CHAPTER 1

Introduction

The Western world is a world of written numbers. One can hardly imagine an industrial civilization functioning without the digits 0 through 9 or a similar system. Yet while these digits have pervasive social and cognitive effects, many unanswered questions remain concerning how humans use numerals. Why do societies enumerate? How does the representation of numbers today differ from their representation in the past? Why does the visual representation of number figure so prominently in complex states? What cognitive and social functions are served by numerical notation systems? How do numeral systems spread from society to society, and how do they change when they do so? And, despite their present ubiquity, why have the vast majority of human societies not possessed them at all?

If you look up from this page and examine your surroundings, I am certain that you will encounter at least one instance of numerical notation, probably more. Moreover, unless you have a Roman numeral clock nearby, I am nearly certain that all of the numerals you encounter are those of the Hindu-Arabic or Western[1] system. Numerals serve a wide variety of functions: denotation – "Call George,

[1] The conventional term used in popular literature, "Arabic numerals," and the term used in most scholarly literature, "Hindu-Arabic numerals," can lead to considerable confusion because the scripts used to write the Hindi and Arabic languages use numerical notation systems that differ from those of the West in the shape of the signs. I use the term "Western numerals" to refer to this system because it developed in Western Europe in the late Middle Ages, while fully acknowledging its Indian and Arabic ancestry.

Cambridge University Press
978-0-521-87818-0 - Numerical Notation: A Comparative History
Stephen Chrisomalis
Excerpt
More information

876–5000"; computation – "21.00 × 1.15 = 24.15"; valuation – "25 cents"; ordination – "1. Wash dishes, 2. Sweep floor, 3. Finish manuscript"; and so on. Most of the thousands of numerals we see each day barely register on our conscious minds; regardless, we encounter far more written numbers in our lifetime than we do sunsets, songs, or smiles. Until the past few centuries, the opposite was true for most people.

These ten digits are so prevalent that it is easy to equate our numeral-signs with the set of abstract numbers. In this view, 62 does not merely signify the abstract concept "sixty-two" – it is the raw form of the number itself, the stuff of pure mathematics (or perhaps pure numerology). That these signs are frequently encountered and used in mathematical contexts contributes to the prevalence of such attitudes. According to this view, our numeral-signs constitute abstract number, and other systems (when recognized as such) are simply archaic deviations from the abstract entity comprised by these signs.

This view is erroneous, and rests on the confusion of a mental concept (signified) with its symbolic representation (signifier). Our numerical notation system has an extensive history, as do the more than one hundred systems that have existed over the past five thousand years. Still, the worldwide prevalence of Western numerical notation is undeniable. Most literate individuals worldwide, as well as a sizable number of illiterates, understand them. Nor does any competing system have any reasonable chance of supplanting our system in the near future. This has led many scholars to assert its supremacy solely on the evidence of its near-universality (Zhang and Norman 1995; Dehaene 1997; Ifrah 1998). Nevertheless, this situation does not imply that our system will dominate the whole world forever. The study of numerical notation remains mired in a theoretical framework that has much more in common with late nineteenth-century unilinear evolutionism in anthropology than it does with early twenty-first-century critiques of unfettered scientific progress.

Despite this theoretical weakness, numerical notation as a topic of academic study is a relatively common pursuit, with linguists, epigraphers, archaeologists, anthropologists, historians, psychologists, and mathematicians all making significant contributions to the literature. These studies are mostly restricted to the analysis of one or a few numerical notation systems, although a small number of synthetic and comparative works dealing with numerical notation exist (Cajori 1928; Menninger 1969; Guitel 1975; Ifrah 1998). However, such works rarely consider more obscure numerical notation systems, such as those of sub-Saharan Africa, North America, and Central Asia. Similarly, social scientists such as the anthropologist Thomas Crump (1990), the psychologist David Lancy (1983), and the ethnomathematicians Marcia Ascher (1991) and Claudia Zaslavsky (1973) have undertaken major comparative research on numeracy and mathematics in non-Western societies. Yet numerical notation has not been a primary focus of this body of research.

This study is a comparative analysis of all numerical notation systems known to have existed throughout history – approximately one hundred distinct systems, most of which can be grouped into eight distinct subgroups. By presenting a universal study of such systems and examining the historical connections and contexts in which they are encountered, I will develop a framework that accounts for cultural universals, identifies evolutionary regularities, and yet remains cognizant of idiosyncratic features, seeking to determine, rather than to assume, the amount of intercultural variability among them. I will distinguish several major types of numerical notation, evaluate their efficiency for performing specific functions, link their features to human cognitive capacities, and relate systems to their sociopolitical contexts.

