Properties of Lower Numerals and their Explanation: A Reply to Paweł Rutkowski*

Harald Hammarström
Chalmers University

Abstract

In a recent article (Rutkowski 2003) Paweł Rutkowski argues that the numerals 1-4 are treated specially in their syntax in across languages. Rutkowski wishes to explain this contrast as due to the working memory’s limited capacity, which cognitivists argue is indeed four items. We challenge this claim by presentation typological data to show a decreasing tendency of lower numerals to be more idiosyncratic and that 4 is a soft and arbitrary cut-off point outside Indo-European languages. We therefore argue that frequency, with special attention to cognitive reference points, is a better explanation.

Keywords: numerals, frequency, typology of numerals, subitizing, cognitive reference points

* We wish thank the consulted libraries of Stockholm, Uppsala and Gothenburg and their xerox-machines. A huge debt of gratitude is to Dr. Aarne Ranta and the VINNOVA project for help and financial support.
1. Introduction

In a recent article (Rutkowski 2003) Paweł Rutkowski argues that the numerals 1-4 are treated specially in their syntax in across languages. As a prominent example, Polish 1-4, resembling adjectives, exhibit case agreement with a head noun whereas higher numerals require the genitive plural on the head noun irrespective of the case of the numeral itself\(^1\). Rutkowski wishes to explain this contrast as due to the working memory’s limited capacity, which cognitivists argue is indeed four items. The case is built on the following pillars:

(1) a. Numerals referring to cardinalities in the range of ’one’ to ’four’ differ considerably from higher numerals in “many” languages.
   b. Heine’s explanation (Heine 1997) in terms of frequency is wrong (citing Polish frequency data).
   c. There is independent evidence from neuropsychology (Cowan 2001) that the capacity of the working memory matches the dichotomy on numerals at “the magical number four,” and is, therefore, a plausible cause of it.
   d. Those languages where 1-4 are not special are less conservative (in this respect). They have obscured earlier differences as speakers grew more accustomed to grouping higher magnitudes.

We will try to counter these arguments along the following lines:

(2) a. We present typological data to show that 1-4 aren’t different in the way Rutkowski et al. suggest and especially that there

\(^1\) The full details can be found e.g., in (Przepiórkowski 2001) or any reference grammar of Polish.
is no strict border at 4.

b. The four-limit of the working memory and visualizing is demonstrated well enough but the suggested connection to language is very doubtful. It is much more reasonable that such a connection would also manifest itself on other aspects of numerals in language, notably that of base, but base-4 systems are patently marginal cross-linguistically.

c. The tendency of lower numerals to behave differently can indeed be explained in terms of frequency and so-called cognitive reference points (Rosch 1975, Jansen et al. 2001, Sigurd 1988). Cognitive reference points for numerals would be those numbers which are multiples of the base, e.g., 10, 30, 1000 if base is 10 and also number between two base-multiples, e.g., 5, 15, 50. The addition of special treatment of “round” numbers should cover Rutkowskis objections to Heine.

The typological data is from a database built in connection with a numeral grammars project of the language technology group at Chalmers University. The database contains roughly a 100 languages chosen randomly among reference grammar resources and relevant journal articles of the three main university libraries in Sweden.

Each language is annotated with genetic classification data which is taken into consideration when doing actual counts, so the seeming overrepresentation of the languages of Europe is not really harmful. There are many interesting ways to do discount for genetic bias. Here we have simply only included one random language from each genetic phylum. The full results of the typological investigation will be given in a future publication.

The names and the classification of the languages are given in appendix A. For languages used in the sample it’s best to rely on reference grammars as sources. Other data is also available in general articles on numerals and we shall also make use of such throughout
Properties of Lower Numerals and their Explanation: A Reply to Paweł Rutkowski

this text when the aim is not to show that this or that pattern is more common, but to attest the existence of a certain pattern.

