Giving the wrong impression: Strategic use of comparatively modified numerals in a question answering system

Christoph Hesse
Center for General Linguistics, Berlin
hesse@leibniz-zas.de

Anton Benz
Center for General Linguistics, Berlin
benz@leibniz-zas.de

Abstract

We develop a model simulating real estate question-answer dialogues which reasons strategically in generating answers containing numerals modified by the comparative quantifiers “more than” and “fewer than” by determining the expected utility of answer content based on qualitative differences among alternatives’ numerical attributes. We evaluate this model by successfully testing its ability to deceive users into perceiving a qualitative difference where there is none.

1 Introduction

In natural language, questions are often answered indirectly. Indirect answers are not only often more informative than a simple ‘yes’ or ‘no,’ they also communicate to the addressee that the speaker has reasoned about what their underlying motivation for asking the question might have been. Comparatively modified numerals such as “more than one hundred” can be used to signal such an understanding. The answer in (1) communicates that the speaker has understood that “nearby” signals an underlying decision problem where ‘closer’ is ‘better.’

Comparatively modified numerals can be used as indirect answers to yes/no questions. As an indirect yes:

(1) Q: Is there a bus stop nearby?
A: There’s a bus stop less than 100m away.

As an indirect no:

(2) Q: Is there a bus stop nearby?
A: There’s a bus stop more than 4 miles away.

At the same time, modified numerals carry connotations which are sensitive to discourse context and user preferences:

(3) The bus stop is more than ... away. ⇝ far
The bus stop is less than ... away. ⇝ near

(4) The apartment has more than ... sqm. ⇝ big
The apartment has less than ... sqm. ⇝ small

(5) The rent is more than .... ⇝ expensive
The rent is less than .... ⇝ cheap

The connotations of scalar quantifiers can be used strategically to make indirect answers more informative.

(6) Q: Is there a subway station nearby?
A1: No. → no subway
A2: Well, there’s a bus stop. ⇝ no subway
⇝ alternative
A3: There’s a bus stop, but it’s more than ... away. ⇝ no subway
⇝ alternative
⇝ distance estimate
⇝ alternative not close to apt.

A direct negative answer as in A1 in (6) is neither very informative nor helpful because although it answers the question it leaves open whether there are alternative options. The user would thus have to follow up with questions about potential alternatives. Whether there is a bus stop has no logical bearing on whether there is a subway station, yet A2 in (6) is perfectly natural and more informative than A1 because it names an alternative public transport and thereby indirectly negates the existence of a subway station. A3 in (6) is even more informative than A2 because not only does it provide an alternative to the non-existent subway station, but it also specifies which criterion the choice of an alternative was based on (distance), how far away the bus stop roughly is, and by choosing “more than” rather than “less than” as quantifier, the answer implies that although the bus stop is an alternative, it is not ‘nearby’.

By implementing the model within a question answering system and simulating real estate dialogues, we are able to evaluate it in terms of dialogue efficiency, perceived coherence of answers, and users' ability to draw natural pragmatic inferences. Users of the system have the ability to pose questions to it, for instance, "Is there a bus stop nearby?" meaning near the property, and the system should answer the question with either a direct positive or negative answer (containing no modified numeral) or an indirect positive or negative answer (containing a modified numeral). If there is an alternative which is as good as what the user asked about, the system should present both the original object and the alternative. In case of an indirect answer, the system needs to decide two things: First, which level of precision to round the numeral to. Second, which comparative modifier to use.

2 Model

2.1 Overview

The model builds on work by Jameson (1983); Frank and Goodman (2012); Benz (2015); Stevens et al. (2015); Zeevat and Schmitz (2015); Potts et al. (2016); Qing et al. (2016); Stevens et al. (2016) and extends on them by reinterpreting an alternative's expected utility of a cost/benefit ratio as complex coefficients of numerical attributes. It models an exchange of questions and answers where a user poses a question to be answered by a real estate agent (the system) who alone has access to a database of information about real estate properties. Users ask for a range of attributes of the properties (public transports in the vicinity, supermarkets, restaurants, coffee shops, etc.). The realtor does not know a user's intentions for asking about a specific attribute. The user, on the other hand, does not know the attributes of a particular property.

The user's goal is to find out whether a given property satisfies a set of requirements (e.g., the user is told to look for a property which has a balcony, a supermarket and a subway station nearby, and which is in a particular price range). The goal of the real estate agent is to help the user find a suitable property as efficiently as possible. Given this goal, the realtor should provide alternatives to the attributes specifically mentioned in a user's question only when those alternative attributes are relevant to the user's underlying requirements.

