Language and Cognition (2026), 18, e42, 1–25
doi:10.1017/langcog.2026.10090

ARTICLE

Approximate round number use depends on magnitude‡

Greg Woodin[1], Alexandra Lorson[2] and Bodo Winter[3]

[1]Experimental Psychology, UCL, UK; [2]Semantics and Cognition, University of Groningen, Netherlands and [3]Linguistics and Communication, University of Birmingham, UK
Corresponding author: Greg Woodin; Email: gawoodin@gmail.com

(Received 12 November 2025 UTC; Revised 07 May 2026 UTC; Accepted 08 May 2026 UTC)

Abstract
Numbers have mathematically defined, formal meanings, which afford precision and objectivity. In practice, however, we show that round numbers are frequently used approximately, and this approximate use depends on magnitude. Across three analyses of British and American English, we demonstrate the following: (1) people use round numbers more often, and round to a greater extent, at higher magnitudes; (2) the distributional semantics of larger round numbers resemble those of indefinite hyperbolic numbers such as 'gazillion', which lack a precise value; and (3) larger jigsaw puzzles have greater discrepancies between round advertised piece counts and the actual values (e.g., advertising '13,200 pieces' when the actual count is 13,224). We argue that the relationship between roundness, magnitude and approximation derives from the base-10 structure of English numerals, which renders powers of 10 structurally salient, creating hotspots for communicative functions such as approximation. We discuss how these communicative patterns align with the approximate number system, which cognitively represents larger quantities with increasing imprecision, making them harder to estimate and compare.

Keywords: ANS; corpus linguistics; fictitious numbers; hyperbole; mathematical cognition; numerical cognition; vagueness; word2vec

1. Introduction

If someone states that 'a hundred people' attended an event, as in (1), they may not intend this number as a precise headcount of exactly 100 individuals.

(1) There were a hundred people at that ceremony watching me freeze on that stage. (COCA, TV, *Station 19*, 2019)

‡This article has been updated since original publication. A notice detailing the change has been published.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

In such cases, the round number 'hundred' is often used as an estimate; the actual number could very well have been 98 or 103. This tolerance for nearby numbers reflects what Lasersohn (1999) calls a *pragmatic halo*, a set of values that are acceptable substitutes for a number. By comparison, non-round or 'sharp' numbers tend to be used exactly. Swapping the 'hundred' in (1) for 'ninety-seven', as in (2), strongly suggests that the speaker is specifying an exact headcount.

(2) There were ninety-seven people at that ceremony watching me freeze on that stage.

The difference between (1) and (2) highlights the link between round numbers and approximation (Krifka, 2007, 2009). Another factor driving the approximate use of numbers is their magnitude, as highlighted by the contrast between (3) and the modified example (4).

(3) It was a group of ten people. (COCA, blog, auliffe.blogspot.com, 2012)
(4) It was a group of a million people.

In (3), 'ten people' could be used either precisely or approximately – perhaps it was actually a group of nine or eleven people – while in (4), an imprecise usage is much more probable. Larger round numbers, particularly ones in the thousands or above, often function more like approximate markers of scale than specific quantities. This vague usage is especially clear in hyperbolic statements like (5).

(5) A trillion things happen to a billion people. No one is ever the first one to be anything (...) (COCA, blog, junctivecollapse.com, 2012)

Swapping 'trillion' and 'billion' in (5) with overtly vague expressions, such as what are known as *indefinite hyperbolic numerals* (e.g., 'gazillion' and 'bazillion') (Chrisomalis, 2016; Lavric, 2010), does little to change the numerical meaning of the original statement, as demonstrated by (6).

(6) A gazillion things happen to a bazillion people. No one is ever the first one to be anything (...)

In this paper, we present a series of quantitative analyses that investigate the link between roundness, approximation and magnitude. We find that people use round numbers more often, and round to a greater degree, at higher magnitudes and that larger round numbers like 'million' and 'billion' are used in similar linguistic contexts to indefinite hyperbolic numerals like 'squillion' and 'bajillion'. We also analyse a dataset of jigsaw puzzles that affords calculating the pragmatic halo around numbers directly, thereby demonstrating that larger numbers have wider pragmatic halos.

Theoretically, we propose that this interaction between rounding, magnitude and approximation directly results from the base-10 structure of English numerals: within a base-10 system of representing numbers, certain numbers stand out from others by virtue of occupying salient points on the number line, and these salient points are used for special communicative functions, including approximation. We discuss these ideas in light of the approximate number system (ANS), the cognitive system that is thought to underlie our 'number sense' (Dehaene, 1997). Within the

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

ANS, large-magnitude numbers are cognitively represented in a more approximate fashion than small-magnitude numbers (DeWind et al., 2015; Shepard et al., 1975a). This magnitude-dependent component of numerical cognition aligns with our findings that the use of numbers becomes more approximate at higher magnitudes.

2. Background

2.1. Round numbers are special

From a mathematical perspective, the number 1,000 is no more special than any other number. It is simply another point on an infinite number line – one integer greater than 999 and one fewer than 1,001. However, in practice, we never engage with mathematical abstractions directly; rather, we interact with symbolic representations of numbers (Zhang & Norman, 1995), whether as numerals (e.g., 1,000) or number words (e.g., 'thousand'). Within representational systems, some numbers stand out by virtue of their relation to the system's structure – specifically, numbers that are directly related to the base are 'special'. For example, in the Babylonian sexagesimal (base 60) system, numbers mathematically related to 60 – such as 3,600 or 216,000 – were salient (cf. Chrisomalis, 2020; Ch. 1). By contrast, in the widely used Arabic or 'Western' (Chrisomalis, 2020) numeral system, the base is 10, and numbers related to this base stand out accordingly. The numeral 1,000 is the first four-digit number and consists of a single leading digit followed by three zeros, thereby marking a transition to a new order of magnitude. This structural distinctiveness is mirrored by the English number word system, where 1,000 is designated a succinct lexical item, 'thousand', in contrast to its more grammatically complex neighbours 'nine hundred and ninety-nine' and 'one thousand and one' (Hurford, 1975).

Precisely because the decimal system is structured around the number 10, numbers that are directly related to this base stand out, including the powers of 10, such as 100 (hundred), 1,000,000 (million) and 1,000,000,000 (billion), as well as numbers derived from these powers through multiplication and division (e.g., 20, 500, 7,500). In this paper, we use the term *round number* to refer to any number that is structurally salient within a given representational system by virtue of it being directly related to the base, its powers, or multiples and divisors of the base and its powers. This definition entails that the notion of 'roundness' is graded because numbers differ in the number of relations they have with the base and its powers. For example, the number 310 is round because it is a multiple of 10, but 300 is even rounder because it is both a multiple of 10 and a multiple of 10^2 (3×100). Furthermore, because Pollmann and Jansen (1996) argue that humans have a natural propensity for doubling and halving, the number 500 can be thought of as even rounder than 300: both are multiples of 10 and 10^2, but 500 is also half of 10^3 (see also Sigurd, 1988).

Pollmann and Jansen (1996) proposed what they call 'the principle of favourite numbers' (p. 225), according to which certain numbers are preferred in approximate expressions. This proposal was based on an analysis of Dutch number range expressions (compare English 'ten to fifteen' or 'about ten or fifteen'), where they observed that only particular combinations of numbers were acceptable, specifically those related to the base through simple operations such as doubling and halving. These observations were later extended by Jansen and Pollmann (2001), who showed that numbers conforming to this pattern are also overrepresented across Dutch, English, German and French corpus data. On this basis, they refined the principle of favourite

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

Table 1. Roundness properties adapted from Jansen and Pollmann (2001)

Property	Definition of set	Example numbers
10-ness	$1 \times (k \times 10^n)$, where $k \in \{1, \ldots, 9\}$	80, 300, 3,000
2-ness	$2 \times (k \times 10^n)$, where $k \in \{1, \ldots, 9\}$	60, 200, 1,400
2.5-ness	$2.5 \times (k \times 10^n)$, where $k \in \{1, \ldots, 9\}$	50, 125, 50,000
5-ness	$5 \times (k \times 10^n)$, where $k \in \{1, \ldots, 9\}$	100, 4,500, 350,000

numbers based on four numerical properties: 10-ness,[1] 2-ness, 5-ness and 2.5-ness, as shown in Table 1. A number that falls within any of these sets is round, and the more roundness properties a number has, the rounder it is (Jansen & Pollmann, 2001; see also Woodin et al., 2024). For example, the number 500 is characterised by 10-ness ($1 \times 5 \times 10^2 = 500$), 5-ness ($5 \times 1 \times 10^2 = 500$) and 2.5-ness ($2.5 \times 2 \times 10^2 = 500$) and is thus rounder than 300, which is only characterised by 10-ness ($1 \times 3 \times 10^2 = 300$). These roundness properties provide a more precise operationalisation of the idea that numbers related to the base and its powers are special.

2.2. Functions of round numbers

Due to their structural and psychological salience, round numbers serve diverse functions beyond precisely indexing numerical values. We broadly differentiate between *approximate* and *non-approximate* functions (Table 2). On the approximate side, round numbers are used to 'round' values when the exact value is deemed communicatively irrelevant, such as when rounding 3:08 to 3:10 when stating the time (Gibbs Jr. & Bryant, 2008; Van der Henst & Sperber, 2004; Van der Henst et al.,

Table 2. Diverse approximate and non-approximate functions of round numbers

Function	Use	Example references
Approximate	Knowing the exact value, but choosing to report a rounded value for considerations of communicative relevance	Gibbs Jr. and Bryant (2008), Van der Henst et al. (2002), Van der Henst and Sperber (2004)
Approximate	Knowing the exact value, but choosing to report a rounded value to not appear pedantic	Beltrama et al. (2023)
Approximate	Not knowing the exact value, and reporting a rounded value due to uncertainty	Ruud et al. (2014)
Approximate	Numerical hyperbole	Kao et al. (2014), Lavric (2010)
Non-approximate	Benchmarks, goal posts	Pope and Simonsohn (2011), Ozcan et al. (2024)
Non-approximate	Structuring cultural artefacts and customs	Cummins (2015, p. 36), Jansen and Pollman (2001)
Non-quantitative	Naming and branding	Gunasti and Ozcan (2016)

Note: This is not intended to be an exhaustive list.

