Interpretation of approximate numerical expressions: Computational model and empirical study

Sébastien Lefort, Marie-Jeanne Lesot, Elisabetta Zibetti, Charles Tijus, Marcin Detyniecki

Interpretation of approximate numerical expressions: Computational model and empirical study

Sébastien Lefort[a,*], Marie-Jeanne Lesot[a], Elisabetta Zibetti[b], Charles Tijus[b], Marcin Detyniecki[a,c]

[a]*Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris, France*
[b]*Laboratoire CHArt-LUTIN, EA 4004, Université Paris 8, 2 rue de la liberté, 93526, Saint-Denis - Cedex 02, France*
[c]*Polish Academy of Sciences, IBS PAN, Warsaw, Poland*

Abstract

Approximate Numerical Expressions (ANEs) are linguistic expressions involving numerical values and referring to imprecise ranges of values, illustrated by examples such as "*about 100*". In this paper, a general principle is proposed to interpret uncontextualised ANEs as intervals of denoted values. It is based on an empirically justified combination of characteristics of numerical values, both arithmetical and cognitive, and in particular, taking into account the cognitive salience of numbers. This general principle is instantiated in two computational models that can be extended so as to take into account the applicative context. An empirical study is conducted to assess the performances of the two models, comparing them to state-of-the-art methods, on real interpretations collected through an on-line questionnaire. Results validate the proposed characteristics used to build the models and show that they offer the best performances in estimating the median interval chosen as representative of the collected intervals.

Keywords: Approximate Numerical Expressions, Imprecision, Median Estimation, Cognitive Number Salience

*Corresponding author
Email addresses: sebastien.lefort@lip6.fr (Sébastien Lefort), marie-jeanne.lesot@lip6.fr (Marie-Jeanne Lesot), ezibetti@univ-paris8.fr (Elisabetta Zibetti), tijus@univ-paris8.fr (Charles Tijus), marcin.detyniecki@lip6.fr (Marcin Detyniecki)

1. Introduction

Approximate numerical expressions (ANEs) are vague linguistic expressions of the general form "*about* x" where x is a number. They are used in daily life to denote an imprecise range of values, for instance to give imprecise pieces of knowledge regarding space (e.g., Berlin is located at *about 900km* from Paris), time (e.g., the meeting will last *around 2 hours*) or any numerical evaluation (e.g., the cost will be *about 200 euros*; the audience gathered *around 200* participants). In the field of Human-Computer Interfaces and approximate reasoning, ANEs raise the issue of their interpretation.

Indeed, as intelligent systems whose interaction mode rely on natural language become more and more present in daily life, interpreting such vague expressions is a relevant task. Application domains include database querying, such as Geographic Information Systems [**?** **?**] (e.g., looking for an area whose surface is *about* $100m^2$) or more generally flexible queries or expert systems, such as medical expert systems [**?**] (e.g., interpreting the information of patient saying he has fever since *approximately 48 hours*).

In the general domain of vague expressions, beyond the ones involving numerical values, Lasersohn [**?**] introduces the notion of pragmatic halo as a formalisation of vagueness. While the denotation of precise expressions corresponds to the entity that is explicitly given, the denotation of vague expressions also includes entities of the same semantic type, that are implicitly conveyed by the expression. The pragmatic halo of an expression is defined as the union of the entities explicitly denoted and the implicit ones. For instance, in the sentence "*I left home at 7:56am*", the precise expression "*7:56am*" exactly denotes 7:56 and no other moment (e.g., 7:57 or 7:58). On the contrary, in the sentence "*I will be there at around 6pm*", the expression "*around 6pm*" is vague and not only denotes 6:00pm exactly, but also an implicit range of moments around 6pm (e.g., [5:45, 6:15]).

From a logical perspective, it is considered [**?**] that a proposition involving a vague expression is satisfied (i.e., its truth value equals true) if the actual

entity belongs to the expression pragmatic halo. Therefore, if I actually arrive at 5:49pm, the proposition "*I will be there at around 6pm*" is true because 5:49pm is included in the pragmatic halo of the expression "*around 6pm*". In this theoretical framework, interpreting an ANE is equivalent to determining the range of values that makes a proposition including this ANE true, i.e., the range of values explicitly and implicitly denoted by the ANE. For instance, the task of determining the range of distances that should be considered for the query of a hotel located at *about 100km*.

Three approaches have been proposed to represent the interpretation of ANEs: as fuzzy numbers (see e.g., [**?** **?**]); as probability distributions of possible intervals, as suggested by Ferson et al. [**?**]; as intervals (see e.g., [**?** **?** **?**]). This paper belongs to the latter framework and proposes a general principle instantiated in two computational models to interpret ANEs as intervals.

Approximate Numerical Expressions (ANEs) are a specific kind of vague expressions related to numerical expressions, named *imprecise expressions* by Solt [**?**]. Three parts can be distinguished in an ANE: its semantic and pragmatic contexts, and the reference value. To illustrate them, let us consider the example of a car seller saying "*this car is worth about 10,000 euros*": the semantic context corresponds to what is evaluated, the car, and to what is counted, money. The pragmatic context is the car seller trying to sell a car. Another pragmatic context might be the buyer trying to negotiate the price. In these two cases, the interpretation of the ANE "*about 10,000 euros*" may be different. Finally, the reference value, or nominal value, of the ANE is the number itself: 10,000. Although contexts of ANEs affect their interpretation [**?** **?** **?**], i.e., the intervals they denote, the models proposed in this paper are only based on their reference value and do not consider the factors related to their contexts. The aim here is to provide core models based on context-independent factors, i.e., the numerical properties of the expression, that can be extended to be instantiated in specific contexts.

The aim of this paper is to model the intervals of denoted values, corresponding to pragmatic halos, of ANEs. More specifically, the goal is to determine the

values of the endpoints of the intervals $I(x) = [x^-, x^+]$ denoted by ANEs of the form "*about* x", where x is a non-zero natural number, i.e., $x \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$, in an uncontextualised framework, i.e., not considering factors related to semantic and pragmatic contexts.

This paper proposes two computational models, the Log-Linear Model (LLM) and the Rank Model (RKM), based on an original and empirically justified combination of characteristics of numerical values. Both models rely on a general principle whose first originality is to take into account a cognitive characteristic of ANEs, besides the arithmetical ones, capturing the notion of number salience. A second one comes from the proposed combination method, based on Pareto frontiers. A third one is the performed experimental validation, comparing the estimation performances of both proposed models to the ones of the models from the literature, obtained on real data collected from an on-line questionnaire.

The paper is structured as follows: Section 2 is dedicated to the description of related works and existing models. The general principle and the two proposed models are presented in Section 3. The experimental setup is described in Section 4. Section 5 presents the results of the experimental study. Finally, conclusions and future works are discussed in Section 6.

2. Related works

This section first presents the cognitive bases of number representations in human beings as it is exposed in the cognitive psychology literature. Three computational models from the literature are then presented: the ratio model (RM) [**?**], the scale-based model (SBM) [**?** **?**], and the regression model (REGM) [**?**].

2.1. Cognitive number representation in human beings

Insofar as ANEs involve numbers, it is first relevant to examine the literature dedicated to the way the human cognitive system encodes and represents numbers and quantities.

Based on clinical evidences, Dehaene and Cohen [**?**] propose a triple code functional model of numerical cognition. They argue that three categories of mental representations are used to manipulate numbers, depending of the task: the visual Arabic form, involving strings of digits (e.g., 36), the verbal word frame, involving sequences of words (e.g., thirty-six), and the analogical magnitude representation, corresponding to the numerical value. In healthy persons, all three categories of representations can be transcoded to any other [**?**].

The type of representation used depends on the nature of the executed task. Those involving arithmetic facts (e.g., 242+34) rely more heavily on the verbal representation, while rounding or approximation tasks involve the analogical magnitude representation [**?**].

The authors also argue that the visual and verbal forms are not semantic, the meaning of numbers derive from the the analogical magnitude representation, through a mapping process. This observation implies that putting two numbers in relation, for instance in comparison tasks, can only occur at the analogical magnitude representation.

Further, the analogical magnitude representation can be represented as a logarithmically compressed mental line, where quantities are encoded according to the Weber-Fechner law [**?**]. Formally, two quantities x_1 and $x_2 \in \mathbb{N}$ are distinguished if their difference is greater than a fraction c, called Weber fraction, of the larger of both, i.e., if:

$$\frac{|x_1 - x_2|}{\max(x_1, x_2)} > c \tag{1}$$

The indistinguishability between close quantities results in imprecision when estimations are performed. The value of the Weber fraction c varies from one individual to another and with age [**?**]. The average value of the Weber fraction would be around 0.11 [**?**] or 0.12 [**?**].

From this triple code functional model of numerical cognition, two ANE interpretation perspectives can be postulated. Firstly, interpreting an ANE consists in intuitively estimating the imprecision implied by its analogical mag-

nitude representation. This perspective leads to consider the reference value of the ANE as the relevant factor since it defines the analogical magnitude representation. Secondly, as linguistic numerical expressions, ANEs involve the verbal or the visual Arabic representations, depending on the presentation modality. This perspective leads to consider the ANE interpretation issue as a formal problem as in the case of the Scale-Based Model [**?** **?**] (see Section 2.3).