Numerical attributes (e.g., the distance of the nearest subway station) need to be in a class separate from binary attributes (e.g., whether a property has a balcony or not). And, for instance, although, say, subway stations, bus stops, tram stations, supermarkets, farmers' markets, restaurants, and coffee shops all share certain numerical attributes (e.g., distance from the property), public transports need to be in a separate class from groceries, and separate from places for eating out.

Let r be a requirement the user has, a an attribute which is considered part of the sets of alternatives, and q the questions the user poses to the system. The user knows r, but not a. The system knows a, but has to reason probabilistically which r best applies to q in order to match it with a suitable a. The realtor is utility-maximizing under game-theoretic assumptions, and should aim to look up attributes for which benefit outweighs cost (Stevens et al., 2015).

The cost of considering attribute a is $\mathrm{C}(a)$. $\mathrm{B}(a|r)$ is the benefit of a given r (e.g., a balcony is not as good for gardening as a garden). $\mathrm{P}(r|q)$ is the likelihood that question q implies requirement r (e.g., "Is there a garden?" implies more strongly $r =$ gardening than "Is there a balcony?"). In order to find how useful an attribute a is as a potential answer to q, the realtor needs to consider the weighted average of $\mathrm{B}(a|r)$ and $\mathrm{P}(r|q)$, where $\mathrm{P}(r|q)$ is calculated using Bayes' rule.

$$\mathrm{EU}(a|q,\{r_1 \dots r_n\}) \propto \sum_{i=1}^{n} \mathrm{B}(a|r_i)\mathrm{P}(r_i|q) - \mathrm{C}(a) \qquad (1)$$

Based on their requirements, users may only consider certain numerical attributes but not others. For instance, when asking "Is there a supermarket near by?" one user wants to find the closest one or one with a certain quality of products. Another person might ask about a supermarket because they want a compromise of both distance and quality of products. Over the coarse of the conversation, the realtor should identify which numerical factors are important to a client's decision problem. If we think of weights associated with each component of the coefficient, then updating amounts to shifting weights. The expected utility is then the weighted average over the numerical factors, as long as we map factors to benefits and costs according to their proportionality relations,

$$\begin{aligned}\mathrm{EU}(a|q,\{r_1 \dots r_n\}) &\propto \log\left(\frac{R^{\alpha} \cdot Q^{\beta}}{D^{\gamma} \cdot P^{\delta}}\right) \qquad (2)\\ &\propto \underbrace{(\alpha R + \beta Q)}_{\text{benefits}} - \underbrace{(\gamma D + \delta P)}_{\text{costs}}\end{aligned}$$

Suppose we are looking at a supermarket which is D away from the real estate property, has a range of products of size R and quality Q and is in a price range P. The proportionality relations among them are: The further away, the lower expected utility, EU $\propto 1/D$. The more expensive, the lower expected utility, EU $\propto 1/P$. However, the bigger the range and quality of products the more useful, EU $\propto R$ and EU $\propto Q$. The arity (complexity) of the quality coefficient is the dimensionality of the space of numerical attributes.

A quality coefficient captures natural intuitions when comparing the expected utilities of two alternatives. Suppose there are two supermarkets, one small but close, the other large but farther away. The small supermarket has a limited range of products; the big one a large range. If the range of products of the large supermarket is now twice as big as that of the small one while one is also twice the distance than the other, then range R and distance D will lead to the same value of the coefficient for both, and so they will be equally useful options for the user.

The domain of real estate offers lots of opportunities to generate modified numerals. When generating numerals, the model needs to be aware that expectations about rounding precision vary among numerical attributes and users. Strong expectations about the rounding precision exist for price, floorspace, and also for the relationship among certain attributes. For instance, people have expectations about the size of a property given the number of rooms and vice versa, and these expectations tie in with their expectations about prices. For other attributes prior expectations about precision are less strong (e.g., distance to public transports, restaurants, etc.).

In evaluating our system, we use the example of the distance of a subway station from a residence. What we will see is that 'pragmatic' rounding rules differ from mathematical rounding. '*More than 1 mile*' literally means 'any number of miles greater than one mile.' The literal meaning is obviously not how speakers understand the statement: No one would expect an actual distance of 745,957 miles being told '*more than 1 mile*.' According to Geurts and Nouwen (2007), by deliberately saying '*more than 1 mile*' instead of '*more than 2 miles*' the speaker implicates that the true distance cannot be ≥ 2 miles. So it should be adequate for our system to generate '*more than 1 mile*' whenever the true distance is between 1 and 2 miles. Rounding mathematically would mean that every time the true distance is > 1 & < 1.5, the system should round down and say '*more than 1 mile*'; when it is ≥ 1.5 & < 2 the system should say '*less than 2 miles*,' but that is not how our system rounds.

