Reverse-engineering the language of thought: A new approach

Milica Denić (milica.z.denic@gmail.com)
Institute for Logic, Language and Computation
University of Amsterdam

Jakub Szymanik (jakub.szymanik@gmail.com)
Institute for Logic, Language and Computation
University of Amsterdam

Abstract

A foundational hypothesis in cognitive science is that some of human thinking happens in a *language of thought* (LoT), which is universal across humans (Fodor, 1975). According to this hypothesis, words in different natural languages are labels for primitive concepts or their combinations in LoT. What are LoT's primitives? This is a major challenge because LoT is not directly observable, and thus needs to be inferred or *reverse-engineered*. We put forward a novel approach to reverse-engineering LoT, capitalizing on *the existing knowledge about the optimization of the trade-off between complexity and informativeness* in natural languages.

Keywords: language of thought; numerals; number; complexity/informativeness trade-off

Introduction

What are cognitive representations like? Different answers to this question have been explored (Fodor (1975); Gärdenfors (2014); Rumelhart, McClelland, and The PDP Research Group (1986); Van Gelder (1995); see Piantadosi (2021) for a recent discussion). In the present work, we focus on the hypothesis that at least some of human thinking happens in a *language of thought* (LoT), which is universal across humans (Fodor, 1975). According to this hypothesis, words in different natural languages such as English are labels for primitive concepts or their combinations in LoT.

How can we study LoT? This is a major challenge because LoT is not directly observable, and thus needs to be inferred or *reverse-engineered*. Existing approaches include making inferences about LoT from how people use language (Hackl, 2009; Knowlton, Pietroski, Halberda, & Lidz, 2021; Lidz, Pietroski, Halberda, & Hunter, 2011; Pietroski, Lidz, Hunter, & Halberda, 2009), from how they learn concepts (Piantadosi, Tenenbaum, & Goodman, 2016), from linguistic universals (Züfle & Katzir, 2022), and from language acquisition data (Piantadosi, Tenenbaum, & Goodman, 2012).

We propose a novel approach to reverse-engineering LoT and asking what its primitive components are, capitalizing on *the existing knowledge about the optimization of the trade-off between complexity and informativeness* in natural languages (Kemp & Regier, 2012; Kemp, Xu, & Regier, 2018).

Why do we need another approach? What LoT is like is a central problem for cognitive science and the field is still far from resolving it. All existing approaches, including the one we propose, incorporate non-trivial assumptions which may ultimately prove to be wrong, and disqualify the approach from the set of methods for studying LoT. Furthermore, different approaches may be more or less easy to apply in practice to different semantic domains. For instance, the approach we propose requires gathering cross-linguistic data.

We choose the semantic domain of number as a case study, asking *what LoT primitives underlie numbers concepts 1-99*. In the Discussion section, we relate our findings to two earlier studies investigating LoT primitives underlying number concepts (Piantadosi et al., 2012; Xu, Liu, & Regier, 2020).

Cross-linguistic data

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 collect 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).[1] We analyze only recursive numeral systems, i.e., systems which can construct numerals for all natural numbers.[2] Out of 172 recursive numeral systems in Comrie (2013), 41 were excluded due to challenges with data collection or data interpretation.[3] 131 languages were thus included in the analysis.[4] *WALS* language samples are compiled with an aim to maximize genealogical and areal diversity of languages in them (Comrie, Dryer, Gil, & Haspelmath, 2013) — we can thus have some confidence that we are analyzing a representa-

[1] 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.

[2] Restricted ($N = 20$) and extended-body part numeral systems ($N = 4$) were not included in the analysis, cf. Comrie (2013).

[3] For some of the languages from the sample in Comrie (2013), no appropriate description of the numeral system was accessible to us. Furthermore, a small number of languages were excluded due to difficulties with data interpretation, in particular 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 (2011)).

[4] The list of analyzed languages can be found in Appendix at: https://github.com/milicaden/numerals-lot-cogsci2022.

tive sample of world's languages' recursive numeral systems.