[1]Based on Jansen and Pollman's (2001) naming convention for the other roundness properties, one might expect 10-ness to be termed 1-ness, since 1 is the number multiplied by a power of 10 in the definition of 10-ness, just as 2.5 is the number multiplied by a power of 10 in 2.5-ness. However, for consistency with their terminology, we continue to use 10-ness in this paper.

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

2002), or when reporting the exact figure would seem pedantic (Beltrama et al., 2023). Round numbers are also used approximately when the language user is uncertain of the precise value but wants to convey a general sense of magnitude, as when estimating that 'about 1 million' people live in a city (Ruud et al., 2014). This function of round numbers, reflecting lack of knowledge, is also demonstrated by economics studies that investigate the phenomenon of 'age heaping': the over-representation of multiples of 5 in historical census data suggests that, in the past, people may have reported round numbers because they did not know their exact age (A'Hearn et al., 2009; Turner, 1958). Another approximate function of round numbers is numerical hyperbole, as in phrases like 'I've told you a million times', where 1 million represents a large, unspecified quantity (Kao et al., 2014), or in words like 'millipede', which literally means 'a thousand feet' in Latin, even though all known species had far fewer than 1,000 legs (Lavric, 2010) until the discovery of *Eumillipes Persephone* in 2020 (Marek et al., 2021).

As for non-approximate functions, round numbers can be used as benchmarks or reference points for goals (Ozcan et al., 2024; Pope & Simonsohn, 2011), such as when dieters set themselves a weight loss goal of 50 lb, rather than 51.7 lb. The round number in this instance is a target, and the actual number is evaluated relative to this target. Round numbers are also used to structure cultural artefacts and customs: items are often sold in multiples of 10 (Cummins, 2015, p. 36), and jubilee celebrations happen every 5 or 10 years (Jansen & Pollmann, 2001). Used for this function, round numbers actively *determine* the quantity of items, or the regularity of events, with people organising both the physical world and the temporal structure of their lives around such numbers. Moreover, round numbers are used in contexts where they do not denote quantities at all, but are used in a nominal sense (see Woodin & Winter, 2024), such as when naming products (e.g., Sony DSC-RX100 camera), fictional characters (e.g., T-1000 in the film *Terminator 2: Judgment Day*) and artefacts within fictional universes (e.g., Nimbus 2000 in *Harry Potter*). Such naming choices may be partially motivated by the perception of round numbers as 'whole' and 'complete' (Gunasti & Ozcan, 2016).

These diverse functions of round numbers may explain why round numbers are overrepresented in corpus data (Coupland, 2011; Dehaene & Mehler, 1992; Dorogovtsev et al., 2006; Jansen & Pollmann, 2001; Woodin & Winter, 2024; Woodin et al., 2024). Indeed, the fact that round numbers are used and thus encountered so frequently may reinforce their psychological salience, creating a feedback loop that encourages yet more frequent use (Chrisomalis, 2020, p. 35; Cummins, 2015, p. 36). Frequency data thus provide further evidence for the privileged status of round numbers in the decimal system. However, frequencies alone do not allow us to directly reverse-engineer which of the functions summarised in Table 2 is driving the increased use of these numbers. In other words, a number like 1,000 could be more frequent than nearby numbers such as 986, 998 or 1,112 due to any one or a combination of these functions, whether approximate or non-approximate. In a balanced corpus, all of these functions are likely to be in evidence, albeit not necessarily equally often.

2.3. Rounding and magnitude

Crucially, rounding is intrinsically linked to magnitude, both structurally and cognitively. From a structural perspective, the fact that the decimal system is organised

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

around powers of 10 (e.g., 10, 100, 1,000) means that the roundest numbers – those most closely tied to these powers – also become more widely spaced at higher magnitudes. This structural property of the decimal system gives rise to a natural 'fanning out' of round numbers: while 10 and 100 differ by 90, 100 and 1,000 differ by 900 and so on. Numbers derived from these powers via doubling and halving (e.g., 50, 500, 2,500) exhibit the same fanning pattern. Therefore, roundness is not only graded (see Section 2.1) but also inherently scale-dependent.

This structural property has direct implications for approximation. As discussed above (Section 2.2), when language users approximate, they tend to use round numbers (Ruud et al., 2014; Van der Henst et al., 2002). Since there are fewer round numbers at higher magnitudes, the degree of rounding also tends to increase. For instance, people may round 56 to 60, a difference of 4, but round 556 to 600, a difference of 44. When rounding 556, the number 560 is also available as a target for rounding, but it lacks the 10-ness ($1 \times 6 \times 10^2 = 600$) and 2-ness ($2 \times 3 \times 10^2 = 600$) of 600. From this perspective, round numbers can be seen as being in competition with each other, with rounder numbers being selected more frequently than less round numbers. This competition can be understood via the concept of a *pragmatic halo* (Lasersohn, 1999), where numbers can stand for a range of values beyond their exact mathematical meaning (Cummins, 2015; Krifka, 2009). At higher magnitudes, the pragmatic halo of round numbers widens: 600 could plausibly be used to mean 556, but 100 is unlikely to mean 56.

In corpus data, the idea that language users produce round numbers with an intended pragmatic halo that widens with magnitude leads to the expectation that round numbers absorb the frequencies of an increasingly larger set of numbers at higher magnitudes. This notion must be understood relative to a broader and well-established trend, which is that number frequencies tend to *decrease* with magnitude (e.g., Coupland, 2011; Dehaene & Mehler, 1992): people more often refer to relatively small numbers. Round numbers do not overturn this trend (1,000 is less frequent than 100) but rather create local peaks against it, becoming more frequent at higher magnitudes relative to the surrounding sharp numbers (Woodin et al., 2024).

These patterns observed for language use align closely with findings from numerical cognition. When humans judge or compare numbers above three to four items, they rely on the *approximate number system* (ANS), which represents quantities in an analogue fashion (Cheyette & Piantadosi, 2019; McCrink & Wynn, 2007; Xu & Spelke, 2000), with increasingly approximate representations at higher magnitudes (DeWind et al., 2015; Shepard et al., 1975a). This increasing imprecision is described by the Weber-Fechner law (e.g., Dehaene, 2003; DeWind et al., 2015; Shepard et al., 1975b), which maintains that people's ability to discriminate quantities is linearly related to their ratio, rather than their absolute difference. This cognitive phenomenon provides another explanation, compatible with the notion of a pragmatic halo (Lasersohn, 1999), for why people might preferentially round 56 to 60, but round 556 to 600: both rounding operations involve similar ratios and hence proportional differences (6.7% vs. 7.3%).

2.4. Overview of analyses

In this paper, we triangulate evidence from three analyses that shed light on the relationship between round number use, approximation and magnitude. Table 3

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

Table 3. Overview of the three analyses in this paper, including the datasets used and the research questions addressed

Analysis	Datasets	Research questions
1	British National Corpus (BNC Consortium, 2007), Corpus of Contemporary American English (Davies, 2008)	Do people round more often, and to a greater extent, at higher magnitudes?
2	'word2vec' word embeddings (Mikolov et al., 2013) pre-trained on the Wikipedia corpus (Mikolov et al., 2017)	Are the distributional semantics of larger round numbers more similar to those of indefinite hyperbolic numerals than smaller round and non-round numbers?
3	Jigsaw puzzle dataset	Are differences between the stated and actual number of puzzle pieces greater for larger puzzles?

presents an overview of each analysis. The first analysis is a replication-extension of Woodin et al. (2024), who discussed the fact that in data from the British National Corpus, larger round numbers are *relatively* more frequent than smaller round numbers compared to the overall magnitude-dependent decline in frequencies (see Dehaene & Mehler, 1992). This pattern is a key prediction of an approximation-based perspective, but it was only discussed descriptively in Woodin et al. (2024). Analysis 1 provides a statistical model to formally test the relationship between increased frequency, rounding and magnitude with the same British National Corpus data. It also replicates the pattern for another corpus, the Corpus of Contemporary American English.

Analysis 2 uses a distributional semantics approach to compare the use of round numbers such as 'thousand' and 'million' with indefinite hyperbolic numbers such as 'umpteen' and 'squillion' (Chrisomalis, 2016; see also Burnham, 1951; Pound, 1949). Distributional semantics represents word meanings as vectors derived from their usage contexts: words with more similar vectors occur in more similar contexts and are, therefore, likely to have more similar meanings (Lenci, 2008; Lenci & Sahlgren, 2023), in line with Firth's (1957, p. 179) assertion that 'you shall know a word by the company it keeps'. Using this methodology, inspired by Ariel and Levshina's (2025) distributional semantic analysis of the numerals 2–99, we compare the distributional properties of round and non-round numbers of different magnitudes to those of indefinite hyperbolic numbers. We use these pseudo-numbers for comparison because they are unambiguously inexact – 'bazillion', for example, cannot be assigned a precise location on the number line (Rips, 2013). Vector similarity between certain large round numbers (e.g., 'million', 'billion') and indefinite hyperbolic numerals would suggest a shared tendency towards imprecise use.