For instance in the case of "*about* 8150", the imprecision related to the analogical magnitude representation of 8150 might be high, close to the one of 8000, leading to a wide interval of denoted values. One the contrary, with regards to the verbal or the visual representations and as a numerical fact, 8150 is a more precise description of what is counted or evaluated than 8000. Therefore, the imprecision corresponding to *about* 8150 might be lower than the one corresponding to *about* 8000, leading to a narrower interval of denoted values. Because interpreting an ANE involves both the analogical magnitude and the verbal representation, it is reasonable to think that the interval of denoted depends on a compromise between the outputs of these representations.

The two perspectives of ANE interpretation are supported by the findings of Ferson et al. [**?**]. Indeed, they showed that the order of magnitude, as a function of the ANE reference value, and the roundness, related to the number of zeros at the right of the last significant digit, are good predictors of the intervals corresponding to ANEs (see Section 2.4).

2.2. Ratio model (RM)

The first approach in modelling the interpretation of ANEs as intervals consists in defining the width of the interval of an ANE as a fixed percentage of its value [**?** , p. 116]. Denoting s the considered percentage, it is formally defined as:

$$I_{RM}(x) = [x - x \cdot s, x + x \cdot s] \tag{2}$$

For instance, using $s = 10\%$, "*around 300*" is interpreted as $[270, 330]$.

The ratio model (RM) is derived from the Weber-Fechner law involved in the human number cognition (see Section 2.1). Indeed, as a consequence of this law, for a given reference value x, all values whose difference to x is lower than $s \cdot x$, i.e., v such that $|v - x| \leq s \cdot x$, where s is related to the Weber fraction c, are indistinguishable from x.

It can be observed that the ratio model is solely based on the reference value of the ANE. Therefore, it can provide counter-intuitive results for numbers with multiple significant digits. For instance, in the case of $x = 8150$, with $s = 10\%$, one gets $I_{RM}(8150) = [7335, 8965]$: the width of this interval (1630) can be considered as too high with respect to the relative magnitude of x (50). This issue, that does not occur for numbers with a single significant digit, is due to the fact that the information conveyed by numbers with multiple significant digits is more precise. The scale-based model, presented in the next subsection, addresses this issue.

2.3. Scale-based models (SBM)

The scale-based model [**?** **?** **?**] depends on defined levels of granularity through the use of a scale system $S = \{s_1, \ldots, s_n\}$, where s_i are granularity levels such that $s_i < s_{i+1}$. S can be complex, when the factor between two granularities is not constant, such as the time scale-system (e.g., $S = \{1', 5', 1/4h, 1/2h, 1h, \ldots\}$) or simple, such as the decimal system ($S = \{1, 10, 100, \ldots\}$), where $s_{i+1} = 10 \cdot s_i$.

The scale system represents different granularity levels for the interpretation of a numerical expression. For instance, in the decimal system, the numerical expression "*1000*" can be interpreted at the 1, 10, 100 or 1000 granularity levels. At the 1 granularity level, the expression is precise, "*1000*" exactly denotes 1000, whereas at the 1000 granularity level, "*1000*" denotes a larger range of values (e.g., $[500, 1500]$).

The relevant granularity level for interpreting an ANE is context-dependent and is linguistically expressed using approximators [**?**]. For instance, the approximator "*exactly*" refers to the lowest granularity level, usually the unit,

while "*about*" refers to the largest granularity level the expression belongs to (e.g., the level of thousands for "*about 2000*"). The latter, denoted $Gran_C(x)$, is defined as:

$$Gran_C(x) = \sup_{\{s_i \in S | x \bmod s_i = 0\}} s_i \tag{3}$$

The values denoted by an ANE *about* x are the ones that are closer to the reference value x than to any other number on the largest granularity level of the ANE. For instance, the granularity level of 2000 in the decimal system is 1000. The values denoted by "*about 2000*" thus range from 1500 to 2500. Consequently, the width of the interval equals the largest granularity level the ANE belongs to, $Gran_C(x)$. Because this model consider intervals centered around the ANE reference value, the corresponding interval is defined as:

$$I_{SBM}(x) = \left[x - \frac{Gran_C(x)}{2}, x + \frac{Gran_C(x)}{2} \right] \tag{4}$$

For instance, $I_{SBM}(300) = [250, 350]$ and $I_{SBM}(8150) = [8145, 8155]$. This approach has the advantage of taking into account the granularity of the ANE through a set of scales. However, it does not address the issue of the value of the last significant digit: all ANEs at the same granularity level result in the same interval width. Yet, one may expect that the interval of "*about 100*", for instance, would be narrower than the interval of "*about 800*".

2.4. Regression model (REGM)

The two previous models are theoretical and have not been, to the best of our knowledge, experimentally tested for the approximator "*about*". Unlike this theoretical approach, the empirical one proposed by Ferson et al. [**?**] relies on data collected using Amazon Mechanical Turk. Participants of this study have been asked to evaluate the endpoints of intervals corresponding to semantically contextualised ANEs (e.g., "*Bats make up* ***about 20****% of all classified mammal species globally.*"). The aim of the authors is to test the relevance of predictors of the interval width. Several approximators, all analysed through the same model,

are considered, including "*about*". The intervals are estimated according to the following equation:

$$I_{REGM}(x) = \left[x - \frac{10^{L(x)}}{2}, x + \frac{10^{L(x)}}{2}\right] \tag{5}$$

where $L(x)$ is computed as:

$$\begin{aligned} L(x) = \ & \omega_1 + \omega_2 \cdot O_m(x) + \omega_3 \cdot R(x) + \omega_4 \cdot f(x) \\ & + \omega_5 \cdot O_m(x) \cdot R(x) + \omega_6 \cdot O_m(x) \cdot f(x) + \omega_7 \cdot R(x) \cdot f(x) \\ & + \omega_8 \cdot O_m(x) \cdot R(x) \cdot f(x) \end{aligned} \tag{6}$$

where ω_1 to ω_8 are parameters empirically set by performing a linear regression on the collected dataset. $O_m(x)$ is the ANE order of magnitude ($O_m(x) = \log_{10}(x)$), $R(x)$ its roundness, i.e., the decimal place of the last significant digit (e.g., $R(13) = 1$ and $R(130) = 2$), and $f(x) \in \{0, 1\}$ its "*fiveness*". A number has the *fiveness* property if its last significant digit is 5.

The three variables $O_m(x)$, $R(x)$ and $f(x)$ have been empirically selected as the variables that make the linear regression statistically significant with the collected data [**?**].

According to the authors, the regression model REGM can also be applied to test the relevance of the predictors regarding the endpoints of intervals and not only their width, although they do not report the results [**?**].

This model presents the advantage of allowing the adaptation to different contexts by learning parameters on a dataset. However, it can be noted that in the conducted experiments, the semantic context is not controlled. Indeed, several semantic contexts are mixed, which may result in interactions between the factors related to the context and the ones related to the ANE reference number.

3. Proposed model

This section first describes the proposed general principle that provides a pool of candidates for the interval endpoint values. Its originality is twofold. Firstly, we propose to anchor the choice of the ANE characteristics to take into account, introduced in Section 3.1, on a cognitive ground and to introduce a numerical measure of number complexity to capture the notion of number cognitive salience. Secondly, the aggregation consists in a compromise between the selected characteristics, based on Pareto frontiers, as described in Section 3.2.

Two models are then proposed to select, among the candidate values provided by the general principle, the endpoints of the intervals denoted by ANEs: a discretised log-linear model, presented in Section 3.3, and a rank model, described in Section 3.4. Both models are designed to be extended so as to adapt to specific semantic and pragmatic contexts.

3.1. Characteristics of ANEs

The ANEs considered in this study are of the form "*about* x", where $x \in \mathbb{N}^*$. The aim is to estimate $I(x) = [x - \Delta^-(x), x + \Delta^+(x)]$, the interval corresponding to *about* x, which is the range of values denoted by *about* x. The width of $I(x)$ is therefore defined as $|I(x)| = \Delta^-(x) + \Delta^+(x)$.

3.1.1. Arithmetical characteristics

In the decimal system, x can be written as $x = \sum_{i=0}^{q} a_i \cdot 10^i$, where $a_i \in [\![0, 9]\!]$. Several arithmetical characteristics can be attached to a positive integer, as defined in Table 1. The magnitude is the actual value of x; its number of significant digits is noted $NSD(x)$. The way the human analogical magnitude representation encodes quantities (see Section 2.1) leads to consider the notion of magnitude as a key factor in ANE interpretation.

The granularity $Gran(x)$ is the order of magnitude of x, i.e., the power of ten x belongs to. This definition of granularity can be compared to the one from the scale-based model, $Gran_C(x)$ (see Section 2.3). Indeed, $Gran_C(x) = Gran(x)$ when the selected scale system is the decimal one. The granularity as

Characteristic	Formal definition	Example
Magnitude	x	8150
Granularity	$Gran(x) = 10^{i^*}$ where $i^* = \min\{i \mid a_i \neq 0\}$	10
Last significant digit	$LSD(x) = a_{i^*}$	5
Relative magnitude	$RelMag(x) = a_{i^*} \cdot 10^{i^*}$	50
Number of significant digits	$NSD(x) = q - i^* + 1$	3
Complexity	$Cpx(x) = NSD(x) - B(x)$	2.5

Table 1: Characteristics of a positive natural number $x = \sum_{i=0}^{q} a_i \cdot 10^i$, illustrated in the case of $x = 8150$ in the last column. $B(x)$, used in the definition of complexity, is defined in Eq. (7).

we define it is also related to the notion of *roundness* proposed by [**?**] in the regression model (see Section 2.2): $Gran(x) = 10^{i^*}$ and $R(x) = i^*$ (see Table 1).