For each of the 131 studied numeral systems, for each numeral, its morphosyntactic components and their denotations were identified (cf. Table 1 for a few examples of numerals in Ainu). These numeral systems differ in terms of morphosyntactic rules according to which numerals are generated. These morphosyntactic rules reveal that addition, multiplication, subtraction and division are involved in the composition of numerals[5] — these are sometimes, but not always, morphosyntactically overt. Furthermore, certain number-denoting morphemes play a special role in morphosyntactic rules (the so-called *bases*). For instance, English is a 'base-10 language': this means that its numerals for numbers 10-99 are in general constructed according to the morphosyntactic pattern $x \cdot \mathit{10} + n$ (e.g., in English, the numeral for 67 is composed of morphemes denoting 6, 10 and 7). On the other hand, Ainu is a 'base-20 language': this means that its numerals for numbers 20-99 are in general constructed according to the morphosyntactic pattern $x \cdot \mathit{20} + n$ (e.g., in Ainu, the numeral for 67 is composed of morphemes denoting 3, 20 and 7). Finally, many languages behave as 'base-5' languages when it comes to the composition of numerals for numbers 6-9 (e.g., in Fulfulde, the numeral for 6 is composed of morphemes denoting 5 and 1). Furthermore, Ainu also exemplifies the use of subtraction in the composition of numerals (e.g., in Ainu, the numeral for 6 is composed of morphemes denoting 4 and 10). Danish exemplifies the use of division in the composition of numerals (e.g., in Danish, the numeral for 50 has as one of its morphosyntactic components the morpheme denoting $\frac{1}{2}$).

Table 1: Ainu numerals for numbers 6, 30 and 42

Denoted number (numeral)	**Morphosyntactic make-up**
6 (*iwan*)	10 (*-wan*) $-$ 4 (*i-*)
30 (*wanetuhotne*)	2 (*-tu-*) $\cdot$ 20 (*-hotne*) $-$ (*-e-*) 10 (*wan-*)
42 (*tuikashimatuhotne*)	2 (*-tu-*) $\cdot$ 20 (*-hotne*) $+$ (*-ikashima-*) 2 (*tu-*)

LoT hypotheses

We assume that the LoT representations underlying numerals are composed from the elements of the set of primitive number concepts PRIM and arithmetic operators for addition, subtraction, multiplication and division ($+$, $-$, $\cdot$, $/$). The interpretation of an LoT expression underlying a numeral provides the denotation of the numeral. For instance, if a numeral's underlying LoT expression is $\mathbf{1}+\mathbf{1}$, the denotation of the numeral will be 2 (primitive number concepts will henceforth be written in bold font to distinguish them from semantic denotations). Our research question is what PRIM contains, that is, which number concepts are LoT primitives. We explore 48 hypotheses for what PRIM contains, summarized in (1):

(1) PRIM $= X \cup Y$, for any X, Y s.t.:
$X \in \{\{\mathbf{1},\ldots,n\} \mid n \in \{\mathbf{1},\ldots,\mathbf{9}\}\}$
$Y \in \mathcal{P}(\{\mathbf{5}, \mathbf{10}, \mathbf{20}\})$

In other words, we explore the hypotheses according to which the first n numbers are LoT primitives, together with some subset — including $\emptyset$ — of $\{\mathbf{5}, \mathbf{10}, \mathbf{20}\}$. The consideration of the hypotheses according to which the first n numbers are LoT primitives is well-motivated by previous work on number cognition: Xu et al. (2020) and Piantadosi et al. (2012) assume that PRIM $= \{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$. The consideration of the hypotheses according to which **5**, **10** and/or **20** may be LoT primitives is motivated by the typology of numeral systems, in which number concepts 5, 10 and 20 play a prominent role (cf. Cross-linguistic data section).

Complexity and informativeness

In this paper, we put forward a novel approach to reverse-engineering LoT: we will use the cross-linguistic data described in the Cross-linguistic data section to empirically evaluate the 48 LoT hypotheses. In order to do this, we will capitalize on *the existing knowledge about the optimization of the trade-off between complexity and informativeness* in natural languages (Kemp & Regier, 2012; Kemp et al., 2018). In this section, we summarize this existing knowledge.

