphosyntactic components in recursive numeral systems from the sample of languages in the *Numeral bases* chapter in *The World Atlas of Language Structures (WALS)* (Comrie 2013). *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 will be analyzing a representative sample of world's languages' recursive numeral systems.

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 Chan. This website is a collective effort of language scholars to document world's language's numeral systems. Out of 172 recursive numeral systems in Comrie (2013), 44 were excluded due to: (i) challenges with data collection (when no complete description of the numeral system was accessible to us) or (ii) challenges with data interpretation. The challenge with data interpretation that led to the exclusion of certain languages was, in all but two cases (Danish and Ainu), that morphosyntax and semantics of numerals were sometimes misaligned (for instance in Zoque, the numeral for number 9 is morphologically 6+4; this is dubbed 'correct misinterpretation' in Hurford (1975)). The challenge with data interpretation for Danish is explained in footnote 4 and the challenge with data interpretation for Ainu is explained in footnote 6. 128 languages were thus included in the analysis: their list is in Appendix A. Descriptions of the numeral systems of the 128 languages and the sources used for each language can be found in online Appendix at: https://drive.google.com/file/d/1kHm1vwf9bMYZcSgyj41tm1PjlgX3ovb_/view?usp=sharing.

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 and Table 2 for a few examples of numerals in Georgian and Fulfulde respectively). These numeral systems differ in terms of which number meanings they lexicalize. Some recurring options are listed in (1). The generalization seems to be that (in most cases) they lexicalize the following numbers from the range 1-99: (i) the first n numbers, with n varying across languages, often the first five or the first 10, and (ii) a couple of additional numbers such as 10 and/or 20 as in (1b) or (1c).

(1) Lexicalized numbers:

a. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (74 languages)
b. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20 (20 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 (1c); 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 (1b); 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[4] (cf. Hurford 1975, 2007). Both number-denoting morphemes and morphemes denoting arithmetic operators can in principle be phonetically overt or covert: in practice, number-denoting morhemes are very rarely covert, while morphemes denoting arithmetic operators often are.[5] 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 $x \cdot 10 + n$ (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 $x \cdot 20 + n$ (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).

Finding morphosyntactic components of numerals required studying the morphosyntactic patterns for numerals in each of the 128 languages. In most languages, the relevant morphosyntacic patterns were explicitly described in the descriptive grammars we consulted, or were otherwise easy to deduce from the list of numerals in the language. Some numerals may however be an exception to an established morphosyntactic pattern in a language; we will now describe how different types of exceptions were treated. We can categorize exceptions into 3 types. Type 1: These are numerals which bear no connection whatsoever to morphemes we may expect to find in them had the morphosyntactic patterns been respected. For instance, English is a base-10 language, and we may thus expect to find morphemes for 1 and 10 in the numeral for 11. The numeral 'eleven' however doesn't seem to bear any relation to morphemes denoting 1 or 10. In such cases, we consider the numeral to be monomorphemic. Type 2: These are numerals in which some of the morphemes are clearly identifiable, while others are not; we can however rely on the clearly identifiable morphemes to deduce a full morphological analysis. For instance, in the English numeral *twenty*, the morpheme *-ty* is clearly identifiable (it occurs also in *thirty, forty, fifty* etc.), and we can thus conclude that *twen-* is also a morpheme, in

[4]In the Danish numeral system, a morpheme denoting the fraction *half* is used to construct certain 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 (cf. Hurford 1975, 2007), we assume for simplicity that division is not available, and exclude Danish from the corpus of natural languages.

[5]We postulate covert morphemes only when this is necessary to account for the semantics of numerals; for instance, to account that the meaning of the Fulfulde numeral *jowe go'o* (6) is computed from the meaning of *jowe* (5) and *go'o* (1) (cf. Table 2), we need to assume that these are combined by an addition operator, which is introduced by a covert morpheme.

Table 1: Georgian numerals for numbers 6, 30 and 62. ∅ in the 'Morphosyntactic make-up' column indicates that a morpheme is covert, i.e., it doesn't have a phonetic realization.

Denotation (numeral)	**Morphosyntactic make-up**	**Number of morphemes**
6 (*ekvsi*)	6 (*ekvsi*)	1
30 (*otsdaati*)	20 (*ots-*) + (*-da-*) 10 (*-ati*)	3
62 (*samotsdaori*)	3 (*sam-*) · (∅) 20 (*-ots-*) + (*-da-*) 2 (*-ori*)	5

Table 2: Fulfulde numerals for numbers 6, 30 and 62. ∅ in the 'Morphosyntactic make-up' column indicates that a morpheme is covert, i.e., it doesn't have a phonetic realization.

Denotation (numeral)	**Morphosyntactic make-up**	**Number of morphemes**
6 (*jowe go'o*)	5 (*jowe*) + (∅) 1 (*go'o*)	3
30 (*chappande tatti*)	10 (*chappande*) · (∅) 3 (*tatti*)	3
62 (*chappande jowe go'o i didi*)	10 (*chappande*) · (∅) (5 (*jowe*) + (∅) 1 (*go'o*)) + (*i*) 2 (*didi*)	7

particular, that it is an allomorph (i.e., a variant phonetic realization) of the morpheme *two*. Type 3: These numerals are in between Type 1 and Type 2: i.e., we cannot divide them into substrings for which we can argue based on independent evidence that they are morphemes, but we can nonetheless identify in them (phonetic or orthographic, depending on the available data) elements of morphemes we may expect to find in them had the morphosyntactic patterns been respected. For instance, in English *twelve*, we seem to find some elements of the morpheme for number 2 (i.e., (*twe-*?) may be another allomorph of 2). We will refer to the cases of the Type 3 as unclear cases. Such unclear cases were encountered in 34 languages, listed in Appendix B. In most of these 34 languages, only 1 or 2 numerals were concerned (e.g., in English, the only concerned numeral is *twelve*). We opted to not exclude these 34 languages from the main analysis (however, the supplementary analysis with these 34 languages excluded can be seen in Appendix B), but rather analyze their unclear cases (very simplistically) as morphologically complex: i.e., when a numeral is an exception to an established morphosyntactic pattern in a language, but partially overlaps in its phonetic or orthographic elements with some of the morpheme(s) which should have been there if the morphosyntactic pattern was respected, we will assume that the numeral follows the morphosyntactic pattern but with phonetic variants not seen elsewhere. For instance, we will assume that English *twelve* incorporates the morpheme for number 2 (*twe-*?) and 10 (*-lve*?). As we explain in footnote 8, potential errors due to this simplification wouldn't qualitatively alter our conclusions; cf. also Appendix B.

2.2 Lexicon size and average morphosyntactic complexity

We consider that lexicon size of a language is the number of lexicalized meanings — rather than the number of form-meaning pairs, e.g., we consider that English has one lexicon entry for 10 which can be phonetically realized in multiple ways (i.e., the morpheme *ten*

has allomorphs, in particular *-teen* and *-ty*). This is important to keep in mind as it affects how our results should be interpreted: we will argue that languages optimize the trade-off between the number of lexicalized meanings and average morphosyntactic complexity of utterances, rather than the trade-off between the number of form-meaning pairs and average morphosyntactic complexity of utterances. Of course, an important question that remains open is why languages sometimes have multiple phonetic realizations of a single meaning; we will come back to this point in Section 2.4.4.