The value of the last significant digit of x is noted $LSD(x)$. The relative magnitude $RelMag(x)$ is the product of granularity and the value of the last significant digit. It is meant to reflect the relevant part of x for ANE interpretation from the point of view of the verbal representation of numbers (see Section 2.1). Indeed, one may expect that the width of the interval of "*about 30.000.050*" is comparable to that of "*about 150*", 50 being the common part.

3.1.2. Cognitive characteristic

Beyond arithmetical characteristics, we propose to introduce a complexity measure, $Cpx(x)$, to capture the cognitive salience of numbers, i.e., the fact that some numbers are more easily evoked than others. Indeed, it has been observed from corpus analyses than some numbers occur more frequently than others [**? ?**].

In the Arabic number form, two characteristics of numbers influence their frequency of occurrence. Firstly, the more significant digits a number has, the more complex it is to generate and the less frequently it occurs. Indeed, from a cognitive perspective, frequent numbers would rely on lexically stored infor-

mation, while infrequent numbers would require a step-by-step generation [**?**]. Therefore, at constant number of digits, the frequency of round natural numbers is observed to be higher than the frequency of non-round natural numbers [**?**]. Secondly, numbers whose last significant digit is 5 and, to a lower extent, 2, occur more frequently [**?**].

Taking into account these observations, we propose to define a complexity measure, a posteriori justified by the experimental results reported in Section 5, for natural numbers expressed in the decimal scale system, depending on the number of significant digits $NSD(x)$ and the value of the last significant digit $LSD(x)$, to capture the salience of numbers. We propose to formalise the complexity of a natural number as its number of significant digits minus a bonus if its last significant digit is 2 or 5 and its number of significant digits is at least 2. Therefore, the complexity of all natural numbers with a single significant digit is the same and equal to one.

For symmetry reasons around multiples of 10, we propose to consider the case where the last significant digit is 8 the same way as the case where the last significant digit is 2: any number whose last significant digit is 2 is of the form $(10 \cdot k + 2) \cdot 10^{i^*}$ (e.g., $320 = (30 + 2) \cdot 10$); we propose to stress its symmetrical with respect to $(10 \cdot k) \cdot 10^{i^*}$, i.e.,the numbers whose last significant digit is 8 $(10 \cdot k - 2) \cdot 10^{i^*}$ (e.g., $280 = (30 - 2) \cdot 10$). This symmetry argument applied to a number whose last significant digit is 5 does not require special care as $LSD((10 \cdot k - 5) \cdot 10^{i^*}) = LSD((10 \cdot k + 5) \cdot 10^{i^*}) = 5$. The experimental results described in Section 5 a posteriori justify this proposition.

The bonus function thus distinguishes three categories, depending on the value of the last significant digit $LSD(x)$ and respecting the order of frequency of appearance: $B(x_1) > B(x_2) > B(x_3)$, for $x_1, x_2, x_3 \in \mathbb{N}^*$ such that $LSD(x_1) = 5$, $LSD(x_2) \in \{2, 8\}$ and $LSD(x_3) \notin \{2, 5, 8\}$. Because the proposed exploitation of complexity measure is ordinal (see Section 3.2), the chosen values are only constrained by the induced order. We arbitrarily propose to set these values at 0.5, 0.25 and 0. The bonus function is therefore formalised as:

$$B(x) = \begin{cases} 0.5 & \text{if } LSD(x) = 5 \text{ and } NSD(x) > 1 \\ 0.25 & \text{if } LSD(x) = 2 \text{ or } LSD(x) = 8 \text{ and } NSD(x) > 1 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Figure 1 illustrates the complexity value $Cpx(x)$ for all integers between 500 and 600 as pluses. We consider a number as more salient than another if its complexity is lower. For instance, 500 is more salient than 501 because $Cpx(500) = 1$ and $Cpx(501) = 3$.

3.2. The Pareto frontiers based principle

This subsection presents the general principle on which the two proposed models, described in the next subsections, are based. Its aim is to provide a pool of relevant candidate values to be the endpoints of the interval denoted by an ANE.

3.2.1. Principle

The rationale of the proposed model is that a good interval endpoint candidate simultaneously satisfies two constraints that can be contradictory: they must be both salient (i.e., have a low complexity) and close to the ANE reference value. The possible incompatibility of these requirements can be, for instance, illustrated in the case of "*about* 500": 499 is closer to 500 than 450, but it is less salient ($Cpx(499) = 3$ and $Cpx(450) = 1.5$).

Therefore, optimising both distance and salience at the same time is not achievable and a trade-off must be considered: the general principle we propose relies on the assumption that when interpreting and ANE "*around* x", a compromise is made between the salience of the endpoint values and their distance to x. This assumption implies that for a given distance, the salience of the endpoint values is maximised; similarly, at a given salience, the distance of the endpoint value is minimised. For instance, given the ANE "*about 500*", intervals such as $[499, 501]$, $[490, 510]$ or $[450, 550]$ are good candidates because

their endpoints are the closest to the reference value $x = 500$ when the considered complexity of the candidates is respectively 3, 2 and 1.5. On the contrary, $[497, 503]$ or $[460, 540]$ are not considered as good candidates because they are dominated by other intervals, i.e., intervals that represent better compromises between distance and salience of endpoint values. For instance, $[497, 503]$ is dominated by $[499, 501]$ whose endpoints are as salient as the former but closer to the reference value 500. For the same reason, $[460, 540]$ is dominated by $[490, 510]$, whose endpoints are as salient, but closer to the reference value.

The values that optimise the compromise between salience and distance to the reference value, i.e., that are not dominated by other values, are better candidates to be the endpoint values. To select these good candidates, two Pareto frontiers, $P^-(x)$ related to the left endpoint, and $P^+(x)$ related to the right endpoint, are built [**?**], as detailed below.

3.2.2. Building the Pareto frontiers

To build one Pareto frontier, we consider as candidates all values $v \in V^e(x) \subset \mathbb{N}$, with $e \in \{-, +\}$ and $V^-(x) = [1, x)$ and $V^+(x) = (x, +\infty)$. These values are compared on two criteria: (i) their absolute distance from the ANE reference value: $d_x(v) = |v - x|$; (ii) their complexity $Cpx(v)$, capturing their salience.

The selected values are the ones that are not dominated by any other value. The notion of dominance is a partial order: formally v' dominates v (denoted $v' \prec v$) iff $d_x(v') < d_x(v) \wedge Cpx(v') \leq Cpx(v)$.

They constitute the Pareto frontier, formally defined as $P^e(x) = \{v \in V^e(x) / \forall v' \in V^e(x) \quad v' \not\prec v\}$. The Pareto frontier $P^e(x) = [y_1, \ldots, y_n]$, with $e \in \{-, +\}$, is a list ordered by increasing distance to the reference value x, $\forall i d_x(y_i) < d_x(y_{i+1})$. Note that $P^e(x)$ is a finite set because the complexity of the candidate values cannot be below 1, when they have a single significant digit, and cannot be above the number of significant digits of the ANE reference value. For instance, in the case of "*about 560*", as illustrated on Figure 1, the minimum complexity of candidates is 1 (e.g., 500 or 600) and the maximum is 3

(e.g., 559 or 561). Although there exists an infinite number of candidates with these complexity values, only one of them, the nearest to x, is selected because it dominates all the others.

Figure 1 illustrates the two Pareto frontiers corresponding to the ANE "*about* 560": the black plus signs represent the distribution of $(v, Cpx(v))$ couples for all values v in the interval $[500, 600]$ (indeed, values lower than 500 or greater than 600 are not relevant as they are all dominated, by 500 and 600 respectively). Note that the horizontal axis representing the candidates v also indirectly represents their distance $d_x(v)$: the latter increases linearly from the vertical green line representing $x = 560$. Thus, the part to the right of the vertical line can also be seen as showing the distribution of $(d_x(v) + x, Cpx(v))$; the part to the left shows the the distribution of $(x - d_x(v), Cpx(v))$ $P^-(560)$ and $P^+(560)$ are represented by the connected points circled in red, respectively below and above 560, denoted by the green vertical line. As a result, $P^-(560) = [559, 558, 555, 550, 500]$ and $P^+(560) = [561, 562, 565, 570, 580, 600]$.

The values in the left Pareto frontier are in decreasing order because they are ordered according to their distance to the reference value.

In the case of $x = 560$, 540 is not considered as a good left endpoint candidate because it is dominated by 550, a value which is both closer and more salient than 540: 550 represents a better compromise between distance and salience.

One can notice that this model, taking into account the complexity of numbers, naturally captures the asymmetry experimentally observed (see Section 5.1). Indeed, in the case where x have several significant digits, salient numbers are not symmetrically distributed around x. This is illustrated on Figure 1 for $x = 560$, where 500 is farther from 560 than 600.