The *complexity* of a language measures how difficult it is to represent the language in LoT. The *informativeness* of a language measures how precisely its expressions allow its users to communicate the intended meanings, and it is formally defined using information-theoretic notions (Kemp & Regier, 2012; Kemp et al., 2018). For instance, a language which has non-ambiguous expressions for each number in the range 1-99 would allow for a maximally precise communication about those numbers (i.e., it would be maximally informative), but it would be complex to mentally represent. On the other hand, a language which only has an expression for the number 1 would be simpler, but not very informative. Complexity and informativeness are in a tension: languages cannot both be minimally complex and maximally informative! This tension is known as the *complexity/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 it is not possible to modify it to obtain a language that has both lower complexity and higher informativeness. Remarkably, computational modeling of cross-linguistic semantic data has demonstrated that natural languages are at or very near the Pareto frontier — natural languages are (in the proximity of) one of the optimal solutions to the trade-off problem (Denić, Steinert-Threlkeld,

[5] Some authors assume that the power function is available as well: for instance, Hurford (2011) assumes that power function is involved in the composition of numerals *billion*, *trillion* etc. in English. We do not find these data convincing: assuming that *bi-* denotes 2, and *tri-* denotes 3, there is no x that *-llion* may denote such that $x^2 = 1000000000$ and $x^3 = 1000000000000$. In other words, if power function is involved in the composition of e.g. *billion* and *trillion*, one would need to assume that *bi-* denotes 3 and *tri-* 4.

& Szymanik, 2021, 2022; Kemp & Regier, 2012; Kemp et al., 2018; Steinert-Threlkeld, 2019, 2021; Uegaki, 2022; Xu et al., 2020; Zaslavsky, Kemp, Regier, & Tishby, 2018; Zaslavsky, Maldonado, & Culbertson, 2021). Importantly, that natural languages optimize the complexity/informativeness trade-off is not inconsistent with cross-linguistic diversity found in many semantic domains (number included): 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.

Novel approach

Given that LoT cannot be observed directly, there are multiple candidate hypotheses for the underlying LoT in different semantic domains (cf. our 48 LoT hypotheses for the number domain). Importantly, the measure of complexity of a language will depend on the hypothesized LoT. For instance, a language with expressions for numbers 1 and 2 will have a different measure of complexity under the hypothesis that **1** and **2** are LoT primitives, as opposed to the hypothesis that **1** is an LoT primitive, but **2** is not. In this project, we evaluate the empirical adequacy of the candidate LoT hypotheses based on how well they explain cross-linguistic semantic data on numerals in terms of the optimality of complexity/informativeness trade-off. We will have evidence in favor of an LoT hypothesis if under that hypothesis the 131 languages lie close to the Pareto frontier.[6] To put it differently, we ask the following question: which of the 48 LoT hypotheses result in a close fit of the 131 natural languages to the Pareto frontier?

Simplifying somewhat, the distance between a language and the Pareto frontier can be defined as the minimum Euclidean distance between the language and a point on the Pareto frontier in the complexity-informativeness space. In general, in order to compute the minimum Euclidean distance between a language and the Pareto frontier, one would need to know the coordinates (complexity and informativeness) of all the points of the Pareto frontier. For the present case study on numerals, however, this will not be necessary — only one point on the Pareto frontier will be needed, namely the point whose measure of informativeness is maximal. For a language to have the maximal level of informativeness when it comes to communicating about numbers in the range 1-99, the language needs to have non-ambiguous numerals for each number in that range. Let us refer to the point on the Pareto frontier whose measure of informativeness is maximal as the *max-info Pareto point*. The reason why only this point is needed, and not the entire Pareto frontier, is the following. By definition of the Pareto frontier, the *max-info Pareto point* is the (possibly artificial) numeral system with the lowest level of complexity necessary to achieve the maximal level of informativeness. The 131 natural languages all have non-ambiguous numerals for each number in the range 1-99: they thus all have the maximum level of informativeness. Consequently, their complexity must be equal or higher than the complexity of the *max-info Pareto point*, which entails that the *max-info Pareto point* will be the closest point on the Pareto frontier for each of the 131 natural languages.