Num	$D_<$	$D_>$
20	5	5
30	5	10
40	10	10
50	5	10
60	10	10
70	10	10
80	10	10
90	5	10
100	10	10
110	10	10
120	10	10
130	10	10
140	10	10
150	10	10
160	10	10
170	10	10
180	10	5
190	10	10
200	10	10

Table 1: Experimental results. Assumed deviation of true value from modified numeral.

In a series of three experiments (Hesse and Benz, Submitted 2018), 1,270 participants on Amazon Mechanical Turk were shown modified numerals of different degrees of roundness (93 = fine, 90 = medium, 100 = coarse), relative roundness differences (e.g., 100 vs. 110 and 1,000 vs. 1,100), absolute magnitudes (110 > 90, but both are similarly less round than 100), and orders of magnitude (< 100, hundreds, thousands, tens of thousand). Participants saw for instance "Toby just went to the movies to see a blockbuster. How long do you think the movie was? Between ___ and ___, most likely ___." In the case of movie length, they have strong expectations, but we found that when speakers have weak expectations, there is a fixed distance between the modified numeral and how far off they think this 'rounded' numeral is from the true value relative to the numeral's order of magnitude (see the excerpt of our dataset in Tab. 1 from which we extrapolate). At the same time, experimental participants show a strong preference for fractal division points (see Fig. 1) as previously found by Jansen and Pollmann (2001) and Dehaene and Mehler (1992). Rounding granularity is chosen such that the implied deviation from the modified numeral comes closest to the true value.

For example, suppose there is a bus stop in the database which is 276 meters away from the property and the system has to decide whether to generate "less than 300" or "less than 290." On the one hand, there is the deviation from the modified numeral which is assumed to be about 10 for both 300 and 290. If the system generates "less

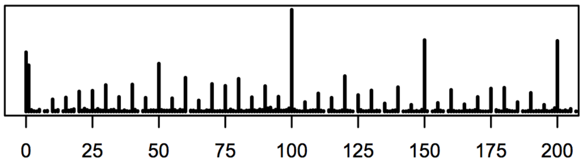


Figure 1: Experimental results. Frequency of true value estimates of "more than n" with $0 \leq n \leq 200$.

than 300," users should assume the true value is somewhere around $300-10=290$; if it generates "less than 290" they should assume it is around $290-10=280$. 280 is closer to the true value than 290. On the other hand, from Fig. 1 we can see fractal steps of 100 (as in 300) are about two to three times as likely reference points as steps of 10 (as in 290). Cognitive effort trumps precision in this case because distances to bus stops are a context with weak expectations. The system should therefore generate "There's a bus stop less than 300 m away."

Scalar quantifiers are chosen with the context-sensitive nature of their connotations in mind. For instance, for distances we can think of as 'nearby' we characterize a radius within which everything is considered 'within walking distance' (or by car). It is therefore prudent to use "less than" for all objects within this radius since "more than" would make them seem 'far' from the property. Associated with this radius is threshold κ_1 on the expected utility of objects that lie on it. We also need a second radius, an exclusion zone, so that when objects lie beyond they are not considered for the set of alternatives. This exclusion radius is the physical counterpart to requiring a *minimum* expected utility κ_2.

2.2 Algorithm and example

Assume the user asks 'Is there a $x=$ subway station nearby?' For illustration, we assume the quality coefficient only hinges on distance with $\mathrm{B}(a \mid r=\text{dist.})=\frac{1}{1+\text{dist.}}$. Fig. 2 illustrates the decisions necessary to arrive at one of seven possible replies to the question (two direct answers and five indirect answers). Left branches at each node are a 'yes' to the conditions of that node, right branches a 'no.' First, we have to decide whether the public transport x explicitly mentioned in the user's question is within the first radius,

$$\mathrm{EU}(x=\text{subway} \mid q, p=\text{dist.}) > \kappa_1. \tag{3}$$

If this holds true, we next look if there are alternative modes of public transport y beyond the first radius,

$$\mathrm{EU}(y=\text{bus, tram, } \ldots \mid q, p=\text{dist.}) < \kappa_1. \tag{4}$$

If this holds true, then other public transport y will not be good alternatives to x, so the system should generate a direct positive answer. If it does not hold, then we next check whether those alternative public transports y would be more useful than x,

$$\begin{aligned} &\mathrm{EU}(x=\text{subway} \mid q, p=\text{dist.}) \\ &> \mathrm{EU}(y=\text{bus} \mid q, p=\text{dist.}). \end{aligned} \tag{5}$$

If they are closer to the property than x, they will be more useful than x, so the system should generate "There's an x less than n_x away." because this answer implicates that x is close to the property and that there are no better alternatives. If there are better alternatives, the system should mention them alongside x as "There's an x and a y less than n_x away" (since y will be closer to the property than x, y's distance falls within x's distance from the property).