The final analysis takes a more direct approach to quantifying the extent of rounding by investigating how round numbers are used on the packaging of jigsaw puzzles. Many puzzles advertise neatly round figures such as 500 or 1,000 pieces, yet they do not actually contain the exact number stated on the packaging. In contrast with the corpus analyses (Analyses 1 and 2), where the true value underlying a round expression is unknown, jigsaw puzzles provide a 'ground truth' in the form of the actual piece count, which allows direct comparison with the advertised round number. This comparison makes it possible to assess how far the exact value deviates from the round one. Here, too, we demonstrate that magnitude matters: larger puzzles tend to deviate more from their advertised counts.

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

3. Modelling roundness and magnitude

In this first analysis, we replicate and extend the work of Woodin et al. (2024), who reported that round numbers occur disproportionately often at higher magnitudes, which they interpreted as being consistent with increased frequency and degree (i.e., coarseness) of rounding. However, Woodin et al. (2024) made this observation descriptively. Here, we test this pattern formally in two corpora, the British National Corpus and the Corpus of Contemporary American English. Specifically, we assess whether the interaction between roundness properties and magnitude predicts number frequency.

3.1. Methodology

All data, analysis scripts and additional information about this study can be found at https://osf.io/ct7f8.

3.1.1. Corpora

We analysed numbers in two corpora: the British National Corpus (BNC; BNC Consortium, 2007) and the Corpus of Contemporary American English (COCA; Davies, 2008). Both corpora sample a range of texts and genres, intending to be generalisable across different types of language use in their respective varieties of English.

The BNC comprises 100 million words of written and spoken British English compiled during 1991–1994 and revised for rerelease in both 2001 and 2007. Written texts comprise 90% of the BNC and include both fiction and non-fiction writing, including extracts from newspapers, academic books and popular novels. Spoken texts comprise 10% of the BNC and include unscripted informal conversations in addition to more formal contexts, such as radio interviews, university lectures and government meetings.

COCA, a corpus of written and spoken American English, was released in 2008 and has received annual updates since. The version of the corpus to which we had access comprises 440 million words and was released in 2014. COCA is divided roughly equally – 20% in each subcorpus – between unscripted conversations on TV and radio programmes, fiction writing, popular magazines, newspapers and academic journals. In contrast with the BNC, COCA does not include an informal conversation register.

3.1.2. Software

To extract numbers from the BNC and COCA, we used Python (version 3.7; Python Software Foundation, 2021) in the PyCharm-integrated development environment (version 2022.2.3; JetBrains, 2022). Extraction involved the built-in libraries Re (version 2.2.1), IterTools, OS and Time (all by Van Rossum, 2020), along with the external libraries NLTK (version 3.7; Bird et al., 2009), Pandas (version 1.1.5; McKinney, 2010), Word2Number (version 1.1; Batorsky et al., 2021) and Num2-Words (version 0.5.12; Dupras et al., 2021). To assist with number extraction and perform the main statistical analyses, we used the R programming language (version 4.3.2; R Core Team, 2023) in the RStudio integrated development environment (version 2023.6.0.421; Posit Team, 2023), with the packages 'tidyverse' (version 2.0.0; Wickham, 2019), 'car' (version 3.0.11; Fox & Weisberg, 2019), 'brms' (Bürkner, 2017, 2018), 'ggmcmc' (version 1.5.1.1; Fernández-i-Marín, 2016),

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

'ggridges' (version 0.5.4; Wilke, 2022), 'ggpubr' (version 0.6.0; Kassambara, 2023), 'tidybayes' (version 3.0.6; Kay, 2023), 'patchwork' (version 1.3.0; Pedersen, 2024), 'fmsb' (version 0.7.6; Nakazawa, 2024), 'png' (version 0.1.8; Urbanek, 2022), 'magick' (version 2.8.5; Ooms, 2024)', 'lsa' (version 0.73.3; Wild, 2022) and 'usedist' (version 0.4.0; Bittinger, 2020).

3.1.3. Number extraction

We extracted numbers from the BNC and COCA using the procedures described in Woodin et al. (2024) and Woodin and Winter (2024), respectively. These procedures aimed to capture numerical expressions from 0 to 1 billion that were transcribed (speech) or written (writing) as numerals (e.g., '42', '1.14'), number words (e.g., 'forty-two', 'one point one four') or a mix of both (e.g., '500 million', '7.5 thousand'). We extracted morphologically unmarked numbers only (e.g., '3' and 'three', but not '3rd' or 'third').

We parsed and part-of-speech tagged each corpus using 'NLTK' (version 3.7; Bird et al., 2009) to identify strings tagged as numbers. Previous studies employed a top-down approach to identify numbers, searching specifically for '20', 'seven hundred', etc. (Coupland, 2011; Dehaene & Mehler, 1992; Dorogovtsev et al., 2006; Jansen & Pollmann, 2001). In comparison, our approach was bottom-up: we tagged each individual number string in both corpora. We then treated adjacent numerical strings as a single unit, also incorporating the words 'and' and 'point', to capture multi-word expressions (e.g., 'one hundred and two', 'two point seven'). This bottom-up method allowed us to capture a greater variety of numerical language. However, it also necessitated more data cleaning: we excluded 146,797 strings from the BNC and 627,596 strings from COCA that contained a mixture of numerals and letters or other characters, including morphologically marked numbers (e.g., '10th'), decade ranges (e.g., '1970s') and digital times (e.g., '06:00'). We also excluded 9,407 strings from the BNC and 11,010 strings from COCA with a leading 0 that resembled numerical codes (e.g., '01760'). For more information about these procedures and all data exclusions, see the OSF repository associated with this paper (https://osf.io/ct7f8) and the descriptions for the BNC in Woodin et al. (2024) and for COCA in Woodin and Winter (2024).

Overall, these procedures identified 1,739,398 occurrences of numbers in the BNC[2] and 8,158,113 occurrences of numbers in COCA that met our search criteria. These occurrences comprised the supersets from which we extracted subsets for statistical modelling.

3.1.4. Statistical models

To model the effect of magnitude on round number frequencies, we implemented Bayesian generalised models using the package 'brms' (Bürkner, 2017, 2018). Two parallel models were fitted, one for each corpus (BNC and COCA). Frequency was the dependent variable, with the predictors Log_{10} Magnitude (continuous) and the roundness properties identified by Jansen and Pollmann (2001) (10-ness, 2-ness, 2.5-

[2]Note that this number of occurrences is slightly higher than the 1,739,343 occurrences reported in Woodin et al. (2024), reflecting a minor amendment to the number identification procedure that was implemented after the reported figure was calculated.

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

ness and 5-ness), in addition to the properties Multiple of 5 and Multiple of 10, which were also found to predict number frequencies by Woodin et al. (2024). Each roundness property was implemented as a categorical variable, where 'does not have property' is the reference level and 'has property' is the comparison level. In contrast to the analysis of the BNC reported in Woodin et al. (2024), the model fitted here includes terms for the interactions between all roundness properties and magnitude, which is replicated for the novel analysis of COCA.

Number frequencies are unbounded count data (i.e., positive integers with no upper limit), for which Poisson or negative binomial regressions are candidate model types. The latter, negative binomial, is preferred in the case of possible overdispersion, when counts exhibit more variance than what would be expected under the Poisson distribution. For example, some specific numbers may occur hundreds of thousands of times, while others occur just once or not at all. A check of the data indicated high overdispersion (see OSF repository: https://osf.io/ct7f8). We, therefore, fitted Bayesian negative binomial regressions (see Winter & Bürkner, 2021) which, unlike Poisson regressions, estimate both the mean and variance in counts.

We modelled number frequencies for integers between 1 and 1 million, excluding decimals (BNC: 70,197 occurrences, COCA: 300,864 occurrences) as well as numbers outside this range (BNC: 24,873 occurrences, COCA: 360,677 occurrences). Zero was excluded as the model included the predictor Log_{10} Magnitude, and the logarithm of 0 is undefined. Imposing an upper limit of 1 million mitigated the computational demands of fitting Bayesian models, which take more time with increasing dataset size: the dataset per corpus contained 1 million rows, including numbers that are used in each corpus (frequency ≥ 1) in addition to those that are not (frequency = 0).

As negative binomial regressions use the log link function, they model the natural logarithm of the mean frequency. Therefore, the slope estimates reported below reflect changes in log-transformed number frequencies ($\log_e$). We used weakly informative priors (normal distribution centred on 0, SD = 0.5) for these slope coefficients. These priors mildly bias coefficient estimates towards zero, especially when there are little data, which makes estimates more conservative, thereby incorporating 'mild skepticism' (McElreath, 2016, p. 186) into the model. For all other parameters, we used default priors from the R package 'brms' (Bürkner 2017, 2018).

3.2. Results

Figure 1 displays the posterior distributions for interactions between the different roundness properties and Log_{10} Magnitude (interaction coefficients). Numerical values and 95% credible intervals of slope estimates are reported in Table 4. All slopes are positive, which indicates that round numbers are used more frequently at higher magnitudes in both the BNC and COCA. The lowest slope estimate is for Multiple of 5, which has a 95% credible interval that is only marginally above 0 in the BNC, and includes 0 in COCA, indicating considerable uncertainty. For all interaction terms other than Multiple of 5, the 95% credible intervals are well above 0 in both corpora, increasing in the following order: Multiple of 10, 2-ness, 5-ness, 2.5-ness and 10-ness.