The two models described in the next subsections exploit the values from the Pareto frontiers as candidates to estimate the endpoint values of the denoted interval.

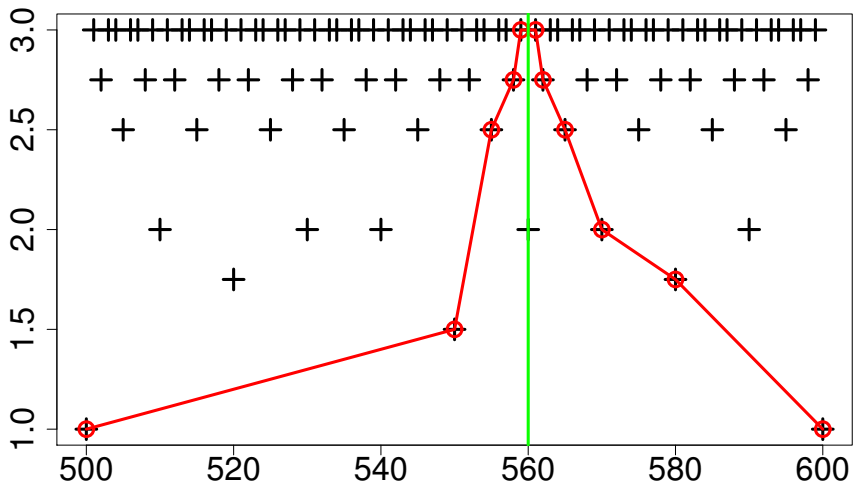

Figure 1: Complexities $Cpx(v)$ (pluses) for integers between 500 and 600, and Pareto frontiers (red lines) for left (left from vertical green line) and right (right from vertical green line) endpoints of the ANE "*about 560*": $P^-(x)(560) = [559, 558, 555, 550, 500]$ and $P^+(x)(560) = [561, 562, 565, 570, 580, 600]$.

3.3. Log-linear model (LLM)

Like the regression model REGM [**?**] (see Section 2.4), the proposed log-linear model (LLM) is based on a linear regression. Its originality lies in the use of a final discretisation step which maps the estimations provided by a regression function to integer values, using the previous Pareto frontiers $P^-(x)$ and $P^+(x)$.

In contrast to the scale-based models (see Section 2.3), and like REGM, this empirical model is adaptable to different semantic contexts because the coefficients of the regression are learned from collected intervals. The principle of the model is first introduced, before the learning phase is presented.

3.3.1. Principle

The model estimates the interval endpoint values by means of a regression function. As suggested in Section 3.1 and in Section 2, the relative magnitude

$RelMag(x)$ and the magnitude x of an ANE are involved in its interpretation. We propose to use the logarithms of these dimensions as variables in the regression function. The obtained real values are then discretised to the nearest integer on the Pareto frontiers corresponding to the considered ANE.

The real value is provided by a regression function that represents the estimation of the absolute distance of the interval endpoints from the ANE reference value x, noted $\Delta_C(x)$ and computed as:

$$\Delta_C(x) = \exp(\alpha_1 \cdot \log(RelMag(x)) + \alpha_2 \cdot \log(x) + \beta) \tag{8}$$

where α_1, α_2 and β are regression parameters. Their learning phase is described in the next subsection.

At this stage, a symmetrical interval whose endpoints are real values can be constructed as $I = [x_R^-, x_R^+] = [x - \Delta_C(x), x + \Delta_C(x)]$. Because salient endpoints are more relevant than real ones, we propose to discretise them to the closest values in the Pareto frontiers, according to the following function:

$$D(v) = \arg \min_{y_i \in P(x)} |v - y_i| \tag{9}$$

where $P(x)$ is the union of the two Pareto frontiers $P(x) = P^-(x) \cup P^+(x)$. The final interval corresponding to the ANE *about* x is then:

$$I_{LLM}(x) = [D(x_R^-), D(x_R^+)] \tag{10}$$

where $x_R^- = x - \Delta_C(x)$ and $x_R^+ = x + \Delta_C(x)$.

Note that the resulting interval $I_{LLM}(x)$ maybe asymmetric because the distribution of the values on the Pareto frontiers may be different when considering the right or the left endpoint (see Figure 1).

3.3.2. Regression parameters

The regression parameters α_1, α_2, and β are learned from an experimentally collected database of intervals corresponding to ANEs, e.g., as the one described in Section 4.1. The data are first transformed to a set of training

triples $(x, RelMag(x); \Delta^e(x))$, where $e \in \{-, +\}$. So as to distinguish between the left and right endpoints, we define two training triples per ANE, $(x, RelMag(x); \Delta^-(x))$ and $(x, RelMag(x); \Delta^+(x))$, where $\Delta^-(x)$ is distance between the ANE reference value x and the representative of the left endpoints of its corresponding interval and $\Delta^+(x)$ the distance between x and the representative of the right endpoints. These two triples per ANE are obtained by aggregating all the learning intervals corresponding to the considered ANE. The choice of the aggregation operator is specified in Section 4.5.

In a second step, a linear regression according to the least square algorithm is performed to find optimal values of α_1, α_2, and β, adjusting for the function deriving from Equation (8):

$$\log(\Delta^e(x)) = \alpha_1 \cdot \log(RelMag(x)) + \alpha_2 \cdot \log(x) + \beta \tag{11}$$

3.4. Rank model (RKM)

The second model, RKM, exploits the Pareto frontiers values in a different way: it consists in estimating directly the position of the endpoint values on the Pareto frontiers. To that aim, it selects $y_{r_P(x)} \in P^e(x) = [y_1, \ldots, y_n]$ endpoint value and the issue is to estimate the relevant rank $r_P(x)$. The latter depends on the relative magnitude $RelMag(x)$ and the number of significant digits $NSD(x)$ of the considered ANE *about* x. The use of the number of significant digits in Equation (12) relies on the observation that, in the collected data, the interval width increases with the number of significant digits (e.g., $|I(8150)| > |I(50)|$ and $NSD(8150) > NSD(50)$, see Section 5.1).

The position of the selected endpoint is computed according to the following equation:

$$r_P(x) = round\left(\log(RelMag(x)) - 1 + \sum_{k=1}^{NSD(x)} k\right) \tag{12}$$

The estimation of the interval corresponding to the ANE is then $I_{RKM}(x) = [y^-_{r_P(x)}, y^+_{r_P(x)}]$. $r_P(x)$ is limited in $[1, n^e]$, i.e., $r_P(x) = \min(\max(1, r_P(x)), n^e)$,

with $e \in \{-, +\}$, where n^- and n^+ respectively are the length of $P^-(x)$ and $P^+(x)$, in the case where $r_P(x) < 1$ or $r_P(x) > n^e$.

4. Experimental settings

An empirical study has been conducted to collect experimental data in the form of intervals corresponding to ANEs. The latter allow to: (i) validate the relevance of the characteristics of ANEs used in both proposed models; (ii) assess, through an experimental study, the performances of the two models proposed in this paper, and to compare them to the models from the literature. Although the content of the questionnaire is different from the one used by [**?**] to collect intervals, the collection methods is similar. Indeed, ANEs are presented to the participants, who explicitly give the endpoints values of the corresponding intervals.

The first subsection presents the material and methods used to collect and process the data. The next two subsections discuss the proposed quality criteria: endpoint prediction accuracy, relative distance, median prediction accuracy and score of median error. The last two subsections are dedicated to the experimental procedure designed to compare the models and their parameterisation.

4.1. Data collection and analysis

Material. 24 ANEs, whose reference values are listed in Table 2, are considered in the study. Their reference values are selected in order to cover different combinations of characteristics, to avoid biases towards a specific one: several values of the last significant digit at a granularity level (e.g., 20/30/40/50/80), several granularity levels at a same last significant digit (e.g., 80/800/8000), several numbers of significant digits at the same relative magnitude (e.g., 50/150/8150).

ANEs are not semantically contextualised, no cues are given to participants as to what is measured or counted.

An online questionnaire that includes the 24 selected ANEs has been designed to collect data. Instructions, given in French to the participants, are

"*Selon vous, entre quelles valeurs (MINIMALE-MAXIMALE) se trouve 'environ x'?*", which can be translated as "*In your opinion, what are the MINIMUM and MAXIMUM values associated with 'about x'?*". Each participant fills the 24 items corresponding to the 24 ANEs. The order in which they are presented is randomly set for each participant.

Finally, each participant is asked to answer the two following questions, meant to control inter-individual variability:

- Use of mathematics: to examine whether a daily use of mathematics at works or in the studies may affect interpretation of ANEs, participants are asked if their work or study involves mathematical skills.
- Subjective level of mental arithmetic skills: as complementary information to the previous question, participants are asked to auto-evaluate, on a five-point Likert scale, whether they think they are good at mental arithmetic.

Population. Participants were recruited through an announcement diffused on mailing-lists. 146 adults volunteered to freely take part in the study, 102 women and 44 men, aged 20 to 70 (mean= 38.6; standard deviation= 14.2). All were native French speakers.