Average morphosyntactic complexity of numerals in a language L is computed according to the formula in (2). In (2), $ms\_complexity(n, L)$ is the morphosyntactic complexity 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. For instance, average morphosyntactic complexity of English is obtained as $P(1)$ × the number of morphemes in 'one' + $P(2)$ × the number of morphemes in 'two' +...+ $P(99)$ × the number of morphemes in 'ninety-nine'. We assume that the probabilities that different numbers need to be communicated follow a power-law distribution as in (3) (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.

(2)
$$average\_ms\_complexity(L) = \sum_{n \in [1,99]} P(n) \cdot ms\_complexity(n, L)$$

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

The distribution in (3) is an idealization of a more complex cross-cultural reality: while previous work motivates that probabilities that different numbers need to be communicated generally follow such a distribution, there are also some deviations from this distribution across languages (Dehaene and Mehler 1992). Firstly, numerals which can get round (i.e., approximate) interpretation (e.g., numerals for 10, 20 in English) are used more frequently than expected given (3) (Dehaene and Mehler 1992). Importantly for our purposes, however, Krifka (2007) presents empirical evidence that the frequencies of numerals for 10, 20 etc. vary across languages, and that they depend on the specifics of the numeral system, in particular, on what base(s) the numeral system has. In other words, which number meanings are lexicalized as bases (multiplicatives) may influence which expressions can get a round interpretation, and thus influence their frequency. Because of this, and as our goal is to explain which meanings are lexicalized across languages in the first place, using a prior which incorporates peaks on numbers such as 10, 20 etc. would lead to circularity — we thus use the idealized prior distribution in (3) without such peaks. Secondly, certain numbers may be referred to more or less frequently than expected given (3) in a language because of culture-specific reasons (e.g., number 13 may be referred to less frequently than expected given (3) in cultures in which this number is considered unlucky). While these culture-specific variations should ideally be taken into account when evaluating average morphosyntactic complexity of languages, we are not in a position to evaluate per-language probability distributions for each of the 128 studied languages within the scope of this study. We thus assume that the simplified/idealized prior in (3) holds for all languages, noting that future research may refine this aspect of our work.

2.3 Approximating the Pareto frontier

In order to evaluate whether natural languages with a recursive numeral system optimize the trade-off between lexicon size and average morphosyntactic complexity of numerals, we need to establish how close they are to the optimal solutions to the lexicon size/average morphosyntactic complexity of numerals trade-off problem. The set of optimal languages — called *the Pareto frontier* — is a set of (theoretically possible) languages for which there is no other (theoretically possible) language which is better on one of the two dimensions (lexicon size, average morphosyntactic complexity of numerals) without being worse on the other.

The space of theoretically possible languages is too large for an exhaustive search of optimal languages. We thus use an evolutionary algorithm to approximate the Pareto frontier (cf. Steinert-Threlkeld 2019, 2021, Denić et al. 2021, 2022), which involves generations of many artificial languages (i.e., artificial recursive numeral systems). Before we explain how the evolutionary algorithm works, we explain how artificial languages are generated.

We use the grammar proposed by Hurford (1975, 2007) for natural language numerals generation to generate numerals of artificial languages. Hurford (1975, 2007) in his seminal work proposes that number-denoting morphemes across languages 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 grammar in (4). 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.[6]

(4) NUMBER $\longrightarrow$ D | PHRASE | PHRASE + NUMBER | PHRASE − NUMBER
PHRASE $\longrightarrow$ M | NUMBER · M

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 (4)), and (b) a set of numerals for numbers 1-99 generated according to the morphosyntactic rules in (4) from the lexicon, where each numeral is the shortest expression (or one of them, in case of a tie) for a given number. For example, an artificial language L may have in its lexicon morphemes for 1, 2, 4, 6, 8, 10, such that 1, 2 and 4 are of category D and 6, 8 and 10 are of category M. Numerals of L for other numbers in range 1-99 would be generated from these morphemes as explained above. For instance, the numeral for number 16 in L would be randomly selected from one of the three morphosyntactically simplest options: $10 + 6$, $2 \cdot 8$, $8 + 8$. These can be generated by the grammar in (4) and are of equal morphosyntactic complexity (each has 3 morphemes). On the other hand, $4 \cdot 4$ is not an option for numeral 16 in L because, even though it would also have three morphemes, it cannot be generated by the grammar in (4) (because 4 is of category D in L). Similarly,

[6] All numerals in all natural languages in this study can be constructed using the grammar in (4). 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 (4) 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 (4) to generate artificial languages and to estimate the Pareto frontier (cf. Section 2.3), we excluded Ainu from the corpus of natural languages.

$2 \cdot 6 + 4$ is not an option for number 16 in L because, even though it can be generated by the grammar in (4), it is not one of the morphosyntactically simplest options (it has 5 morphemes).

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 (extending the search to depths greater than 7 was impractical because of computational complexity considerations).

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 no more than 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 morphosyntactic complexity 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 include 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 can be represented as a point in a two-dimensional (lexicon size and average morphosyntactic complexity of numerals) space; we do spline interpolation of these points to form a Pareto frontier.

2.4 Results

We will start by presenting the main result (Section 2.4.1). To preview, we find that natural languages' recursive numeral systems are indeed (near-)optimal solutions to the lexicon size/average morphosyntactic complexity of numerals trade-off problem. In Sections 2.4.2 and 2.4.3, we present supplementary analyses which allow to better understand the role of different properties of natural languages' recursive numeral systems in the aforementioned trade-off optimization.

2.4.1 Trade-off results

The natural languages and the artificial languages generated through the 100 generations of the evolutionary algorithm are plotted in Figure 1.[7] The approximated Pareto frontier

[7]Jittering was applied to points corresponding to artificial and natural languages in Figures 1, 2, 3 and 5 to facilitate visualizing individual points (languages).

