Language Sciences 105 (2024) 101656

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

Diachronic changes of number use in written American English from 1923 to 2008

Gui Wang [a], Jing Shu [b], Li Wang [c,d,*]

[a] School of International Studies, Zhejiang University, Hangzhou, China
[b] School of International Chinese Language Education, Beijing Normal University, Beijing, China
[c] XianDa College of Economics and Humanities, Shanghai International Studies University, Shanghai, China
[d] Foreign Languages College, Shanghai Normal University, Shanghai, China

ARTICLE INFO

Article history:
Received 15 December 2023
Received in revised form 24 May 2024
Accepted 7 June 2024
Available online xxx

Keywords:
Number use
TIME corpus
Number frequencies
Diachronic changes
Cultural salience

ABSTRACT

This study examines diachronic changes in number use in written American English from 1923 to 2008 in TIME magazine, analyzing a 170-million-token corpus. We investigate the effects of magnitudes, roundness, and representational formats on number frequency, along with the evolution of culturally salient numbers reflecting societal shifts. Utilizing Bayesian negative binomial regression for in-depth corpus analysis, our findings demonstrate a consistent influence of magnitudes and roundness, with smaller magnitudes and rounder numbers appearing more frequently. We observe a significant standardization in portraying large numbers, marked by a shift from numerical to mixed forms (e.g., “6,000,000,000” to “6 billion”) around 1940s. This reflects changes in both formal writing conventions and editorial practices of numerical representation. Our research further identifies distinct culturally significant numbers for each decade, linked to social, economic, and technological trends, underscoring the role of numerical analysis in media to decode complex cultural and societal patterns. This study contributes significantly to understanding the dynamic interplay between language, culture, and media in the context of numerical representations.

1. Introduction

There seems to be a well-established intellectual practice of separating mathematics from language-based disciplines, which might lead to the intuitive belief that “linguists don’t do numbers” (Coupland, 2011). However, the reality is that numbers permeate our daily language far more than we might initially perceive. From the number of likes on a social media post to the billions in government budgets, numbers shape our discussions and decisions. Interestingly, the way we use numbers has evolved over time. For example, the term “million” once indicated a vast, almost unfathomable quantity, but in today’s context of global population and economics, it has become a much more commonplace term. Yet, the study of numbers in linguistics remains surprisingly underexplored, presenting a rich vein of research yet to be fully mined. Indeed, as Coupland (2011: 27) asserts, linguists not only engage with numbers but must do so, as they are crucial meaning-bearing elements within our discourse, reflecting and shaping societal norms and values.

* Corresponding author. 390 Dongtiyu Hui Road, Hongkou District, Shanghai, China.
E-mail address: wanglily22@shnu.edu.cn (L. Wang).

https://doi.org/10.1016/j.langsci.2024.101656

There are two main strands of research related to numbers. The first focuses on the social meaning of numbers (e.g., Lakoff and Núñez, 2000; O'Halloran, 2005; Porter, 1995; Roeh and Feldman, 1984; Woodward, 1999). For instance, Porter (1995) conducted a historical analysis of the use of statistics to achieve objectivity in professional and academic practice, arguing that we turn to numbers when our faith in wisdom and community experience fails. He suggested that this shift towards quantitative analysis is not inherent in science, but rather a response to political and societal pressures.

The second strand emphasizes more general aspects of number use such as differences across text genres and factors influencing use like magnitude and roundness (e.g., Coupland, 2011; Dehaene and Mehler, 1992; Jansen and Pollmann, 2001; Woodin et al., 2023). For example, Coupland (2011) found a declining frequency of use with increasing size of numbers in both British National Corpus (BNC) and the world-wide web. In addition, Woodin et al. (2023) noted a higher occurrence of round numbers compared to non-round numbers of similar magnitudes. They posited that this phenomenon could be attributed to the unknown or irrelevant nature of the exact value (Ruud et al., 2014; Van Der Henst et al., 2002).

These studies have mainly focused on the synchronic aspects of number use. Few attentions were paid to the diachronic evolution and cultural significance of numbers. As Woodin et al. (2023) noted, as cultures evolve over time, so may preferences for using and communicating different numbers.

To address this gap, this study explored diachronic changes in number use in written American English. We utilized the TIME magazine corpus (Xu, 2015), a massive dataset of 170 million tokens spanning the period from 1923 to 2008. While TIME magazine represents a specific genre (i.e., news media), it covers a wide range of topics of general interest, including national and international affairs, business, education, science, and the arts. By examining number use across decades within this corpus, we might gain valuable insights into how societal changes and global events are reflected in language. Specifically, we examined the effects of magnitude, roundness, and representational formats on number use. Additionally, our research investigated the diachronic changes in culturally salient numbers to detect potential social movements and changes in each decade, thus offering new insights into the dynamic relationship between numerical language and social change.

2. Literature review

2.1. Factors influencing number use

Previous studies found that the number frequency in use was affected by magnitude (e.g., Coupland, 2011; Cummins, 2015; Dehaene and Mehler, 1992; Dorogovtsev et al., 2006; Woodin et al., 2020), roundness (e.g., Dehaene and Mehler, 1992; Jansen and Pollmann, 2001; Woodin et al., 2023), and cultural salience (Woodin et al., 2023).

Numerical magnitude refers to the size or quantity of a number. By probing into the number frequency of numerals and ordinals in seven different languages, Dehaene and Mehler (1992) found a decrease of the frequency with magnitude for the numerals 1–9 and 10–90, as well as for the ordinals 1st to 9th. In other words, people tend to use smaller numbers than large ones. These results are consistent with those of Jansen and Pollmann (2001) and Coupland (2011). Jansen and Pollmann (2001) found a similar trend for numbers 2–1000 by examining a 27-million-word Dutch corpus, while Coupland (2011) demonstrated a scalar decease in occurrence frequencies across the number words one to seventeen in the BNC.

Some explanations have been proposed for the effect of magnitude on number frequency. Cummins (2015) suggested that people use small numbers more frequently, as numerals of lesser magnitude exhibit greater frequency in everyday life than those of greater magnitude. For instance, there are many more sets of 6 objects than there are sets of 767 objects in daily life.

From a psychological perspective, Dehaene and Mehler (1992) argued that small numbers are discussed more because these numbers receive an expanded and more accurate mental representation relative to large numbers. This idea has been confirmed by psychological tasks. For instance, Buckley and Gillman (1974) and Dehaene (1989) found that small numbers were processed faster than large numbers in a larger-number comparison task. Similarly, Coupland (2011) noted that big numbers might require more 'effort', presumably cognitive processing effort. In addition, Coupland (2011) suggested that this phenomenon could be attributed to the frequent scale transformation in daily life. For instance, instead of saying 1000 g, we often prefer to use 1 kg.

Roundness is another factor that affects number frequency. Round numbers typically consist of multiples of 10 and occasionally 5 in decimal number systems (Coupland, 2011; Dorogovtsev et al., 2006; Sigurd, 1988). Dehaene and Mehler (1992) denote a more general definition of round numbers, which is a group of numbers or reference numbers most suited for use in estimates.

Dehaene and Mehler (1992) demonstrated that there are localized sharp increases in frequency for round numbers that serve as important reference points, such as 10, 12, 15, 20, 50 or 100, despite the overall downward trend of smaller numbers being more frequent than larger numbers. They posited that such an increase may arise from the dual representation of round numbers (e.g., 10), which can signify both the number itself and values proximate to it (e.g., 9 and 11).

Based on Dehaene and Mehler (1992), Jansen and Pollmann (2001) further explored the properties that make a round number. They identified four numerical properties closely associated with the degree of roundness: 10-ness, 2-ness, 5-ness, and 2.5-ness. According to Jansen and Pollmann (2001), a number possesses 5-ness property if, when divided by 5 multiplied by a power of 10, the result is an integer between 1 and 9. For example, when the number 50 is divided by 5 multiplied by 10^1 (i.e., $50/(5 * 10)$), the result is 1, an integer between 1 and 9. Therefore, we can say that the number 50 possesses 5-ness property. Similarly, a number possesses 2.5-ness property if, when divided by 2.5 multiplied by a power of 10, the result is an integer between 1 and 9. The roundness property for 2-ness can be determined in a similar manner. It is worth noting that 10-

ness is defined differently. A number exhibits 10-ness if it equals an integer between 1 and 9 when divided by 1 (not 10) multiplied by a power of 10. The more of these properties a number shares, the rounder it is and the higher its frequency. Jansen and Pollmann (2001) argued that this phenomenon could be attributed to the fact that doubling and halving (sometimes followed by halving again) are basic means for humans to manipulate quantities.

Expanding on the findings of Jansen and Pollmann (2001), Woodin et al. (2023) determined the effect of each roundness property on number frequencies. Specifically, by adopting Bayesian negative binomial regression with six roundness properties as independent variables and number frequency as the dependent variable, they probed into the individual effect of each of the six roundness properties. Woodin et al. (2023) found that the order of effects from the largest to the least is 10-ness > 2.5-ness > 5-ness > 2-ness > Multiple of 10 > Multiple of 5, thus confirming that there are differences in the effect of roundness properties on number frequency.