To make the pattern captured by the models more concrete, we can examine the raw frequencies of powers of 10 across magnitudes. In the BNC, for example, the frequency of 100 is 14,819, compared to only 6,975 for 1,000, reflecting the general

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

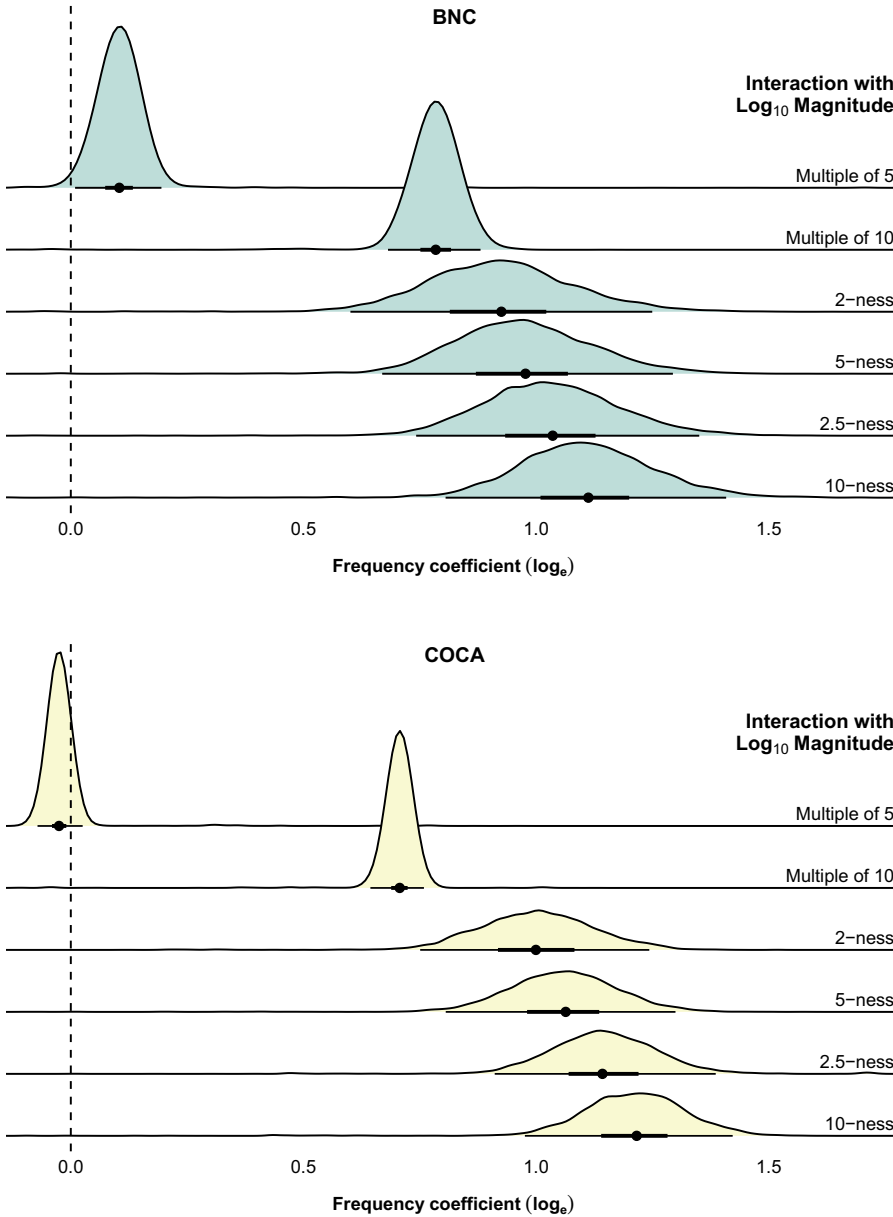


Figure 1. Posterior distributions of $\log_e$ coefficients for interactions between Log_{10} Magnitude and roundness properties Multiple of 5, Multiple of 10, 2-ness, 5-ness, 2.5-ness and 10-ness in the BNC (top row) and COCA (bottom row). Dots indicate slope estimates (posterior means), thick error bars indicate 50% credible intervals, and thin error bars indicate 95% credible intervals.

trend for absolute frequencies to decline with magnitude (see Section 2.3). We can also compare these frequencies to those of numbers within ±10% of each power – 90 and 110 for 100 and 900 and 1,100 for 1,000 – as a rough approximation of the

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

Table 4. Slope estimates (β) and 95% credible intervals (95% CI) for interactions between roundness properties and $\log_{10}$ magnitude in the BNC and COCA

Interaction with $\log_{10}$ magnitude	BNC coefficient ($\log_e$)		COCA coefficient ($\log_e$)	
	β	95% CI	β	95% CI
Multiple of 5	+0.10	[+0.02, +0.19]	+0.03	[−0.07, +0.02]
Multiple of 10	+0.78	[+0.69, +0.88]	+0.71	[+0.65, +0.76]
2-ness	+0.93	[+0.63, +1.25]	+1.00	[+0.77, +1.24]
5-ness	+0.98	[+0.72, +1.28]	+1.06	[+0.85, +1.29]
2.5-ness	+1.04	[+0.78, +1.78]	+1.14	[+0.94, +1.35]
10-ness	+1.11	[+0.86, +1.41]	+1.22	[+1.01, +1.43]

respective pragmatic halos of each number. When we do this, the relative differences between the powers and surrounding numbers become larger: the mean frequency of numbers between 90 and 110 is 849, equivalent to 5.7% of the frequency of 100 itself, while the mean frequency of numbers between 990 and 1,100 is 25, equivalent to only 0.4% of the frequency of 1,000.

3.3. Discussion

The first analysis demonstrates that round numbers are used more frequently at higher magnitudes in both the BNC and COCA, extending results reported descriptively by Woodin et al. (2024) for the BNC. This increasing reliance on round numbers with magnitude may reflect a greater tendency towards approximation. The trends observed are consistent across the two corpora: numbers with 10-ness show the greatest frequency increase, followed by 2.5-ness, 5-ness, 2-ness and then Multiple of 10. Multiple of 5 shows a weaker interaction with magnitude in both corpora, particularly COCA, where there is little evidence of an effect. This result hints that, while divisibility by 5 is sufficient for numbers to be used in lower-magnitude approximate expressions (e.g., 'about 15 people'), at higher magnitudes, speakers increasingly use numbers more closely associated with the base 10 (e.g., those with 10-ness or 2.5-ness). In other words, the set of numbers speakers use for approximation become more restricted as magnitude increases.

4. Indefinite hyperbolic numbers

Another window into the approximate nature of round numbers is to compare them with numbers that we know to be imprecise. Indefinite hyperbolic numbers like 'kazillion' and 'bajillion' provide an ideal comparison because they look like numbers and behave like them grammatically, but lack an exact value. In this second analysis, inspired by Ariel and Levshina (2025), we compare the distributional semantics of round and non-round numbers across different magnitudes with those of indefinite hyperbolic numbers. To the extent that certain larger round numbers are used in similar linguistic contexts as indefinite hyperbolic numerals, this pattern would suggest that they are similarly imprecise.

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

4.1. Methodology

We used 'word2vec' word embeddings (Mikolov et al., 2013) pre-trained on the Wikipedia corpus (Mikolov et al., 2017). We opted to use word2vec due to its accessibility as a pre-existing database of word embeddings derived from a broad, general-purpose corpus. Additionally, comparative research indicates that these embeddings tend to perform marginally better than several other readily available alternatives across various applications (Baroni et al., 2014; Beekhuizen et al., 2021; Pereira et al., 2016).

We extracted word embeddings for English number words, including 1–99, powers of 10 ('hundred', 'thousand', 'million', 'billion', 'trillion', 'quadrillion', 'quintillion', 'sextillion') and indefinite hyperbolic numbers ('umpteen', 'zillion', 'gazillion', 'bazillion', 'bajillion', 'squillion', 'skillion', 'gajillion', 'kazillion'). We focused on single words (e.g., 'hundred') and hyphenated numbers (e.g., 'seventy-two') and not multi-word expressions (e.g., 'two thousand'). Hence, the embeddings for single words (e.g., 'three' and 'million') include their use in multi-word expressions (e.g., 'three million'). We only analysed number words (e.g., 'two' rather than '2') because these are most comparable to indefinite hyperbolic numbers, which lack numeral representations.

The word embeddings extracted are numerical representations of words in a continuous vector space with 300 dimensions. Words with similar embeddings, represented as closer points in this high-dimensional space, tend to appear in similar contexts and thus are semantically related (Lenci, 2008; Lenci & Sahlgren, 2023; see also Firth, 1957). Using these embeddings, we calculated cosine similarities between all the words in our dataset across all 300 dimensions. The cosine similarity for a pair of numbers tells us to what extent a number is used in similar sentence and discourse contexts to the other. To visualise the similarities, we used multidimensional scaling (MDS) (see Shepard, 1962) to reduce the 300 dimensions to 2 dimensions while preserving the high-dimensional similarities as much as possible. As an example of how MDS works, if two words are far from each other in one dimension but close in many others, the MDS algorithm prioritises maintaining their overall closeness in two-dimensional space, at the expense of losing information about the conflicting dimension.