Our question thus reduces to: under which of the 48 hypotheses are the 131 natural languages the closest to the *max-info Pareto point*? As the 131 natural languages have the same level of informativeness as the *max-info Pareto point*, the Euclidean distance of a natural language from the *max-info Pareto point* reduces to their difference in complexity.

Computing complexity

Complexity of natural languages

Most work so far in the complexity/informativeness trade-off framework has analyzed semantic systems such as kinship terms or quantificational determiners which are fundamentally different from numeral systems in that the expressions of the latter are clearly not all memorized — for recursive numeral systems memorization would not be possible even in principle — but are rather generated by a (language-specific) set of morphosyntactic rules (call it grammar G). This means that complexity measures developed for systems such as kinship terms or quantificational determiners, which are typically the sum of lengths of LoT representations of expressions of the system, would not be appropriate for numeral systems. How to measure cognitive complexity of numeral systems?

To our knowledge, there is no definite answer to this question. In an existing proposal by Xu et al. (2020), the complexity of a numeral system is measured as the complexity of G needed to generate it, with rules of G written in LoT. However, there is an empirical problem with that approach that is best illustrated with an example. Imagine a language which has a single morpheme x denoting number 1, and which builds expressions denoting a number n by concatenating x n times. This G is extremely simple, and the resulting language is maximally informative when it comes to communicating about numbers. However, there is no known natural language numeral system that works like this — why? Intuitively, the reason seems to be that, even though such G would be simple, the expressions built by it wouldn't be. In other words, languages seem to care not (only) about the complexity of G, but about the complexity of expressions generated by G.

We thus propose a different perspective. Specifically, we propose to connect the measure of cognitive complexity of a numeral system to how often language users need to communicate about specific numbers, and consequently construct, using their grammar G, LoT representations of numerals denoting those numbers. More precisely, we propose to mea-

[6] This approach relates to an analysis reported in Zaslavsky et al. (2021), albeit their goal is not reverse-engineering LoT. They show that two different hypotheses about how important it is to convey different conversational roles (speaker vs. non-speaker) when using personal pronouns (e.g., *I*, *you*) result in different complexity/informativeness trade-off for personal pronoun systems across languages: languages are closer to the Pareto frontier when conveying the speaker role is given more importance compared to other conversational roles, which is in line with previous work on personal pronouns (cf. discussion in Maldonado and Culbertson (2020)).

sure the complexity of a numeral system as the expected LoT complexity of its numerals, defined in (2). In (2), $p([\![numeral]\!])$ is the probability that the number denoted by *numeral* needs to be communicated; we assume that these probabilities 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. On the other hand, $c(numeral)$ models the complexity of LoT representation of the numeral, and it is defined in what follows.

(2) **Complexity of a language L:**

$$Comp(L) = \sum_{numeral \in L} p([\![numeral]\!])c(numeral)$$

(3) **Prior over numbers:**

$$p(n) \propto n^{-2}$$

Morphemes are the smallest parts of words that add their own distinct meaning component to the word. For instance, the numeral *seven* consists of a unique morpheme, while the numeral *seventy* consists of two morphemes, *seven-* and *-ty*. We assume that the LoT representation of each morpheme is *the shortest LoT formula* (the length of the LoT formula being the number of LoT primitives in the formula, with LoT primitives being the elements of PRIM, $+, -, \cdot$ and $/$) which results in the denotation of the morpheme. For instance, if PRIM = $\{\mathbf{1},\ \mathbf{2},\ \mathbf{3}\}$, the LoT representation of *five* would be $\mathbf{3}+\mathbf{2}$, and not $\mathbf{3}+\mathbf{1}+\mathbf{1}$ or $\mathbf{2}+\mathbf{2}+\mathbf{1}$.