Cultural salience is another factor that influences the frequency of use (Woodin et al., 2023). Specifically, culturally salient numbers tend to be used frequently. For instance, Pollmann (1998) found that recent years are discussed more often in conversion than less recent years, as they are more relevant to the present discussion. In addition, years referring to important historical events like the Fall of the Berlin Wall (1989), the French Revolution (1789–1799), and the Great Depression (1929–1939) are much referred to in our conversation or written texts. Furthermore, Coupland (2011) found that numbers featuring repeated numerals are often integrated into product names, as seen in the card game ‘Ninety-nine’, which may attribute to their commercial appeal, phonological alliteration, and perceived ‘coolness’. Moreover, numbers carrying numerological importance are also expected to be used frequently such as the numbers 911 and 110, which indicate emergency services in the USA and China. In short, despite sporadic references in previous studies as mentioned above (e.g., Coupland, 2011; Pollmann, 1998), the influence of cultural salience on number use remains underexplored quantitatively.

Woodin et al. (2023) are the most relevant to our area of interest. When examining the factors affecting number frequency in use, Woodin et al. (2023) not only took magnitude and roundness into account but also the potential effect of cultural salience. They attempted to depict a comprehensive view of number use from a more than 100-million-word corpus BNC. Specifically, they probed into the effect of cultural salience on number use through the residuals of Bayesians binominal regression model. Woodin et al. (2023) argued that larger residuals indicate a higher likelihood that the numbers cannot be explained by the model’s independent variables (i.e., roundness and magnitude). Thus, the numbers with high residuals may reflect numbers with potential cultural salience. By examining the top residual numbers, Woodin et al. (2023) identified a significant presence of dates in the BNC and other culturally salient numbers such as 80,486 (microprocessor) and 999 (the emergency services number in the UK). These results confirmed the effect of culture salience on number frequency in use by adopting a large-scale quantitative perspective.

Woodin et al. (2023) have offered a comprehensive depiction of number use by taking both magnitude, roundness, and especially cultural salience into investigation. However, Woodin et al. (2023) mainly focused on the synchronic aspects of number use (or a short period of time of number use from 1991 to 1994, the collation of the BNC). Thus far, limited studies have investigated the diachronic changes in number use, even though it may serve as a revealing lens through which to observe and analyze shifts or movements in the writing system or within a society.

2.2. Diachronic studies on number use

As a case study to demonstrate that there is a variety of strategies in expressing numbers in writing, Chrisomalis (2020) investigated how the written forms used to express the number 1,200,000 have changed over time. He examined the frequency of its various notations in the Google Books corpus from 1800 to 2000. Chrisomalis (2020) found that the format “twelve hundred thousand” was the most frequently used throughout the nineteenth century. However, people dominantly used the format “1.2 million” from 1940s, while other potential formats of this number almost vanished.

Chrisomalis (2020) proposed several factors that could account for this change. One such factor is the substantial inclusion of scientific texts in the Google Books corpus from the 1900s (Pechenick et al., 2015). Scientific writing may prefer the concise form of numbers due to word limits in journal articles. Additionally, changes in formal style guides may influence the format of numbers presented in books. These guides evolve to meet the changing needs of society. Compared to a century ago, numerical values exceeding 100 million might become more common in daily life. For instance, with the rapid population growth in the past century, especially in urban centers, there are more references to large numbers in discussions about population statistics, city planning, and infrastructure needs. As a result, there is a pressing need for updated formal style guides to reflect these evolving societal needs, thereby influencing the formatting of numerical data in books.

Instead of examining the format changes of one number, Berg and Neubauer (2014) focused on the diachronic changes in structural patterns of number words. They examined four patterns of numbers 21 to 99, namely the unit-and-ten pattern, the ten-and-unit pattern, the ten-before-unit pattern, and the unit-before ten patterns, across six corpora from 1100 to 1914. For instance, the number 33 could be represented as *three and thirty* in the unit-and-ten pattern, *thirty and three* in the ten-and-unit pattern, *thirty-three* in the ten-before-unit pattern, and **three thirty* in the unit-before-ten patterns. The six corpora span from Middle English to Late Modern English and include the Innsbruck Corpus of Prose (1100–1500), Helsinki Corpus (1150–1500), Innsbruck Corpus of Letters (1386–1688), Helsinki Corpus (1500–1710), Lampeter Corpus (1640–1740), and Penn Parsed Corpus (1700–1914).

Berg and Neubauer (2014) identified a change trajectory from the unit-and-ten to the ten-and-unit and eventually to the ten-before-unit pattern. They suggested that this change may be a contact-induced phenomenon, particularly influenced by language contact with Norman French. They proposed that the ten-and-unit pattern in Middle English could be viewed as directly influenced by the Old French pattern, where the tens preceded the units with a connecting element used (e.g., *vint et trois* for 'twenty-three'; Kibler, 1984: 193).

Furthermore, Berg and Neubauer (2014) argued that this change may have been facilitated by a reduction in mental cost. For a listener, a consistent higher-before-lower order (e.g., ten-before-unit pattern) allows for a more efficient comprehension process compared to a lower-before-higher order (e.g., unit-and-ten pattern) in multidigit numbers (Greenberg, 1978: 274). This efficiency stems from the listener's ability to derive a reasonable approximation of the number value based on the beginning of the number word, thereby reducing the overall mental processing cost.

Berg and Neubauer's (2014) study is significant in inquiring into the nature of the change in the numeral system. However, similar to Chrisomalis (2020), such a study primarily focused on the changes of individual numbers or specific patterns (such as the unit-and-ten to the ten-before-unit pattern). Neither study provided a comprehensive overview of the diachronic changes in number patterns. In addition, neither study focused on the cultural salience of numbers from a diachronic perspective. By examining the use of culturally salient numbers from a diachronic perspective, we may expect to reveal the changes of cultural or social movements.

Thus, the present study aims to depict the diachronic aspect of number use in written American English. Specifically, this study probes into the effect of roundness, magnitude, and cultural salience on number use by examining the frequency data of numbers ranging from 1 to 1 billion from 1920s to 2000s. The corpus adopted was the TIME Magazine corpus (Xu, 2015). Moreover, the present study investigated the use of different formats of numbers, including numerals (e.g., '11', '999'), number words (e.g., 'nine', 'one hundred'), and mixed words (e.g., '10 million', '1 billion'), across various magnitudes and decades. Specifically, we are going to answer the following four questions:

1. What are the most frequently used numbers across each decade from 1920s to 2000s in the TIME Magazine corpus?
2. Does the effects of magnitude and roundness on the number frequency change across decades? If so, how?
3. How do culturally salient numbers change over time?
4. Does representational format of numbers (i.e., numerals, number word, and mixed word) change across magnitude and decades? If so, how?

3. Methodology

3.1. Corpus data

To probe into diachronic changes in number use in written American English, we adopted the TIME Magazine Corpus (Xu, 2015). There are mainly three reasons for choosing this corpus.

Firstly, the TIME Magazine Corpus is large, encompassing over 170 million words, nearly twice the size of the BYU TIME magazine corpus (Davies, 2007), which comprises about 100 million words. The sheer size of the corpus enables a more thorough examination of number use in written American English.

Secondly, this corpus contains all issues of TIME Magazine issues, from its inception in March 1923 to October 2008. This comprehensive coverage provides a continuous diachronic framework for observing the evolution of number use over distinct periods in time.

Thirdly, TIME Magazine, as a prominent news outlet in America, addresses a broad array of subjects of widespread significance and interest, ranging from national and international affairs to education, business, religion, and culture (Granath and Ullén, 2019). Consequently, the TIME Magazine Corpus offers valuable insights into the shifts in society and significant social events, particularly within the American context, by examining diachronic changes in number use.

Table 1 provides the statistical overview of this corpus. To explore changes in number use over decades, we combined the texts from the same decade. For instance, the texts from 1931 to 1940 were aggregated into the subcorpus of the 1930s. The inaugural issue of TIME was published on Mar. 3, 1923. Thus, there are 26 fewer texts in the first decade compared to subsequent decades (24 texts for 1921 and 1922, and 2 texts for 1923). Additionally, the corpus compilation concluded in October 2008. Accordingly, there were also 26 fewer texts in the 2000s than in other decades (24 texts for 2009 and 2010, and 2 texts for 2008).

It is worth noting that the TIME Magazine Corpus exclusively encompasses the content of TIME Magazine, which possesses some specific characteristics. For instance, it comprises written rather than oral texts, reflecting highly edited, literate work adhering to a style guide, and represents a specific genre (news media). Therefore, the number use in this corpus might also be influenced by these conditions. Furthermore, the target readership of TIME Magazine is predominantly educated individuals from the middle and upper classes in America. Accordingly, the cultural salience reflected in number use in this magazine may also mirror the culture of this particular demographic to some extent.

Table 1
Descriptive statistics of the TIME Magazine Corpus.

Decades	Tokens	No. of texts
1920s	11,559,014	94
1930s	18,560,904	120
1940s	19,744,345	120
1950s	22,871,494	120
1960s	24,499,012	120
1970s	22,657,436	120
1980s	21,866,860	120
1990s	18,875,032	120
2000s	13,705,186	94
	174,339,283	**1028**

3.2. Data processing

In the present study, we replicated the pioneering work conducted by Woodin et al. (2023: 10–13) with regard to number extraction, number cleaning, and number translation. Woodin et al.'s (2023) original scripts can be accessed on the OSF repository: https://osf.io/ze9vk/. In addition, we made some adjustments to these scripts, especially in terms of number cleaning and number translation. Specifically, we covered more classifications in reidentifying valid numbers within NA data and in removing invalid numbers.