2. The Special Status of 1-4

Rutkowski claims that typological research can confirm that “there is a distinction between the four lowest numerals and the higher ones”. For this he cites a landmark paper by Greenberg (Greenberg 1978) saying that “in some languages (e.g., in New Guinea) numerals other than ’one’ to ’four’ have not developed at all”. This is not strictly false but it is quite uninteresting because there is nothing special about four here. It is namely the case that for some languages three is the limit (Green 1997), some two (McGee 1898) and there is also Pirahã, a language that lacks exact number words altogether, about to be properly documented (Everett 2004a; Everett 2004b). And sometimes the last number is 5, 6 or 7 (Roux 1950). What is more interesting than existence is thus how common one system is over another. In our sample, considering the languages with a numeral system smaller than ten, !Xù (Snyman 1970), Wardaman (Merlan 1994) and Gooniyandi (McGregor 1995) have 3, Kayardild (Evans 1995) has 4 and Kwaza (van der Voort 2000) has 5.

Continuing, Rutkowski notes that in many languages, 5–9 are transparent compounds of lower numerals and cites a few examples where ’four’-forms occur in such compounds. However, it has been known for 150 years (Pott 1847) in the typology of numerals that the overwhelmingly most common member of such compounds is 5. In our sample, the number 4 occurs as a member of some compounded 5-9 in about one tenth as many languages as 5 does. And 3 is as common as 4 in this respect. Most languages in our sample have forms below ten that show no sign of being composite. But we cannot conclude that they were not compounds when initially formed since the secret of their formation is almost always too remote in time. The
longest inventory of synchronically monomorphemic numerals known to us is 1-12, exemplified in Jere (Shimizu 1982), a base-12 language in Plateau Nigeria.

Ifrah (1985) is given as the next source, but Ifrah only argues for four as the limit of perception and does not state that there is such a tendency in language (except for Indo-European; see below).

Nevertheless, the subsequent reference to Hurford (2001) is the most important one since it gives empirical data for the proposed dichotomy. Hurford gives 8 ways in which numerals 1-4 are said to have the tendency to differ from the higher ones and backs it up with tables on the incidence of each property in about 16 languages in Europe.  

The 16 languages are some distinct branches of Indo-European, Basque, Finnish and Hungarian (distantly related) in the Uralic family, the Nakho-Dagestanian Archi, Avar, Lezgian and Godoberi as well as Adyghe and Kabardian which are Circassian.

Let’s have a completely sober look at Hurford’s evidence. Below, “more often” should be interpreted as “more often than higher numerals,” not as “more often than not at all.”

(3) a. Lower numerals more often have a distinct counting form (as opposed to the attributive)
   b. Lower numerals more often have suppletive or irregular ordinal forms
   c. Lower numerals more often exhibit a different word order when used attributively
   d. Lower numerals more often have distinct forms for counting different objects

The features simply don’t show a strict specialness of 1-4 but rather a tendency for lower numerals. For in all these cases, the number 2

\[2\] More details on the same tables are given in (Hurford 2003).
shows the feature more often than 3 and so on (and it does not necessarily stop at 4 at all). Hurford also notes the well-known hierarchy of grammaticalized number i.e., that trial > dual > singular/plural distinction. This isn’t a feature of number names but nevertheless an excellent parallel. My key point here is that there’s no particular cut-off point at 4 readable from the tables – it’s more like a downwards continuum and the only reasonable cut-off point would to be 2.

The other four properties are also of the kind that: the lower the numeral the more prominently it exhibits the property. But, they do have a clear cut-off point at 4:

(4) a. Definite/indefinite marking
    b. Distinct case-forms
    c. Governing of case-marking on head noun
    d. Different gender (a special case of the last point of the previous paragraph)

However, the languages for which there is such a cut-off are precisely those that are Indo-European! In fact, the idiosyncratic behaviour of 1-4 can be reconstructed for proto-Indo-European (Gvozdanović 1992) (at least non-Anatolian proto-Indo-European) because they clearly show the signs of shared retention and not independent innovation in the daughter languages (except possibly the definite/indefinite distinction in Albanian). This idiosyncrasy of 1-4 seemingly existing in “many” languages is only a multiplication of the fact that it was so in the parental Indo-European. In a typological comparison it can only be counted once. The reader may object that the bona fide dichotomy in Indo-European merits something because at least there’s one language(-family) that professes it. But such reasoning is idle exercise because it’s trivial to select a case with special properties when looking at a large number of languages. For instance, Inari Sami (Nelson et al. 2000) and
Quimbundo (da Silva Maia 1964) have an equally bona fide syntactic dichotomy at 6!