If x, the public transport explicitly mentioned in the question, is not within the first radius, we still need to check if it is within the second radius. The second radius is an exclusion perimeter. Anything beyond the second radius is too far away to be a reasonable alternative, but those between radius one and two might still be preferable compared to those beyond radius two. We therefore check

$$\mathrm{EU}(x=\text{subway} \mid q, p=\text{dist.}) > \kappa_2. \tag{6}$$

If it holds, x lies between the first and the second radius, and we next check for alternatives y. Since x lies between radius one and two, alternatives y within the first radius would be preferable to those which also happen to be between radius one and two. Alternatives outside the second radius are of no interest here because they cannot be preferable to x,

$$\mathrm{EU}(y=\text{bus, tram, } \ldots \mid q, p=\text{dist.}) > \kappa_1. \tag{7}$$

If the condition holds, there is a y within the first radius and the system should generate "Well, there's a y less than n_y away." "Less than" implies that y is close to the property. If the condition does not hold, y is no better than x and so the system should generate "There's an x more than n_x away." "More than" implies that x is not close to the property, and the fact that no alternatives are mentioned implies that there are no alternatives better than x.

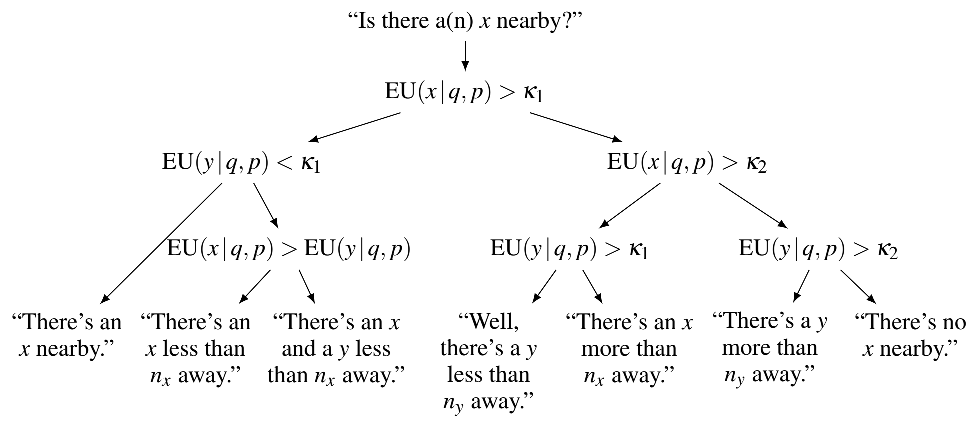

"Is there a(n) x nearby?"
$\mathrm{EU}(x\,|\,q,p) > \kappa_1$
$\mathrm{EU}(y\,|\,q,p) < \kappa_1$
$\mathrm{EU}(x\,|\,q,p) > \kappa_2$
$\mathrm{EU}(x\,|\,q,p) > \mathrm{EU}(y\,|\,q,p)$
$\mathrm{EU}(y\,|\,q,p) > \kappa_1$
$\mathrm{EU}(y\,|\,q,p) > \kappa_2$
"There's an x nearby."
"There's an x less than n_x away."
"There's an x and a y less than n_x away."
"Well, there's a y less than n_y away."
"There's an x more than n_x away."
"There's a y more than n_y away."
"There's no x nearby."

Figure 2: Decision tree for generating modified numerals.

If x is beyond the second radius, it should not appear in the set of alternatives, but there might still be better alternatives y. We therefore need to check whether there are some public transports within the second radius,

$$\mathrm{EU}(y = \text{bus, tram, } \ldots\,|\,q, p = \text{dist.}) > \kappa_2. \quad (8)$$

If there is a y for which this condition holds true, this y is still better than nothing, and so the system should generate "There's a y more than n_y away." Only if both x and all alternatives y are outside the second radius, should the system generate a direct negative answer. In other word, in this case the set of alternatives is empty.

The system makes strategic use of the connotations of scalar quantifiers. So when an object is in the vicinity of the property, it should generate answers with quantifiers that imply it is 'close' rather than 'far away.' "Less than" implies that it is 'close'; "more than" implies that it is 'far away.' Therefore, when an object is in the vicinity of the property (within the first radius), the system rounds distances up to the nearest fractal reference point and generates "less than." When an object is outside the first radius, the system rounds distances down to the nearest reference point and generates "more than."