3.2.1. Number extraction

There are three types of numbers: cardinal, ordinal, and nominal (Nieder, 2005; Wiese, 2003). Cardinal numbers express quantities or numerosity, for example, the count of oranges placed on a table. Ordinal numbers denote a position or rank within a sequence, such as ranking the top five students in a class. Nominal numbers serve as identifiers or labels that do not convey numerical value or sequence, such as telephone numbers or bus route numbers. The present study primarily focuses on cardinal numbers because they represent the most basic and prototypical form of number sense by denoting precise quantities (Greenberg, 1978). Specifically, we examined numbers greater than or equal to 0, including both integers and decimals. Fractions were excluded from our investigation due to the difficulty in distinguishing them as cardinal numbers. For instance, expressions like '09/12' could be interpreted as a date or a fraction. Moreover, fractions could not be translated into numerals using the Word2Num library in Python, thus rendering them unsuitable for further data analysis.

We utilized the NLTK Python library to tag words and identify numbers within texts. This tool enabled us to distinguish between pronominal uses of the number word 'one' (for example, in the phrase 'as one does') and its numerical applications (as in 'one man') (Woodin et al., 2023). However, NLTK faced limitations in recognizing multi-word numbers like "twenty five", as it tags words separately. Following Woodin et al. (2023), we combined words tagged as numbers that appeared consecutively within the text. Additionally, we treated hyphenated numbers (e.g., 'thirty-one') as single units and recognized compound numbers involving conjunctions (such as "and" or "point") in forms like "three hundred and seven" and "five point three". This number identification process yielded 3,665,127 potential numbers.

3.2.2. Number cleaning

After an initial review of the identified potential numbers, we found several incorrectly identified number groups that needed to be excluded. For instance, there were mixtures of numerals and words or characters such as ordinals (e.g., '13th' and '51st'), times (e.g., '5am', '5s', '1960s', and 'June 11'), temperature degrees ('2500DEG') that were not numerals. Accordingly, we removed 57,910 such relevant items. In addition, words with ambiguous meanings across different types of numbers were excluded. For example, the number word "ten thirty" could refer to a time (e.g., '10:30') or price (e.g., '$10.30'). These ambiguous number words amounted to 46 items. We also excluded 2869 numbers beginning with 0, like '0304' and '0437–5439', as they may represent dates or telephone numbers. Furthermore, we removed 9158 erroneously coded string items such as 'learner' and 'alone'. Finally, we obtained a count of 3,592,518 numbers after the number cleaning process.

It should be noted that despite the meticulous number cleaning process, the dataset may still contain non-cardinal numbers like ordinal and nominal numbers. Due to the extensive size of the dataset, it is practically impossible to exclude every non-cardinal number. This challenge has also been encountered in prior studies on number word and numeral frequencies (e.g., Jansen and Pollmann, 2001; Woodin et al., 2023). Similar to Woodin et al. (2023), the present study acknowledges the inclusion of non-cardinal numbers in our dataset and utilizes them to examine the effect of cultural salience on number frequency.

3.2.3. Number translation

For data analysis purposes, it was necessary to convert number words into numerals. We adopted the Word2Num library in Python for this translation process. Then, we utilized the Word2Num library to back-translate. If the back-translated word matched the original number word, we considered the translation accurate. It is important to note that Word2Num cannot process mixed numbers that include a numeral combined with a multiplier word, such as '5 million'. Thus, we manually translated these mixed numbers into numerals in R (e.g., '5 million' became '5,000,000').

In addition to the mixed numbers, some number groups still could not be translated. As suggested by Woodin et al. (2023), these were often not single numbers but combinations of several numbers (e.g., ‘9.50 and 9.50’ = ‘9.50’ and ‘9.50’). These multi-number items occurred due to our previous steps on tagging sequences of words as numbers when they appeared consecutively in the text. To cover as many valid numbers as possible, we coded the structure patterns of these numbers in R. Specifically, within each multi-number item we coded numeral as “DIG” and number word as “NUM”. For instance, the multi-number ‘three and 17’ was coded as ‘NUM and DIG’, while ‘25 and 50 million’ is coded as ‘DIG and DIG million’. We focus on patterns with a frequency greater than or equal to 20, as manual translation of all patterns was infeasible. Accordingly, there are 42 relevant structure patterns covering an additional 21,869 multi-number items. The full list of these structure patterns has been provided in Table A.1. After the manual translation, 58,197 items still could not be translated and were excluded. In addition, we removed 58 numbers larger than a billion from our data, as very large numbers are difficult to handle computationally and less meaningful for this study.

Finally, we obtained 3,500,771 number words, accounting for 2% of the total words (174,339,283) in the TIME Magazine Corpus.

3.3. Data analysis

In addressing the first research question to identify the most used numbers across each decade from 1920s to 2000s in the TIME Magazine corpus, we initially calculated the frequency of numbers within each decade. Subsequently, we determined the normalized frequency of each number by dividing its raw frequency by the corpus size of its decade.

For the second research question, we used Bayesian negative binomial regression to explore the influences of magnitude and roundness on number frequencies across the decades. This approach mirrored the data analysis methodology outlined by Woodin et al. (2023), a choice guided by the shared characteristics between our study and theirs. Specifically, our count data displayed more variability than expected under a Poisson distribution.

Following the procedures outlined by Woodin et al. (2023), we incorporated numbers absent from the TIME magazine corpus into the regression. For these numbers, we assigned a frequency of 0, as they indeed possess a frequency of zero within the corpus. Additionally, we excluded the number 0 from our analysis, as logarithmically transforming zero is unfeasible. To reduce computational complexities and processing time, we set an upper limit to the number 1 million. Furthermore, we limited our analysis to integer numbers in our dataset, as non-integer values do not possess roundness properties, which is one of the dependent variables under investigation.

Regarding statistical priors, we adhered to the default settings provided by the R package brms (Bürkner, 2017) for intercept and standard deviation. For the independent variable slopes, we opted for weakly informative priors, utilizing a normal distribution with a mean of 0 and a standard deviation of 0.5. This choice introduces a degree of “mild skepticism” into our analyses (McElreath, 2016), which gently biases the slope estimates towards zero. As a result, our findings lean towards a more conservative interpretation compared to those derived from a corresponding frequentist model. These decisions are consistent with the methodology employed by Woodin et al. (2023).

In our model, the dependent variable was the frequency of numbers, while the independent variables included Log 10 Number Magnitude and six roundness properties (i.e., Multiple of 5, Multiple of 10, 10-ness, 2-ness, 2.5-ness, and 5-ness). Notably, Variance Inflation Factors (VIFs) for the roundness properties in a linear regression model were all under 3, suggesting that collinearity was not a concern (Winter, 2019). The following outlines the standards for determining the six roundness properties.

Specifically, numbers that are divisible by 5 or 10 without remainder are considered to possess the roundness properties of Multiple of 5 or Multiple of 10, respectively. It is evident that some numbers can concurrently exhibit both traits, as exemplified by numbers such as 10 and 20.

To determine the other four roundness properties (i.e., 10-ness, 2-ness, 2.5-ness, and 5-ness), we adopted the formulas provided by Jansen and Pollmann (2001), which are presented in Formula 1. Take the third formula about 5-ness as an example, it indicates that a number has the 5-ness property if it can be expressed as $5 * (10^n) * x$, where n is a non-negative integer, and x is an integer between 1 and 9 inclusive. In other words, the number is considered to have 5-ness if when divided by 5 multiplied by a power of 10, the result is an integer between 1 and 9. For instance, the number 3000 has 5-ness, as 3000 equals 6 when divided by $5 * (10^2)$. The formulas for 10-ness, 2-ness, and 2.5-ness can be interpreted similarly.

Formula 1

10-ness: ‘is the number contained in the set $[1 * (x * \mathbf{10^n})]$?’
2-ness: ‘is the number contained in the set $[2 * (x * \mathbf{10^n})]$?’
5-ness: ‘is the number contained in the set $[5 * (x * \mathbf{10^n})]$?’
2.5-ness: ‘is the number contained in the set $[2.5 * (x * \mathbf{10^n})]$?’
Notes: $n \geq 0$; $1 \leq x \leq 9$; both x and n are integers (Jansen and Pollmann, 2001: 198)

As for the third research question to identify the potential culturally salient numbers, we examined the residuals of the binominal regression model utilized in our second research question. These residuals represent the discrepancy between the actual frequencies of numbers and the predictions of the model. A high residual for a number in this model indicates a significant deviation that cannot be accounted for by either magnitude or roundness. Thus, numbers with high residuals might be indicative of cultural salience.

To calculate the residuals for each number, we first obtained the predicted frequency of each number using the binominal regression model. The residual of a number was then calculated as the absolute difference between the predicted frequency and the actual frequency. We subsequently ranked the top 10 numbers in residuals within each decade. According to Woodin et al. (2023), culturally salient numbers related to year information have frequently appeared in the BNC. To further explore other potential culturally salient numbers, we excluded numbers that might be related to years, specifically those between 1000 and 2010. This range was chosen because the TIME Magazine Corpus used in the present study includes texts published from 1923 to 2008. After this exclusion, we listed the top 10 potential culturally salient numbers with normed frequencies exceeding 10 in each decade. We set the threshold at 10, as we believe that a frequency too small may introduce too much noise into the analysis. This approach allows us to focus on the most significant numbers while minimizing the impact of outliers or infrequent occurrences.