In our sample we have not encoded syntactic idiosyncracies of numerals since an open set of languages has such rich syntactic variation that it would utterly complicated to do so in a complete manner. We can however point out informally that we have failed to note the “magical number four,” especially not in the terms of Hurford and Rutkowski, in going through the descriptive grammars of hundreds of languages. Sometimes there is syntactic variation conditioned by magnitude, sometimes the border is 4, sometimes it’s 2, 3, 5, 6, 9, 20 or when the numeral goes composite.3

It must rest completely clear that it is illegal to claim a cross-linguistic special limit of 4, based on the evidence and methodology of Hurford and Rutkowski. However, the weaker reading of Hurford’s claim i.e., that the lower the numeral the more likely it is to behave differently, has some merit.4 Of course, it is not a sufficient criterion on it’s own since not all languages show it.

### 3. The Neuropsychological Basis

We shall now reflect upon the idea that the limit of perception, termed the subitizing limit (van Oeffelen et al. 1982), or limit of working memory at 4 (Ifrah 1985, Cowan 2001) might have repercussions in language. We do not challenge the claim that there is such a neuropsychological configuration.

A limit at 4 means that higher magnitudes must be broken into chunks of 2, 3 or 4. The first place we’d expect this kind of chunking

---

3 Romanian quantifies thousands with a “de” if the quantifier is ≥ 20 (Murrell et al. 1970).

4 The exact same conclusion on Hurford’s evidence is drawn independently by Dehaene (Dehaene 2001).
to turn up would reasonably be that of the (first) base of a numeral system of higher capacity. But it is universally agreed that systems with a first base other than 5 or 10 are very rare cross-linguistically. There are of course a considerable number of languages with rudimentary number systems in which 2 is the point after which adding starts i.e., 1, 2, 2+1 and so on (Curr 1887). However, it is significantly much rarer that 3 or 4 plays this role. We cannot give any hard figures on commonness because none of the languages in our sample are of this type.

It can of course be objected that 5 and 10 are so universal as the (first) base only because we have five fingers on each hand, and that Rutkowski’s theory is, although shadowed, alive and kicking too. Nevertheless, if this is so, we should expect chunk forming under 5 very often. Such is attested (Pilhofer 1933) but does not occur at all in our sample – even in languages where the formations are transparent. Of course, for the vast majorities that are not transparent Rutkowski’s theory might be correct but unprovable. But theories that are epistemologically empty are dispreferred.

Thus, the reason why a neuropsychological “magical number four” should show up in syntax with other linguistic elements and not in the composition of number words is obscure to us. Constraints on the working memory show up almost trivially in other linguistic expressions. For example, the number of repetitions of identical elements in a compound (usually additive) seems constrained almost everywhere. I.e we see compounds like 2+2 for 4 a lot, 2+2+2 for 6 sometimes and 2+2+2+2 for 8 very rarely. Similarly, reduplication is attested abundantly in the world’s languages but not so for triplication or quadruplication.

5 The only well-documented cases with base-4 known to us are Kewa (Franklin et al. 1962), Chumash (Beeler 1967), Biem (Ossart 2004) and perhaps Lule (Ibarra Grasso 1939).

6 See (Green 1997) for some attestations and examples.
To sum up, we can note that Rutkowski’s hypothesis of the neuropsychological constraint is of course consistent with the fact that lower numerals tend to be more idiosyncratic, but it is not the tightest explanation nor does it explain the other numbers’ specialness. Surely, to say that 4 is special and explain it is clearly inadequate if 3, 5 and 6 are equally or more special but unexplained.