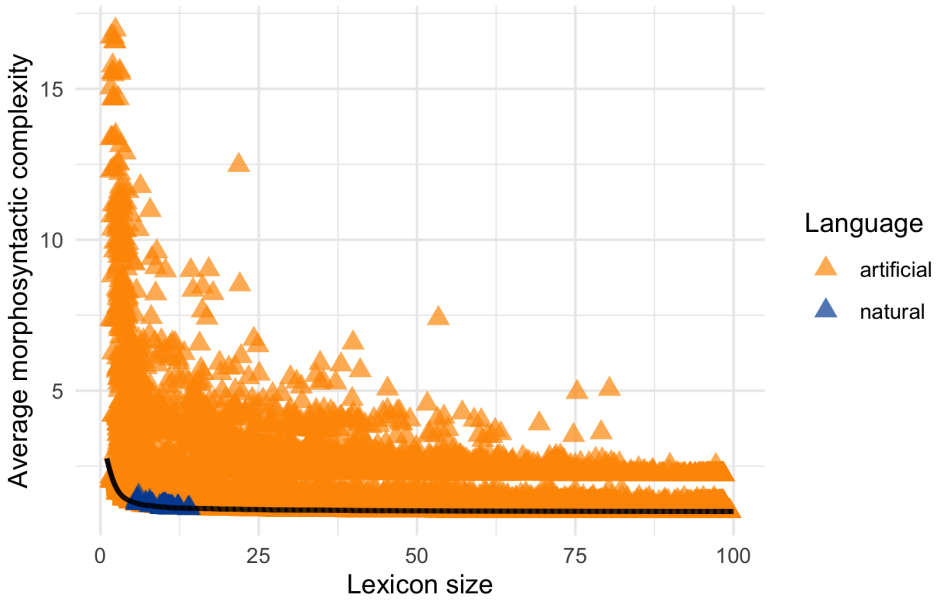

Figure 1: Experiment: Lexicon size and average morphosyntactic complexity of natural languages and of artificial languages generated via an evolutionary algorithm.

— i.e., the set of (nearly-)optimal solutions to the problem of trading off between lexicon size and average morphosyntactic complexity of numerals — is plotted as the black curve in Figure 1. Figure 2 represents a zoom into the region of Figure 1 surrounding the natural languages. Two natural languages — Fulfulde and Georgian — are labeled in Figure 2 for illustration.

Critically, natural languages all lie along or very close to the Pareto frontier in Figure 2. This speaks in favor of natural languages' recursive numeral systems optimizing the trade-off between lexicon size and average morphosyntactic complexity of numerals.[8]

Interestingly, there is some variation in terms of where along the Pareto frontier natural languages lie. In other words, natural languages differ in terms of which optimal solution to the lexicon size/average morphosyntactic complexity of utterances trade-off problem they are approaching. Take for example cases of Georgian and Fulfulde: a few examples of their numerals are in Tables 1 and 2 respectively. We have seen that Georgian lexicalizes numbers in (1b), while Fulfulde lexicalizes those in (1c). The lexicon size of Georgian is thus 11, and that of Fulfulde 6. The average morphosyntactic complexities of

[8] Recall that, when morphologically analyzing unclear cases of natural languages' numerals, we adopted the following simplification: if a numeral is an exception to an established morphosyntactic pattern in a language, but partially overlaps in its (phonetic or orthographic, depending on the available data) elements with 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 (cf. Section 2.1). Of course, at least some of these unclear cases, which will be analyzed as morphologically complex, may be monomorphemic in reality. In other words, for some of the 34 languages with (typically just 1 or 2) unclear cases, we may be somewhat underestimating their lexicon size and somewhat overestimating their average morphosyntactic complexity. This means that the correct position of some of these 34 languages would be slightly down-right-wards from where they are plotted in Figures 1 and 2. It is easy to see that, even if all natural languages plotted in Figures 1 and 2 were to shift slightly down-right-wards, they would nonetheless remain relatively close to the Pareto frontier. It turns out therefore that the error introduced by our simplification in practice wouldn't qualitatively alter our conclusions. For completeness, however, see Appendix B with the supplementary analysis in which the 34 languages with unclear cases are excluded.

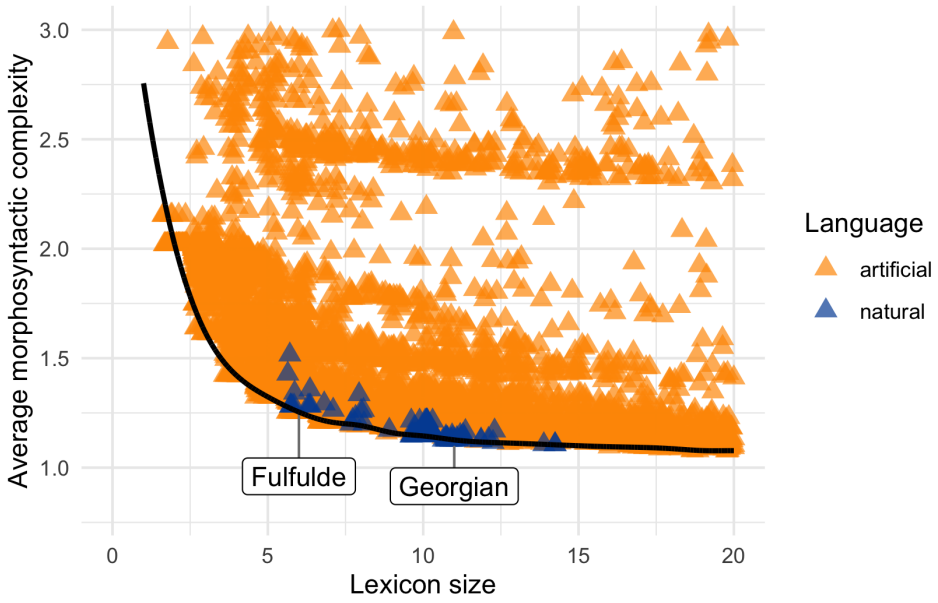

Figure 2: Experiment: Lexicon size and average morphosyntactic complexity of natural languages and of artificial languages generated via an evolutionary algorithm (zoom into the area surrounding natural languages). Two natural languages — Fulfulde and Georgian — are labeled for illustration. Natural languages lie at or very close to the Pareto frontier (black curve).

numerals of Georgian and Fulfulde (computed according to the formula in (2)) are 1.13 and 1.28 respectively. In other words, Fulfulde has a smaller lexicon size than Georgian, but Georgian has a lower average morphosyntactic complexity of numerals than Fulfulde.

Finally, it is interesting to notice another property of the data in Figure 2. Natural languages lie along the region of the Pareto frontier where both average morphosyntactic complexity of numerals and lexicon size are relatively low. There are (at least) two ways to interpret this finding. The first is that there are levels of average morphosyntactic complexity of numerals and lexicon size above which natural languages are, in a sense, not willing to go, even if it is still possible to achieve Pareto-optimality with such higher levels of average morphosyntactic complexity of numerals and lexicon size. In other words, in addition to optimizing the trade-off between lexicon size and average morphosyntactic complexity of numerals, natural languages' recursive numeral systems may also be under the pressure to keep lexicon size and average morphosyntactic complexity of numerals below a certain threshold. Another possibility is that natural languages are not only optimizing the trade-off between lexicon size and average morphosyntactic complexity, but in fact minimizing the sum of (some function of) lexicon size and (some function of) average morphosyntactic complexity of numerals (with 'functions' modulating relative importances of lexicon size and average morphosyntactic complexity). If this option is on the right track, one may propose that languages care to minimize their overall complexity, which may have multiple components (such as lexicon size and average morphosyntactic complexity of utterances). We leave for future work a more detailed exploration of these various refinements of our proposal.