Then, we utilized WordSmith Tools (v8; Scott, 1996) to extract the concordance lines associated with the identified potential culturally salient numbers. This corpus tool was employed due to its capacity to handle extensive corpora. Its convenient feature of producing concordance lines significantly contributed to the efficiency of our research process. Guided by the insights gleaned from the concordance lines, we performed a comprehensive classification of the culturally salient numbers. Simultaneously, we cross-verified the identified numbers against the concordance lines to exclude those lacking evident cultural significance, despite exhibiting high residuals. Notably, instances such as the code number for TIME's customer service '8463' and misspelled year numbers like '19,303' (with the correct representation being '1930/3'), were eliminated through this rigorous validation process.

As for the fourth research question to investigate the effects of magnitude and decade on the use of number formats, we calculated the proportion of each number format (i.e., number word, numeral, mixed) within each order of magnitude across the nine decades from 1920s to 2000s. Specifically, we categorized all numbers examined in this study (ranging from 0 to 10 billion) into nine orders of magnitude or Log_{10} number ranges. These ranges are as follows: 1–9 as the first order of magnitude (i.e., $\text{Log}_{10}\ 0 - \text{Log}_{10}\ 1$), 10–99 as the second order of magnitude (i.e., $\text{Log}_{10}\ 1 - \text{Log}_{10}\ 2$), 100–999 as the third order of magnitude (i.e., $\text{Log}_{10}\ 2 - \text{Log}_{10}\ 3$), and so on, up to 1 billion–10 billion (i.e., $\text{Log}_{10}\ 8 - \text{Log}_{10}\ 9$) as the ninth order of magnitude.

4. Results

4.1. Most frequently used numbers across decades

To depict the most used numbers in TIME magazine from 1920s to 2000s, we conducted a ranking based on frequency across the nine decades. Fig. 1 shows the top 10 numbers in frequency for each decade. It is worth noting that we used the normed frequency as the y axis to ensure comparability across decades. Furthermore, the points in Fig. 1 are color-coded to represent different types of numbers: red denotes numbers in the range of 1–9, blue designates round numbers like 10, 20, and 30, while black signifies other numbers.

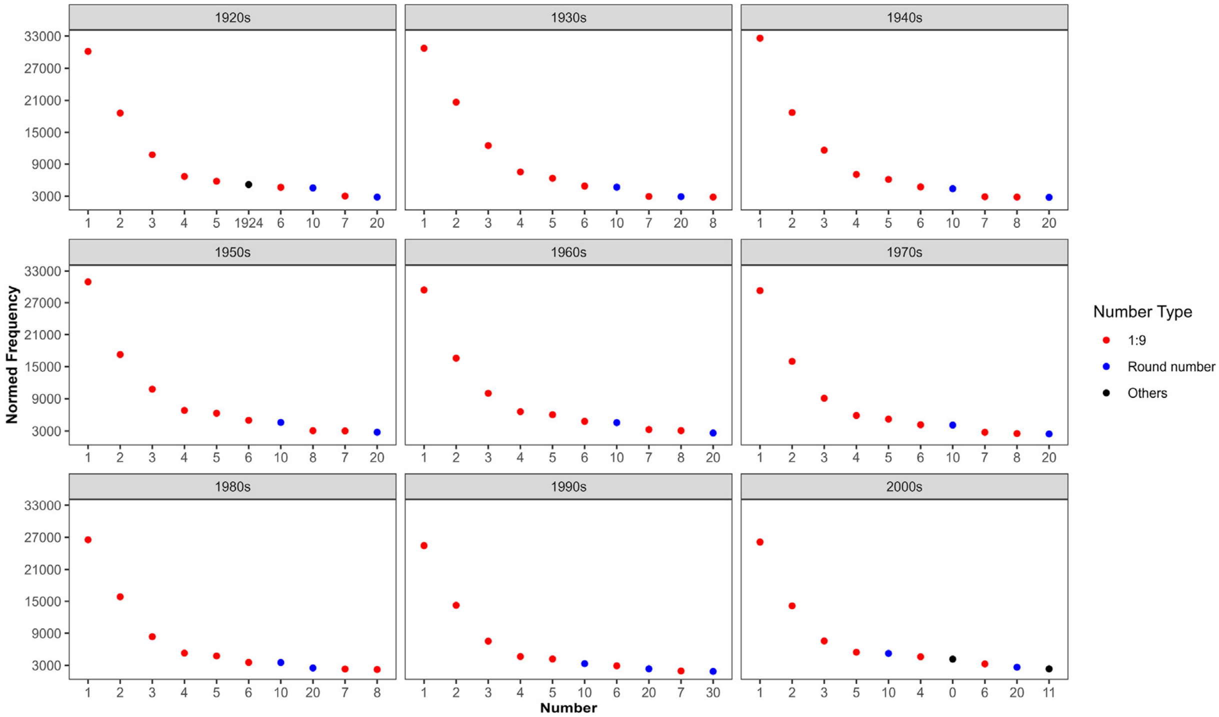

Fig. 1. Top 10 numbers in normed frequency across each decade.

Fig. 1 illustrates that compared with numbers in larger magnitudes, those in smaller ranges, particularly 1 to 8, consistently rank among the top 10 most frequently used across the nine decades. Specifically, numbers 1 to 3 have consistently held the top three positions from 1920s to 2000s. These findings indicate that numbers of lower magnitude are generally more commonly used than their higher counterparts.

Additionally, Fig. 1 shows that numbers with roundness properties (i.e., 10, 20, 30) consistently maintain positions in the top 10 most frequently used list, especially for numbers 10 and 20. In contrast, there is a noticeable absence of numbers in close proximity to these round numbers. For example, numbers 19 and 21 do not appear in the top 10 list, which are near the round number 20. This finding suggests that compared with unround numbers, round numbers of a similar magnitude are employed more frequently. It is worth noting that numbers 10 and 20, the two smallest multiples of 10, are frequent, which might also be attributed to their low magnitude.

4.2. Statistical models of roundness and magnitude on number frequency over decades

Table 2 shows the estimates for Log 10 Number Magnitude and the six roundness properties on predicting number frequency across the nine decades from 1920s to 2000s. For more detailed information, including estimates, standard errors, and credible intervals, please refer to Table A.2.

Table 2
Estimates for Log 10 Number Magnitude and six roundness properties in predicting number frequency across decades.

	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
(Intercept)	15.07	14.82	14.66	15.59	15.75	15.9	15.58	13.51	16.58
NumberLog	−4.50	−4.40	−4.42	−4.68	−4.75	−4.80	−4.67	−4.05	−4.95
TenNessyes	4.29	4.18	4.48	4.32	4.49	4.28	4.30	4.17	4.42
TwoPointFiveNessyes	3.91	3.86	3.91	3.57	3.86	3.68	3.56	3.41	3.36
FiveNessyes	2.98	3.07	3.23	3.02	3.20	2.99	2.97	3.13	3.04
TwoNessyes	2.16	2.23	2.29	2.06	2.19	2.01	2.11	2.36	2.30
Multiple5yes	0.39	0.38	0.42	0.67	0.45	0.42	0.32	−0.06	0.33
Multiple10yes	4.70	5.00	4.93	5.21	5.42	5.60	5.15	3.86	5.05

Table 2 shows a consistent negative correlation between Log_e Frequency of numbers and Log 10 Number Magnitude across all nine decades investigated in the present study, as all the estimates for Log 10 Number Magnitude are negative as well as confidence interval well below zero. In contrast, there is a general positive correlation between Log_e Frequency of numbers and roundness properties. Specifically, numbers with roundness properties tend to be used more frequently than those without, as most estimates for roundness properties are well above zero.

It is worth noting that the roundness property Multiple of 5 is the least predictive one among all six roundness properties. In eight of the nine decades, the estimates of Multiple of 5 are slightly above zero, while the estimate for 1990s is even below zero ($\beta = -0.06$, 95% CI = [−0.28, 0.15]).

4.3. Culturally salient numbers across decades

To investigate the potential culturally salient numbers, we calculated the residuals of all numbers adopting the binominal regression model. A high residual value for a number indicates a significant deviation that cannot be accounted for by either magnitude or roundness properties. Thus, numbers with high residuals might be indicative of cultural salience. Table 3 displays the top numbers based on residuals in each decade from 1920s to 2000s. Values in parentheses refer to their normed frequencies.

As shown in Table 3, the top residual numbers for each decade predominantly correspond to year numbers around that specific period. For instance, the largest residuals in 1970s all fall between 1968 and 1976. This result indicates a potential recency effect that people are inclined to discuss dates (e.g., recent years) that are pertinent to present discussions (Pollmann and Baayen, 2001).

Table 3
Numbers with top 10 residuals in each decade.