4. Explanation through Frequency and Cognitive Reference Points

Heine (Heine 1997) had given the tendency for lower numerals to be different as conditioned by frequency and put in the frame of grammaticalization (where frequency plays an important role in general). A typical number word frequency-curve for a European language is given in Figure 4 (BNC 2004). The curve is similar for other base-10 languages (Dehaene et al. 1992).

Figure 1. Frequency of English Number Words According to the British National Corpus
Note that the sloping frequencies are exactly what is needed to govern the increased tendency for lower numerals to be constructed differently that we see cross-linguistically.

A particular language, Polish, has a super-sharp syntactic distinction between four and five. Rutkowski uses a Polish frequency dictionary (Kurcz et al. 1990) to demonstrate that this does not correlate with frequency. First, 5 was more frequent than 4 and there is no observable syntactic distinction in Polish to parallell the frequency drop between say 3 and 4. Moreover, e.g., 100 and 1000 had very exceptional frequencies compared to other numbers of similar magnitude like 99 and 1001.

All that is needed to make sense out of frequency curve like that of figure 4 are two factors:

(5) a. The lower the numeral the higher the frequency (roughly logarithmic on 1-9)

b. Cognitive reference points (Rosch 1975) have a higher frequency (than (16) would give them). The cognitive reference points for numerals are multiples of the base and numbers in-between them. See e.g., (Sigurd 1988) for an elaboration and an explicit formula. The reason for their higher frequency can be that they are used for approximation (Dehaene et al. 1992, Jansen et al. 2001).

It would be very interesting to know how the second principle is affected if the base is other than 10 or 20. But big-enough systems with other bases are very rare—let alone finding a corpus. The only

---

7 A short excursion reveals that 5 is slightly more frequent than 4 in three Russian (Zasorina 1977, Shteinfel’d 1963, Löngren 1993) and one Spanish frequency dictionary (Juilland et al. 1964). But the other way around in one French (Juilland et al. 1960), one Latvian (Jakubaitis et al. 1973) and one Italian frequency dictionary (Juilland et al. 1973).
case known to us that comes close is Birom, a base-12 language in Nigeria, which has a volume of texts in print (Bouquiaux 1970). Despite it being 400 pages (including phonetic transcription and French translation), neither 5, 6 nor 10, 12 occur at all (count by the present author).

The frequencies so obtained are sufficient to explain a cross-linguistic tendency for differing syntax in lower numerals. The disharmony of this law on Polish does not matter at all unless the situation in Polish is typical at large. As argued above, it is not.

The frequency curve can also be used to correlate with length of the number word. On the assumptions that length can be heuristically measured by the number of ascii characters in transcription, and (more controversially) that the frequency-curve is similar across languages, we have performed experiments on the number-words on all the languages in our sample. That frequency correlates with the inverse of length, receives overwhelming support. The data will be published in the near future.

References


Cowan, N. 2001. The Magical Number 4 in Short-term Memory: A

---

8 I say a and not the because it remains to be conclusively shown. Our typological data does not include it but we very good faith in it.
Reconsideration of Mental Storage Capacity. *Behavioral and Brain Sciences* 87-114.


Ibarra Grasso, D. 1939. Las Numeraciones Cuaternarias. *Boletín de la
Academia Argentina de Letras VII 28, 585-606.
Dissertation, Tübingen: University of Tübingen.
Appendix A: The Languages of the Sample

Afro-Asiatic Chadic Biu-Mandara A A.2 2:
Margi

Afro-Asiatic Chadic West B, B.2.:
Miya

Afro-Asiatic Cushitic East Oromo:
Oromo

Afro-Asiatic Cushitic East Somali:
Af Tunni

Afro-Asiatic Omotic North Gonga-Gimojan Gimojan Ometo-Gimira
Ometo:
Maale

Afro-Asiatic Semitic Central South Arabic:
Maltese

Afro-Asiatic Semitic North-West Semitic:
Biblical Hebrew

Afro-Asiatic Semitic South-Semitic Ethiopian North:
Ge’ez

Afro-Asiatic Semitic South-Semitic Ethiopian Transversal
Amharic-Argobba:
Amharic

