International Journal of Approximate Reasoning 57 (2015) 19–39

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Natural language of uncertainty: numeric hedge words

Scott Ferson [a,*], Jason O'Rawe [a,b], Andrei Antonenko [a,c], Jack Siegrist [a], James Mickley [a,d], Christian C. Luhmann [e], Kari Sentz [f], Adam M. Finkel [g]

[a] Applied Biomathematics, 100 North Country Road, Setauket, NY 11733, USA
[b] Genetics Program, Stony Brook University, 100 Nicolls Road, Stony Brook, NY 11794, USA
[c] Linguistics Department, Stony Brook University, Stony Brook, NY 11794, USA
[d] Ecology & Evolutionary Biology, The University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, CT 06269-3043, USA
[e] Psychology Department, Stony Brook University, Stony Brook, NY 11794, USA
[f] Los Alamos National Laboratory, Post Office Box 1663, MS F609, Los Alamos, NM 87545, USA
[g] The University of Pennsylvania Law School, 3501 Sansom Street, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 1 January 2014
Received in revised form 2 August 2014
Accepted 10 November 2014
Available online 14 November 2014

Keywords:
Approximator
Linguistic expression of uncertainty
Hedge
Amazon Mechanical Turk
Elicitation
Uncertainty communication

ABSTRACT

An important part of processing elicited numerical inputs is an ability to quantitatively decode natural-language words that are commonly used to express or modify numerical values. These include 'about', 'around', 'almost', 'exactly', 'nearly', 'below', 'at least', 'order of', etc., which are collectively known as approximators or numerical hedges. Figuring out the quantitative implications of these expressions for the uncertainty of numerical quantities is important for being able to understand, for example, what is actually being reported by a patient who says a headache has lasted for "about 7 days", and how we should translate the patient's report into uncertainty about the duration. We used Amazon Mechanical Turk to empirically identify the implications of various approximators common in English. To evaluate the numerical range implied by each approximator, we analyzed paired statements differing only in the approximator used in numerical expressions. Despite often considerable diversity, there were several statistically significant findings, but far less quantitative variation implied by the approximators than might have been expected. The numerical implication of different approximators interacts with the magnitude and roundness of the nominal quantity. This investigation strategy generalizes easily to languages other than English.

1. Introduction

The innate numerical acuity of humans is remarkably poor [16]. Although recognizing and thinking with numbers may involve multiple cognitive systems that are not yet fully understood (cf. [11]), laboratory and field observations show that without tutoring people typically have a number sense that can distinguish only up to about four items [29]. Human societies that have not developed number systems or at least finger counting have difficulty discerning the quantity four [26], and humans seem to innately distinguish only the quantities one, two and many, which represents any quantity more than two.

* Corresponding author. Tel.: +1 631 751 4350.
E-mail address: sandp8@gmail.com (S. Ferson).

http://dx.doi.org/10.1016/j.ijar.2014.11.003
0888-613X/

Number systems were invented repeatedly in human history [30], apparently to facilitate commerce, to bring clarity to ideas otherwise expressed by words like 'some', 'many', 'more', 'less' and 'fewer'. These systems permit the description of quantities with expressions consisting of a numeral and units. The units specify the scale of measurement, which is either a count noun (e.g., dollars, days, chairs, bushels, people, acres), or what linguists call a measure word used with a mass noun (e.g., kernels of corn, bushels of wheat, liters of water, tanks of gas, rooms of furniture). The numeral represents an integral count or real-valued measurement revealing the multiplicity or fractionality of the unit equivalent to the quantity being described. The numeral expresses a magnitude, possibly spelled out in words ('one', 'two', 'sixty-eight', 'three quarters', etc.) or expressed with numerical digits ('1', '256', '0.5', etc.).