3 Implementation

A simple interactive question answering system was built using Javascript with a database back-end. The system emulates the behavior of a real estate agent answering customer questions such as "Is there a subway station nearby?" out of a set of nine attributes pertaining to a real estate property (three basic questions about price, floorspace, and the number of bedrooms; three questions about the availability of public transports; three questions about the availability of grocery shops). The system generates answers by considering all alternatives within a minimal distance from the property, rounding the distance estimate to a contextually appropriate level of precision, modifying it by the comparative quantifier "more than." The system then translates this information into natural language by using simple sentence templates like "There's a(n) X [more than n [unit]] away." Participants in the evaluation study interact with the system through a text-based web application over an internet browser.

4 Evaluation

We test our dialogue system's ability to suggest qualitative differences via choice of the rounding precision of a numeral. We start from the assumption that a system which can communicate qualitative differences is one which can make things that are objectively speaking not different *seem* like they are. A real estate sales dialogue is a common example of a situation where deception may take place, but where interlocutors are also very alert to signals that such deception is taking place.

Participants in the evaluation study are led to believe they will view five different properties for a friend who is looking to buy a house, but in actuality two of the houses are identical. Participants are misled by the real estate agent (our system) which will generate a different numeral for the two identical houses with respect to one attribute (the distance to the nearest subway station) so as to make it seem like there is a qualitative difference between the two. For one, the system will generate a vague expression using a comparatively modified numeral ("more than 1 mile"), for the other, it will generate an exact unmodified numeral ("1.2 miles" or "1.7 miles"). The unmodified numeral is the objective distance rounded to one decimal. The critical distance of the subway station is just outside the inner radius (thus EU $< \kappa_1$).

The experiment has two parts: In the first part, participants are deceived into thinking the subway station is further away from one of the two houses than the other. Since the two houses are identical, participants should not prefer one over the other. Finding such a preference would be evidence that the deception was successful. In the second part of the experiment, we reveal the true distance of the subway station and ask participants whether they feel they have been misled by the realtor. We would expect that if the true distance is within the range where participants expect potential values, they will not feel misled although the perceived qualitative difference between the properties is only apparent.

4.1 Methods

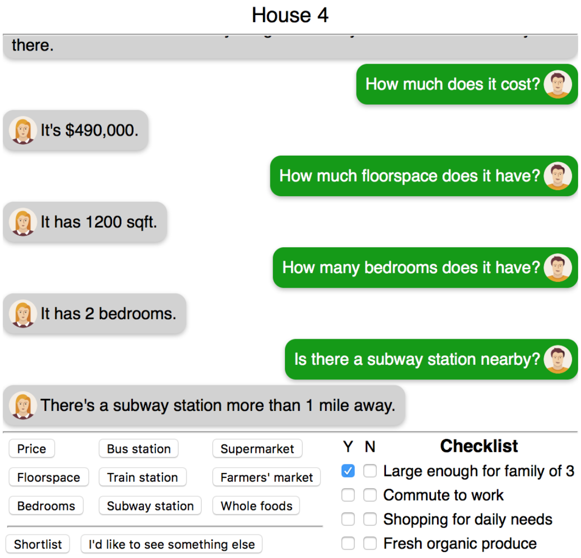
4.1.1 Participants

We recruited 100 participants (43 female, 55 male, 2 who indicated no gender, mean age 35.9 years) with U.S. IP addresses via Amazon's Mechanical Turk and required them to provide information about their native language, foreign language skills, gender, age, and level of education. Out of the 100, 24 participants failed to properly engage in posing questions to the realtor, failed to shortlist any houses, failed to choose their favorite house or failed to indicate whether they felt deceived or not. The remaining 76 native English speakers were included in the final analysis. Participants were assigned to one of two versions of the evaluation study (50 participants each). Participants received 30 cents for their participation.

4.1.2 Material and procedure

At the beginning of the study, participants were required to answer questions concerning their age, gender, native and non-native languages, and educational level. Then they are instructed to imagine they are helping a friend buy a house for his family of three in Brooklyn, New York. This cover story includes four requirements for the new home. They need two bedrooms, one master bedroom and one for their daughter. The first requirement is not subject to gradient qualitative differences (either a house has at least two bedrooms or not) whereas the next three are. They would like a supermarket nearby. The friend would like to use public transport to get to work. This requirement can be met by at least one of three options: subway, bus or train. A property with all three types of public transport in the vicinity would be qualitatively superior to a property with only two or only one type. The friend would also like an organic shop, a requirement which would be met by a whole foods store or a farmers' market. Qualitative differences are also expressed by proximity to the potential home. A property with a subway station close-by is more attractive than one with a subway further away.