We assume that an LoT representation of a complex expression is composed from LoT representations of its parts (LoT-level compositionality). In other words, if a numeral consists of multiple morphemes denoting numbers (e.g., *seventy* consists of morphemes *seven-* and *-ty*, denoting 7 and 10 respectively), we assume that the LoT representation of each morpheme feeds into the LoT representation of the numeral; these number-denoting morphemes are combined via some of the primitive LoT arithmetic operators ($+, -, \cdot$ and $/$). We measure the *complexity of a numeral*, *c(numeral)*, as the number of LoT primitives in its LoT representation. For instance, the complexity of the numeral *seventy* in English would be the sum of the numbers of LoT primitives in the LoT representations of the morphemes *seven-* and *-ty*, plus the number of (covert) LoT arithmetic operators via which LoT representations of these morphemes are combined (1 in this case: *seven-* and *-ty* are combined via $\cdot$).

It is important to keep in mind however that it is conceivable that neither the complexity measure as in (2) nor the complexity measure in Xu et al. (2020) is on the right track, and that future research may establish a more accurate measure of cognitive complexity of numeral systems (perhaps a measure integrating the complexity of the grammar, as in Xu et al. (2020), and the expected complexity of generated expressions, as in (2)). Our results should thus be taken as preliminary, to be re-visited if/when a more appropriate complexity measure is developed.

Complexity of the *max-info Pareto point*

The *max-info Pareto point* is the (possibly artificial) language which has the lowest level of complexity necessary to reach the maximal level of informativeness. Under each of the 48 LoT hypotheses, each of the 99 numerals of the *max-info Pareto point* is assigned the shortest LoT formula which results in its denotation and the complexity of the *max-info Pareto point* is computed according to the formula in (2).

Correcting the measure of distance

How can we evaluate how close the 131 natural language are to the *max-info Pareto point*?

We have assumed so far that the measure of distance of a natural language x from the *max-info Pareto point* is their Euclidean distance in the complexity-informativeness space, which, when x has the maximal degree of informativeness, reduces to the difference in complexity between x and the *max-info Pareto point*.

Under this assumption, for each of the 48 LoT hypotheses, one could compute the average distance of the 131 natural languages from the *max-info Pareto point* and compare those averages to evaluate different LoT hypotheses. This would however be problematic for the following reason. The 48 LoT hypotheses differ among themselves in the number and types of elements in PRIM. These different hypotheses will result not only in the difference in distances of natural languages to the *max-info Pareto point*, but also in the difference in the measure of complexity of the *max-info Pareto point* itself. For instance, under some of these hypotheses, the complexity of the *max-info Pareto point* may be 5, while under some other it may be 15. The difference in complexity of a natural language x from the *max-info Pareto point* in, e.g., 1 unit suggests more important differences between the two languages when the complexity of the *max-info Pareto point* is 5 than when it is 15: a greater % of the complexity of x would need to disappear through language evolution for x to reach optimality in the former case; in that sense, x is further from optimality in the former case. Because of this, when we evaluate under which of the 48 hypotheses the 131 natural language are the closest to the *max-info Pareto point*, we relativize the distances to the complexity measure of the *max-info Pareto point* as in (4).

(4) **Relativized distance (RD) measure of language L from the *max-info Pareto point (MIPP)*:**

$$RD(L, MIPP) = \frac{Comp(L) - Comp(MIPP)}{Comp(MIPP)}$$

Results

For each of the 48 LoT hypotheses, we compute the *average* relativized distance $\overline{RD}$ of the 131 natural languages from

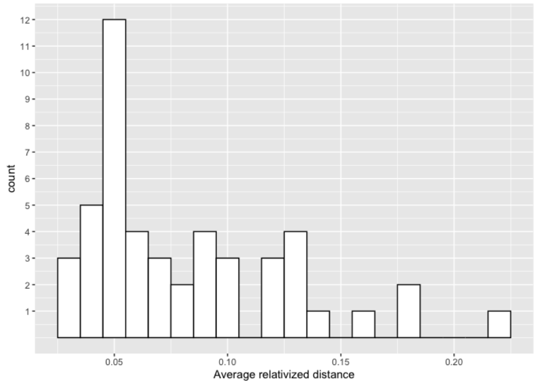

Figure 1: Distribution of $\overline{RD}$s of the 131 natural languages from the *max-info Pareto point* for 48 LoT hypotheses.

the *max-info Pareto point*.[7] The distribution of $\overline{RD}$s can be viewed in Figure 1. As can be seen in Figure 1, there is a lot of variation in terms of how close natural languages are to the *max-info Pareto point* under different LoTs.