1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
1924 (5145)	1932 (1556)	1941 (1367)	1952 (1368)	1964 (1313)	1972 (1598)	1982 (1569)	1992 (1325)	2008 (1498)
1928 (1385)	1929 (1554)	1942 (1268)	1956 (1343)	1962 (1132)	1973 (1328)	1981 (1516)	1994 (1263)	2001 (1460)
1923 (1340)	1933 (1537)	1940 (1330)	1954 (1254)	1961 (1108)	1974 (1163)	1984 (1484)	1993 (1053)	2004 (1268)
1925 (1256)	1934 (1420)	1939 (981)	1953 (1150)	1963 (1044)	1976 (1150)	1980 (1678)	1996 (1047)	2002 (1263)
1926 (1122)	1935 (1309)	1943 (919)	1955 (1131)	1965 (945)	1971 (1093)	1983 (1331)	1990 (1029)	2003 (1249)
1922 (1052)	1931 (1270)	1944 (885)	1951 (1044)	1968 (849)	1968 (1067)	1985 (1200)	1991 (968)	1999 (960)
1927 (1043)	1936 (1255)	1946 (876)	1948 (981)	1960 (1250)	1975 (1036)	1979 (1142)	1995 (852)	2005 (899)
1929 (925)	1930 (1362)	1948 (812)	1957 (939)	1966 (843)	1969 (946)	1986 (1084)	1989 (844)	1998 (825)
1920 (1126)	1937 (1002)	1945 (767)	1949 (926)	1967 (790)	1970 (1283)	1978 (822)	1988 (794)	2006 (782)
1921 (908)	1928 (898)	1947 (729)	1958 (914)	1959 (709)	1967 (768)	1988 (782)	1997 (649)	1996 (581)

To further explore culturally salient numbers, we excluded those numbers that might be associated with years from the high residual numbers and then ranked the top 10. By resorting to the concordance lines of these numbers, we further determined their cultural meanings.

Table 4 shows the major culturally related meanings associated with each number in the top 10 potential culturally salient list without potential year-related numbers, all of which exhibit a normed frequency surpassing 10. Here by the "major cultural meanings", we indicate that the frequency of that particular meaning should not be the pure quantity meaning and the frequency of that meaning should be above 5. In addition to the major meanings, we have also provided their raw frequencies followed by a "*N*" in Table 4.

Table 4
Potential culturally salient numbers with meanings in each decade.

Decades	Culturally salient numbers
1920s	999: Lease duration $N = 7$
1930s	999: Miscellaneous $N = 7$; Lease duration $N = 5$; Ford 999 (racer) $N = 8$
	776: Section 776 of the Bankruptcy Act $N = 23$
	534: *Queen Mary* (ship) $N = 38$
1940s	777: *Call Northside 777* (movie) $N = 18$
1950s	707: Boeing 707 jet $N = 142$
1960s	707: Boeing 707 jet $N = 184$
	727: Boeing 727 jet $N = 66$
	747: Boeing 747 jet $N = 56$
1970s	747: Boeing 747 jet $N = 268$
	707: Boeing 707 jet $N = 190$
	727: Boeing 727 jet $N = 91$
	767: Boeing 767 jet $N = 20$
1980s	747: Boeing 747 jet $N = 200$
	737: Boeing 737 jet $N = 86$
	727: Boeing 727 jet $N = 81$
	847: Trans World Airlines Flight 847 (Hijacking) $N = 68$
	707: Boeing 707 jet $N = 70$
	572: 572 new nuclear missiles (1979/12) $N = 59$
	767: Boeing 767 jet $N = 47$
1990s	401: 401(k) plan $N = 131$
	911: Emergency service number $N = 95$
	486: RU-486 medication for medical abortion $N = 104$; microprocessor $N = 11$
	747: Boeing 747 jet $N = 58$
	737: Boeing 737 jet $N = 21$
2000s	401: 401(k) plan $N = 346$
	911: Emergency service number $N = 98$
	747: Boeing 747 jet $N = 60$
	527: Political committees $N = 18$
	512: Computer memory $N = 9$
	787: Boeing 787 jet $N = 34$
	529: 529 college savings plan $N = 51$
	757: Boeing 757 jet $N = 26$

The culturally salient numbers in Table 4 can be categorized into four groups. Firstly, there are numbers that signify general social services, such as the well-known U.S. emergency service number '911', which was frequently referred to in TIME magazine during 1990s and 2000s.

The second group encompasses certain social entities that were quite popular at that time, such as cars, movies, and ships. Specifically, the surge in use of Ford 999 (racer) and *Queen Mary* (ship) with 534 in 1930s stands out. The *Queen Mary*, launched in 1934, was a British response to express superliners built by German, Italian, and French companies in the late 1920s and early 1930s. Another example is the movie *Call Northside 777*, which gained cultural salience in the 1940s. This 1948 reality-based newspaper drama tells the story of a Chicago reporter who proved a man jailed for murder was wrongly convicted 11 years prior.

The third group of culturally salient numbers pertain to significant social events or economic plans, for instance, the Hijacking event that happened in the Trans World Airlines Flight 847 in 1985, the deployment of 572 new nuclear missiles in Western Europe in 1979, 401(k) plan for pension, and 529 plan for education. Especially for 401(k) plan, it surged as the top culturally salient number in both 1990s and 2000s, indicating a heightened societal interest in pension-related matters. These culturally significant numbers related to economic plans may reflect the primary cultural values of TIME magazine's target readership—educated, middle and upper-class Americans, who place significant emphasis on retirement security and educational opportunities.

The fourth group signifies technological advancements, such as the introduction of 512 RAM of computer storage in 2000s and the 486-microprocessor in 1980s, the latter marking the first $\times 86$ chip family to exceed one million transistors.

Worth noting that, in addition to the four groups mentioned above, numbers related to certain Boeing jet consistently ranked among the top potential culturally salient numbers from 1950s. After closer examination in the concordance, it was found that the initial Boeing 707 first flew on December 20, 1957, which aligns with the first decade (i.e., 1950s) it appears in Table 4. The emergence of these airplane numbers may not be directly affected by cultural salience but rather result from the genre of our corpus, news media. On the one hand, the release of new airplane has always been an important source news, given its significant economic impact. The aviation industry plays a vital role in both the global and national economy, contributing to employment, tourism, and transportation. Furthermore, it may also serve as a form of advertisement, particularly for major aircraft manufacturers like Boeing. Reports or news about their new models could enhance brand visibility and customer engagement.

Furthermore, number 999 holds a special cultural significance. While it could have been rounded up to 1000, it emerged as a culturally salient number in both the 1920s and 1930s. After a close check of its concordance lines, we found that there is a shift in meaning from 1920s to 1930s. In 1920s, it was mostly used in phrases like “999-year lease” to emphasize the long-term nature of the agreement. For instance, the lease duration 999 in Example 1a implies that the lease was intended to last for an exceptionally extended period. Such long-term leases are often used in various legal contexts, including real estate and land agreements, to establish extended rights and obligations between parties over a substantial period. However, in 1930s, the number 999 took on additional culturally salient meanings. For instance, it served as a rhetorical device to emphasize the abundance or large quantity of something. As seen in Example 1b, it was used to indicate the large number of individuals who may potentially benefit from the preventive measures against cancer.

Example 1:

a: Together with 73.49 miles of single track, 0.34 miles of double track, a **999**-year lease dating from 1863, the Pennsy acquired one of the freak obligations of railroad finance.

b: Dr. Bloodgood believed with many another wise cancer specialist that it is worth scaring the wits out of **999** people in order to save the thousandth man from death by cancer.

In sum, these results indicate that numbers may serve as a tool for uncovering social movements and changes. Specific numbers in certain decades encapsulate the distinctive characteristics and significant social events of their time. While our research design may not precisely pinpoint the emergence of each culturally salient number across decades listed in Table 4, it effectively demonstrates the surge or burst in their use during specific periods.

4.4. Representations of numerical formats across magnitudes and decades

Fig. 2 illustrates the shifts in the prevalence of different numerical formats (i.e., word, numeral, and mixed) across each Log 10 number range and over nine decades. As displayed in Fig. 2, number 1–9 (Log 10 number range 0–<1) are predominantly represented by number words and less by numerals across all nine decades. Conversely, Numbers 10–99,999 (Log 10 number range 1–<6) are primarily denoted using numerals throughout the observed time span.