Data preprocessing. The answer given by participant p for the ANE "*about* x" is noted $I_p(x) = [I_p^-(x), I_p^+(x)]$. The answers are used to compute two absolute distances $\Delta P_p^e(x)$ between the ANE reference value x and the endpoints $e \in \{-, +\}$ of the interval: $\Delta P_p^e(x) = |I_p^e(x) - x|$. Consequently, $I_p(x) = [x - \Delta P_p^-(x), x + \Delta P_p^+(x)]$ and $|I_p(x)| = \Delta P_p^-(x) + \Delta P_p^+(x)$.

The analyses reported in the following sections are based on the absolute distances. This variable is more relevant than the interval width, because it allows to compare intervals together, without losing the symmetry information. Indeed, two widths may be equal although the values of the endpoints are different (e.g., $|[95, 110]| = |[90, 105]|$).

x	$Gran(x)$	$LSD(x)$	$RelMag(x)$	$NSD(x)$
20	10	2	20	1
30	10	3	30	1
40	10	4	40	1
50	10	5	50	1
80	10	8	80	1
100	100	1	100	1
110	10	1	10	2
150	10	5	50	2
200	100	2	200	1
400	100	4	400	1
440	10	4	40	2
500	100	5	500	1
560	10	6	60	2
600	100	6	600	1
800	100	8	800	1
1000	1000	1	1000	1
1100	100	1	100	2
1500	100	5	500	2
2000	1000	2	2000	1
4700	100	7	700	2
4730	10	3	30	3
6000	1000	6	6000	1
8000	1000	8	8000	1
8150	10	5	50	3

Table 2: Reference values of ANEs used in the questionnaire (see Section 4.1) and their characteristics: magnitude (x), granularity ($Gran$), value of the last significant digit (LSD), relative magnitude ($RelMag$) and number of significant digits (NSD), as defined in Section 3.1.

Data cleaning. In order to exclude outlier pairs (minimum and maximum) from the set, data are processed according to the following three steps procedure:

Step 1. 84 intervals are considered as outliers because either:

- they are inadequate ([0, infinity])
- the right endpoint is below the reference value or the left endpoint is above the reference value, formally: $I_p^+(x) < x$ or $I_p^-(x) > x$ (e.g., $I(800) = [700, 750]$ or $I(800) = [810, 850]$)
- an endpoint is greater than ten times or lesser than one tenth the reference value, formally $I_p^+(x) > 10 \cdot x$ or $I_p^-(x) < \frac{x}{10}$ (e.g., $I(100) = [9, 1101]$)

Step 2. For both endpoints of each ANE, the mean and standard deviation of the data remaining after step 1 are computed. Any endpoint value beyond three standard deviations of the mean is considered as an outlier. This step leads to exclude 180 intervals.

Step 3. 10 participants are considered as untrustworthy and are excluded because more than 70% of their interval endpoints are missing values or outliers.

From 3504 intervals I_p in the original corpus, 3177 (91%) are used as dataset for the experimental validation of the two proposed models, LLM and RKM. The next subsections describe the quality criteria we propose to perform this validation.

4.2. Endpoint prediction criteria

We propose two quality criteria to assess whether the intervals estimated by interpretation models fit the data provided by the participants of the empirical study: endpoint prediction accuracy (PA) and relative distance (RD).

$\mathcal{X}$ represents the set of considered ANEs and $\mathcal{P}(x)$ the set of participants whose interval is taken into account (i.e., not considered as an outlier) for ANE "*about* x". We note $\Delta Med^e(x)$ the median of the distances $\Delta P_p^e(x)$ over all

participants $p \in \mathcal{P}(x)$ and $\Delta M_m^e(x)$ the absolute distance of the endpoint value (for $e \in \{-,+\}$) from the ANE reference value x, estimated by a model m.

4.2.1. Endpoint prediction accuracy

The endpoint prediction accuracy is a criterion designed to assess how well a model correctly estimates the participants' answers.

A good prediction occurs when the estimated absolute distance from the ANE reference value is equal to the one from the participant: $\Delta M_m^e(x) = \Delta P_p^e(x)$. A relative error is not penalised. Two rates of relative error were tested: 10% and 5%. We observed that a rate of 5% leads to a heavy penalisation of the models that produce real numbers as estimations: the ratio model and the regression model. Therefore, we propose to set the rate of relative error at 10%.

Formally, for model m, and endpoint e of the ANE *about* x, the prediction accuracy, to be maximised, is defined as:

$$PA(m,x,e) = \frac{1}{|\mathcal{P}(x)|} \cdot \left| \left\{ p \in \mathcal{P}(x) \,\middle|\, \frac{|\Delta M_m^e(x) - \Delta P_p^e(x)|}{\min(\Delta M_m^e(x), \Delta P_p^e(x))} \leq 0.1 \right\} \right| \quad (13)$$

Because the models considered in this paper produce a single estimation per endpoint, the maximum score a model can reach corresponds to the relative frequency of the endpoint values mode over the participants.

The global score of model m, to be maximised, is obtained by averaging PA over the two endpoints $e \in \{-,+\}$ of all ANEs *about* $x \in \mathcal{X}$:

$$PA(m) = \frac{1}{2 \cdot |\mathcal{X}|} \cdot \sum_{x \in \mathcal{X}} \sum_{e \in \{-,+\}} PA(m,x,e) \quad (14)$$

4.2.2. Relative distance

A distance measure is proposed as another way to assess the proximity of the estimated endpoint values with the ones given by the participants.

To take into account the error with regards to the characteristics of the ANE, we propose to define this distance as relative to its relative magnitude

$RelMag(x)$. Indeed, an error of 10 units, for instance, appears as more significant for $x = 100$ than for $x = 1000$. On the other hand, the same error is considered to be as significant for 8150 as for 50. This implies that the normalisation factor should not depend on the magnitude but on the relative magnitude of the considered ANE, leading to consider $|\Delta M_m^e(x) - \Delta P_p^e(x)|/RelMag(x)$.

Averaging over the two endpoints $e \in \{-, +\}$ and all participants $p \in \mathcal{P}(x)$, as for PA, we propose to define RD for model m, to be minimised, as:

$$RD(m) = \frac{1}{2 \cdot |\mathcal{X}|} \cdot \sum_{x \in \mathcal{X}} \sum_{e \in \{-,+\}} \frac{1}{|\mathcal{P}(x)|} \cdot \sum_{p \in \mathcal{P}(x)} \frac{|\Delta M_m^e(x) - \Delta P_p^e(x)|}{RelMag(x)} \qquad (15)$$

4.3. Median prediction criteria

As detailed in Section 5.1, the empirical study shows that no consensus can be observed regarding the interpretation of ANEs, and that a high variability occurs in the collected intervals. It is thus relevant to discuss the definition of the interval an ANE interpretation model should aim at. The first subsection details the motivation of the proposed approach, that focuses on the median of the observed endpoint values. The next two subsections introduce the quality criteria designed to assess whether this goal is reached.

4.3.1. Rationale of the median interval

As vague expressions, the interval denoted by an ANE has no sharp endpoints: the transition between acceptable intervals and unacceptable ones is progressive and might be variable.

One can distinguish between objective and subjective aspects of variability in endpoints. Firstly, from an objective point of view, variability in interpretation across participants can be observed (see Section 5.1): one participant may set endpoint values to $[90, 110]$ while another one may set them at $[80, 120]$ or $[95, 105]$.

Secondly, from a subjective point of view, variability in interpretation within participants, called tolerance by Wright [**?**], appears: one participant might

set endpoint values of "*about 100*" at $[90, 110]$. However, it might be the case that for him/her $[89, 111]$ or $[88, 112]$ are admissible as well. Therefore, when a participant gives intervals corresponding to ANEs, the endpoints should be considered as indications because close values may also be good candidates.

To overcome these issues, we propose to model the endpoints of the interval corresponding to an ANE as a statistical distribution. However, the quantity of training data to learn relevant models of this type might be huge. Therefore, we propose to estimate parameters of this distribution, such as the mean or the median.

Estimating the median interval appears to be relevant in both aspects of progressiveness and variability: indeed, it corresponds to a value given by one or several participants and is therefore not an arbitrary value. As observed in the empirical study (see Section 5.1), participants tend to give salient numbers as endpoint values: estimating the median leads to naturally produce such values that make sense to the users. Moreover, the median is a robust parameter with regards to extreme values. On the contrary, the mean may result in precise values, less natural to human users, while being less robust with regards to extreme values.

The next subsections present the quality criteria we propose to assess whether the estimations provided by interpretation models fit the observed medians.

4.3.2. Median prediction accuracy

We first propose to use the accuracy score of the median prediction as a quality criterion of media estimation.

We consider a median estimation $\Delta M^e_m(x)$ of model m for endpoint $e \in \{-, +\}$ of ANE "*about* x" as correct if the error, as for PA, does not exceed 10% of the actual median. This quality criterion, to be maximised, can be formalised as:

$$MA(m) = \frac{1}{2 \cdot |\mathcal{X}|} \sum_{x \in \mathcal{X}} \left| \left\{ e \in \{-, +\} \;\middle|\; \frac{|\Delta M^e_m(x) - \Delta Med^e(x)|}{\Delta Med^e(x)} \leq 0.1 \right\} \right| \quad (16)$$

4.3.3. Score of median error

In case of incorrect median estimation, one should be able to evaluate the degree of error, i.e., the distance between the prediction and the actual median. To address this issue, we propose a criterion that assesses the balance between participants' answers that are above and below the estimated endpoint value.