Table 3: Lexicalized meanings in optimal artificial languages whose lexicon size < 10. Lexicalized meanings in bold are of category M (bases).

Lexicon size	**Lexicalized meanings**
2	1, **3**
3	1, **2**, **3**
4	1, **2**, **3**, **10**
5	1, **2**, 3, **4**, **10**
6	1, **2**, 3, 4, **5**, **15**
7	1, **2**, 3, 4, **5**, **6**, **15**
8	1, **2**, 3, 4, **5**, 6, 10, **19**
9	1, 2, 3, **4**, 5, 6, **7**, 10, **25**

2.4.2 Natural languages vs. optimal artificial languages

In the previous section, we have seen that natural languages are similar in how they trade-off between lexicon size and average morphosyntactic complexity to the languages which optimally solve this trade-off problem (i.e., the languages on the approximated Pareto frontier). The languages on the Pareto frontier are the dominant languages from the union of the last generation found by the evolutionary algorithm and of natural languages. Languages at the Pareto frontier may thus be natural or artificial.

In this section, we focus on the artificial languages on the Pareto frontier (henceforth *optimal artificial languages*). Are there interesting differences between natural languages (optimal or optimizing), and optimal artificial languages? In particular, is there more variety in which number meanings are lexicalized in optimal artificial languages, and/or do they construct their morphosyntactically complex numerals in more diverse ways than natural languages? If so, this would suggest that, while natural languages lexicalize number meanings which allow them to optimize the lexicon size/average morphosyntactic complexity trade-off, there may be additional cognitive or communicative pressures pushing natural languages towards specific types of optimal solutions to the trade-off problem.

Let us start with numbers that optimal artificial languages lexicalize. Table 3 lists lexicalized numbers in optimal artificial languages whose lexicon size < 10. These languages have a lot in common with natural languages when it comes to the contents of their lexicons: both natural languages and optimal artificial languages tend to lexicalize the first n numbers, together with a couple of larger numbers, such as around 10 or around 20. There is nonetheless more diversity in which numbers around 10 or 20 are lexicalized in optimal artificial languages than in natural languages. For instance, some of the optimal artificial languages in Table 3 lexicalize 15, 19 or 25: while there are natural languages which lexicalize 15 in our corpus (e.g., Mixtec, Drehy, Diola; cf. Appendix referenced in footnote 2.1), there is no natural language in our corpus that lexicalizes 19 or 25. In other words, certain lexicalization patterns which could be good solutions to the lexicon size/average morphosyntactic complexity of numerals trade-off are not attested in natural languages (at least not in our corpus of natural languages).

Secondly, optimal artificial languages may differ from natural languages in how they construct their morphosyntactically complex numerals from the elements of the lexicon. To illustrate this difference, let us look into how the optimal artificial language which

Table 4: Numerals for numbers 10-20 in the optimal artificial language which lexicalizes 5 numbers (1, 2, 3, 4, 10)

Numeral's denotation	**Numeral's morphosyntax**
10	10
11	$10 + 1$
12	$3 \cdot 4$
13	$10 + 3$
14	$10 + 4$
15	$10 + 2 + 3$
16	$4 \cdot 4$
17	$4 \cdot 4 + 1$
18	$(10 - 1) \cdot 2$
19	$10 + 10 - 1$
20	$2 \cdot 10$

lexicalizes 5 numbers (i.e., 1, 2, 3, 4, 10, cf. Table 3) constructs numerals for numbers 10-20: these are listed in Table 4.

Numerals in Table 4 differ from numerals in natural languages: intuitively, there is too much variation in bases with which these numerals are constructed (e.g., numerals for 11, 13, 14, 15 and 19 are constructed with base 10, numerals for 12, 16 and 17 with base 4, and numeral for 18 with base 2). In other words, there isn't a unique morphosyntactic pattern according to which numerals in Table 4 are constructed (cf. Section 2.1 for discussion of morphosyntactic patterns in natural languages).

The source of this 'unorderliness' is that, for each number, artificial languages select at random one of the morphosyntactically simplest expressions for that number constructed from the elements of their lexicon. For instance, as far as the artificial language in Table 4 is concerned, the numeral for 12 could be constructed either as $3 \cdot 4$ or as $10 + 2$, as these are both well-formed expressions in this language and they are equally morphosyntactically complex.

Contrary to artificial languages, in most natural languages, there is (typically) exactly one way to express each number. For instance, the French numeral for 40, *quarente*, is morphosyntactically $4 \cdot 10$ (and 40 cannot be expressed as, for instance, equally morphosyntactically complex $2 \cdot 20$ in French), while the French numeral for 80, *quatre-vingt*, is morphosyntactically $4 \cdot 20$ (and 80 cannot be expressed as, for instance, equally morphosyntactically complex $8 \cdot 10$ in French)[9]. This means that each natural language has additional constraints on the application of morphosyntactic rules in (4). For instance, French speakers need to memorize that, in the rule PHRASE $\rightarrow$ NUMBER$\cdot$M (cf. (4)), M is 10 when they are constructing the expression for 40, but 20 when they are constructing the expression for 80. Furthermore, such constraints across languages most often apply not to a single numeral but to a sizeable range of numerals (e.g., French speakers know that, in the rule PHRASE $\rightarrow$ NUMBER $\cdot$ M (cf. (4)), M is 20 when they are constructing expressions for numbers in the range 80-99). This results in natural languages recursive numeral systems exhibiting regular morphosyntactic patterns for larger groups of

[9] These data are restricted to the variety of French spoken in France; numerals in other varieties of French are constructed differently.

numerals, unlike artificial languages (cf. Table 4).

The two identified differences between natural and optimal artificial languages — namely, that natural and optimal artificial languages differ in certain lexical and morphosyntactic properties — suggest that considerations related to lexicon size, average morphosyntactic complexity and their trade-off are not the only pressures shaping natural languages' recursive numeral systems: there are other cognitive and/or communicative pressures (which may be specific to numeral systems or applicable more generally across semantic categories) shaping them. What may these pressures be? We discuss here two natural directions to be pursued in future work.

First, we established above that in natural languages, there is (typically) exactly one way to construct a numeral for any number, which demonstrates that each language has a set of constraints for how the rules in (4) are applied to construct numerals. These additional constraints are something that needs to be memorized by a learner, and it is plausible that a pressure to minimize the number and/or complexity of such constraints is an additional pressure shaping specifically numeral systems. An important direction for future work would be to operationalize how to measure complexity of these constraints in language, and how to integrate the pressure to minimize the complexity of these constraints within the model of lexicon size/average morphosyntactic complexity trade-off optimization presented here, and evaluate whether some of the differences between natural and optimal artificial languages disappear.