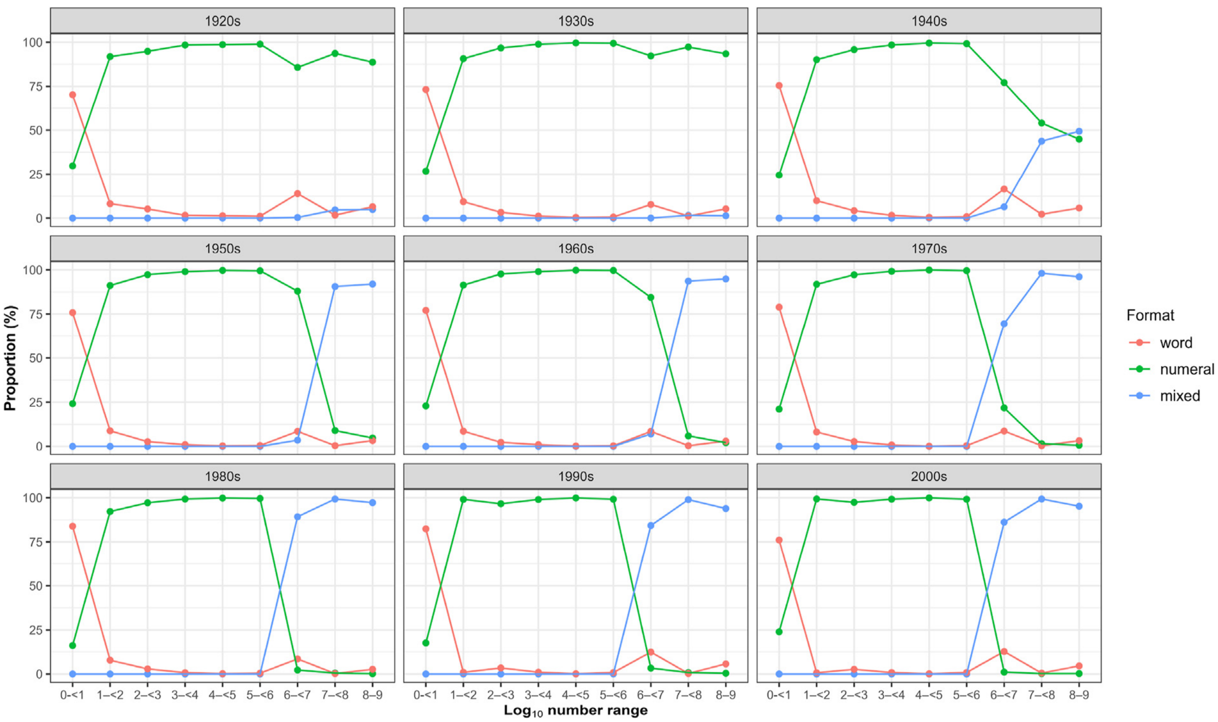

Fig. 2. Proportions of number formats across magnitude and decades.

A notable transition occurs with numbers between 1,000,000 and 9,999,999 (Log 10 number range 6–<7). Until 1960s, numerals were the dominant format. Post-1970s, however, mixed format representations have gained precedence, reaching a peak of 86.17% (3470 instances) in 2000s, followed by words at 12.76% (514 instances), and numerals at a mere 1.07% (43 instances).

For larger values, specifically those from 10 million to 1 billion (Log 10 number range 7–<9), there is also a distinct shift. In 1920s, numerals were the primary format (Log 10 number range 7–<8, 93.64%, 4709 instances; Log 10 number range 8–<9, 88.65 %, 3313 instances). Results in 1930s continued this trend with even higher proportions (Log 10 number range 7–<8, 97.26%, 8587 tokens; Log 10 number range 8–<9, 93.43 %, 5305 tokens). However, in 1940s, mixed format usage became more prevalent, especially for numbers in the 100 million to 1 billion range, where mixed formats surpassed numerals (49.39% vs. 44.88%). Furthermore, since 1950s the proportion of mixed surged, while the proportion of numerals decreased greatly to almost zero from 1970s. These results suggest that the use of different formats to represent numbers has experienced changes across decades to form a standard custom.

5. Discussion

5.1. Effects of magnitude, roundness, and cultural salience on number use

The present study demonstrates a consistent negative correlation between the magnitude of numbers and their frequency of use regardless of decades in American written English. This extends previous synchronic studies on number use (Coupland, 2011; Dehaene and Mehler, 1992; Jansen and Pollmann, 2001; Woodin et al., 2023) by adding a longitudinal perspective, revealing enduring patterns over a longer timeline. Their analyses, while insightful, were confined to shorter, specific periods. Our broader analysis thus offers a more comprehensive understanding of this phenomenon.

Moreover, our findings indicate that round numbers were more frequently used, consistent with previous studies (e.g., Dehaene and Mehler, 1992; Jansen and Pollmann, 2001; Woodin et al., 2023). Specifically, among the six roundness properties investigated in this study (i.e., 10-ness, 2.5-ness, 5-ness, 2-ness, Multiple of 10, and Multiple of 5), Multiple of 5 is the least predictive one concerning number frequency, which aligns with the findings of Woodin et al. (2023). This challenges the traditional view in numeracy research that all rounded numbers are equally preferred, suggesting a more nuanced interaction between number perception and usage. The lower effect of Multiple of 5 compared to other roundness properties may result from various reasons.

Firstly, we propose that the general rounding principles applied to decimal numbers might transfer to integer numbers ending with 5. Consequently, this could lead to a reduction in the frequency of numbers possessing Multiple of 5, affecting its predictability in terms of number frequency. Numbers with the roundness property, Multiple of 5, can be categorized into two groups: those ending with 0 (e.g., 10, 20, and 30) and those ending with 5 (e.g., 5, 15, and 25). The latter category may be influenced by general rounding principles such as round half up and round half to even.

Round half up is a common method taught in schools and often used in general arithmetic. According to this method, if the number to be rounded is exactly halfway between two numbers, it is rounded up to the next higher number. For example, 2.5 rounded to the nearest whole number using round half up would be 3. In contrast, round half to even method, often employed in financial calculations, rounds a number that is exactly at the halfway point to the nearest even number. For example, 2.5 rounded to the nearest whole number using round half to even would be 2, while 3.5 would be rounded to 4.

These two rounding methods might affect how people express integer numbers ending with five (e.g., 5, 15, and 25). With both methods, numbers like 15 might be rounded to either 10 or 20, leading to a lower occurrence of numbers ending in 5 than might be expected. Accordingly, such a rounding transfer to integers may influence the predictability of Multiple of 5 on number frequency, thus making its estimate on number frequency lower than other roundness properties.

Secondly, the calculation of Multiple of 5 property differs markedly from that of the other five roundness properties. Properties such as 10-ness, 2-ness, 2.5-ness, and 5-ness are derived from powers of ten, because their calculation formulas all contain $10^\wedge n$ with the integer n larger than 0. Multiple of 10 can be expressed similarly: $x*10^\wedge n$, with both integer x and n larger than 0. However, not all numbers possessing Multiple of 5 can be expressed using powers of ten. For instance, while the number 25 possesses Multiple of 5, it cannot be represented as $x*10^\wedge n$ with both x and n as integers larger than 0. If we assign the smallest possible integer to n (i.e., 1), we get $x = 2.5$, which is not an integer.

Jansen and Pollmann (2001) suggest that numbers associated with the base of a mathematical system (in this case, powers of ten) are more frequently used. This tendency is attributed to the natural human inclination to double or halve a base unit for the purposes of quantification and estimation. Thus, numbers possessing Multiple of 5 but unable to be expressed using powers of ten might be used less frequently than numbers possessing other roundness properties that can all be expressed with powers of ten.

As for culturally salient numbers, our results reveal a strong tendency for numbers associated with years to be frequently used within their corresponding decade. For instance, during 1940s numbers falling within the range of 1940 and 1949 were notably prevalent. This observation aligns with prior studies (Pollmann 1998; Pollmann and Baayen 2001; Woodin et al., 2023). Woodin et al. (2023) reported that the top 10 culturally salient numbers in the BNC were concentrated between 1984 and 1993, which might correspond to the years closest to the completion of the corpus (i.e., 1994). The inclination towards recent years likely stems from their relevance to present discussions (Pollmann and Baayen, 2001). Furthermore, cognitive limitations make it more convenient for individuals to establish connections between the present and the recent past, as opposed to the distant past (Pollmann and Baayen, 2001).

Additionally, different from Woodin et al. (2023), the prevalence of numbers associated with specific years within their respective decades may also be attributed to the genre of our corpus, namely news media. Timeliness is a core value in news media (Bednarek, 2016; Caple and Bednarek, 2013), which underscores the importance of promptly delivering information, particularly regarding current events and developments. Accordingly, the prominence of year-related numbers in each decade may directly correlate with significant issues or events reported during that decade in TIME magazine.

In addition to years, our study identified other groups of culturally salient numbers including references to social entities (e.g., the movie *Call Northside 777*), economic plans (e.g., 401(k) plan), and technology related entities (e.g., 486-microprocessor). These numbers emerged and faded over time alongside societal advances. As Woodin et al. (2023: 24) argue, cultural evolution may lead to shifts in the inclination to communicate about various numbers, as well as changes in the ways these numbers are communicated. Our findings suggest, however, that the prevalence of numbers is not solely dictated by cultural dynamics but is also affected by economic, technological, and political factors. For instance, new economic policies, innovations, and significant political events may introduce novel salient numbers that reflect broader societal changes. In this way, both the cultural evolution and the significant events within other societal domains play a pivotal role in shaping the salience of specific numbers and the way they are utilized.

It is worth noting that given the diachronic corpus we utilized, which is based on TIME magazine, a representative of the news media genre, the culturally salient numbers outlined in this study are largely related to this genre. In other words, these numbers are closely associated with the principal subjects that news media typically cover or disclose, such as pressing societal issues and novel products during specific periods. Additionally, these numbers also mirror the cultural values of the target readership of TIME magazine, namely middle and upper-class Americans at specific times. However, considering the wide range of topics covered by TIME magazine and its significant influence in American society, we contend that it still provides a unique and valuable lens for exploring cultural dynamics.