Definitions

A **numerical notation system** is a visual, relatively permanent, and primarily non-phonetic structured system for representing numbers. Signs such as 9 and 68, IX and LXVIII, are part of numerical notation systems, but numeral words such as *nine* and *achtundsechzig* are not. Though there are ties between numeral words and numerical notation, a **lexical numeral system**, or the sequence of numeral words in a language (whether written or spoken), has a language-specific phonetic component. Every language has a lexical numeral system of some sort, while numerical notation is an invented technology that may or may not be present in a society.[2] Some numerical notation systems contain a small phonetic component, as in *acrophonic* systems whose signs are derived from the first letters of the appropriate number-words in a language. However, since such systems are still comprehensible without having to understand a specific language, they are numerical notation systems.

Numerical notation systems must be *structured*. Simple and relatively unstructured techniques, such as marking lines on a jailhouse cell to count one's days or piling pebbles in a basket, are largely or entirely unstructured. They rely on **one-to-one correspondence**, in which things are counted by associating them with an equal number of marks or other identical objects. A numerical notation system, by contrast, is a system of different discrete **numeral-signs**: single elementary symbols, or, in the terminology used in writing systems, *graphemes*, which are then

[2] I will leave aside for the moment discussions of counterevidence questioning the assumption of the universality of lexical numeral systems (Hurford 1987: 68–78; Gordon 2004; Everett 2005).

used in combination to represent numbers.[3] A **numeral-phrase** is a group of one or more numeral-signs used to express a specific number (e.g., MMDXXV); numeral-phrases such as 8 or Roman L are nonetheless complete even though they only use one sign apiece.

All numerical notation systems (and most lexical numeral systems) are structured by means of **powers** of one or more **bases**. A **power** is a number X multiplied by itself some number of times (its power); $10^1 = 10$, $10^2 = 100$, $10^3 = 1000$, etc. By mathematical definition, a number raised to the power 0 equals 1. A **base** is a natural number B in which powers of B are specially designated. While mathematicians normally require that a base be extendable to an infinite number of powers of B (e.g., 10, 100, 1000, 10,000, … ad infinitum), most numerical notation systems are not infinitely extendable. It is sufficient that *some* powers of B are specially designated within a numerical notation system. Western numerals and many other systems use a base of 10, but this is not universal. In addition to its base, a numerical notation system may have one or more **sub-bases** that structure it. The Roman numeral system has a primary base of 10 with a sub-base of 5. Unlike bases, the powers of sub-bases are not specially designated; there are no special Roman numerals for 25 or 125. It is, rather, the *products* of a sub-base and the powers of the primary base that are specially designated – for the Roman numerals, 50 (5 × 10) and 500 (5 × 100).

Two topics that I will present only peripherally are number and mathematics. **Number** is an abstract concept used to designate quantity. For the purposes of my study, a simple (if philosophically naïve) definition will suffice. Questions such as whether numbers are "real" or Platonic entities, or the connection of the set of natural numbers to formal logic, are beyond its scope. The distinction between **cardinal** numbers – denoting quantity but not order – and **ordinal** numbers – designating ordered sequences – is extremely important for lexical numerals, where many languages use different series of words (e.g., *two* versus *second*) for the two concepts. This distinction also has implications for our understanding of the origin of numerals and numerical concepts in humans (Crump 1990: 6–10), but has little influence on numerical notation. In defining **mathematics** as the science that deals with the logic of quantity, shape, and arrangement, I am consciously employing a simple definition for a complex term. In order to understand numerical notation, one needs no mathematical ability save some knowledge of basic arithmetic. While some parts of mathematics make frequent

[3] A few numeral-signs are more complex in that they graphically combine two or more signs into one in order to represent multiplication, but they are treated as elementary numeral-signs because their use is identical to that of all other simple signs in the systems in question.

use of numbers (number theory being the most obvious example), large parts of the discipline have only infrequent or peripheral encounters with numerical notation. Numerical notation systems are not necessarily designed with mathematical purposes in mind. Even in contemporary industrial societies, where mathematical ability is more extensive than in any other historical or modern society, most numerical notation is nonmathematical.

Universal Comparison

The present study is, as far as possible, a universal one. I have not excluded any numerical notation system intentionally save where data are not plentiful enough to undertake a reasonable analysis. Most comparative research in anthropology aims to discover generalizations and patterns in human behavior, but using the universe of cases is neither possible nor desirable in most cross-cultural studies. In order to use most analytical statistics on cross-cultural data, each case must be independent of the others, which requires that each case may not be historically derived or diffused from any other case. This issue, known as *Galton's problem*, is the thorniest methodological issue in statistical cross-cultural research (Naroll 1968: 258–262). The establishment of correlations between traits among historically independent societies is enormously useful, and is the basis for most cross-cultural research in modern anthropology.