We report the top three LoT hypotheses, i.e., those with three lowest $\overline{RD}$s, in Table 2 (for each of these three hypotheses, $\overline{RD} \approx 0.03$). Recall that, the lower $\overline{RD}$, the closer on average natural languages are to the Pareto frontier, that is, to the *max-info Pareto point*. If languages optimize the complexity/informativeness trade-off, natural languages should be close to the Pareto frontier and thus have low $\overline{RD}$.

Table 2: Top three LoT hypotheses

PRIM
$\{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{5}, \mathbf{10}\}$
$\{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{10}\}$
$\{\mathbf{1}, \mathbf{2}, \mathbf{5}, \mathbf{10}\}$

How should we interpret the results from Table 2? That these LoT hypotheses result in the best fit of natural languages to the Pareto frontier may be taken as suggestive evidence in favor of them, but we cannot be certain that the true set of LoT primitives is among them (cf. Discussion section).

The results may further be informative about hypotheses which do not provide such a good fit to the Pareto frontier. Interestingly, the hypothesis that PRIM = $\{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$, which has been entertained in previous work (Piantadosi et al., 2012; Xu et al., 2020), leads to a worse complexity/informativeness trade-off results than most other hypotheses ($\overline{RD} = 0.1$, ranking 36 (out of 48 hypotheses)).

Discussion

Previous work

We start by discussing how this work relates to two previous studies: Xu et al. (2020) and Piantadosi et al. (2012).

Xu et al. (2020) Xu et al. (2020) argue that natural languages' numeral systems optimize the complexity/informativeness trade-off. To construct their argument, they stipulate the underlying LoT primitives, and show that with the stipulated set of primitives, natural languages optimize the complexity/informativeness trade-off.

While the investigation of the complexity/informativeness trade-off in numeral systems is a common point between our study and that of Xu et al. (2020), our study departs from that of Xu et al. (2020) in a number of important ways.

Firstly, the two studies have different starting assumptions and different aims. While Xu et al. (2020) investigate natural languages' numeral systems' complexity/informativeness trade-off under one specific LoT hypothesis, our study assumes that natural languages optimize the complexity/informativeness trade-off and investigates under which LoT hypotheses this assumption holds. More specifically, Xu et al. (2020) assume that PRIM = $\{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$, while the contents of PRIM are the object of investigation of our study. Xu et al. (2020) motivate their assumption by a phenomenon called *subitizing* whereby sizes of small sets (up to 4 members) are evaluated differently than larger set sizes (Revkin, Piazza, Izard, Cohen, & Dehaene, 2008). However, recent work suggests that subitizing is a consequence of lower level constraints on perception rather than of numerical cognition per se (Cheyette, Wu, & Piantadosi, 2021): if this is correct, subitizing is not an argument in favor of number concepts 1, 2, 3 being LoT primitives. In fact, according to our results, the hypothesis that PRIM = $\{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$ fares worse than most other explored hypotheses (cf. Results section).

Secondly, our study uses a different corpus of natural languages. While Xu et al. (2020) analyzed 6 recursive and 24 restricted numeral systems, we have analyzed 131 recursive numeral system and no restricted ones. The reason for excluding restricted numeral systems from our study is the following. Restricted numeral systems don't have numerals for all numbers 1-99: most of them have numerals for only the first few numbers. For instance, the language Krenak only has numerals for numbers 1-3 (Hammarström, 2010; Xu et al., 2020), and the language Rama only has numerals for numbers 1-5 (Grinevald, 1990). While considering such languages would in principle be valuable for the goals of the present study, it would require making additional assumptions. Specifically, as restricted languages are not maximally informative, it is not possible to know which point at the Pareto frontier would be the closest one to them without knowing the coordinates (complexity and informativeness measures) of all the points on the Pareto frontier. This would in turn require spelling out exactly how to measure informativeness of different languages, knowing that multiple ways

[7] Scripts and data files used for the analysis can be found at: https://github.com/milicaden/numerals-lot-cogsci2022.

to measure informativeness have been discussed in the literature (Denić et al., 2022; Kemp & Regier, 2012; Steinert-Threlkeld, 2019).