Second, it may be that numbers differ in how difficult they are to cognitively represent independently of the numeral system used by the language. There are several ways to make this hypothesis more concrete, depending on one's approach to cognitive representations of numbers. For instance, according to the language-of-thought approach (Fodor 1975), some numbers may be cognitive primitives, while others may be cognitively complex (even if they were labeled by a single morpheme in some languages). For instance, 1, 2 and 3 may be cognitive primitives, and the cognitive representation of 4 may be constructed as 3+1 or 2+2. One could then hypothesize that languages are not exactly under the pressures to minimize lexicon size and average morphosyntactic complexity, but rather under the cognitive-representations versions of these pressures: namely, under the pressures to minimize the size of cognitive representations of lexicalized items, and to mimimize average complexity of cognitive representations of expressions. For instance, if 1, 2 and 3 were cognitive primitives, but 4 wasn't, lexicalizing 1, 2 or 3 would be less costly than lexicalizing 4, and if 4 was lexicalized, constructing the numeral for five as (morpheme for) 3 + (morpheme for) 2 would be cognitively less costly than constructing it as (morpheme for) 4 + (morpheme for) 1. If one had a language-independent method to evaluate cognitive complexities of various numbers, one could explicitly test this hypothesis, and evaluate whether some of the differences between natural and optimal artificial languages disappear.

Critically for our purposes, the fact remains that natural languages are trading off nearly optimally between lexicon size and average morphosyntactic complexity (cf. Section 2.4.1). In other words, while exploring which additional cognitive and communicative pressures push recursive numeral systems towards specific types of optimal solutions to the lexicon size/average morphosyntactic complexity trade-off, and how they do so, are important directions for future work, our results demonstrate that the pressure to minimize lexicon size, the pressure to minimize average morphosyntactic complexity, and their trade-off, play a role in shaping language's lexicons.

2.4.3 Importance of lexicalized larger numbers

According to the results in Section 2.4.1, natural languages' recursive numeral systems trade off optimally between lexicon size and average morphosyntactic complexity of numerals. This suggests that the reason why languages lexicalize number meanings they do — i.e., the first n numbers, in addition to a few larger numbers such as 10 and 20 (and outside of the 1-99 range, 100, 1000, 1000000), cf. Section 2.1 — is because these allow to construct numerals using very few morphemes on average, in a way that optimizes the trade-off between lexicon size and average morphosyntactic complexity of numerals.

Given the probability distribution in (3), according to which low numbers need to be communicated frequently, it is unsurprising that it may be a good strategy for languages to lexicalize the first n number meanings if their goal is to optimize the trade-off between lexicon size and average morphosyntactic complexity. However, one may wonder whether lexicalizing larger number meanings such as 10 and/or 20 plays an important role in the trade-off optimization, given the probability distribution in (3), according to which large numbers rarely need to be communicated. In other words, is lexicalizing the first n numbers — independently of which larger numbers are lexicalized, if any at all — sufficient for languages to trade off nearly optimally between lexicon size and average morphosyntactic complexity? Or is the choice of which larger numbers they lexicalize important — i.e., do they lexicalize some of the larger numbers which significantly help with the trade-off optimization?

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[10] and analyzed their lexicon size/average morphosyntactic complexity of numerals trade-off. We plot them in Figure 3 (*artificial_first_n* languages) against the natural and artificial languages from the main analysis (cf. Section 2.4.1). Critically, *artificial_first_n* languages are not all clustered close to the Pareto frontier (black curve) in Figure 3. Here are three examples of these far-from-optimal languages which lexicalize the first few number meanings, simply to illustrate how these languages may look like. (i) A language which lexicalizes 1, 2, 3 and 4 as morphemes of category D, and 23 and 58 as morphemes of category M has average morphosyntactic complexity of 1.98, while the optimal language with the same lexicon size has average morphosyntactic complexity 1.25. (ii) A language which lexicalizes 1, 2, 3, 4, 5 as morphemes of category D, and 51 and 70 as morphemes of category M has average morphosyntactic complexity of 1.8, while the optimal language with the same lexicon size has average morphosyntactic complexity 1.21. (iii) A language which lexicalizes 1, 2, 3, 4, 5, 6 as morphemes of category D, and 45 and 65 as morphemes of category M has average morphosyntactic complexity of 1.56, while the optimal language with the same lexicon size has average morphosyntactic complexity 1.19.

This analysis demonstrates that lexicalizing the first n numbers does not guarantee a nearly optimal lexicon size/average morphosyntactic complexity of numerals trade-off. In other words, natural languages' choices to lexicalize number meanings such as 10 and/or 20 (in addition to the first n number meanings) are important choices from the perspective

[10]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 analysis (cf. Section 2.3).

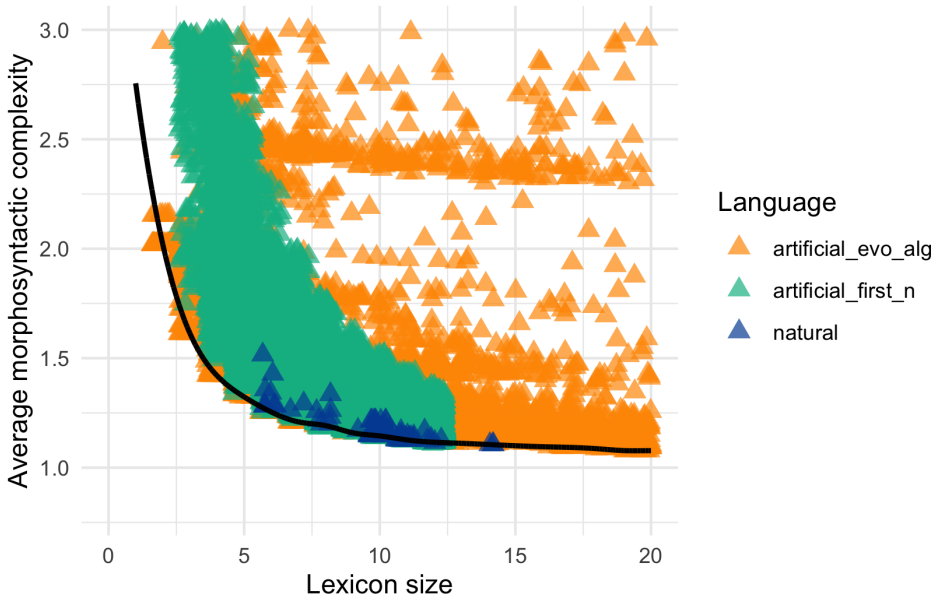

Figure 3: Lexicon size and average morphosyntactic complexity of natural languages compared to artificial languages generated via an evolutionary algorithm (*artificial_evo_lang*) as in Figure 2, with 50 000 artificial languages which lexicalize the first n numbers and at most 2 additional number meanings (*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 morphosyntactic complexity trade-off.

of lexicon size/average morphosyntactic complexity of numerals trade-off optimization.

2.4.4 Results summary and discussion

In Section 2.4.1, we establish that natural languages' recursive numeral systems are indeed (near-)optimal solutions to the lexicon size/average morphosyntactic complexity of numerals trade-off problem.