5.2. Changes of number format across decades

Our results reveal a discernible shift in the preferred formats for representing numbers 10 million-1 billion from 1920s to 2000s in the TIME magazine corpus. Preceding 1930s, these numbers were predominantly represented using numerals, but a notable transition emerged in 1940s, characterized by a balanced use of both numerals and mixed numbers. Subsequently, mixed numbers became the dominant representative format, aligning with the recommendations found in contemporary writing style guides, such as American Psychological Association (2022) and U.S. Government Publishing Office (2017). This trend mirrors findings from the synchronous study by Woodin et al. (2023), which underscored that numbers 10 million–1 billion were commonly expressed by a combination of numerals and multiplier words.

The progression toward a mixed format indicates a gradual evolution in the composition style pertaining to numerical representation. One possible reason for this shift could be attributed to changes in government style guidelines for expressing large numbers across different versions of style guides published over periods. Specifically, before 1950s, Government Publishing Office (1933: 66) stated, “In expressing large numbers, the word *million* (or a similar group term) **may be** spelled out”. However, in the revised edition in 1953, Government Publishing Office (1953) changed the phrase “may be” to “should be” as “In expressing large numbers, the word million (or a similar larger group term) **should be** spelled out”. This shift suggests a transition from possibility or uncertainty to recommendation, expectation, or obligation, which may account for the dominant use of mixed form from 1950s.

Furthermore, the practical constraints of print media may play a critical role. Publications like TIME magazine, with wide readerships and space limitations, would have sought to economize on the physical space used by long numbers. In this context, using a mixed format allows for more information to be conveyed within the same column inch, a consideration that would have been particularly important in an era when print media was a primary source of information dissemination.

Such a shift in format may also stem from changes in editorial practice, notably by T. S. Matthews, who served as managing editor of TIME magazine from 1943 to 1949 and later as its editor until 1953. Matthews has been noted for prioritizing readability over objectivity (Swanberg, 1972). This editorial emphasis likely prompted a transition in the number format used in TIME magazine during 1940s and 1950s, from numeral format to a mixed format. The rationale behind this transition lies in the recognition that presenting large numbers solely in numeral format can pose readability challenges. The cognitive load theory, as proposed by Sweller (1988), suggests that individuals have a limited capacity for processing new information. Consequently, a mixed format, which combines numerals with familiar words, likely reduces the cognitive burden associated with parsing and comprehending large numbers. This reduction in cognitive load enhances readability and improves recall for readers.

6. Conclusion

Our study found consistent effects of magnitude and roundness on number use across decades in American written English; numbers with smaller magnitudes and more roundness properties tended to occur more frequently. Furthermore, consistent with Woodin et al. (2023), we noted that Multiple of 5 is the least predictive one in terms of number frequency among all six roundness properties investigated in the present study. We proposed that this could be due to a potential transfer of general rounding principles from decimal system to integers ending in 5. In addition, it may result from that Multiple of 5 is not directly related to powers of ten, the base of the mathematical system investigated in the present study.

In addition, we found that numbers indicating recent years account for the largest proportion of culturally salient numbers, which aligns with Woodin et al. (2023). We took a further step to identify distinct groups of culturally salient numbers reflecting, for instance, social entities, events, economic plans, and technologies in a certain decade. Accordingly, we proposed that culturally salient numbers serve as a great window for revealing social movements, changes, and broader societal development.

Furthermore, by examining representational formats of numbers across magnitudes and decades, we found that before the 1940s, extremely large numbers were represented as numerals (e.g., 60,000,000,000) rather than the widely accepted mixed forms (e.g., 6 billion) nowadays in the TIME magazine corpus. This indicates standardized number writing conventions were not established early on. In other words, adopting recommendations from style guides like APA 7 emerged gradually over time. Such transformation may potentially be driven by various reasons such as new standards from governments and changes in editorial practices.

Some limitations should be acknowledged. First, the research, based solely on the TIME magazine corpus, which is the product of highly edited, literate work to a style guide and a particular genre (i.e., news media). Therefore, our findings might not represent a wider number use across various genres and registers. Future studies could expand this research by utilizing diverse corpora, such as the Corpus of Historical American English (COHA), to explore general number use patterns. Additionally, our study primarily focused on the use of numbers in American English. Future research could extend this investigation to a wider cross-linguistic and cross-cultural context. This could further enrich our understanding of how culturally salient numbers reflect societal and cultural dynamics.

Funding

This work is supported by the General Project of Shanghai Education Science Research (No. C2024197), Chinese National Funding of Social Sciences (No. 23&ZD320), and Innovation Team for Comparative Linguistics and International Communication of Chinese of Shanghai Normal University (202106140059).

Declaration of competing interest

None.

CRediT authorship contribution statement

Gui Wang: Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Jing Shu:** Software, Methodology, Data curation. **Li Wang:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

Data availability

The authors do not have permission to share data.

Appendices

Table A.1
Full list of structure patterns of multi-number items for manual translation.

Structure patterns	Examples	Frequency
DIG and DIG million	25 and 50 million	20
DIG million and DIG million	10 million and 12 million	25
DIG–DIG million	60–100 million	7,258
NUM and NUM million	Two and three million	538
DIG and DIG	14 and 17	59
DIG and DIG DIG	25 and 26 1000	480
DIG and DIG.	1885 and 1886.	153
DIG DIG and DIG	15 16 and 17	86
DIG DIG DIG and DIG	1916 1917 1918 and 1921	21
DIG DIG.	28 1920.	27
NUM and DIG	Three and 17	362
NUMteen and DIG	Twelve and 20	54
DIG and DIG NUM	1847 and 1848 two	39
DIG and NUM	1922 and three	40

(continued on next page)

Table A.1 (*continued*)

Structure patterns	Examples	Frequency
DIG DIG NUM	10 1922 three	630
DIG NUM	1843 two	76
DIG NUMteen	1898 eleven	40
NUM DIG	Three 12,000	20
NUMteen DIG	Fifteen 38	56
DIG thousand	320 thousand	55
DIG-	1854-	9860
DIG and DIG.DIG	1922 and 25.6	23
DIG DIG DIG.DIG	30 1924 21250812989.49	33
DIG DIG.DIG	1900 2.5	99
DIG DIG.DIG DIG.DIG	1000000 240768.75 439620.00	91
DIG.DIG and DIG.DIG	9.50 and 9.85	73
DIG.DIG DIG	8.8 1923	44
DIG/	1928/	59
DIG|	6|	246
DIG-and	1914-and	38
Point DIG	Point 15	23
DIG and DIG DIG/DIG	13 and 13 1/2	96
DIG DIG/DIG-c-	89 3/4-c-	129
DIG DIG:DIG	42 4:36	391
DIG.DIG DIG/DIG	4.73 1/2	163
DIG DIG million	24000000 230 million	66
DIG DIG.DIG million	1940 30.5 million	64
DIG DIG/DIG DIG.DIG DIG.DIG	58 1/4 7.50 7.8	157
DIG DIG:DIG DIG DIG	201 2:15 1 30	29
DIG DIG DIG/DIG	70 4 1/3	32
DIG DIG/DIG DIG	19 1/4 8	83
DIG:DIG NUM	7:15 one	31

Table A.2
Summary of negative binominal regression models across decades.

Decades	Term	estimate	std.error	CI
1920s	(Intercept)	15.07	0.20	[14.69, 15.47]
	NumberLog	−4.50	0.05	[−4.59, −4.4]
	TenNessyes	4.29	0.24	[3.83, 4.78]
	TwoPointFiveNessyes	3.91	0.24	[3.46, 4.38]
	FiveNessyes	2.98	0.26	[2.47, 3.49]
	TwoNessyes	2.16	0.28	[1.61, 2.71]
	Multiple5yes	0.39	0.13	[0.13, 0.65]
	Multiple10yes	4.70	0.15	[4.41, 4.98]
1930s	(Intercept)	14.82	0.18	[14.47, 15.18]
	NumberLog	−4.40	0.05	[−4.5, −4.32]
	TenNessyes	4.18	0.23	[3.74, 4.65]
	TwoPointFiveNessyes	3.86	0.24	[3.39, 4.33]
	FiveNessyes	3.07	0.25	[2.59, 3.56]
	TwoNessyes	2.23	0.28	[1.72, 2.83]
	Multiple5yes	0.38	0.11	[0.16, 0.6]
	Multiple10yes	5.00	0.13	[4.76, 5.24]
1940s	(Intercept)	14.66	0.19	[14.31, 15.04]
	NumberLog	−4.42	0.05	[−4.52, −4.33]
	TenNessyes	4.48	0.23	[4.01, 4.91]
	TwoPointFiveNessyes	3.91	0.24	[3.46, 4.39]
	FiveNessyes	3.23	0.25	[2.76, 3.73]
	TwoNessyes	2.29	0.27	[1.77, 2.82]
	Multiple5yes	0.42	0.12	[0.2, 0.66]
	Multiple10yes	4.93	0.13	[4.69, 5.20]
1950s	(Intercept)	15.59	0.21	[15.17, 16]
	NumberLog	−4.68	0.05	[−4.78, −4.57]
	TenNessyes	4.32	0.23	[3.88, 4.79]
	TwoPointFiveNessyes	3.57	0.25	[3.09, 4.06]
	FiveNessyes	3.02	0.25	[2.56, 3.54]
	TwoNessyes	2.06	0.28	[1.54, 2.62]
	Multiple5yes	0.67	0.13	[0.42, 0.92]
	Multiple10yes	5.21	0.14	[4.93, 5.47]
1960s	(Intercept)	15.75	0.22	[15.31, 16.15]
	NumberLog	−4.75	0.06	[−4.85, −4.63]