Yet to do so in a study such as this one, in which there are perhaps only seven independently invented numerical notation systems, would be pointless. Firstly, seven cases would be too small a sample to analyze statistically. Secondly, by studying all cases, I am able to show that the total observable variability among numerical notation systems is far greater than has previously been believed. This variability cannot be understood by studying only a fraction of numerical notation systems. To paraphrase the old fable, if we study only the elephant's trunk or tail, we ignore most of the animal. Thirdly, I wish to explain structural variation among historically related systems, which frequently differ considerably from their relations. This would be impossible using a sampling technique that omitted related cases. Finally, were I to omit related cases, I could not analyze how systems change over time or how new systems develop out of existing ones. By taking events of change, rather than static systems, as the units of analysis in my comparisons, I am able to elucidate both synchronic and diachronic patterns among numerical notation systems. It is worth noting that Galton's problem does not apply to events of change of the sort I am analyzing, since every *event* is essentially independent of every other, and can thus be analyzed statistically, where relevant.

I reject as false the dichotomy in anthropology between universalism (Tylor 1958 [1871], White 1949, 1959; Steward 1955; Harris 1968) and relativism (Lowie 1920, Boas 1940, Sahlins 1976, Geertz 1984), both of which presume rather than evaluate the degree of regularity that social phenomena display. While numerical notation systems display remarkable regularities and even universals, historical contingencies also played a major role in shaping the cultural history of numerical notation. Yet the only way to determine which features of numeration are cross-culturally regular and which are idiosyncratic is to undertake cross-cultural comparison. The best way to deal with the messiness of the world – less universal than universalists would like, less relative than relativists prefer – is through a body of theory that deals with constraints.

Most anthropological theory is predicated on the existence of very strong constraints on the forms possible within human societies. Some of these constraints are so strong as to produce cross-cultural universals (Brown 1991). Most cultural relativists dismiss these universals as minimally true, but facile, irrelevant, and useless for understanding humanity (cf. Geertz 1965, 1984). The denial of comparativism on this basis is an overly negative position, given that those who criticize comparativism most harshly are very often those who have not undertaken it. One of the most crucial theoretical contributions of anthropology should be to indicate the degree to which human societies are alike and the degree to which they differ. While some aspects of human existence are truly universal, and others are almost infinitely variable, most of the really intriguing domains of activity fall somewhere in the middle.

In the early 1900s, Alexander Goldenweiser developed his "principle of limited possibilities," which stated that for any social or cultural phenomenon, there are a limited number of possible forms that can be expressed in human societies (Goldenweiser 1913). Goldenweiser was particularly interested in the limitations imposed by human psychology on the expression of cultural traits, although, given the inchoate nature of psychological theory at the time, he was unable to describe these mechanisms precisely. Bruce Trigger (1991) has rejuvenated the idea of constraints, proposing that anthropologists should use the concept of constraint to describe the limitations on human sociocultural variation – whether those constraints are biological, ecological, technological, informational, psychological, or historical – in order to analyze statistical regularities among cultures without implying determinism. We must be cautious, with both the "limited possibilities" and the "constraint" approaches, not to restrict our formulations and assume the restricting influence of various factors to be more important than positive (enabling) effects. A very strong propensity in favor of some trait is not the same thing as a very strong constraint against all other possibilities. Constraints and inclinations can and do coexist, and the negative limitations of one variable must be weighed against the positive inclinations of another. Despite this caveat, I find a constraint-based approach to be the most promising theoretical perspective for explaining the regularities found in numerical notation systems, something to which I will return in Chapters 11 and 12.

In much of my analysis, I follow Joseph Greenberg (1978), whose analysis of significant regularities in lexical numeral systems presents a list of fifty-four generalizations. Unlike much of his later work, Greenberg's study of numerals is universal and cognitive in orientation rather than phylogenetic. It is synthetic, based on the detailed empirical work of earlier scholars, such as the German linguist Theodor Kluge, who spent years compiling sets of numeral terms in languages throughout the world (Kluge 1937–42). While many of Greenberg's regularities are extremely complex[4] or have some exceptions, others reveal truly universal and nontrivial features of every natural language; for instance, every numeral system contains a complete set of integers between one and some upper limit – each system is finite[5] and has no gaps (Greenberg 1978: 253–255). Similarly, no natural language expresses "two" as "ten minus eight" or "twenty" as "one-fifth of one hundred."