These quantities can be formally defined as

$$N_+(x,e) = |\{p \in \mathcal{P}(x) | \Delta P_p^e(x) > \Delta M_m^e(x)\}|$$

and

$$N_-(x,e) = |\{p \in \mathcal{P}(x) | \Delta P_p^e(x) < \Delta M_m^e(x)\}|.$$

The model m should be such that $N_+(x,e) = N_-(x,e)$ for all x, e.

However, because participants' intervals endpoints are not uniformly distributed, but rather distributed on few values, a perfect balance may not be possible. For instance let's consider the case of the ANE "*about 100*", the distribution, over 60 intervals, of the left endpoint values might be 20 times 40, 30 times 45 and 10 times 48. The median of this distribution is $\Delta Med^-(100) = 45$. Even if the model correctly predicts this median value, a perfect balance is not achieved because 20 endpoint values are below and 10 are above. The score would therefore not reflect the fact that the median is actually correctly estimated. To overcome this issue, we propose that the balance score of estimation takes into account the balance of the actual median, i.e., the two quantities

$$N_+^*(x,e) = |\{p \in \mathcal{P}(x) | \Delta P_p^e(x) > \Delta Med^e(x)\}|$$

and

$$N_-^*(x,e) = |\{p \in \mathcal{P}(x) | \Delta P_p^e(x) < \Delta Med^e(x)\}|.$$

The score of the model then depends on the absolute difference between N_+ and N_+^* and the absolute difference between N_- and N_-^*. Averaging over the two endpoints and all considered ANEs, the median error, to be minimised, can be defined as:

$$MErr(m) = \frac{1}{2 \cdot |\mathcal{X}|} \cdot \sum_{x \in \mathcal{X}} \sum_{e \in \{-,+\}} \frac{|N_+(x,e) - N_+^*(x,e)| + |N_-(x,e) - N_-^*(x,e)|}{|\mathcal{P}(x)|} \tag{17}$$

4.4. Experimental procedure

Using the four quality criteria PA, RD, MA and $MErr$ described above and the empirically collected intervals as data, we compare the performances of the two models using the Pareto frontier principle we propose (see Section 3): the log-linear model (LLM) and the rank model (RKM), to the ones of the three models from the literature discussed in Section 2: the ratio model (RM) [**?**], the scale-based model (SBM) [**?** **?**] and the regression model (REGM) [**?**].

A cross-validation procedure is performed on benchmarks consisting in 1000 runs of learning and test steps, where all models, whether they require a learning step or not, are evaluated against the same test dataset:

1. Participant: the learning step of REGM and LLM is performed on the intervals given by 75% of randomly selected participants, the intervals from the remaining 25% constitute the test dataset. This benchmark is meant to assess whether the models that need a learning step are able to generalise across participants.
2. ANE: the learning step of REGM and LLM is performed on the intervals given by all participants on 17 (66.7%) randomly selected the ANEs. The intervals of all participants corresponding to the 7 remaining ANEs are used as test dataset. This benchmark is designed to measure the robustness of the log-linear model because the regression parameters are estimated on only $17 \cdot 2 = 34$ points. It is also designed to evaluate the generalisation ability of models that need learning across ANEs.

The ratio model takes into account only the magnitude of the ANEs, one can expect that it deals betters with numbers with a single significant digit, for which $RelMag(x) = x$ (e.g., $x = 8000$, $RelMag(8000) = 8000$), than numbers

with multiple significant digits, for which $RelMag(x) \neq x$ (e.g., $x = 8150$, $RelMag(x) = 50$). Similarly, scale-based model uses the relative magnitude but not the magnitude. One can therefore expect that this model better deals with multiple significant digits than with a single significant digits. To avoid biases towards models that are advantaged by reference values with a single significant digit only, or reference values with multiple significant digits only, a constraint is included in the random selection: the learning and test sets must include a mix of both types.

In order to determine which model shows the best performances in each benchmark, statistical analyses using ANOVA tests with model as factor, and Tukey's HSD post-hoc tests [**?**] are performed. The significance threshold is set at $p = .01$.

4.5. Parameterisation of considered models

Several combinations of parameters are tested for the parameterised models. The results reported in Section 5 are the ones obtained with the parameters that give the best scores on the four criteria. Indeed, it has been observed that changing the parameters of a model results in a change in the same direction for all four scores.

RM gives the best performances with parameter $s = 5\%$. For SBM, the decimal system $S = \{1, 10, 100, \ldots\}$ is the one that better fits the data.

REGM only provides the size of the intervals and no information about their locations or symmetry around the ANEs. We make the assumption that they are centered around the considered ANEs.

In the learning step, LLM requires the choice of a representative of the intervals corresponding to an ANE. In order to account for the variability observed in the collected intervals (see Section 5.1), we propose use the median as aggregation operator of the intervals to constitute the training triples used be the proposed Log-Linear Model (see Section 3.3).

Parameter	Reported by [**?**]	Obtained on collected data
ω_1	**-0.208**	0.375
ω_2	**0.428**	**0.220**
ω_3	**0.281**	**-0.00485**
ω_4	**0.0940**	0.977
ω_5	**0.0147**	**0.0807**
ω_6	**-0.0640**	-0.244
ω_7	**-0.0102**	**-0.496**
ω_8	**0.0404**	0.138

Table 3: Coefficients of the REGM linear regression (see Eq.(6), Section 2.4), reported by Ferson et al. [**?**], and obtained on the collected data. Bold values are the significant ones in the regression performed on the collected data.

4.6. Preliminary validation of the regression model

As a preliminary validation of REGM, the linear regression proposed by [**?**] (see Eq.(6), Section 2.4) was performed on the overall set of the collected data. Table 3 presents the coefficients obtained and the ones reported by the authors [**?**]. Four variables of the regression on seven appear to be statistically significant on the considered data: order of magnitude ($O_m(x)$, associated to ω_2), roundness ($R(x)$, associated to ω_3), product of order of magnitude by roundness ($O_m(x) \cdot R(x)$, associated to ω_5), and product of roundness by fiveness ($R(x) \cdot f(x)$, associated to ω_7). The coefficient of determination is much lower on the collected data ($R^2 = 0.272$) than the one reported by the authors ($R^2 = 0.741$). These results suggest that the REGM proposed by [**?**] does not fit well the data. This may be due to the difference in the contexts or in the language used to collect the data. Indeed, while Ferson et al. [**?**] embed ANEs in several semantic contexts, the ANEs of this study are not contextualised. Moreover, the participants of this study are native French speakers while the workers of the Ferson et al. [**?**] study are English speakers.

Further, some random selections in the ANE benchmark result in aberrant

linear regression coefficients for REGM. This can be explained by the *fiveness* variable $f(x)$, a property which only occurs five times in the collected corpus: some learning sets do not offer enough ANEs with this property to correctly estimate coefficients related to this variable. Runs in which such aberrant coefficients are observed, resulting in estimated interval endpoints located at at least 10.000 units from the ANE reference value, are not included in the analyses. On 1000 runs, 88 lead to aberrant *fiveness* coefficients. Therefore, the means and standard deviations of REGM performances in the ANE benchmark are based on 912 runs.

5. Experimental results

This section presents the results of the experimental study. The first subsection briefly introduces the analyses performed on the collected data in order to test the hypotheses concerning the relevant characteristics of ANEs and the distributions of the intervals endpoints. The next subsection deals with the results of the performance assessment of the five interpretation models.

5.1. Data description and analyses

Statistical analyses were performed on the data in order to check whether the hypotheses on which the two proposed models are designed are relevant. The main conclusions are reported here.

Characteristics of ANEs. Results of statistical analyses, not detailed here for space reasons, show that the magnitude, the granularity, the value of the last significant digit and the number of significant digits influence the width of the intervals. More specifically, the greater the magnitude, the granularity, the value of the last significant digit or the number of significant digits of an ANE reference value, the wider its interval.

Variability in the intervals. It is observed that participants tend not to agree on the intervals denoted by ANEs: on average, the 136 respondents give 15.4

different answers per endpoint. The number of different values for each endpoint ranges from 9 ("*about 20*" for left endpoint) to 22 ("*about 8150*" for right endpoint).

Symmetry of intervals. The symmetry of the collected intervals largely depends on the considered ANE. Indeed, even if on the whole dataset, 74.2% of the intervals are centered around the reference number, intervals of some ANEs, such as 440 or 4730, are significantly less often symmetric (63% and 50% respectively). This observation can be explained by the fact that these ANEs are less round, i.e., have more significant digits, than others (e.g., 500, 8000) and do not satisfy the fiveness property. Indeed, the collected data show that participants tend to give salient numbers whose complexity is low (e.g., 400 and 450 for 440; 4700 and 4750 for 4730), as endpoint values. Because these salient numbers may be at different distances from the ANE reference value, the resulting intervals may be asymmetric.

However, most endpoint values (84.4%) are included in the Pareto frontiers as defined in Section 3.4, providing an empirical validation of our general principle.