4.2. Results

Figure 2 shows the word embeddings projected into two-dimensional MDS space. The numbers from 2 to 99 cluster towards the right side of the space, with magnitude accounting for their vertical distribution – larger numbers are positioned further upwards. The crescent shape of this distribution is more challenging to explain. On the left side of the figure, the indefinite hyperbolic numbers form a distinct semantic cluster, separate from the cloud of smaller numbers towards the right side of the space. Larger powers of 10 – 'million', 'billion', 'trillion', 'quadrillion', 'quintillion' and 'sextillion' – also appear in this left region of the MDS space. Their proximity to indefinite hyperbolic numbers, which have no precise value, suggests that large round numbers like these are similarly used imprecisely. In contrast, smaller powers of 10 – 'hundred' and 'thousand' – appear in a transitional position between the two clusters. This midway placement suggests that these numbers are more likely to be used imprecisely than smaller numbers, but less so than the larger powers of 10. Therefore,

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

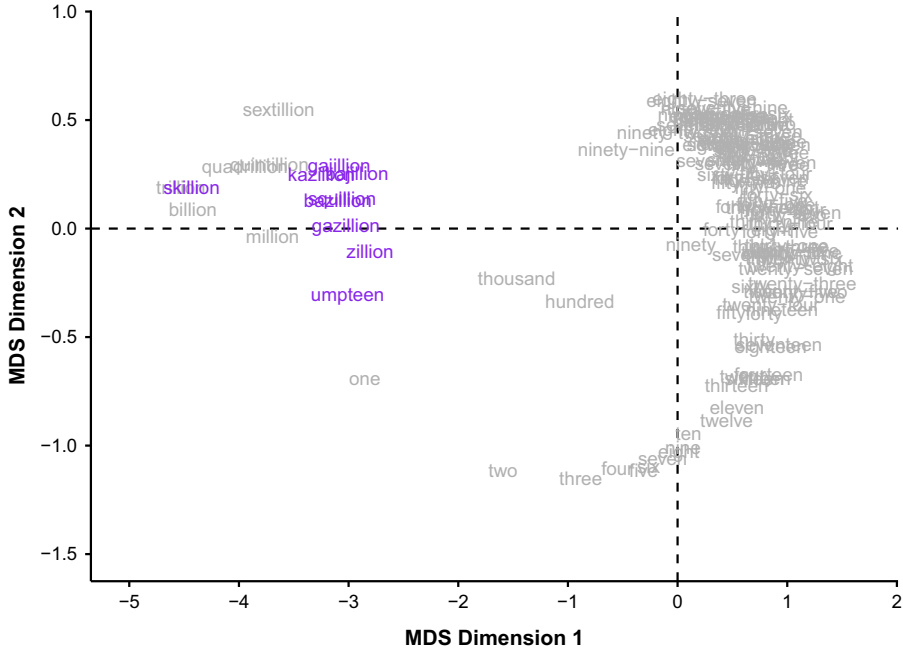


Figure 2. Multidimensional scaling (MDS) based on word2vec cosine similarity: number words (grey) and indefinite hyperbolic numbers (purple). Indefinite hyperbolic numbers occupy the same region of MDS space as larger powers of 10, like 'trillion' and 'million'; smaller numbers from 2 to 99 occupy a separate region of MDS space; and 'thousand' and 'hundred' appear between these two main clusters.

the horizontal axis of the MDS space may be loosely interpretable as a semi-continuous dimension, reflecting the probability of a number being used imprecisely, where this probability increases from right to left in a magnitude-dependent fashion. The number word 'one' appears isolated in MDS space, presumably because the word vector for this string also reflects the non-numerical use of 'one' as a pronoun (e.g., 'as one does', see Woodin et al., 2024).

4.3. Discussion

The second analysis indicates that large round numbers like 'billion' and 'trillion' are used in similar contexts to indefinite hyperbolic numbers like 'bajillion' and 'zillion'. Following the foundational idea of distributional semantics – words used in similar contexts have similar meanings – we can, therefore, conclude that these large round numbers are semantically similar to number-like words that we *know* to be imprecise. This interpretation suggests that, in many usage contexts, large round number words such as 'million' do not denote a precise number, but instead only give a general impression of magnitude, similar to words such as 'zillion' and 'umpteen'. In contrast, smaller round numbers such as 'hundred' and 'thousand' occupy a position in semantic space that is in between the indefinite hyperbolic numerals and small

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

numbers. This pattern of results is consistent with the idea that approximate uses of round number words (at least the ones analysed here) are magnitude-dependent, with the distributional behaviour of larger round numbers being more similar to that of pseudo-numbers. Conceptually, these results are consistent with Ariel and Levshina (2025), who reported that the relatively more precise numerals 2–99 were more densely clustered in semantic space than words from other lexical fields, such as cooking verbs (e.g., 'boil') and precision adjectives (e.g., 'fixed'), which they argue have less precisely defined meanings. Here, we show a similar pattern, but compare numbers to a more closely related lexical field: fictitious numbers. Our results show that while small numbers are densely clustered together, larger round numbers are closer in semantic space to fictitious numbers.

5. Rounding in jigsaw puzzles

The third and final analysis investigates rounding by comparing the advertised number of pieces in jigsaw puzzles to their actual counts, with the latter serving as a 'ground truth' allowing us to assess rounding directly. Because puzzle manufacturers often use rounded values on puzzle boxes, testing whether deviations from the stated number increase with puzzle size can reveal whether the degree of rounding grows with magnitude.

5.1. Methodology

5.1.1. Puzzle counting

We counted the pieces in an opportunity sample of 34 jigsaw puzzles with advertised piece counts ranging from 12 to 13,200 (*median* = 1,000). These advertised counts were mostly round numbers: all but the two smallest piece counts (12, 49) were round (e.g., 150, 1,500, 6,000; see Table 4). By far the most common advertised piece count was 1,000 ($N = 19$). Puzzle pieces were counted manually, twice to check for possible miscounts. The puzzles were produced by 14 brands, some of which were frequent in the dataset (e.g., Clementoni, Ravensburger, each: $N = 9$) and others which were infrequent (e.g., Marks & Spencer, Martin Schwartz, each: $N = 1$). Crucially, the observed differences were not random: they were usually large, even numbers (e.g., 30, 24, 16) and often were consistent within brands (in fact, Clementoni puzzles include a note specifying how much exact piece counts deviate from cover values). These facts strongly suggest that the differences observed are not mere artefacts of measurement error or missing pieces. For more details about the dataset, see the OSF repository: https://osf.io/ct7f8.

5.1.2. Statistical model

We used Bayesian negative binomial regression to model the effect of magnitude (puzzle size) on the absolute difference between the stated and actual counts, disregarding whether the actual count was lower or higher than the stated value (e.g., both a 994-piece puzzle and a 1,006-piece puzzle that are advertised as having 1,000 pieces would have a deviation of 6). In a negative binomial regression, the shape parameter estimates overdispersion; thus, we can also model this shape parameter as a function of magnitude, allowing us to assess whether variability in the absolute

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

differences changes as puzzle size increases.[3] For this model, to mitigate against miscounts or missing pieces, we made the conservative choice of recategorising actual counts that differ by only a single piece from the advertised value as having no difference (i.e., 0).

5.2. Results

Fewer than half the jigsaw puzzles had the exact number of puzzle pieces stated on the packaging (44.1%, $N = 15$). The majority (55.9%, $N = 19$) differed from the stated value, with discrepancies ranging from 1 to 30 puzzle pieces. As shown in Figure 3, the size of these discrepancies varied by puzzle size: the width of the yellow distribution, representing larger puzzles (≥1,000 pieces, i.e., the median piece count advertised or above), is much wider compared to that of the green distribution, representing smaller puzzles (<1,000 pieces). In other words, larger puzzles tended to differ by a greater amount from the advertised value than smaller puzzles. Table 5 shows the mean discrepancies for each puzzle size in our dataset.

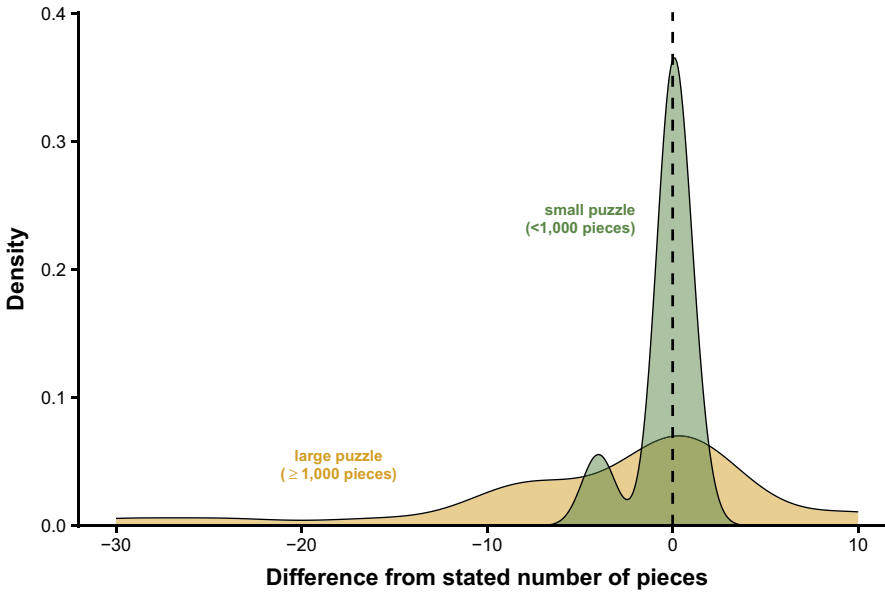


Figure 3. Density plot demonstrating that the actual count of larger puzzles (≥1,000 pieces, yellow distribution) tends to differ more from the advertised count, compared with smaller puzzles (<1,000 pieces, green distribution). Note that puzzle size was continuous, rather than discretised, in the regression analysis.