Afro-Asiatic Semitic South-West Semitic:
Classical Arabic

Aracaunian:
Mapuche

Australian Bunaban:
Gooniyandi

Australian Gunwingguan Yangmanic Yibwan:
Wardaman

Australian Pama-Nyungan Tangic:
Kayardild

Austro-Asiatic Mon-Khmer Eastern Mon-Khmer Bahnaric South
Bahnaric Stieng-Chrau:
Stieng

Austro-Asiatic Mon-Khmer Eastern Mon-Khmer Khmer:
Khmer

Austro-Asiatic Munda North Munda Kherwari Mundari:
16 Properties of Lower Numerals and their Explanation: A Reply to Paweł Rutkowski

Ho
Austronesian Malayo-Polynesian Central-Eastern Central
Malayo-Polynesian Bima-Sumba:
Kambera
Austronesian Malayo-Polynesian Central-Eastern Eastern
Malayo-Polynesian Oceanic Central-Eastern Oceanic Remote Oceanic
Central Pacific East Fijian-Polynesian Polynesian Nuclear
Samoic-Outlier Samoan:
Samoan
Austronesian Malayo-Polynesian Central-Eastern Eastern
Malayo-Polynesian Oceanic Central-Eastern Oceanic Remote Oceanic
North and Central Vanuatu Northeast Vanuatu-Banks Islands West
Santo:
Araki
Austronesian Malayo-Polynesian Western Malayo-Polynesian Sulawesi
Muna-Buton Tukangbesi-Bonerate:
Tukang Besi
Austronesian Malayo-Polynesian Western Malayo-Polynesian Sundic
Malayic Malayan Local Malay:
Malay
Basque:
Basque
Burushaski:
Yasin Burushaski
Central Khoisan Nama:
Khoe-Kowap
Creole Ngbandi based:
Sango
Dravidian Southern Tamil-Kannada Tamil-Kodagu Kodagu:
Kodagu
Dravidian Southern Tamil-Kannada Tamil-Kodagu Tamil-Malayalam
Tamil:
Tamil
Irula
Eskimo-Aleut Eskimo Siberian Yupik:
Siberian Yupik Eskimo
Guahiban:
Guahibo
Hokan Washo:
<table>
<thead>
<tr>
<th>Language Type</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indo-European Albanian:</td>
<td>Albanian</td>
</tr>
<tr>
<td>Indo-European Balto-Slavic Baltic:</td>
<td>Latvian, Lithuanian</td>
</tr>
<tr>
<td>Indo-European Balto-Slavic Slavic East Slavic:</td>
<td>Russian, Ukrainian, Old Church Slavonic, Bulgarian</td>
</tr>
<tr>
<td>Indo-European Balto-Slavic Slavic South Slavic:</td>
<td>Croatian, Czech</td>
</tr>
<tr>
<td>Indo-European Balto-Slavic Slavic West Slavic:</td>
<td>Polish, Slovak, Czech</td>
</tr>
<tr>
<td>Indo-European Celtic Insular Goidelic:</td>
<td>Irish</td>
</tr>
<tr>
<td>Indo-European Germanic North-West Germanic:</td>
<td>English, German, Dutch</td>
</tr>
<tr>
<td>Indo-European Germanic Scandinavian:</td>
<td>Icelandic, Bokmål Norwegian, Swedish</td>
</tr>
<tr>
<td>Indo-European Hellenic Attic-Ionic:</td>
<td>Classical Greek</td>
</tr>
<tr>
<td>Indo-European Hellenic Demotic:</td>
<td>Modern Greek</td>
</tr>
<tr>
<td>Indo-European Indo-Iranian Indo-Aryan Central zone Rajasthani:</td>
<td>Unclassified: Lamani, Indo-European Indo-Iranian Indo-Aryan Central zone Western Hindi Hindustani: Hindi</td>
</tr>
<tr>
<td>Indo-European Indo-Iranian Indo-Aryan Northwestern zone Dardic Chitral:</td>
<td>Khowar</td>
</tr>
</tbody>
</table>
Properties of Lower Numerals and their Explanation: A Reply to Paweł Rutkowski