Next participants saw a tutorial describing the elements of the user interface. In the top two thirds of the screen, they see their conversation with the real estate agent and in the bottom third are buttons to pose questions to the agent, a checklist of the friend's requirements, an option to shortlist properties which meet the requirements, and a button to look at another listing. At the beginning of each new offering, the realtor gives some general information about the location of the property. Participants are free to ask as many or as few questions as they like in order to find out whether a particular property meets the requirements. See Fig. 3 for a screenshot.



House 4

there.

How much does it cost?

It's $490,000.

How much floorspace does it have?

It has 1200 sqft.

How many bedrooms does it have?

It has 2 bedrooms.

Is there a subway station nearby?

There's a subway station more than 1 mile away.

Y N **Checklist**

☑ □ Large enough for family of 3
□ □ Commute to work
□ □ Shopping for daily needs
□ □ Fresh organic produce

Figure 3: Screenshot of the dialogue system.

After they have viewed five listings, participants are required to review the properties they have shortlisted. For each property, the shortlist highlights attributes of it that were part of the previous conversation. So for instance, if they talked to the realtor about subway stations near, say, house 4, the shortlist will include this information for house 4. Participants are asked to choose their favorite among the shortlisted properties and to write a note to their friend motivating their choice. The study is set up in a way that only two out of the five offers met all requirements (house 4 and 5). All attributes of house 4 are identical to house 5 except one: how the real estate agent phrases the proximity of a subway station. When participants ask about a subway station near house 4, they are told it is "more than 1 mile away." When they pose the same question for house 5, the agent will give them an exact distance. Participants in the first version of the study are told the subway station is "1.2 miles away" from house 5; those in the second version are told the station is "1.7 miles away."

What participants in both groups do not know is that the database entries of house 4 are identical to those of house 5. It is only the realtor's phrasing that makes the subway station seem further away from one house than the other. Those participants who are told the subway is "1.7 miles away" from house 5 should perceive a greater apparent difference in distance to the subway station of house 4 than those participants who are told the subway is "1.2 miles away" from house 5.

After they submitted the shortlist to their fictitious friend, they enter the evaluation phase of the study where they are asked to give a verbal description of their satisfaction with the generated answers. Then they are asked to indicate whether distance expressions should be more precise or less precise on 7-points Likert scales, first in general and then for each mode of public transport and grocery options individually. On the next page, they are confronted with those distance expressions they received from the realtor for the house they marked as their favorite in the shortlist earlier. They see the realtor's exact phrasing. Next, the true distances are revealed, and participants must indicate whether they feel they have been misled by the realtor's phrasing or whether it was an appropriate simplification. Suppose, for instance, someone favored house 5 and was told the nearest subway station is "1.7 miles away." When the true distance is revealed to be 1.741 miles they might feel that the realtor merely rounded it off in order to make it easier to grasp (following Grice's maxim of Manner, Grice, 1989). Someone who favored house 4 over house 5 when they were also told that the subway was "1.7 miles away" from house 5, on the other hand, might feel misled when they learn that "more than 1 mile away" actually meant "1.7 miles away." Hence, the study used a 2 (comparatively modified numeral versus unmodified numeral) × 2 (a subway station 1.2 miles or 1.7 miles away from house 4 and 5) design. The order of requirements and question buttons was pseudo-randomized and participants only saw one version of the study.

4.2 Results

Fig. 4 shows the proportion of participants who favored a particular house. Out of the 100 participants only 11 favored houses 1, 2 or 3. About two thirds of participants favored house 4 and about a quarter favored house 5. Notice that among those who favor house 4, the difference between those who were told the subway was "1.2 miles away" (blue) from house 5 versus those who were told it was "1.7 miles away" (red) is proportional to the difference in preference for house 5 between the two groups of participants. 27 out of 39 participants (69.2%) choose house 4 when they are told that the subway station is "1.7 miles away" from house 5; 21 out of 37 participants (56.8%) choose house 4 when they are told the station is "1.2 miles away" from house 5. Only 7 out of 39 participants (18%) choose house 5 when they are told the nearest subway station is "1.7 miles away," but 10 out of 37 (27%) prefer house 5 when they are told the subway is only 1.2 miles away. Overall, the preference distribution in the two groups of participants is very similar to one another ($r = .98, p < .01$).

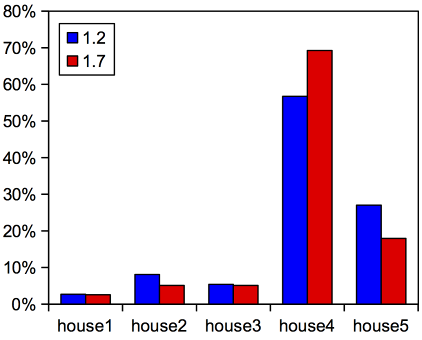


Figure 4: Portion of participants (%) who favored a particular house. Blue are participants who were told the subway was "1.2 miles away" from house 5; red those who were told it was "1.7 miles away."