[3]We selected this model by fitting different models with candidate distributions (negative binomial, negative binomial with its shape parameter dependent on magnitude, zero-inflated negative binomial, Poisson, zero-inflated Poisson). We then compared their respective fits to the data using posterior predictive simulations and Leave-One-Out Cross-Validation (LOO-CV) (see Vehtari et al., 2017). Including random effects in the final model caused convergence issues, and did not reliably improve the model fit, as many levels of the random effects had only one data point (e.g., some brands only appeared in the dataset once), making them difficult to estimate. Hence, we report the model without random effects.

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

Table 5. Mean absolute differences from advertised values for each puzzle size in the dataset

Advertised number of pieces	Number of puzzles	Mean absolute difference from advertised count
12	1	0.0
49	1	0.0
150	2	0.0
200	2	2.0
500	2	0.5
1,000	19	3.5
1,500	2	17.0
2,000	2	1.0
3,000	1	8.0
6,000	1	16.0
13,200	1	24.0

Note: For counts represented by only one puzzle, the mean absolute difference represents the exact absolute difference for that puzzle.

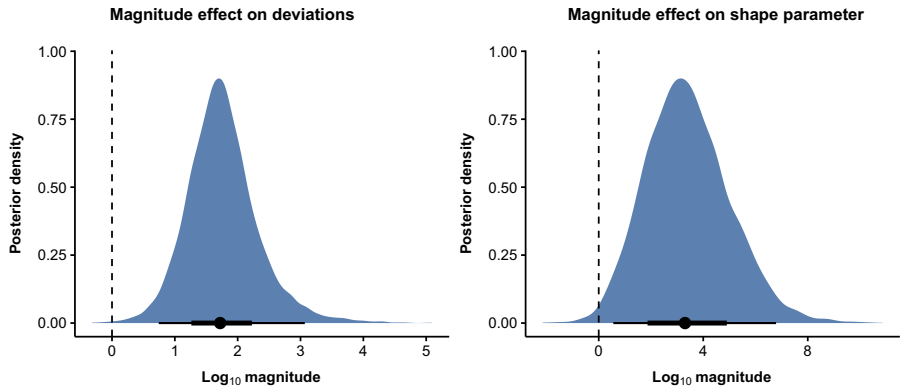


Figure 4. Posterior distributions of $\log_e$ coefficients for the effect of $\log_{10}$ Magnitude on differences between the actual and stated jigsaw piece counts (left), and the variability in these deviations, represented by the shape parameter of the model (right). Dots indicate posterior medians, thick error bars indicate 50% credible intervals, and thin error bars indicate 95% credible intervals. Both 95% credible intervals are above zero, providing substantial evidence that rounding increases and becomes less variable when jigsaw puzzle size increases.

Figure 4 shows the coefficients of the effect of Log_{10} Magnitude on differences between the stated and actual piece counts (left) and the variability of these differences (right). The coefficients are both positive, demonstrating that deviations between actual and stated counts increase for larger puzzles (β +1.76, 95% CI = [+0.74, +3.07]), and that the variability in these deviations decreases (β +3.40, 95% CI = [+0.56, +6.79]), given the inverse relationship between the shape parameter and variability.

This analysis has concentrated on absolute deviations from the advertised value. Importantly, when we look at *proportional* deviations, we do not see any clear systematic magnitude-related patterns. For example, the largest proportional deviation (2.0%) was obtained for a puzzle advertised as having 200 pieces, but also for one advertised as having 1,500 pieces. Some of the larger puzzles – advertised as having 3,000, 6,000 and 13,200 pieces – had proportional deviations of only 0.3%,

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

0.3% and 0.2%, respectively. The difference between absolute and proportional deviations is expected given the Weber-Fechner law (e.g., Dehaene, 2003; DeWind et al., 2015; Shepard et al., 1975b), which predicts tolerance for imprecision to remain approximately constant in proportional terms even as absolute deviations increase systematically with magnitude.

5.3. Discussion

The third analysis shows that, for larger jigsaw puzzles, the stated – typically round – number of pieces on the packaging deviates by more pieces from the actual count. When a round number such as 1,000, or number word such as 'thousand', is considered in the frequency analysis (Section 3) or the distributional semantic analysis (Section 4), we ultimately do not know whether it was used precisely or approximately. Moreover, insofar as rounding has occurred, we have no access to the underlying value from which the round number was derived. As argued above, the patterns we observe are consistent with an approximate function of round numbers, but these analyses lack a clear 'ground truth' for what has been approximated. Our analysis of jigsaw puzzles provides more direct evidence of increased approximation at higher magnitudes. The finding that larger advertised counts stand for an increasingly wide range of values is in line with the idea that larger numbers have wider pragmatic halos.

6. General discussion

Numbers are often seen as precise and objective. Our results show, however, that their use in everyday communication is systematically shaped by approximate or even wholly vague use. Across three analyses of British and American English, we identified patterns that link round numbers to approximation or imprecision in a magnitude-dependent fashion. First, corpus analyses show that round numbers occur with increasing frequency at higher magnitudes, relative to nearby values. This increase is strongest for numbers most closely related to powers of 10, suggesting not only more frequent approximation but also coarser rounding at higher magnitudes. Second, larger round numbers (e.g., 'million', 'trillion') exhibit distributional profiles similar to those of indefinite hyperbolic numbers (e.g., 'skillion', 'bazillion'), which indicates that they similarly do not imply exact values. Third, using jigsaw puzzles as a test case, we show that the discrepancy between advertised (typically round) and actual piece counts increases with magnitude, which provides direct evidence that larger round numbers are used to represent a wider range of values – in other words, they have wider pragmatic halos (Krifka, 2007; Lasersohn, 1999). Taken together, these findings show that round number use becomes increasingly approximate as magnitude increases, and in some cases lacks a precise value altogether.

In this respect, round number usage often resembles more explicitly vague ways of communicating about quantities (see Winter & Marghetis, 2023), such as vague quantifiers (e.g., 'many', 'a lot') (Cummins, 2015; Moxey & Sanford, 2023), expressions involving nouns (e.g., 'masses of', 'oodles of') (Channell, 1994, Ch. 5), vertical or size-based terms (e.g., 'skyrocketing', 'falling', 'growing', 'shrinking' numbers) (Lakoff & Johnson, 1980; Winter et al., 2013), numerical range expressions (e.g., '10–15 people') (Pollmann & Jansen, 1996) and hedge words (e.g., 'about', 'give or

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

take') (Ferson et al., 2015). The fact that the English language has so many linguistic devices for approximate or vague reference to numerical information suggests a strong communicative need for this type of information (see Regier et al., 2016; Winter et al., 2018). Indeed, vague quantitative reference appears to dominate even in contexts where precision might be expected, such as academic writing, political discourse, and business communication (Banks, 1998; Cutting, 2012); political discourse and news programmes (Alcaraz-Carrión et al., 2022; Woodin et al., 2020); and business contexts (Koester, 2007). Our results show that this preference for approximation (Solt et al., 2017) extends to the use of numbers: despite their capacity for precision, numbers are often used in approximate or vague ways, especially at higher magnitudes. Seen from this perspective, there is a tension between the cultural status of numbers as hallmarks of objectivity (Porter, 1995) and their actual use in practice.

This tension is most directly evidenced by the existence of indefinite hyperbolic numerals, which show that people are drawn to the *form* of numbers, but often do not care about precise reference (Chrisomalis, 2016; see also Burnham, 1951; Pound, 1949). These fictitious numbers are not limited to English: for example, 'ørten' in Norwegian (Universitetet i Bergen & Språkrådet, n.d.), 'milhentos' in Portuguese (Dicionário Priberam, n.d.), 'tropocientos/tropocientas' in Spanish (Real Academia Española, n.d.) and 'عشرطعش' ('ashartaash') in Palestinian Arabic are similar in etymology and function to the English 'umpteen'. Whereas these indefinite hyperbolic numerals stand out as informal and playful (Chrisomalis, 2016), our distributional semantics analysis suggests that they may be interchangeable with round numbers such as 'million' and 'billion' in certain contexts. For instance, it may make no pragmatic difference whether the car was moving at a 'million' or 'bazillion' miles an hour – the important thing is that it was moving very fast. In communication, then, if not in mathematics, the boundary between round and fictitious numbers seems to be permeable, with both types of number being capable of conveying a vague sense of magnitude.

While the analysis of fictitious numbers shows that their distributional semantics resemble that of larger round numbers, it does not show a continuous magnitude-related pattern, as do the other analyses, because we only looked at the numbers 1–99 and words corresponding to other powers of 10 up to sextillion. In relying on pre-trained word2vec embeddings, this method captured not just isolated uses of 'billion', 'quadrillion' and so on, but also their use in multi-word numerical expressions, such as 'one billion and one'. Consequently, this analysis did not differentiate between '1 million' and '10 million' – indeed, the latter may not behave any more like 'squillion' than the former, despite its larger magnitude. It is plausible that '10 million' is more often used for relatively more precise numerical reference, given that the lexicalised 'million' can flexibly indicate an order of magnitude, while the compositional meaning of '10 million' implies a specific calculation (10 × 1 million). Relatedly, because distributional semantics is based on word co-occurrence, the fact that lexicalised number words can occur in multi-number expressions introduces a source of difference between them and the indefinite hyperbolic numerals. For example, 'umpteen' cannot felicitously be modified (e.g., 'twenty umpteen'), and indeed, it does not occur in the corpus this way. The observation that lexicalised magnitude words still exhibit distributional resemblance to pseudo-numbers therefore suggests that it is their flexible, often imprecise use as indicators of magnitude –

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

rather than their participation in compositional numerical expressions – that contributes most strongly to this resemblance.