Summary. The results empirically support our choices concerning the characteristics of the ANEs used in the Pareto frontiers model and the two estimators LLM and RKM. Indeed, relative magnitude, as a combination of granularity and the value of the last significant digit, and the number of significant digits appear to be key factors in the interpretation of uncontextualised ANEs. The notion of complexity seems relevant as participants tend to minimise the complexity of the numbers corresponding to the endpoint values they give. Finally, the variability observed in the data suggests that a single interval per ANE cannot satisfy all participants. This variability therefore supports the view of a distribution representation approach of intervals endpoints and the relevance of estimating the median interval as a meaningful parameter of the distribution.

Scores on Participant benchmark				
Model	Endpoint accuracy PA (%)	Relative distance RD	Median accuracy MA (%)	Median error $MErr$
RM	8.9 (1.1)	0.78 (0.05)	18.4 (7.2)	0.76 (0.11)
SBM	19.5 (2.4)	0.43 (0.08)	28.0 (6.9)	0.76 (0.08)
REGM	7.9 (1.7)	**0.39 (0.09)**	20.0 (7.2)	0.67 (0.18)
LLM	**22.4 (2.6)**	**0.38 (0.08)**	54.1 (9.6)	0.38 (0.11)
RKM	**22.2 (2.6)**	**0.39 (0.08)**	**58.3 (8.9)**	**0.35 (0.12)**

Table 4: Score means and standard deviations in parentheses of the four criteria for each model on the Participant benchmark. Bold scores are statistically the best ones according to the ANOVA tests. Several scores are bold if no significant difference is revealed by post-hoc analyses.

5.2. Results on the benchmarks

Tables 4 and 5 present the performances of the five models, Figures 2 and 3 graphically illustrate them. Results are similar in both the Participant and the ANE benchmarks, in which the proposed RKM offers the best performances. The next subsections discuss the results of each criterion in turn.

Endpoint prediction accuracy. The scores obtained by LLM and RKM are the best ones and do not statistically differ from each other according to the ANOVA tests performed. These scores can be compared to the baseline, which is calculated as the mean relative frequency of the endpoint values mode (see Section 4.2): 28.1% ($\sigma = 6.9$). The high scores of LLM and RKM can be due to the fact that the intervals they provide can be asymmetric while the three other models provide intervals centered around the ANE reference value.

RM and REGM perform poorly on endpoint prediction accuracy. This can be explained by the fact that they tend to make real-numbered estimations while participants tend to give integers as interval endpoints. Moreover, RM is well suited for numbers with a single significant digit only, resulting in poor

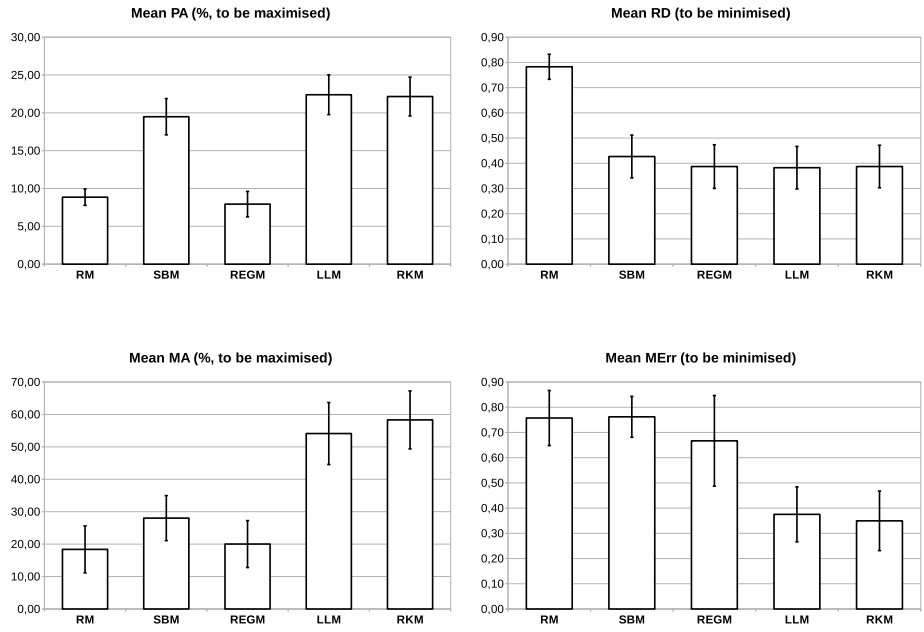

Figure 2: Mean scores of models in Participant benchmark: endpoint prediction accuracy PA (top, left), relative distance RD (top, right), median prediction accuracy MA (bottom, left), and median error $MErr$ (bottom, right). It can be observed that the two proposed models, RKM and LLM, show the best performances while the Ratio model (RM) offers the worst, in each criterion.

performances concerning numbers with multiple significant digits.

Relative distance. The ANOVA tests reveal a significant effect of the model factor on the RD scores. More precisely, post-hoc analyses show that this difference is due to RM and that scores of other models do not differ significantly.

Like for endpoint prediction accuracy, because RM does not take into account the relative magnitude of the ANE reference value, it shows poor performances for ANEs with multiple significant digits. These performances are all the lower as the difference between the magnitude and the relative magnitude of an ANE reference value is high. Indeed, the highest relative distances occur for "*about 4730*" ($RD = 6.92$) and "*about 8150*" ($RD = 7.46$), the two ANEs

Scores on ANE benchmark				
Model	Endpoint accuracy PA (%)	Relative distance RD	Median accuracy MA (%)	Median error $MErr$
RM	8.7 (3.0)	0.84 (0.65)	24.4 (13.0)	0.70 (0.17)
SBM	19.3 (2.7)	**0.44 (0.23)**	24.9 (14.5)	0.79 (0.18)
REGM	8.1 (4.1)	**0.45 (0.64)**	15.7 (13.2)	0.65 (0.14)
LLM	**21.9 (3.2)**	**0.41 (0.25)**	56.6 (15.3)	0.32 (0.14)
RKM	**22.0 (3.1)**	**0.40 (0.22)**	**63.8 (14.0)**	**0.27 (0.16)**

Table 5: Score means and standard deviations in parentheses of the four criteria for each model on the ANE benchmark. Bold scores are statistically the best ones according to the ANOVA tests. Several scores are bold if no significant difference is revealed by post-hoc analyses.

with the highest difference between relative magnitude and magnitude, while the scores for other ANEs range from 0.06 to 0.73.

A twofold explanation can account for the statistically non-significant difference in the relative distance scores of the four other models. Firstly, REGM and LLM are regression-based models. Although not based on the same ANE characteristics, they both aim at minimising the distance between observed data and the linear regression, leading to estimations that are close to the collected intervals. Secondly, for SBM and RKM, results reveal differences for specific ANEs, depending on the characteristics taken into account by the model: SBM shows poor performances in high relative magnitude and high magnitude ANEs (e.g., $x = 8150$), because it does not consider the ANE magnitude. The non significant difference between the score of this model and the others can thus be explained by the low number of such ANEs in the corpus.

Median prediction accuracy and median error. Statistical analyses reveal that RKM shows the best performances, both in median prediction accuracy, and in median error, providing an empirical validation of our proposed model.

REGM offers a poor prediction score but an average error score. As for

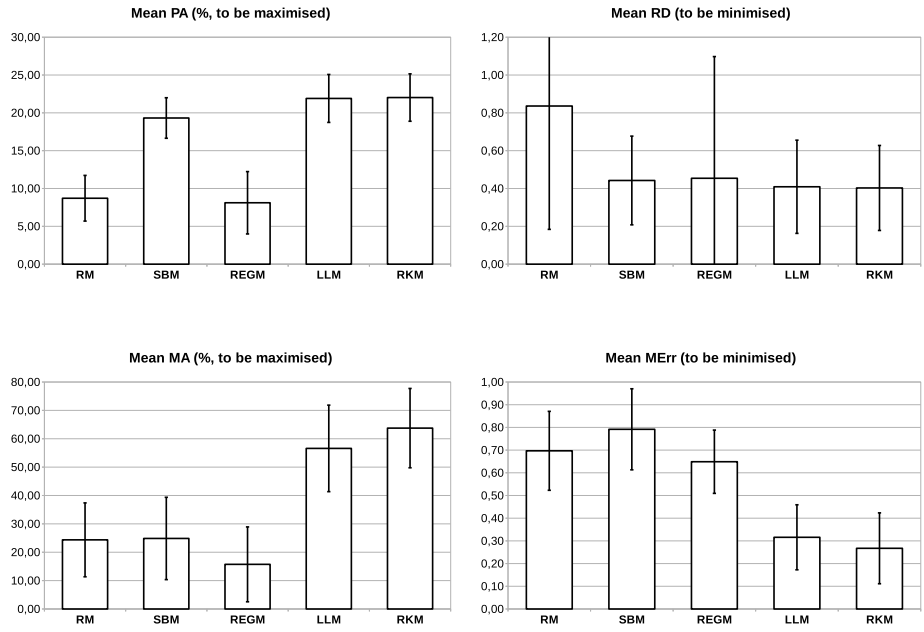

Figure 3: Mean scores of models in ANE benchmark: endpoint prediction accuracy PA (top, left), relative distance RD (top, right), median prediction accuracy MA (bottom, left), and median error $MErr$ (bottom, right). The two proposed models, RKM and LLM, show the best performances in each criterion.

endpoint prediction accuracy, this can be due to the fact that the model provides real-numbered estimations while participants tend to give round numbers. However, the error score reveals that these real-numbered estimations are closer to the actual medians than the estimations provided by SBM and RM. On the contrary, SBM performs better than REGM on prediction accuracy while the prediction errors are more important.