While every language has a set of lexical numerals, most pre-modern societies functioned quite well without numerical notation. It is possible to conceive of a world in which there are many regularities in lexical numerals, but in which numerical notation systems are highly specific and unique responses to local needs. We do not live in such a world. There is considerable uniformity among the world's numerical notation systems, and they display many synchronic and diachronic regularities.

In fact, the number and variety of conceivable numerical notation systems is far greater than what is attested historically. To take only a very limited example, a numerical notation system can very easily be imagined that is just like the Western system but – instead of being a decimal system – having a base of any natural[6] number of 2 or higher. Yet most numerical notation systems have a base-10 structure (and those that do not use multiples of 10). This does not preclude the existence of binary and hexadecimal numerical notation for specialized computing purposes. Similarly, while there are only five basic principles of numerical notation systems found historically (as described earlier), it is easy to imagine other types that could have existed: a system where the size of a numeral-sign is relevant to its

4 For instance: "37. If a numeral expression contains a complex constituent, then the numerical value of the complex constituent itself in isolation receives either simple lexical expression or is expressed by the same function and in the same phonological shape, except for possible automatic phonological alternations, stress shifts, or overt expressions of coordination" (Greenberg 1978: 279–280).

5 This is not true of numerical notation systems, some of which (like our own) are truly infinite.

6 Or even, as discussed in some aspects of number theory, having a fractional or negative base!

value, or where all nonprime numbers are expressed multiplicatively using prime number numeral-signs. Several modern writers, abandoning traditional principles of numerical notation, have created new systems ex nihilo that rely on rather different principles than do the systems discussed in this study (Harris 1905; Pohl 1966; Dwornik 1980–81). Explaining regularities from a constraint-based perspective allows us to speculate about why certain numerical notation systems flourish while others do not. Instead of denying the existence of exceptions, I use general rules to explain why special cases are special, and why some imaginable systems are unattested in the ethnographic and historical records.

Yet one might wish to contend that comparison of any sort, much less the universal type of this study, is misleading because each culture, and hence each numerical notation system, is a product of unique historical circumstances. If so, comparing Egyptian hieroglyphic numerals to Shang oracle-bone numerals and Inka *khipus* might be misleading. At best, even if there is a core of features common to all numerical notation systems, I would be labeling oranges apples in order to compare them to other apples. At worst, if these systems are entirely different phenomena, I am trying to make apples out of abaci. Yet the relative ease of intercultural communication refutes the claim that all cultures are incommensurable. The intercultural transmission of ideas relating to numerical notation systems is frequent and poses a serious challenge to this degree of relativism. Prior to comparing phenomena among multiple societies, one cannot assume either that the phenomenon is cross-culturally regular or that it is not. Having compared numerical notation systems on a worldwide basis, I regard the systems as being sufficiently similar to warrant their theoretical analysis as variations on a single theme.

I regard numerical notation as translatable cross-culturally without significant loss of information or change of meaning. The number 1138 is practically identical in referent to MCXXXVIII or 𓆼𓍢𓎆𓎆𓎆𓏺𓏺𓏺𓏺𓏺𓏺𓏺𓏺 or any other representation. These systems have very different structures, but, in Saussurean terms, the various signifiers refer to the same signified (Saussure 1959). Although the linguistic and symbolic signifiers for numbers may differ greatly (23, *dreiundzwanzig,* XXIII, *viginti tres,* etc.), the correlation of both numeral-phrases and lexical numerals with natural numbers is not culturally relative. Yet, while seemingly uncontroversial in the exact sciences, the cross-cultural universality of number concepts has been criticized recently by relativistic anthropologists and sociologists. In his recent work on Quechua number and arithmetic, Urton (1997) asserts that Western concepts such as "odd/even" are not appropriate to the Quechua arithmetical experience, and that the Quechua use a fundamentally different ontology of numbers than the Western one. Yet Quechua numbers can be understood in the same way as any others, and the Inka numerical notation used by Quechua speakers (Chapter 10) can be compared to others without any particular difficulty. Relativist philosophers such as Restivo (1992) claim that 1 + 1 cannot equal 2 in any absolute manner, because if one were to take a cup of popcorn and add a cup of milk to it, the result would not be two of anything, but somewhat more than a cup of pulpy mush. Resisting the temptation to describe such casuistry as pulpy mush, I simply point out that addition is an arithmetical function that can only represent adding discrete objects of a like nature. Such evidence does not convince me that the number concepts of non-Western societies are incommensurable with our own. On the contrary, my own research suggests that these differences are relatively inconsequential in comparison to the commonalities observed in all societies.