The clarity of number systems often implies greater precision than is practically achievable in many situations. This fact requires some scheme to relax or discount this precision. In linguistics, a *hedge* is a word or phrase that modifies the force or precision of ideas or statements [37]. Hedges serve several purposes in language, including expressing uncertainty or transience, stipulation, responsibility focusing, and obfuscation. Prince et al. [50] recognized two kinds of hedges: shields and approximators. Shields, such as 'I think' or 'probably' modify propositions, whereas approximators modify numerals to alter the magnitude or precision implied by the expression. The latter function of approximators is to convey that the quantity is either less precise or more precise than the meaning of the corresponding numerical quantity without the hedge word. For instance, the sentence 'About 105 people came to the party' may mean that any number of people between 100 and 110 came to the party. In contrast, the sentence '105 people came to the party' has a smaller range of possible values for the implied number of people attending the party. In this case, the approximator 'about' introduces more uncertainty into the interpretation of the sentence.

Sometimes uncertainty is implicit in a numerical expression because of the roundness of the number even though no explicit hedge words may be present at all. For example, a reasonable interpretation of the phrase '1000 people came to the protest' would infer that the number of people who attended the protest is somewhere in the neighborhood of 1000, but not necessarily exactly 1000. Comparing that example to a phrase 'Exactly 1000 people came to the protest', one can see that the hedge 'exactly' reduces the uncertainty of the statement: the latter example means that there were exactly 1000 people at the protest, no more, no less.

In English, quantities are described with expressions generally involving three elements: *unit*, *numeral*, and *approximator*. Grammatically, the approximator is an adverb that modifies the numeral which is an adjective which in turn modifies the unit which is a noun. The order in which the three elements appear is not fixed in English. For example, the written phrase '$100 or so' is unit–numeral–approximator, but 'nearly 5 pounds' is approximator–numeral–unit, and '35 years or more' is numeral–unit–approximator. Sometimes elements may be elided when context or convention allows. The phrase 'three coffees' omits the unit (measure word) 'cup'. Mathematicians discuss abstract quantities which are pure, dimensionless numbers without units. The idea is not so much that there are no units, but that the numbers represent quantities with *any* units. When the numeral is omitted, it is usually understood to be one, unless context forces another value. Omitting the approximator element—using what we might call the null hedge or *null approximator*—does not usually mean there is no imprecision whatever about the quantity. Instead, the value is understood to have a precision implicitly encoded in the roundness of the number, the discourse environment (e.g., bank statements versus barroom braggadocio), and measurability of the quantity.

There are many approximators in English, including generic hedges such as 'around' and 'nearly', archaic hedges such as 'well-nigh', and idiomatic constructions such as 'in the ballpark of'. Some hedges generally appear before the numeral like 'around' and 'as many as', and some generally appear after the numeral like 'or so' and 'and change'. Some approximators can appear either before or after the numeral like 'approximately', 'almost' and 'at least'. Table 1 lists many approximators in wide use which are distinguished into four categories. Channell [14] asserted that all of the approximators imply a range of possible values for the quantity being described. Sometimes this interval is explicitly indicated with ranging constructions like '5 or 6' and '15 or 20' and 'between 86 and 94', but many hedged numerical expressions refer to a single exemplar number, about which the interval of imprecision is understood to be symmetrically or asymmetrically positioned around this value. For example, 'around 5' is symmetric, whereas 'more than 5' is asymmetric. The null approximator is in a category by itself.

Sadock [55] argued that approximators but also many other factors affect the implied imprecision about a quantity. From introspective linguistic analysis of pairs of natural-language expressions such as

1 million	990 000
about 1 million	1 million
about 990 000	990 000
about 1 million	about 990 000
about a dozen	about 12
about two and a half	about 2.5
six-foot insect	six-foot man

where the value on the left is understood to be less precise than the value on the right, he concluded that the roundness of the number mentioned, its display format, the possible range of the quantity, the relevant standards of precision, and the units themselves including whether they refer to discrete entities or mass nouns all affect the implied imprecision about