In Section 2.4.2, we discuss how optimal artificial languages look like. This allows to highlight two important differences between optimal artificial languages and natural languages. In particular, certain lexicalization patterns which could be good solutions to the lexicon size/average morphosyntactic complexity trade-off problem are not attested in natural languages (e.g., solutions which include lexicalizing numbers such as 19 and 25, cf. Table 3). Furthermore, natural languages have more regular morphosyntactic patterns than optimal artificial languages. These two identified differences between natural and optimal artificial languages suggest that there are additional cognitive and/or communicative pressures pushing natural languages towards specific types of optimal solutions to the lexicon size/average morphosyntactic complexity trade-off problem.

In Section 2.4.3, we zoom into the specific numbers natural languages choose to lexicalize and investigate their role in the lexicon size/average morphosyntactic complexity trade-off optimization. Namely, we have seen in Section 2.1 that natural languages often lexicalize (a) the first n numbers, e.g., first 5 or first ten numbers, and (b) a couple of additional numbers, such as 10 and/or 20. In Section 2.4.3, we show that the lexicon size/average morphosyntactic complexity trade-off optimization is not driven only by property (a), but also by (b). In other words, natural languages make good choices

from the trade-off perspective for which numbers to lexicalize in addition to the first n numbers.

It is important to reiterate that we have operationalized lexicon size in this study as the number of lexicalized meanings, and not the number of form-meaning pairs. This is important to keep in mind because morphemes across languages — including number-denoting morphemes — may have allomorphs (variant phonetic realizations) (e.g., we have seen in Section 2.2 that the morpheme *ten* in English has allomorphs *-teen* and *-ty*). According to our results then, which meanings are lexicalized is shaped by the optimization of the trade-off between the number of lexicalized meanings and average morphosyntactic complexity of utterances (rather than by the optimization of the trade-off between the number of form-meaning pairs and average morphosyntactic complexity of utterances). While in this project we have focused on the question of which meanings are lexicalized across languages, a related but separate question which hasn't been addressed is: which meanings have multiple phonetic realizations in a language, and why (essentially, why is there allomorphy across languages?). This is an important question for future work.

To summarize, we have learnt that natural languages' recursive numeral systems optimize lexicon size (as number of lexicalized meanings)/average morphosyntactic complexity of numerals trade-off, and that the numbers they lexicalize are good choices from the perspective of that trade-off optimization. The trade-off optimization can thus explain in part why natural languages lexicalize the numbers they do, as well as the range of variation we see across languages in numbers they lexicalize.

3 Discussion: Previous work on minimization of utterance complexity

As mentioned in Introduction, several other lines of work incorporate the idea that speakers attempt to minimize complexity of their utterances. We will now discuss these proposals in detail, and explain in what way they differ from the proposal we are pursuing.

Zipf (1949) and related work Multiple studies have found that meanings that are conveyed more frequently tend to be associated to shorter forms (with length often operationalized as the number of phonemes) (Zipf 1949, Piantadosi et al. 2011, Haspelmath 2021, Mollica et al. 2021). For instance, in recent work, Mollica et al. (2021) conduct two types of analyses: the first type of analysis shows that languages lexicalize meanings which allow them to optimize the simplicity/informativeness trade-off; the second type of analysis shows that the forms attached to lexicalized meanings are such that average phonetic length of utterances is minimized. Association of shorter forms to more frequent meanings 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.

The pressure to minimize average morphosyntactic complexity of utterances can be viewed as another instance of the same principle. The novelty of our proposal is in showing that this pressure plays a role in determining which meanings get lexicalized, and not only how long forms are associated with lexicalized meanings. According to our proposal, this pressure on its own doesn't suffice to explain which meanings are lexicalized across languages: we have shown that, in semantic domains in which productive morphosyntax enables precise communication, languages lexicalize meanings which allow them to

optimize the trade-off between this pressure and the pressure to minimize lexicon size.

Carcassi and Sbardolini (2022) Another recent proposal gives a central role to utterance complexity in explaining typological patterns, focusing on the semantic domain of Boolean connectives (Carcassi and Sbardolini 2022). 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). According to the proposal by Carcassi and Sbardolini (2022) (but see Bar-Lev and Katzir 2022, Enguehard and Spector 2021, Incurvati and Sbardolini 2023, Katzir and Singh 2013, Uegaki 2022, a.o. for competing proposals), 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 morphosyntactic complexity*). 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 average morphosyntactic complexity. The second pressure, labeled *conceptual complexity* in Carcassi and Sbardolini (2022), 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 (2022) assume that how context is updated by *A nand B* is more conceptually complex than 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 morphosyntactic complexity and the total conceptual complexity of contextual update procedures of lexicalized connectives.

Our proposal shares with Carcassi and Sbardolini (2022) the idea that languages are under the pressure to minimize average morphosyntactic complexity. 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 (2022) the competing pressure relates to another aspect of language use, namely, how conversational participants proceed to update contextual information when they hear an expression.

Memory/computation trade-off optimization in morphological processing The work on psycholinguistic processing of morphologically complex expressions has studied the trade-off between memory and computation (e.g. Frauenfelder and Schreuder (1992), and much related work, including more recently e.g., Kuperman, Bertram, and Baayen (2010), O'Donnell, Snedeker, Tenenbaum, and Goodman (2011) and O'Donnell (2015)). The central question in that line of work is: which morphologically complex linguistic expressions are retrieved from memory as a chunk, rather than computed morphosyn-

tactically at each use? Of course, language users must store in memory morphologically simple linguistic expressions, such as 'walk', and the past tense morpheme '-ed'. But do they also store in their memory 'walked', so that when they construct a sentence 'John walked', they retrieve 'walked' as a chunk, rather than construct it as 'walk'+'ed'? One response is that complex expressions are stored in memory in a way that optimizes the memory/computation trade-off in language processing (e.g, O'Donnell et al. 2011, O'Donnell 2015). Simplifying a lot, if some morphologically complex expression needs to be used very frequently, we are more likely to store it in memory as a chunk and retrieve it as a whole in language use. At least some English speakers (those who use the linguistic expression 'walked' a lot) will end up storing it as a chunk in their memory.

That line of work has thus established the existence of memory/computation trade-off optimization bias in morphological processing. This connects to our work in an interesting way. What we have shown is that lexicon size/average morphosyntactic complexity of utterances trade-off optimization — which is essentially memory/computation trade-off optimization — ends up shaping the primitives (monomorphemic expressions) of the linguistic system, and explains typological diversity in what these primitives are. It is conceivable that this is an outcome of the accumulated effect in the course of language evolution of the aforementioned morphological processing bias. To explore this possibility, an interesting direction for future work may be to directly model the evolution of numeral systems across generations, with language users biased to optimize memory/computation trade-off in their morphological processing, and examine whether this (possibly together with some additional assumptions about how language transmission proceeds) results in numeral systems which optimize the lexicon size/average morphosyntactic complexity trade-off.