Differences between ANEs with a single and with multiple significant digits. We performed complementary analyses, not detailed here, to check whether the global performances of the models presented above are also valid when distinguishing between ANEs with a single and multiple signifcant digits: results reveal very similar scores of LLM, RKM and REGM, for ANEs with a single

and multiple significant digits in the Participant benchmark.

RM performs well on ANEs with a single significant digit and poorly on ANEs with multiple significant digits while SBM presents the opposite profile. These results are consistent with the design of the two models: RM takes into account only the magnitude of an ANE, while SBM only considers its relative magnitude.

Concerning the ANE benchmark, one can observe a difference between ANEs with a single and with multiple significant digits, for all models although the difference is less marked for RKM. Indeed, they perform more poorly on ANEs with multiple significant digits than with a single one. Moreover, standard deviations are higher for ANEs with multiple than with a single significant digit.

The difference between the Participant and the ANE benchmarks may be due to the random selection of ANEs in the latter. Indeed, in this benchmark, the models are evaluated on only 7 randomly selected ANEs. A single error thus implies a 14% contribution to the score which quickly leads to poor performances.

Moreover, as noted for the Participant benchmarks, RM and SBM are more or less adapted to reference values with a single or multiple significant digits. Their scores therefore highly depends on the random draw of ANEs in the test dataset. Concerning REGM and LLM which require a learning step, the training dataset is reduced in the ANE benchmark to 34 triples while they learn on 48 triples in the Participant benchmark. The regression parameters may therefore be less robust in the ANE benchmark.

One can conclude from these results that the generalisation ability of the models is better across participants than across ANEs. The learning datasets should be carefully constituted to cover a large combination between dimensions of ANEs.

6. Conclusion and future works

To estimate the intervals corresponding to Approximate Numerical Expressions, this paper proposes two computational models, the Log-Linear model and the Rank model. Both are based on a common underlying general principle that takes into account the cognitive salience of numbers and exploit Pareto frontiers.

An experimental study is conducted, using real data collected from an online questionnaire, which supports the proposed general principle: results reveal that the two models exploiting this principle offer the best performances in the four quality criteria we propose, especially in median prediction accuracy.

The performed analyses also show that there is no consensus across participants with regard to the intervals corresponding to an ANE. The approach proposed in this paper therefore consists in estimating the median as the targeted statistical parameter of the endpoints distributions rather than a single interval that should satisfy all participants.

Moreover, the experimental validation of the two proposed models supports the relevance of the cognitive salience, among arithmetical characteristics of ANEs, as a key factor in ANE interpretation.

Future work will include extensions of the proposed model, in particular so as to take into account the two other factors implied by ANE interpretation: the semantic and pragmatic contexts [**?**]. Indeed, the general principle based on Pareto frontiers can be extended so as to include other variables, related to the context in the compromises step. For instance, in a business context, beyond the arithmetical and cognitive dimensions, a third one can be added to represent the role of the user: seller or buyer. Indeed, the right endpoint of the interval corresponding to the ANE "*about 10.000 euros*" should be closer to the reference value than the left endpoint if the user is a buyer, representing the fact that he prefers lower prices. Conversely, the right endpoint of the interval should be farther from the reference value than the left endpoint is the user is a seller, representing the fact that he prefers higher prices. This can be achieved by proposing a third criterion to capture this context-dependent dimension, in

the same way as the complexity dimension that has been proposed to capture the cognitive characteristic. The Pareto frontiers can then be generalised to address three-criteria optimisation problems. A second approach concerns more specifically the Log Linear model: as it relies on a learning step, it can fit to specific contexts by constituting dedicated learning datasets.

Another possible extension of the Pareto frontiers model consists in generalising the complexity of numerical expressions so as to take into account real-numbered ANEs (e.g., "*about 3.14*") and scales of measurement, such as time (seconds, minutes, quarters of hours, etc.) or space (centimeters, meters, kilometers, etc.), to interpret expressions based on more complex scale systems (e.g., "*I will be there in approximately half an hour*" or "*This rope is approximately one meter long*").

Two other representations of ANEs should also be investigated to model the variability in the collected intervals. The first one is proposed by Lakoff [**?**] and consists in using fuzzy numbers as representation of ANEs. From this point of view, the median of the endpoints distribution is of particular interest because it corresponds to the 0.5-cut of the fuzzy numbers in the random set view of elicited membership functions proposed by Bilgiç and Türkşen [**?**]. Moreover, values located on the Pareto frontiers can be used to characterise other α-cuts [**?**]. The second approach consists in representing ANEs as probability distributions. One can either, as suggested by Ferson et al. [**?**], model the residual variation by such distributions, or directly the whole distributions. However, this approach may need a significant number of ANEs and participants to be able to properly characterise the factors and the parameters involved in any ANE interpretation.

Finally, in this study, participants were asked to explicitly give the intervals corresponding to ANEs. It may be that implicit interpretation of ANEs differs from the explicit one. Future work should address this issue, by collecting intervals implicitly, and testing the proposed models, to ensure that there is no bias in collecting data using explicit questionnaires.

Aknowledgments

This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

7. References

[1] B. Bennett, Spatial vagueness, in: R. Jeansoulin, O. Papini, H. Prade, S. Schockaert (Eds.), Methods for Handling Imperfect Spatial Information, Vol. 256 of Studies in Fuzziness and Soft Computing, Springer Berlin Heidelberg, 2010, pp. 15–47.

[2] T. M. Delboni, K. A. Borges, A. H. Laender, C. A. Davis, Semantic expansion of geographic web queries based on natural language positioning expressions, Transactions in GIS 11 (3) (2007) 377–397.

[3] E. Straszecka, Combining uncertainty and imprecision in models of medical diagnosis, Information Sciences 176 (20) (2006) 3026–3059.

[4] P. Lasersohn, Pragmatic halos, Language 75 (3) (1999) 522–551.

[5] M. Krifka, Approximate interpretations of number words: A case for strategic communication, in: G. Bouma, I. Krämer, Z. Joost (Eds.), Cognitive foundations of interpretation, Koninklijke Nederlandse Akademie van Wetenschapen, Amsterdam, 2007, pp. 111–126.

[6] G. Lakoff, Hedges: A study in meaning criteria and the logic of fuzzy concepts, Journal of philosophical logic 2 (4) (1973) 458–508.

[7] S. Lefort, M.-J. Lesot, E. Zibetti, C. Tijus, M. Detyniecki, How much is "about" ? Fuzzy interpretation of approximate numerical expressions, in: Proc. of 16th Int Conf on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'16), Vol. 610, Springer, Eindhoven, Netherlands, 2016, pp. 226–237.

[8] S. Ferson, J. O'Rawe, A. Antonenko, J. Siegrist, J. Mickley, C. C. Luhmann, K. Sentz, A. M. Finkel, Natural language of uncertainty: numeric hedge words, International Journal of Approximate Reasoning 57 (2015) 19 – 39.

[9] S. Solt, An alternative theory of imprecision, in: Semantics and Linguistic Theory, Vol. 24, 2014, pp. 514–533.

[10] S. Solt, Vagueness and Imprecision: Empirical Foundations, Annual Review of Linguistics 1 (1) (2015) 107–127.

[11] J. M. Sadock, Truth and approximations, in: Proceedings of the Annual Meeting of the Berkeley Linguistics Society, Vol. 3, 1977.

[12] S. Dehaene, L. Cohen, Towards an anatomical and functional model of number processing, Mathematical cognition 1 (1) (1995) 83–120.

[13] S. Dehaene, The neural basis of the Weber-Fechner law: a logarithmic mental number line, Trends in Cognitive Sciences 7 (4) (2003) 145–147.

[14] J. Halberda, M. M. Mazzocco, L. Feigenson, Individual differences in non-verbal number acuity correlate with maths achievement, Nature 455 (7213) (2008) 665–668.

[15] J. Halberda, L. Feigenson, Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults., Developmental Psychology 44 (5) (2008) 1457.

[16] P. Pica, C. Lemer, V. Izard, S. Dehaene, Exact and approximate arithmetic in an amazonian indigene group, Science 306 (5695) (2004) 499–503.

[17] U. Sauerland, P. Stateva, Scalar vs. epistemic vagueness: Evidence from approximators, Proceedings of SALT (1995) (2007) 228–245.

[18] C. Jansen, M. Pollmann, On round numbers: Pragmatic aspects of numerical expressions, Journal of Quantitative Linguistics 8 (3) (2001) 187–201.

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

[20] M. Ehrgott, Multicriteria optimization, Vol. 491, Springer Science & Business Media, Berlin, 2013.

[21] C. Wright, On the coherence of vague predicates, Synthese 30 (3-4) (1975) 325–365.

[22] B. Winer, D. Brown, M. Michels, Statistical Principles in Experimental Design, McGraw-Hill, Boston, 1991.

[23] T. Bilgiç, I. Türkşen, Measurement of membership functions: Theoretical and empirical work, in: D. Dubois, H. Prade (Eds.), Fundamentals of Fuzzy Sets, Vol. 7 of The Handbooks of Fuzzy Sets Series, Springer US, 2000, pp. 195–227.