Thirdly, while we assume that the elements of PRIM can be combined via the arithmetic operations $+, -, \cdot$ and $/$ only, Xu et al. (2020) assume a larger set of options, including, in addition to these four, power and successor functions and the greater-than relation, among others. Our assumption that $+, -, \cdot$ and $/$ are available arithmetic operations for combining the elements of PRIM is supported by morphosyntactic evidence (cf. Cross-linguistic data section). On the other hand, the 131 languages we studied provided no similar morphosyntactic evidence for, say, power function use in the construction of numerals (cf. footnote 2). We acknowledge however that further typological investigation of numeral systems may result in the revision of our assumption.

Finally, for reasons explained in the Computing complexity section, we resort to a different measure of complexity of a numeral system than Xu et al. (2020).

Piantadosi et al. (2012) The goal of Piantadosi et al. (2012) was to provide a computational cognitive model of how children learn a counting routine (i.e., how they select a number word which describes the size of some set of objects). In order to do so, they stipulate the LoT number primitives that are available to the child learner as they begin to learn the counting routine. While a detailed description of their assumed LoT would be outside of the scope of the present paper, what is relevant for our purposes is that they too, like Xu et al. (2020), assume that **1**, **2** and **3** are LoT primitives. They suggest that this assumption is central for their model to accurately predict certain properties of children's learning trajectory: if only **1** was an LoT primitives or, if **1**, **2**, **3**, **4** and **5** were LoT primitives, their model's predictions would match the data less well.

As the hypothesis according to which PRIM = $\{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$ fares worse than most other hypotheses we explored (cf. Results section), a question remains as to how to reconcile our findings with those of Piantadosi et al. (2012). While it would be interesting to explore how well our top hypotheses fit with the learning data from Piantadosi et al. (2012), another possibility is that the relevant patterns in child counting routine learning data, like subitizing, may have an explanation rooted in the lower level constraints on perception (cf. the discussion above of Xu et al. (2020) and Cheyette et al. (2021)). We leave the exploration of these possibilities for future work.

Further assumptions and limitations

This work incorporates several important assumptions which should be highlighted. Assumptions (i) and (ii) below apply to the new method for reverse-engineering LoT, and assumption (iii) applies to the case study on numerals presented here.

(i) There is an LoT, whose primitives are common to all humans.

(ii) Natural languages indeed optimize the complexity/informativeness trade-off, whereby complexity is rooted in LoT representations.

(iii) LoT has among its primitives a set of numbers and arithmetic operations $+, -, \cdot, /$. Importantly, however, alternative proposals for how LoT may look like exist. For instance, Piantadosi (2021) proposes that LoT has very few, perhaps as few as two, primitives, and that all our concepts, including all number concepts and arithmetic operations, are composed of them. If that approach is correct, our hypotheses about what PRIM contains would need to be revised. The method could then be re-applied to evaluate competing LoT hypotheses which are in line with Piantadosi (2021).

Finally, the present approach has two important limitations.

The first limitation stems from the observation that, while languages may be optimizing the complexity/informativeness trade-off, it isn't necessarily the case that all natural languages have converged on one of the optimal solutions. In other words, natural languages may deviate somewhat from the Pareto frontier (indeed, under no LoT hypothesis we explored is the $\overline{RD} = 0$). This means that different LoT hypotheses remain viable candidates as long as they don't give rise to large deviations from the Pareto frontier. Ultimately, a criterion for what counts as a large deviation from the Pareto frontier should be developed — what is the maximum deviation from optimality that natural languages tolerate? We leave this important problem for future work.