This preference for approximation in number use answers a fundamentally different question than that of their *interpretation*. Indeed, Ariel and colleagues argue that in comprehension, numerals are interpreted as *precise* by default (e.g., Ariel, 2021; Ariel & Levshina, 2025). Katzir and Ariel (2026) present a series of experiments suggesting that interlocutors are aware that speakers often use round numbers imprecisely, but that this awareness does not necessarily translate across to interpretation. Importantly, the largest number included in Katzir and Ariel's numerical stimuli was 90, and likewise, the numbers included in Ariel and Levshina's (2025) corpus analyses were much smaller than those analysed in this paper. Our results suggest that magnitude should be explicitly considered in future studies, particularly in experiments testing whether comprehenders would interpret references to round numbers in the order of millions or billions in a similarly precise fashion.

A central claim of this paper is that the numbers deemed 'round', and thus are used for approximation, derive from the structure of the numeral system in operation. In English and many other languages (Comrie, 2013), where the decimal system is dominant, powers of 10 occupy structurally salient positions on the number line, and other 'special numbers' are mathematically related to these powers. This way of thinking about roundness develops the graded characterisation of roundness presented by Woodin et al. (2024), which suggests that certain round numbers can be 'rounder' than others. Because the roundest numbers in the system are related to powers of 10, they grow further apart with magnitude (Jansen & Pollmann, 2001). This magnitude-dependent aspect of roundness provided us with an opportunity to infer approximate number use from frequency data, even though such data inherently underspecify a number's function. Round numbers can be used for both approximate functions (e.g., estimating unknown quantities; Ruud et al., 2014) and non-approximate ones (e.g., benchmarking; Ozcan et al., 2024; Pope & Simonsohn, 2011). The fact that numbers possessing Jansen and Pollman's (2001) roundness properties show particularly pronounced frequency increases with magnitude, above and beyond other multiples of 5 and 10, provides stronger evidence for magnitude-relative effects being attributable to approximation. As these round numbers become sparser, the pragmatic halo (Lasersohn, 1999) around them widens, which means that they are used to represent a broader range of values (Cummins, 2015; Krifka, 2009).

Whereas our claims about pragmatic halos in relation to aggregated number frequencies were inferred, our analysis of jigsaw puzzles permitted us to measure the pragmatic halo of round numbers more directly. We showed that when the actual value that was rounded is known, it indeed deviates more from the communicated round number at higher magnitudes. An outstanding question in regard to jigsaw puzzles is from where this imprecision originates. During manufacturing, jigsaw puzzles are typically cut out from cardboard using a custom die that produces pieces of roughly uniform size, minimising cues that could reveal where pieces should fit based on their size or shape. It is likely to be more difficult to cut out puzzle pieces in such a way that an exact value of, say, 10,000 is reached. Notwithstanding this explanation for the *actual* number of pieces not being round, puzzle makers still report rounded values, and the numbers reported diverge further from the real figure at higher magnitudes. The source of imprecision in jigsaw puzzles is thus identical to

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

other instances of rounding: a discrepancy between actual values in the world and the round numbers we preferentially use when communicating. Because our preferred round numbers (e.g., those with 10-ness) become sparser at higher magnitudes, this discrepancy tends to increase as magnitude grows.

Overall, our results suggest a resolution to the apparent contradiction between accounts of numbers as exact by default, which appeal to formal semantic analyses (Ariel, 2021; Ariel & Levshina, 2025), and competing accounts that assume that most numbers cannot have exact representations, which rely more on cognitive explanations based on experimental research on the approximate number system (Dehaene, 1997). These two positions may correspond to different facets of meaning that can be true simultaneously: numbers do have precise mathematical meanings, but in practice, our results suggest that round numbers, in particular, are often used in an approximate fashion. The dual nature of round numbers is not a paradox, but rather a direct consequence of their role in bridging the precise, abstract world of mathematics with the fuzziness of human cognition and communication.

Data availability statement. All data, analysis scripts and additional information about this study can be found at https://osf.io/ct7f8.

Acknowledgements. Bodo Winter was supported by UK Research and Innovation (Future Leaders Fellowship MR/T040505/1).

References

A'Hearn, B., Baten, J., & Crayen, D. (2009). Quantifying quantitative literacy: Age heaping and the history of human capital. *The Journal of Economic History*, 69(3), 783–808. https://doi.org/10.1017/S0022050709001120.

Alcaraz-Carrión, D., Alibali, M. W., & Valenzuela, J. (2022). Adding and subtracting by hand: Metaphorical representations of arithmetic in spontaneous co-speech gestures. *Acta Psychologica*, 228, 103624. https://doi.org/10.1016/j.actpsy.2022.103624.

Ariel, M. (2021). Why it's hard to construct ad hoc number concepts. In C. Mauri, I. Fiorentini, & E. Goria (Eds.), *Building categories in interaction: Linguistic resources at work* (pp. 439–462). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.220.16ari.

Ariel, M., & Levshina, N. (2025). The counting principle makes number words unique. *Corpus Linguistics and Linguistic Theory*, 21(1), 173–199. https://doi.org/10.1515/cllt-2023-0105.

Banks, D. (1998). Vague quantification in the scientific journal article. *ASp: La Revue Du GERAS*, 1922, 17–27. https://doi.org/10.4000/asp.2666

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 238–247). https://doi.org/10.3115/v1/P14-1023.

Batorsky, B., Ledvosky, A., Yarkoni, T., & Groove, B. (2021). *Word2Number* [Computer software]. https://w2n.readthedocs.io/en/latest/

Beekhuizen, B., Armstrong, B. C., & Stevenson, S. (2021). Probing lexical ambiguity: Word vectors encode number and relatedness of senses. *Cognitive Science*, 45(5), e12943. https://doi.org/10.1111/cogs.12943.

Beltrama, A., Solt, S., & Burnett, H. (2023). Context, precision, and social perception: A sociopragmatic study. *Language in Society*, 52(5), 805–835. https://doi.org/10.1017/S0047404522000240.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit* (1st ed.). O'Reilly Media.

Bittinger, K. (2020). *usedist: Distance matrix utilities* (Version 0.4.0) [Computer software]. https://CRAN.R-project.org/package=usedist

BNC Consortium. (2007). *British National Corpus* (Version 3) [Computer software] . BNC Consortium. http://www.natcorp.ox.ac.uk/

Bürkner, P.-C. (2017). Brms: AnRPackage for Bayesian multilevel models UsingStan. *Journal of Statistical Software*, 80(1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411. https://doi.org/10.32614/RJ-2018-017.

Burnham, J. M. (1951). Observations on indefinite numbering. *American Speech*, 26(1), 65–67. https://doi.org/10.2307/453320.

Channell, J. (1994). *Vague language*. Oxford University Press.

Cheyette, S. J., & Piantadosi, S. T. (2019). A primarily serial, foveal accumulator underlies approximate numerical estimation. *Proceedings of the National Academy of Sciences*, 116(36), 17729–17734. https://doi.org/10.1073/pnas.1819956116.

Chrisomalis, S. (2016). Umpteen reflections on indefinite hyperbolic numerals. *American Speech*, 91(1), 3–33. https://doi.org/10.1215/00031283-3509480.

Chrisomalis, S. (2020). *Reckonings: Numerals, cognition, and history*. The MIT Press. https://doi.org/10.7551/mitpress/13381.001.0001.

Comrie, B. (2013). Numeral bases. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Zenodo. https://doi.org/10.5281/zenodo.13950591.

Coupland, N. (2011). How frequent are numbers? *Language and Communication*, 31(1), 27–37. https://doi.org/10.1016/j.langcom.2010.09.001.

Cummins, C. (2015). *Constraints on numerical expressions*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199687909.001.0001.

Cutting, J. (2012). Vague language in conference abstracts. *Journal of English for Academic Purposes*, 11(4), 283–293. https://doi.org/10.1016/j.jeap.2012.05.004.

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)* [Computer software]. https://www.english-corpora.org/coca/

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press.

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, 7(4), 145–147. https://doi.org/10.1016/S1364-6613(03)00055-X.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1), 1–29. https://doi.org/10.1016/0010-0277(92)90030-L.

DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (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.

Dicionário Priberam. (n.d.). *Milhentos*. https://dicionario.priberam.org/milhentos

Dorogovtsev, S., Mendes, J. F., & Oliveira, J. (2006). Frequency of occurrence of numbers in the world wide web. *Physica A: Statistical Mechanics and its Applications*, 360(2), 548–556. https://doi.org/10.1016/j.physa.2005.06.064.

Dupras, V., Ortiz, E. R., Szalaï, I., Lodato, M., & Harper, B. (2021). *Num2Words computer software*. Savoir-faire Linux Inc.

Fernández-i-Marín, X. (2016). ggmcmc: Analysis of MCMC samples and Bayesian inference. *Journal of Statistical Software*, 70(9). https://doi.org/10.18637/jss.v070.i09.

Ferson, S., O'Rawe, J., Antonenko, A., Siegrist, J., Mickley, J., Luhmann, C. C., Sentz, K., & Finkel, A. M. (2015). Natural language of uncertainty: Numeric hedge words. *International Journal of Approximate Reasoning*, 57, 19–39. https://doi.org/10.1016/j.ijar.2014.11.003.

Firth, J. (1957). *Papers in linguistics, 1934–1951*. Oxford University Press.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression*. (3rd ed.). Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Gibbs Jr., R. W., & Bryant, G. A. (2008). Striving for optimal relevance when answering questions. *Cognition*, 106(1), 345–369. https://doi.org/10.1016/j.cognition.2007.02.008.