4 Discussion: 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 — in other words, the use of morphosyntax for numeral construction is often limited. 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. 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 languages' numeral systems optimize the sim-

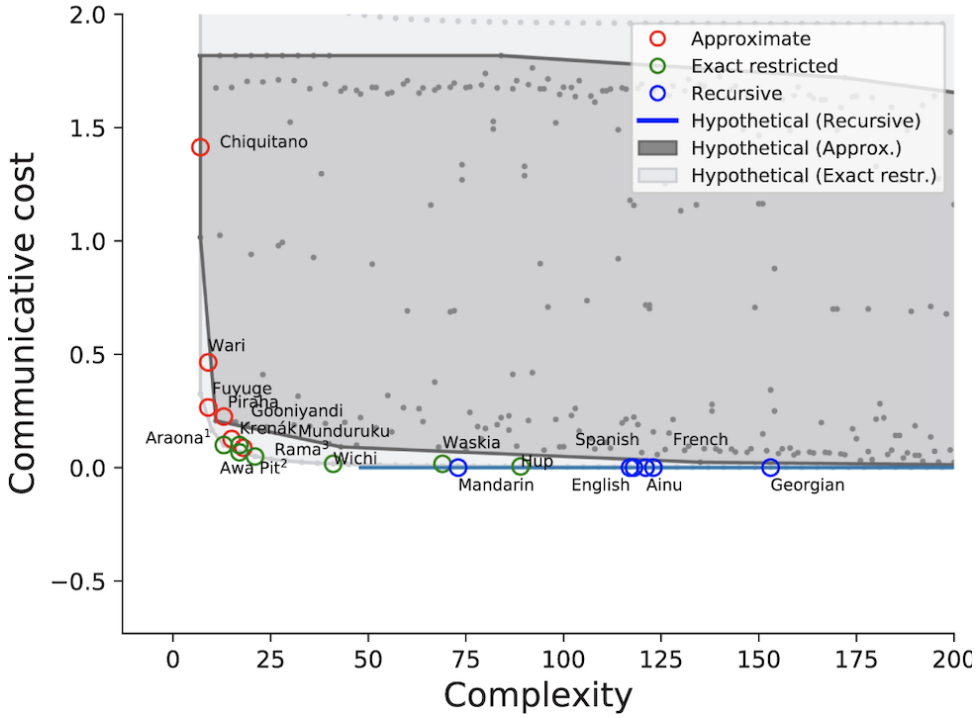

Figure 4: Reproduction of Figure 4b from Xu et al. (2020) plotting complexity and communicative cost (the opposite of simplicity and informativeness) measures of natural languages' numeral systems (red, green and blue circles) and artificial (hypothetical) numeral systems. Blue circles represent the 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.

plicity/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. According to Xu et al.'s (2020) results (cf. Figure 4b in Xu et al. 2020, reproduced here as Figure 4), 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 do not support the hypothesis that recursive numeral systems optimize the simplicity/informativeness trade-off — in other words, there is no tension between the results of Xu et al. (2020) and our findings. Importantly, however, 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 4). We will now discuss the implications of their result for restricted systems and our result for recursive systems for the larger question: *what pressures determine which meanings get lexicalized across semantic domains and languages?*

5 Discussion: Three pressures

Recursive and restricted numeral systems are, in a sense, two extremes. In recursive numeral systems, languages allow to unambiguously express (practically) any number meaning if we consider both morphosyntactically simple and morphosyntactically complex expressions. Such languages lexicalize those meanings which allow them to optimize the trade-off between lexicon size and average morphosyntactic complexity of numerals.

In restricted numeral systems, number meanings are typically conveyed using one of the few (often monomorphemic) numerals, and in most cases it is not possible to unambiguously single out a number meaning with an expression of a language. Such languages lexicalize those meanings which allow them to optimize the simplicity/informativeness trade-off (cf. Section 4).

However, many semantic domains in many languages are arguably in-between these two extremes. 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 (e.g., *blue-green, dark-blue*). What approach should one pursue to explain which meanings get lexicalized in such domains?

The results on restricted and recursive numeral systems provide evidence that (at least) three pressures are shaping which meanings get lexicalized across languages: minimize complexity of the lexicon, minimize average morphosyntactic complexity of utterances and maximize informativeness. It is plausible to expect that these pressures are not specific to the semantic domain of number, i.e., that they are applicable across semantic domains and languages. 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 any meaning can be conveyed precisely and informativeness is maximal (as in recursive numeral systems), this reduces to finding optimal solutions to the lexicon size/average morphosyntactic complexity of utterances 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 (as is arguably most often the case 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).

What semantic domain could be investigated as a case study of the interaction of the three pressures? The aforementioned color domain could be a good starting point, for multiple reasons. (i) The meaning space for colors has been usefully formalized in multiple works (e.g., Zaslavsky et al. 2018, Steinert-Threlkeld and Szymanik 2019), building on cognitive models of human color representations. (ii) In many if not all natural languages, not every element of the color meaning space (i.e., not every color nuance) can be expressed precisely, be it with a morphosyntactically simple or a morphosyntactically complex expression, which paves way for the role of informativeness maximization in lexicalization. (iii) Morphosyntactically complex expressions in many languages improve how precisely we can communicate about colors (e.g., *blue-green, dark-blue*), which paves way for the role of average morphosyntactic complexity minimization in lexicalization. Here is one possible starting point for such a case study. Languages may differ in whether they employ syntactic strategies for color modification, and if so, how complex these strategies are (e.g., if they allow for expressions such as *blue-green, dark-blue, between blue and green*). One could investigate whether and how the meanings of lexicalized color terms

vary with availability and complexity of syntactic color modification across languages. For instance, if simple syntactic modification strategies are available in a language, one may expect the language to divide their color space into fewer basic color categories. If so, that would be initial evidence of morphosyntactic considerations playing a role in the lexicalization of color terms. One could then build on that finding, as well as on the findings about the role of simplicity of the lexicon and informativeness in the lexicalization of color terms (Zaslavsky et al. 2018), to develop and test a three-pressures interaction model. We hope to pursue this avenue in future work.

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 argue for a refinement of what it means for a language to support efficient communication. In particular, the standard approach to communicative efficiency — the simplicity/informativeness trade-off optimization approach — cannot explain which meanings get lexicalized in semantic domains in which lexicon size and informativeness are not in direct competition. Using recursive numeral systems as a case study, we have argued that in such domains languages lexicalize meanings which allow them to optimize the lexicon size/average morphosyntactic complexity of utterances trade-off.

Our work in combination with previous work thus evidences that there are (at least) three pressures shaping lexicons of natural languages: minimizing lexicon size, maximizing informativeness, and minimizing average morphosyntactic complexity of utterances. A more general proposal for how supporting efficient communication shapes which meanings get lexicalized across languages may thus be that languages lexicalize those meanings which allow them to solve the trade-off problem between these three pressures in a (nearly) optimal way.

