How numerical communication reflects cognition: A corpus-based analysis

Greg Woodin, Bodo Winter, Jeannette Littlemore, Marcus Perlman & Jack Grieve
University of Birmingham, gawoodin@gmail.com

Keywords: number frequencies, prototype theory, big data, corpus linguistics, vague language

Numbers are typically characterised as precise and objective. At least at first sight, mathematics is an area of knowledge where classical theories of categorization in terms of necessary and sufficient conditions might still apply, given that there seem to be hard and clearly definable boundaries between different mathematical categories, like integers and decimals (but see Armstrong, Gleitman & Gleitman 1983). However, in this paper, we show that the way people use numbers is vague and subjective. Using Bayesian models, we analyse over 1.7 million occurrences of numbers between 0 and a billion in the 100 million word British National Corpus (BNC Consortium 2007). We find that, rather than numbers exactly quantifying mathematical properties of the world around us, certain numbers are used more frequently than others based on their magnitude and roundness, which may reflect our cognitive processing and representation of these numbers.

First, we replicate the finding that smaller numbers are used more frequently than larger numbers (Dehaene & Mehler 1992; Dorogovtsev, Mendes & Oliveira 2005). This pattern is believed to reflect the fact that our 'mental number line' is logarithmically scaled (Dehaene 1992), and that smaller numbers are easier to mentally process (Dehaene & Mehler 1992). Second, in a model that controls for this small number bias, we show that round numbers are overrepresented. Like small numbers, round numbers are psychologically salient (Van der Henst & Sperber 2004) and more cognitively accessible (Cummins 2015: 32). Third, we find that word frequency distributions reflect the fact that people generally round larger numbers to a greater extent than smaller numbers – for example, rounding 86 up to 90 (the nearest multiple of ten) but rounding 186 up to 200 (the nearest multiple of 100). This finding aligns with the idea that higher numbers have more approximate and fuzzy mental representations than smaller numbers (DeWind et al. 2015; Shepard, Kilpatric & Cunningham 1975). Fourth, we show that round numbers are not created equal, and that 'roundness' can hence be productively viewed as a prototype category with graded membership (see Rosch 1973). Specifically, 10-ness (10, 20, 30, ... 100, 200, 300, ...), 2-ness (20, 40, 60, ... 200, 400, 600, ...), 2.5-ness (25, 50, 75, ... 250, 500, 750, ...), and 5-ness (50, 100, 150, ... 500, 1000, 1500 ...) are features of roundness (Jansen & Pollmann 2001), and numbers with more of these roundness properties are 'rounder' than others (e.g., 100 versus 25), which increases their frequency in the BNC.

Altogether, our results demonstrate that language use about numerical information broadly seems to mimic the way people think about numbers, suggesting deep links between numerical cognition and communication that bear an imprint on word frequency statistics observed in corpora.

References

Armstrong, S. L., L. R. Gleitman & H. Gleitman. 1983. What some concepts might not be. Cognition 13(3). 263–308. https://doi.org/10.1016/0010-0277(83)90012-4.

BNC Consortium. 2007. British National Corpus. University of Oxford, Oxford: BNC Consortium. http://www.natcorp.ox.ac.uk/. (13 July, 2019).

Cummins, Chris. 2015. Constraints on Numerical Expressions (Oxford Studies in Semantics and Pragmatics). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199687909.001.0001.

Dehaene, S. 1992. Varieties of numerical abilities. Cognition 44(1–2). 1–42. https://doi.org/10.1016/0010-0277(92)90049-n.

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

DeWind, Nicholas K., Geoffrey K. Adams, Michael L. Platt & Elizabeth M. Brannon. 2015. Modeling the approximate number system to quantify the contribution of visual stimulus features. Cognition 142. 247–265. https://doi.org/10.1016/j.cognition.2015.05.016.

Dorogovtsev, Sergey, José Fernando Mendes & Joao Oliveira. 2005. Frequency of occurrence of numbers in the World Wide Web. Physica A: Statistical Mechanics and its Applications 360. 548–556. https://doi.org/10.1016/j.physa.2005.06.064.

Jansen, C. J. M. & M. M. W. Pollmann. 2001. On round numbers: Pragmatic aspects of numerical expressions. Journal of Quantitative Linguistics. Routledge 8(3). 187–201. https://doi.org/10.1076/jqul.8.3.187.4095.

Rosch, E. H. 1973. Natural categories. Cognitive Psychology. Netherlands: Elsevier Science 4(3). 328–350. https://doi.org/10.1016/0010-0285(73)90017-0.

Shepard, Roger N., Dan W. Kilpatric & James P. Cunningham. 1975. The internal representation of numbers. Cognitive Psychology. Netherlands: Elsevier Science 7(1). 82–138. https://doi.org/10.1016/0010-0285(75)90006-7.

Van der Henst, Jean-Baptiste & Dan Sperber. 2004. Testing the cognitive and communicative principles of relevance. In Ira A. Noveck & Dan Sperber (eds.), Experimental Pragmatics (Palgrave Studies in Pragmatics, Language and Cognition), 141–171. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230524125_7.