Gunasti, K., & Ozcan, T. (2016). Consumer reactions to round numbers in brand names. *Marketing Letters*, 27(2), 309–322. https://doi.org/10.1007/s11002-014-9337-7.

Hurford, J. R. (1975). *The linguistic theory of numerals*. Cambridge University Press.

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

JetBrains. (2022). *PyCharm* (Version 2022.2.3) [Computer software].

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33), 12002–12007. https://doi.org/10.1073/pnas.1407479111.

Kassambara, A. (2023). *ggpubr: "ggplot2" based publication ready plots* (Version 0.6.0) [Computer software]. https://CRAN.R-project.org/package=ggpubr

Katzir, N., & Ariel, M. (2026). *Don't overthink! Round number expressions are interpreted sharp.*

Kay, M. (2023). *tidybayes: Tidy data and geoms for Bayesian models* (Version 3.0.6) [Computer software]. http://mjskay.github.io/tidybayes/>

Koester, A. (2007). 'About twelve thousand or so': Vagueness in north American and UK offices. In J. Cutting (Ed.), *Vague language explored* (pp. 40–61). Palgrave Macmillan. https://doi.org/10.1057/9780230627420_3.

Krifka, M. (2007). *Approximate interpretation of number words*. Philosophische Fakultät II. https://doi.org/10.18452/9508.

Krifka, M. (2009). Approximate interpretations of number words: A case for strategic communication. In E. Hinrichs & J. Nerbonne (Eds.), *Theory and evidence in semantics* (pp. 109–132). CSLI Publications.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lasersohn, P. (1999). Pragmatic halos. *Language*, 75(3), 522–551. https://doi.org/10.2307/417059.

Lavric, E. (2010). Hyperbolic approximative numerals in cross-cultural comparison. In G. Kaltenböck, W. Mihatsch, & S. Schneider (Eds.), *New approaches to hedging* (pp. 123–164). Brill. https://doi.org/10.1163/9789004253247_008.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1), 1–31.

Lenci, A., & Sahlgren, M. (2023). *Distributional semantics*. Cambridge University Press. https://doi.org/10.1017/9780511783692.

Marek, P. E., Buzatto, B. A., Shear, W. A., Means, J. C., Black, D. G., Harvey, M. S., & Rodriguez, J. (2021). The first true millipede—1306 legs long. *Scientific Reports*, 11(1), 23126. https://doi.org/10.1038/s41598-021-02447-0.

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month-old infants. *Psychological Science*, 18(8), 740–745. https://doi.org/10.1111/j.1467-9280.2007.01969.x.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press/Taylor & Francis Group.

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 445, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* (arXiv:1301.3781). *arXiv*. https://doi.org/10.48550/arXiv.1301.3781.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). *Advances in pre-training distributed word representations* (arXiv:1712.09405). *arXiv*. https://doi.org/10.48550/arXiv.1712.09405.

Moxey, L. M., & Sanford, A. J. (2023). *Communicating quantities: A psychological perspective*. Routledge. https://doi.org/10.4324/9781003429838.

Nakazawa, M. (2024). *fmsb: Functions for medical statistics book with some demographic data* (Version 0.7.6) [Computer software]. https://cran.r-project.org/web/packages/fmsb/index.html

Ooms, J. (2024). *magick: Advanced graphics and image-processing in R* (Version 2.8.5) [Computer software]. https://cran.r-project.org/web/packages/magick/index.html

Ozcan, T., Hair, M., & Gunasti, K. (2024). How reaching numerical roundness on subgoals affects the completion of superordinate goals. *Journal of Business Research*, 177, 114617. https://doi.org/10.1016/j.jbusres.2024.114617.

Pedersen, T. L. (2024). *patchwork: The composer of plots* (Version 1.3.0) [Computer software]. https://cran.r-project.org/web/packages/patchwork/index.html

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, 33(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907.

Pollmann, T., & Jansen, C. (1996). The language user as an arithmetician. *Cognition*, 59(2), 219–237. https://doi.org/10.1016/0010-0277(95)00704-0.

Pope, D., & Simonsohn, U. (2011). Round numbers as goals. *Psychological Science*, 22(1), 71–79. https://doi.org/10.1177/0956797610391098.

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

Porter, T. M. (1995). *Trust in Numbers: The pursuit of objectivity in science and public life* (1st ed.). Princeton University Press.

Posit Team. (2023). *RStudio: Integrated development environment for R* (Version 2023.6.0.421) [R]. Posit Software, PBC. http://www.posit.co/

Pound, L. (1949). Indefinite numbering. *American Speech*, 24(3), 229–230. https://doi.org/10.2307/487053.

Python Software Foundation. (2021). *Python* (Version 3.7) [Computer software]. https://www.python.org/

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Real Academia Española. (n.d.). Tropecientas. In *Diccionario de la Lengua Española*. https://dle.rae.es/tropecientas

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLoS One*, 11(4), e0151138. https://doi.org/10.1371/journal.pone.0151138.

Rips, L. J. (2013). How many is a zillion? Sources of number distortion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(4), 1257–1264. https://doi.org/10.1037/a0031143.

Ruud, P., Schunk, D., & Winter, J. (2014). Uncertainty causes rounding: An experimental study. *Experimental Economics*, 17(3), 391–413. https://doi.org/10.1007/s10683-013-9374-8.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27(3), 219–246. https://doi.org/10.1007/BF02289621.

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975a). The internal representation of numbers. *Cognitive Psychology*, 7(1), 82–138. https://doi.org/10.1016/0010-0285(75)90006-7.

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975b). The internal representation of numbers. *Cognitive Psychology*, 7(1), 82–138. https://doi.org/10.1016/0010-0285(75)90006-7.

Sigurd, B. (1988). Round numbers. *Language in Society*, 17(2), 243–252. https://doi.org/10.1017/S0047404500012781.

Solt, S., Cummins, C., & Palmović, M. (2017). The preference for approximation. *International Review of Pragmatics*, 9(2), 248–268. https://doi.org/10.1163/18773109-00901010.

Turner, S. H. (1958). Patterns of heaping in the reporting of numerical data. *Proceedings of the Social Statistics Section*, 248–251.

Universitetet i Bergen & Språkrådet. (n.d.). Ørten. In *Norsk Ordbok*. https://naob.no/ordbok/%C3%B8rten

Urbanek, S. (2022). *png: Read and write PNG images* (Version 0.1.8) [Computer software]. https://cran.r-project.org/web/packages/png/index.html

Van der Henst, J.-B., & Sperber, D. (2004). Testing the cognitive and communicative principles of relevance. In I. A. Noveck & D. Sperber (Eds.), *Experimental pragmatics* (pp. 141–171). Palgrave Macmillan UK. https://doi.org/10.1057/9780230524125_7.

Van der Henst, J.-B., Carles, L., & Sperber, D. (2002). Truthfulness and relevance in telling the time. *Mind and Language*, 17(5), 457–466. https://doi.org/10.1111/1468-0017.00207.

Van Rossum, G. (2020). *The Python Library Reference, Release 3.8.2* (Version 2.2.1) [Computer software]. Python Software Foundation.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.

Wickham, H. (2019). *tidyverse: Easily install and load the "tidyverse"*. (Version 2.0.0) [R]. https://CRAN.R-project.org/package=tidyverse

Wild, F. (2022). *_lsa: Latent semantic analysis_* (Version 0.73.3) [Computer software]. https://CRAN.R-project.org/package=lsa

Wilke, C. O. (2022). *ggridges: Ridgeline plots in "ggplot2"* (Version 0.5.4) [R]. https://CRAN.R-project.org/package=ggridges

Winter, B., & Bürkner, P.-C. (2021). *Poisson regression for linguists: A tutorial introduction to modeling count data with brms*. OSF Preprints. https://doi.org/10.31219/osf.io/93kaf

Winter, B., & Marghetis, T. (2023). Multimodality matters in numerical communication. *Frontiers in Psychology*, 14. 10.3389/fpsyg.2023.1130777.

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press

Winter, B., Perlman, M., & Matlock, T. (2013). Using space to talk and gesture about numbers. *Gesture*, 13(3), 377–408. https://doi.org/10.1075/gest.13.3.06win.

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, 179, 213–220. https://doi.org/10.1016/j.cognition.2018.05.008.

Woodin, G., & Winter, B. (2024). Numbers in context: Cardinals, ordinals, and nominals in American English. *Cognitive Science*, 48(6), https://doi.org/10.1111/cogs.13471.

Woodin, G., Winter, B., Perlman, M., Littlemore, J., & Matlock, T. (2020). Tiny numbers" are actually tiny: Evidence from gestures in the TV news archive. *PLoS One*, 15(11), e0242142. https://doi.org/10.1371/journal.pone.0242142.

Woodin, G., Winter, B., Littlemore, J., Perlman, M., & Matlock, T. (2024). Large-scale patterns of number use in spoken and written English. *Corpus Linguistics and Linguistic Theory*, 20(1), 123–152. https://doi.org/10.1515/cllt-2022-0082.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, 74(1), B1–B11. https://doi.org/10.1016/s0010-0277(99)00066-9.

Zhang, J., & Norman, D. A. (1995). A representational analysis of numeration systems. *Cognition*, 57(3), 271–295. https://doi.org/10.1016/0010-0277(95)00674-3.

Cite this article: Woodin, G., Lorson, A., & Winter, B. (2026). Approximate round number use depends on magnitude, *Language and Cognition*, 18, e42, 1–25. https://doi.org/10.1017/langcog.2026.10090

https://doi.org/10.1017/langcog.2026.10090 Published online by Cambridge University Press