The fact that participants most often shortlist house 4 and 5, and then favor one of them, indicates that they perceive a qualitative differences between the two house. The realtor suggests a qualitative difference by qualifying the distance to a subway station as either “1.2 miles away” from house 5 or “1.7 miles away” from house 5, while telling participants in both conditions the subway is “more than 1 mile away” from house 4. This perceived difference is an apparent qualitative difference because in actuality house 4 and house 5 have the same database entry. We are interested in whether they will feel misled when they learn what the true distance of the subway station from house 4 is. Consequently, the following analysis is based only on those 48 participants who chose house 4 as their favorite. When participants learned the true distance they indicated on a 7-point Likert scale whether they felt they had been misled (-3) or whether they felt the use of an imprecise numeral was appropriate (+3). Fig. 5 shows that, on average, participants who opted for house 4 when they were told the subway station was “more than 1 mile away” but “1.7 miles away” from house 5 (red), give a rating of -0.370, which is 1.561 lower than the rating given by participants who compared house 4 to a house 5 where the subway was only “1.2 miles away” (blue).

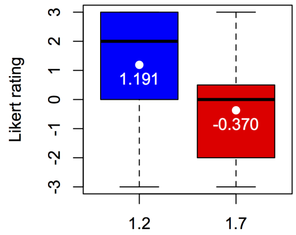


Figure 5: Likert scale ratings of perceived deception (-3) or adequacy ($+3$) of being told a subway station is “more than 1 mile away” when in actuality it is either 1.2 miles away (blue) or 1.7 miles away (red).

We fitted a linear mixed effects model to the Likert ratings with participants’ group membership as fixed effect and by-subject variation as a random effect with random intercepts and random slopes. The mean rating of 1.191 of participants who saw house 5 with the subway “1.2 miles away” is the reference level. Participants who later learned the subway station which they were told is “more than 1 mile away” is actually “1.7 miles away” give significantly lower Likert ratings than participants in the other group (1.561 lower, std. err. = 0.535, $t = -2.917$, $p = 0.0054$). For comparison we fitted a model without the effect group but with by-subject variation to the Likert ratings (effectively a linear regression). The comparison model estimated that by-subject variation only accounted for a lowering in ratings of 0.06 instead of the mixed effects model which estimated that group predicts a lowering of 1.561, the lowering actually observed. Details of the mixed model are presented in Tab. 2.

	Est.	Std. err.	t-value	p-value
“1.2” (Int.)	1.191	0.401	2.967	0.0048 ***
“1.7”	−1.561	0.535	−2.917	0.0054 ***

Table 2: Results of mixed effects model including estimates, standard errors, t-values, and p-values (n = 48, log likelihood = -96.46).

It is of course trivial that experimental participants opt for the house where the subway station is closer, but that is not what we are interested in. We are interested in where they think the true distance of ‘*more than 1 mile*’ most likely falls (is it below 1.5 miles or above 1.5 miles). Our reasoning is this: If the distance which participants think of for ‘*more than 1 mile*’ is the same as the exact distance they are told for the other house, then it should not matter to them which house they choose because they perceive them to be qualitatively identical. We see in Fig. 4 that even when the subway is 1.2 miles away from house 5, twice as many participants opt for house 4 than for house 5. We can therefore deduce that the true value they expect when they read ‘*more than 1 mile*’ is between 1.0 miles and 1.2 miles. In the other group of participants who are told the subway is 1.7 miles away from house 5 (red in Fig. 4), even more people opt for house 4 than in the first group (blue in Fig. 4). So the system should not generate ‘*more than 1 mile*’ whenever the actual distance is between 1 and 1.5 miles, it should only be generated when the actual distance is between 1 and 1.2 miles (from our experiments in Hesse and Benz, submitted 2018, we suspect that the expected distance should be somewhere around 1.1 miles). We see that we cannot simply round mathematically because the tipping point for rounding up or down is not 1.5 miles but much closer to 1 mile.