A 128 studied natural languages

Abkhaz, Abun, Acoma, Albanian, Arabic (Egyptian), Arawak, Archi, Armenian (Eastern), Aymara, Bagirmi, Bambara, Basaa, Basque, Batak (Karo), Bawm, Berber Middle Atlas (Tamazigt), Bribri, Burmese, Burushaski, Cahuilla, Chamorro, Chinantec (Lealao), Chuukese, Chuvash, Comanche, Damana, Diola-Fogny, Drehu, English, Evenki, Ewe, Fijian, Finnish, French, Fulfulde, Garo, Georgian, German, Goajiro, Gola, Greek (Modern), Guarani, Haida, Hausa, Hebrew, Hindi, Hmong Njua, Huave, Hungarian, Hunzib, Hupa, Igbo, Indonesian, Ingush, Iraqw, Irish, Japanese, Jaqaru, Kabardian, Kana, Kannada, Kanuri, Kayah Li (Eastern), Khalaj, Khalkha (Mongolian), Khanty, Kilivila, Kiribati (Gilbertese), Korean, Koromfe, Koyraboro Seni, Kunama, Lak, Lakhota, Lango, Latvian, Lavukaleve, Lega, Lezgian, Malagasy, Mandarin, Mangab-Mbula, Maori, Mapudungun, Mixtec (Atatlahuca), Mixtec (Chalcatongo), Nahuatl (Sierra de Zacapoaxtla), Nama Hottentot (Khoekhoe), Navajo, Ndyuka (Aukan), Nenets (Tundra), Nez Perce, Nivkh, Nkore-Kiga, Noon, Nubian (Dongolese), Oneida, Oromo (Harar), Otomi (Mezquital), Paiwan, Persian, Pohnepeian, Quechua Imbabura, Quileute, Rapanui, Russian, Sahu, Sango, Sapuan, Sorbian Upper, Spanish, Supyire, Swahili, Taba-East Makian, Tagalog,

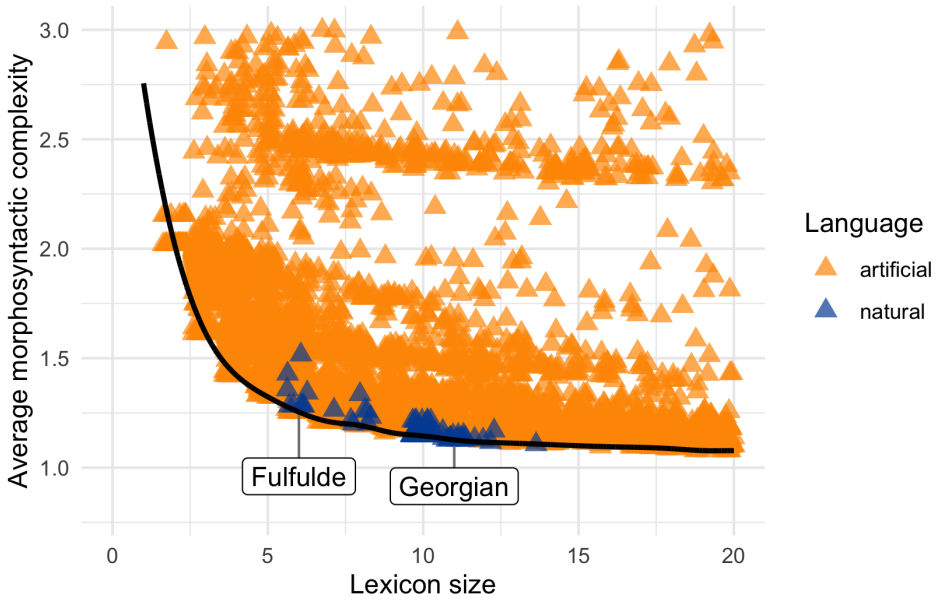

Figure 5: Experiment: Lexicon size and average morphosyntactic complexity of natural languages without unclear cases ($N = 94$) and of artificial languages generated via an evolutionary algorithm (zoom into the area surrounding natural languages). Two natural languages — Fulfulde and Georgian — are labeled for illustration. Natural languages lie at or very close to the Pareto frontier (black curve).

Tarahumara (Western), Telugu, Thai, Tommo So Dogon, Tsez, Tuareg, Tukang Besi, Turkish, Vietnamese, Yagua, Yakut, Yucatec, Yipik, Zulu

B 34 languages with unclear cases

Below is a list of 34 languages in which the morphosyntactic content of some of the numerals was difficult to determine (when a numeral is an exception to an established morphosyntactic pattern in a language, but partially overlaps in its (phonetic or orthographic, depending on the available data) elements with morpheme(s) which should have been there if the morphosyntactic pattern was respected).

34 languages with unclear cases: Arabic (Egyptian), Archi, Armenian, Chinantec (Lealao), Chuukese, Chuvash, English, Ewe, Hausa, Hindi, Hungarian, Ingush, Kannada, Khalkha (Mongolian), Khanty, Koromfe, Kunama, Lak, Lezgian, Ndyuka, Nez Perce, Nkore Kiga, Oneida, Oromo, Otomi, Persian, Quileute, Rapanui, Russian, Spanish, Swahili, Telugu, Tukang Besi, Turkish

For completeness, we exclude these 34 languages from the original 128 plotted in Figure 2, and include the updated plot (with 94 natural languages) in Figure 5.

Acknowledgements

Removed for review.

References

Moshe E. Bar-Lev and Roni Katzir. Communicative stability and the typology of logical operators. *Linguistic Inquiry*, pages 1–42, 2022.

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. Assertion, denial, and the evolution of boolean operators. *Mind & Language*, 2022.

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. Revisiting greenberg's "generalizations about numeral systems"(1978). *Journal of Universal Language*, 21(2):43–84, 2020.

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. 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.

Émile Enguehard and Benjamin Spector. Explaining gaps in the logical lexicon of natural languages: A decision-theoretic perspective on the square of aristotle. *Semantics and Pragmatics*, 14:5, 2021.

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.

Uli H Frauenfelder and Robert Schreuder. Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In *Yearbook of morphology 1991*, pages 165–183. Springer, 1992.

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

J. Greenberg. Generalizations about numeral systems. In K. Denning and S. Kemmer, editors, *On Language: Selected Writings of Joseph H. Greenberg*, pages 271–309. Stanford University Press, nov 1990. doi: 10.1515/9781503623217-014.

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.

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

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.

Luca Incurvati and Giorgio Sbardolini. Update rules and semantic universals. *Linguistics and Philosophy*, 46(2):259–289, 2023.

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 and Raj Singh. Constraints on the lexicalization of logical operators. *Linguistics and Philosophy*, 36:1–29, 2013.

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.

Manfred Krifka. Approximate interpretations of number words: A case for strategic communication. *Cognitive foundations of interpretation*, pages 111–127, 2007.

Victor Kuperman, Raymond Bertram, and R Harald Baayen. Processing trade-offs in the reading of dutch derived words. *Journal of Memory and Language*, 62(2):83–97, 2010.

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.

Timothy O'Donnell, Jesse Snedeker, Joshua Tenenbaum, and Noah Goodman. Productivity and reuse in language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

Timothy J O'Donnell. *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press, 2015.

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.

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, 2019.

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.

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.