Table A.2 (*continued*)

Decades	Term	estimate	std.error	CI
	TenNessyes	4.49	0.23	[4.06, 4.95]
	TwoPointFiveNessyes	3.86	0.25	[3.37, 4.34]
	FiveNessyes	3.20	0.25	[2.74, 3.71]
	TwoNessyes	2.19	0.28	[1.66, 2.73]
	Multiple5yes	0.45	0.13	[0.19, 0.69]
	Multiple10yes	5.42	0.15	[5.11, 5.71]
1970s	(Intercept)	15.90	0.24	[15.44, 16.35]
	NumberLog	−4.80	0.06	[−4.92, −4.68]
	TenNessyes	4.28	0.24	[3.83, 4.76]
	TwoPointFiveNessyes	3.68	0.26	[3.22, 4.20]
	FiveNessyes	2.99	0.27	[2.48, 3.51]
	TwoNessyes	2.01	0.29	[1.45, 2.57]
	Multiple5yes	0.42	0.14	[0.13, 0.69]
	Multiple10yes	5.60	0.17	[5.25, 5.91]
1980s	(Intercept)	15.58	0.23	[15.15, 16.03]
	NumberLog	−4.67	0.06	[−4.78, −4.56]
	TenNessyes	4.30	0.23	[3.86, 4.75]
	TwoPointFiveNessyes	3.56	0.26	[3.07, 4.07]
	FiveNessyes	2.97	0.26	[2.43, 3.45]
	TwoNessyes	2.11	0.29	[1.53, 2.64]
	Multiple5yes	0.32	0.14	[0.06, 0.60]
	Multiple10yes	5.15	0.16	[4.84, 5.47]
1990s	(Intercept)	13.51	0.18	[13.15, 13.85]
	NumberLog	−4.05	0.04	[−4.14, −3.97]
	TenNessyes	4.17	0.24	[3.74, 4.66]
	TwoPointFiveNessyes	3.41	0.27	[2.86, 3.92]
	FiveNessyes	3.13	0.26	[2.61, 3.63]
	TwoNessyes	2.36	0.28	[1.82, 2.90]
	Multiple5yes	−0.06	0.11	[-0.28, 0.15]
	Multiple10yes	3.86	0.13	[3.59, 4.08]
2000s	(Intercept)	16.58	0.25	[16.09, 17.06]
	NumberLog	−4.95	0.06	[−5.08, −4.83]
	TenNessyes	4.42	0.24	[3.96, 4.89]
	TwoPointFiveNessyes	3.36	0.28	[2.81, 3.91]
	FiveNessyes	3.04	0.28	[2.47, 3.58]
	TwoNessyes	2.30	0.28	[1.73, 2.86]
	Multiple5yes	0.33	0.15	[0.03, 0.62]
	Multiple10yes	5.05	0.17	[4.71, 5.37]

References

American Psychological Association, 2022. APA Style numbers and statistics guide. https://apastyle.apa.org/instructional-aids/numbers-statistics-guide.pdf.

Bednarek, M., 2016. Investigating evaluation and news values in news items that are shared through social media. Corpora 11 (2), 227–257. https://doi.org/10.3366/cor.2016.0093.

Berg, T., Neubauer, M., 2014. From unit- *and* -ten to ten-before-unit order in the history of English numerals. Lang. Var. Change 26 (1), 21–43. https://doi.org/10.1017/S0954394513000203.

Buckley, P.B., Gillman, C.B., 1974. Comparisons of digits and dot patterns. J. Exp. Psychol. 103 (6), 1131.

Bürkner, P.-C., 2017. brms: an R package for Bayesian multilevel models using Stan. J. Stat. Software 80, 1–28.

Caple, H., Bednarek, M., 2013. Delving into the Discourse: Approaches to News Values in Journalism Studies and beyond. Reuters Institute for the Study of Journalism, Oxford. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Delving%20into%20the%20Discourse_0.pdf.

Chrisomalis, S., 2020. Reckonings: Numerals, Cognition, and History. The MIT Press.

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

Cummins, C., 2015. Constraints on numerical expressions (Oxford Studies in Semantics and Pragmatics) (Vol. 5). Oxford University Press.

Davies, M., 2007. TIME magazine corpus. https://www.english-corpora.org/time/.

Dehaene, S., 1989. The psychophysics of numerical comparison: a reexamination of apparently incompatible data. Percept. Psychophys. 45 (6), 557–566. https://doi.org/10.3758/BF03208063.

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

Dorogovtsev, S.N., Mendes, J.F.F., Oliveira, J.G., 2006. Frequency of occurrence of numbers in the world wide web. Phys. Stat. Mech. Appl. 360 (2), 548–556. https://doi.org/10.1016/j.physa.2005.06.064.

Government Publishing Office, 1933. Style Manual. Government Publishing Office. https://www.govinfo.gov/app/details/GOVPUB-GP-b6b0b70bec04bb27bdc566025f2f155e.

Government Publishing Office, 1953. Style Manual. Government Publishing Office. https://www.govinfo.gov/app/details/GOVPUB-GP-6eda8715b8f5cb1d8514325b97334d4f.

Granath, S., Ullén, M., 2019. 'The elevation of Sensitivity over Truth': political correctness and related phrases in the time magazine corpus. Appl. Linguist. 40 (2), 265–287. https://doi.org/10.1093/applin/amx019.

Greenberg, J.H., 1978. Generalizations about numeral systems. In: Universals of human language (Vol. 3, pp. 249–295). Stanford University Press, Stanford.

Jansen, C.J.M., Pollmann, M.M.W., 2001. On round numbers: pragmatic aspects of numerical expressions. J. Quant. Ling. 8 (3), 187–201. https://doi.org/10.1076/jqul.8.3.187.4095.

Kibler, W.W., 1984. An Introduction to Old French. Modern Language Association of America.

Lakoff, G., Núñez, R., 2000. Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being (Vol. 6). Basic Books.

McElreath, R., 2016. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781315372495.
Nieder, A., 2005. Counting on neurons: the neurobiology of numerical competence. Nat. Rev. Neurosci. 6 (3), 177–190.
O'Halloran, K., 2005. Mathematical Discourse: Language, Symbolism and Visual Images. A&C Black, New York.
Pechenick, E.A., Danforth, C.M., Dodds, P.S., 2015. Characterizing the Google books corpus: strong limits to inferences of socio-cultural and linguistic evolution. PLoS One 10 (10), e0137041. https://doi.org/10.1371/journal.pone.0137041.
Pollmann, T., 1998. On forgetting the historical past. Mem. Cognit. 26 (2), 320–329. https://doi.org/10.3758/BF03201143.
Pollmann, T., Baayen, R.H., 2001. Computing historical consciousness. A quantitative inquiry into the presence of the past in newspaper texts. Comput. Humanit. 35 (3), 237–253. https://doi.org/10.1023/A:1017563931115.
Porter, T.M., 1995. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press. https://doi.org/10.1515/9781400821617.
Roeh, I., Feldman, S., 1984. The rhetoric of numbers in front-page Journalism: how numbers contribute to the melodramatic in the popular press. Text Interdiscipl. J. Study Discourse 4 (4). https://doi.org/10.1515/text.1.1984.4.4.347.
Ruud, P.A., Schunk, D., Winter, J.K., 2014. Uncertainty causes rounding: an experimental study. Exp. Econ. 17 (3), 391–413. https://doi.org/10.1007/s10683-013-9374-8.
Scott, M., 1996. WordSmith tools (version v8) [Computer software]. https://lexically.net/wordsmith/.
Sigurd, B., 1988. Round numbers. Lang. Soc. 17 (2), 243–252.
Swanberg, W.A., 1972. Luce and His Empire. Scribner.
Sweller, J., 1988. Cognitive load during problem solving: effects on learning. Cognit. Sci. 12 (2), 257–285. https://doi.org/10.1207/s15516709cog1202_4.
U.S. Government Publishing Office, 2017. U.S. Government Publishing Office Style Manual. U.S. Government Publishing Office. https://www.govinfo.gov/app/details/GPO-STYLEMANUAL-2016/.
Van Der Henst, J., Carles, L., Sperber, D., 2002. Truthfulness and relevance in telling the time. Mind Lang. 17 (5), 457–466. https://doi.org/10.1111/1468-0017.00207.
Wiese, H., 2003. Numbers, Language, and the Human Mind. Cambridge University Press.
Winter, B., 2019. Statistics for Linguists: An Introduction Using R. Routledge.
Woodin, G., Winter, B., Littlemore, J., Perlman, M., Grieve, J., 2023. Large-scale patterns of number use in spoken and written English. Corpus Linguist. Linguistic Theory 0 (0). https://doi.org/10.1515/cllt-2022-0082.
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.
Woodward, K., 1999. Statistical panic. Differences J. Feminist Cult. Stud. 11 (2), 177–203.
Xu, J., 2015. TIME Magazine Corpus (1923-2008). http://114.251.154.212/cqp/time/.