The second limitation of the approach is the hypothesis space, which is limited in two ways. First, the 48 LoT hypotheses explored are only a handful of possible hypotheses: in principle, any subset of number concepts can be considered as a viable hypothesis for PRIM. Even if we assumed that PRIM contains a subset of numbers 1-99, there would be $2^{99} - 1$ hypotheses for what PRIM contains and computational constraints prevent us from exploring them all. It is thus conceivable that some of the hypotheses we haven't explored proves to be the best one. Second, we have defined complexity of an LoT expression as its length in LoT primitives, in line with much previous work (Denić et al., 2021, 2022; Kemp & Regier, 2012; Steinert-Threlkeld, 2019, 2021). The underlying assumption of this definition is that all elements of PRIM and the arithmetic operations $+, -, \cdot, /$ have equal complexity. To dispense with this assumption, one could explore various complexity assignments to different primitives and evaluate how the results depend on these — this would allow to make inferences both about LoT primitives and about their relative complexities.

Conclusion

In this work, we have developed a new method for studying cognitive representations using cross-linguistic semantic data. The method was applied to numerals; importantly, it can be applied to other semantic domains for which cross-linguistic semantic data is available. We thus hope that it will be a valuable addition to the toolbox of linguists and cognitive scientists interested in studying cognitive representations.

Acknowledgements

We wish to thank Fausto Carcassi, Roni Katzir, Mora Maldonado, Shane Steinert-Threlkeld as well as the audience of the MLC seminar at the University of Amsterdam for helpful discussion. 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

Bylinina, L., & Nouwen, R. (2020). Numeral semantics. *Language and Linguistics Compass*, *14*(8), e12390.

Cheyette, S. J., Wu, S., & Piantadosi, S. (2021). The psychophysics of number arise from resource-limited spatial memory. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Comrie, B. (2013). Numeral bases. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/chapter/131`

Comrie, B., Dryer, M. S., Gil, D., & Haspelmath, M. (2013). Introduction. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/chapter/s1`

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.

Denić, M., Steinert-Threlkeld, S., & Szymanik, J. (2021, mar). Complexity/informativeness trade-off in the domain of indefinite pronouns. *Semantics and Linguistic Theory*, *30*, 166. doi: 10.3765/salt.v30i0.4811

Denić, M., Steinert-Threlkeld, S., & Szymanik, J. (2022). Indefinite pronouns optimize the simplicity/informativeness trade-off. *Cognitive Science*.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard University Press.

Gärdenfors, P. (2014). *The Geometry of Meaning*. The MIT Press.

Grinevald, C. G. (1990). A grammar of rama.

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

Hammarström, H. (2010). *Rarities in numeral systems*. De Gruyter Mouton.

Hurford, J. R. (2011). *The linguistic theory of numerals* (Vol. 16). Cambridge University Press.

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

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128. doi: 10.1146/annurev-linguistics-011817-045406

Knowlton, T., Pietroski, P., Halberda, J., & Lidz, J. (2021). The mental representation of universal quantifiers. *Linguistics and Philosophy*, 1–31.

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics*, *19*(3), 227–256.

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

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

Piantadosi, S. T. (2021). The computational origin of representation. *Minds and machines*, *31*(1), 1–58.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, *123*(4), 392.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'most': Semantics, numerosity and psychology. *Mind & Language*, *24*(5), 554–585.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008, jun). Does subitizing reflect numerical estimation? *Psychological Science*, *19*(6), 607–614. doi: 10.1111/j.1467-9280.2008.02130.x

Rumelhart, D. E., McClelland, J. L., & The PDP Research Group. (1986). *Parallel Distributed Processing* (Vol. 1). The MIT Press.

Spector, B. (2013). Bare numerals and scalar implicatures. *Language and Linguistics Compass*, *7*(5), 273–294.

Steinert-Threlkeld, S. (2019). Quantifiers in natural language optimize the simplicity/informativeness trade-off. In *Amsterdam Colloquium 2019.*

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

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

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57–70. doi: 10.1162/opmi_a_00034

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the 43rd annual meeting of the cognitive science society.*

Züfle, M., & Katzir, R. (2022). An evolutionary model-based approach to the missing o-corner. In *Proceedings of sinn und bedeutung 26.*