I acknowledge that, by treating all numerical notation systems purely as systems for representing number, I do not do justice to the complex symbolism that complements many of them or to the scholarship on numerology (Hopper 1938, Crump 1990). The arrival of the year 2000 was not simply another cause for celebration (or trepidation); rather, the nature of our numerical notation system and the "rolling over" of the calendrical odometer on 2000/01/01 held great symbolic and even mystical significance for much of the world's population. My decision to underemphasize numerology is based partly on space limitations, but also on my theoretical interest in the comparable core of features underlying all lexical numeral systems and numerical notation systems. These interesting differences do not affect the validity of cross-cultural comparisons, but merely highlight the need to establish, rather than assume, the level of regularity in sociocultural phenomena. It may be true, as Geertz (1984: 276) famously asserted, that "[i]f we wanted home truths, we should have stayed at home," but if we want *human* truths, we must compare.

Structural Typology of Numerical Notation

The systematic classification of numerical notation systems helps to identify their relevant features, distinguish independent inventions from cultural borrowings, and determine how their features relate to their uses. The goal of typology is not simply to develop a scheme into which every case fits, but to do so in a way that allows us to ask and answer questions that could not otherwise be considered. When poorly done, typology is descriptive but nonanalytical, and thus largely useless; when well done, it organizes knowledge in a way that answers inquiries. Any classificatory scheme is inherently theory-laden, and answers only some of the questions that might be asked of a set of data. The typology presented here represents all the major principles by which numbers are represented and emphasizes the features of numerical notation that are cognitively most important. It removes each system from its temporal, geographic, and spatial contexts and examines how numeral-signs are combined to represent numbers.

Any natural number can be expressed as the sum of multiples of powers of some base. In Western numerals, 4637 is $4 \times 1000 + 6 \times 100 + 3 \times 10 + 7 \times 1$ – or, to use exponential notation, $4 \times 10^3 + 6 \times 10^2 + 3 \times 10^1 + 7 \times 10^0$. Because the Western numerals use the principle of **place-value**, the value of any numeral-sign in the phrase is determined by its position – position dictates the power of the base that is to be multiplied by the sign in question. If the order changes, the value changes, so that 6437, 3674, and so on mean different things than 4637. We could also write the number out lexically as *four thousand six hundred and thirty seven*. Instead of using place-value, the powers (except for 1) are expressed explicitly – *thousand, hundred, -ty*. Because each multiplier corresponds to a word for a power, we could in theory move each power and its multiplier to a different spot without introducing ambiguity; German lexical numerals, among others, do exactly that – *viertausend sechshundert sieben und dreizig* "four thousand six hundred seven and thirty."

Some numerical systems, however, do not use multiplication at all. To use the Roman numerals, one simply adds up the values of all the signs: MMMMDCXXXVII – 1000 + 1000 + 1000 + 1000 + 500 + 100 + 10 + 10 + 10 + 5 + 1 + 1. Although there is no logical requirement that systems like the Roman numerals list their powers in order, they almost universally do so. The Roman numeral CCLXXVIII could be unambiguously read even if it were written as VIIICCXXL, or even as XLIVIXCIC, if we omit the slight complexity of the occasional use of subtraction. The fact that such disordered phrases are not valid tells us something about systems that lack place-value, however – they too are structured as the sum of multiples of powers. We can thus interpret the Roman numeral MMMMDCXXXVII as (1000 + 1000 + 1000 + 1000) + (500 + 100) + (10 + 10 + 10) + (5 + 1 + 1). I will return in a moment to the issue of the signs for 500 and 5, and how they affect our understanding of such systems.

The only major systematic attempt to date to classify numerical notation systems is Geneviève Guitel's *Histoire comparée des numérations écrites* (1975).[7] Guitel classifies approximately twenty-five systems (drawn from about a dozen societies) according to whether they use addition alone to form numeral-phrases (Type I, like Roman numerals), addition and explicit multiplication (Type II, like English lexical numerals), or implicit multiplication with place-value (Type III, like Western numerals) – just as I have done here. Each type is further subdivided according to the systems' base(s) and other features. Despite an admirable attempt, Guitel's analysis fails the most basic test of classification, which is that it must classify similar systems together and separate dissimilar ones. It is problematic because its primary division is made only on the basis of the degree to which multiplication

[7] See also Zhang and Norman (1995). Ifrah's (1985, 1998) popular studies on the subject follow Guitel's typology.