Knowledge of these ‘human’ rounding rules also needs to inform how the system decides between quantifiers: If ‘*more than 1 mile*’ means, say, about 1.1 miles, then any distance between 1.1 and 2 miles should not result in generating ‘*less than 2 miles*’ because ‘*less than 2 miles*’ would actually imply a distance of around 1.9 miles. So ‘*more* *than n*' implies $n+0.1$ and '*less than n*' implies $n-0.1$. Similarly '*about n*' would imply $n\pm0.1$. So when the condition for generating '*more than 1 mile*' is not met there seems to be no single quantifier which would cover the range between 1.1 and 2 miles. In our previous experiments (Hesse and Benz, submitted 2018) we found that the deviation of the expected value results from uncertainty assumptions which are a Weber fraction proportional to the appropriate level of granularity. So what the system actually needs to do when the conditions for generating '*more than 1 mile*' are not met is adjust the granularity level until it finds a modified numeral which best implies a value close to the true value.

Our previous experiments (Hesse and Benz, submitted 2018) also show that these rounding rules *only* apply in contexts where speakers have no strong experiential expectations (say, a prior with Laplace ignorance). The cover story in the evaluation study of this paper uses the city of New York. Speakers with some knowledge about New York's metro system might have prior beliefs about the density of subway stations in Manhattan, but in Brooklyn and Queens the density of stations is less uniform (it is high in some parts and low in others). So participants who do know New York would also know that density varies, and so they would find it hard to come up with a rounding tipping point which works for the entirety of New York. Participants who do not know New York may not have strong expectations about the density of subway stations, resulting in fuzzy beliefs and—like those knowledgeable about New York—no central tendency towards a common rounding tipping point.

5 Conclusion

This paper uses a dialogue system in order to evaluate an algorithm for adjusting the rounding granularity of comparatively modified numerals. The focus is on rounding numerals in a human-like manner, in such a way that it gives rise to the right scalar implicatures on the part of the user. We proposed a model which captures the complex interactions among numerical attributes in a real estate sales dialogue, and makes strategic use of the connotations of comparative quantifiers. In this model, modified numerals characterize the qualitative differences among alternatives (e.g., a subway station which is close to the real estate property and a bus stop which is not) and in what relation they stand to one another (e.g., a subway station is closer to the property than a bus stop).

In evaluating this model, we wondered whether we could deceive participants into perceiving a qualitative difference where there is none. We use deception as a test of the system's ability to use numerals strategically. The example we gave here is the distance to a subway station and we saw that participants felt that approximating this distance with "more than 1 mile" was more appropriate when its true distance was 1.2 miles than when it was 1.7 miles. We would expect that between 1.2 miles and 1.7 miles there comes a tipping point where participants switch from feeling deceived to feeling the approximation is a valid simplification. Future research should concentrate on locating tipping points such as this one as they signal a switch between cooperative and uncooperative behavior in the realtor and what factors influence interlocutors' sensitivity to these signals. Our next step will be to implement the vague modifier '*about*' in order to compare it to the comparative quantifiers. We would suspect that '*about n*' is appropriate whenever the true value falls within $n\pm0.1$, the rounding tipping point identified for '*more than*' and '*fewer than*.'

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (Grant Nr. 232196050). We would like to thank the reviewers and editors.

References

Anton Benz. 2015. Causal Bayesian Networks, Signalling Games and Implicature of More Than n. In *Bayesian Natural Language Semantics and Pragmatics*, Language, Cognition, and Mind, pages 25–42. Springer, Cham.

Stanislas Dehaene and Jacques Mehler. 1992. Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1):1–29.

Michael C. Frank and Noah Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084).

Bart Geurts and Rick Nouwen. 2007. 'At Least' et al.: The Semantics of Scalar Modifiers. *Language*, 83(3):533–559.

H. Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Christoph Hesse and Anton Benz. Submitted 2018. Scalar bounds and expected values of comparatively modified numerals. *Cognition*.

Anthony Jameson. 1983. Impression Monitoring in Evaluation-Oriented Dialog: The Role of the Listener's Assumed Expectations and Values in the Generation of Informative Statements. In *Proc. IJCAI-83*, pages 616–620.

Carel J. M. Jansen and M. M. W. Pollmann. 2001. On Round Numbers: Pragmatic Aspects of Numerical Expressions. *Journal of Quantitative Linguistics*, 8(3):187–201.

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C. Frank. 2016. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, 33.

Ciyang Qing, Noah Goodman, and Daniel Lassiter. 2016. A rational speech-act model of projective content. In *Proc. of the 38th annual meeting of the Cognitive Science Society*.

Jon Scott Stevens, Anton Benz, Sebastian Reuße, and Ralf Klabunde. 2016. Pragmatic question answering: A game-theoretic approach. *Data & Knowledge Engineering*, 106.

Jon Scott Stevens, Sebastian Reuße, Anton Benz, and Ralf Klabunde. 2015. A strategic reasoning model for generating alternative answers. In *Proc. of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on NLP*.

Henk Zeevat and Hans-Christian Schmitz. 2015. *Bayesian Natural Language Semantics and Pragmatics*. Springer.