**Indo-European Indo-Iranian Indo-Aryan:**
- Sanskrit

**Indo-European Indo-Iranian Iranian Eastern Southeastern Pashto:**
- Pashto

**Indo-European Indo-Iranian Iranian Western Northwestern Kurdish:**
- Sorani (South Kurdish)

**Indo-European Italic Romance:**
- Swiss French
- French
- Italian
- Portuguese
- Romanian
- Spanish
- Catalan

**Isolate:**
- Nivkh

**Isolate/Unclassified:**
- Kwaza

**Japanese:**
- Japanese

**Korean:**
- Korean

**Mongolian Eastern Dagur:**
- Dagur

**Mongolian Eastern Oirat-Khalkha Khalkha-Buriat Mongolian Proper:**
- Mongolian

**Na-Dene Nuclear Na-Dene Athapaskan-Eyak Athapaskan Canadian Hare-Chipewyan Hare-Slavey:**
- Slave (Bearlake variety)

**Niger-Congo Atlantic-Congo Atlantic Northern Senegambian Fula-Wolof Fulani Eastern:**
- Fulfulde (Adamawa Fulani)

**Niger-Congo Atlantic-Congo Volta-Congo Benue-Congo Bantoid Southern Narrow Bantu Central D Lega-Kalanga (D.20):**
- Kwami

**Niger-Congo Atlantic-Congo Volta-Congo Benue-Congo Bantoid Southern Narrow Bantu Central R South Mbundu (R.10):**
- Quimbundo

**Niger-Congo Atlantic-Congo Volta-Congo Benue-Congo Bantoid**
Southern Narrow Bantu Central Swahili:
Swahili
Nilo-Saharan Eastern Sudanic Nilotic Eastern Lotuxo-Teso
Lotuxo-Maa Lotuxo:
Lotuxo
Nilo-Saharan Saharan Western Tebu:
Tuda-Daza
North Caucasian Northwest Circassian:
Kabardian
North Khoisan:
ǃXu
Quechuan Quechua:
Quechua
Sino-Tibetan Sinitic:
Mandarin Chinese
Sino-Tibetan Tibeto-Burman Himalayish Mahakiranti Kiranti
Eastern:
Kulung
Sino-Tibetan Tibeto-Burman Himalayish Tibeto-Kanauri Tibetic
Tibetan Central:
Tibetan
Sino-Tibetan Tibeto-Burman Lolo-Burmese Burmish Northern:
Zaiwa
Sino-Tibetan Tibeto-Burman Lolo-Burmese Loloish Northern Yi:
Lalo
Totonacan:
Totonac
Tupi Tupi-Guarani Guarani(I):
Guarani
Turkic Southern:
Turkish
Uralic Finno-Ugric Finno-Permich Finno-Cheremisic Finno-Mordvinic
Finno-Lappic Baltic-Finnic Finnic:
Finnish
Uralic Finno-Ugric Finno-Permich Finno-Cheremisic Finno-Mordvinic
Finno-Lappic Baltic-Finnic:
Votic
Uralic Finno-Ugric Ugric:
Hungarian
Uralic Samoyedic Northern Samoyedic:
Nenets (Yurak Samoyed)

Uto-Aztecan Northern Uto-Aztecan Numic Southern:
Kawaiisu

West Papuan Bird’s Head North-Central Bird’s Head Central Bird’s Head:
Maybrat

West Papuan Bird’s Head North-Central Bird’s Head North Bird’s Head:
Abun

Yukaghir:
Kolyma Yukaghir