Approximate Interpretations of Number Words: A Case for Strategic Communication

Manfred Krifka[1]
Humboldt Universität zu Berlin
Zentrum für Allgemeine Sprachwissenschaft (ZAS) Berlin

ABSTRACT. This paper gives an explanation of the well-known phenomenon that round numbers in measure terms (like *one hundred meters*) are interpreted in a more approximate way than non-round numbers (like *one hundred and three meters*). Several possible explanations are considered: First, a preference for short expressions and approximate interpretations; second, a conditional preference for short expressions under approximate interpretations; third, an explanation in terms of strategic communication that makes use of the fact that approximate interpretations, even if not favored initially, turn out to be more likely once the probability of the reported values are factored in. These explanations are shown to be flawed, in particular because the complexity of expressions does not always matter. The theory that is put forward makes use of scales that differ insofar as they are more or less fine-grained, and proposes a principle that a number expression is interpreted on the most coarse-grained scale that it occurs on. This principle can be motivated by strategic communication that factors in the overall likelihood of the message. The emerging theory is refined in various ways. In particular, it will be shown that complexity of expressions is important after all, but mainly on the evolutionary level, where it can be shown to lead to characteristic patterns of language change. The paper ends with the discussion of some surprising facts about the influence that the number system of a language has on which numbers are actually expressed in that language.

Round Numbers and Round Interpretations

In Switzerland, one can find street signs like the one near Zürich airport that tells the car driver that there is a stop sign 103 meters down the road. Visitors are struck by this and consider it typical for the land of bankers and watchmakers because it appears so ridiculously precise. But why? Why would a sign that says that there is a stop sign 100 meters down the road be considered unremarkable? Why is *100 meters* interpreted less precise than *103 meters*?

In the United States, one occasionally finds street signs that give distances in miles and in kilometers, in an effort to make the public familiar with the metric system. A sign of this type might read: *Eagle Pass. 7 miles. 11.265 kilometers*. Evidently, the educational effect of such signs is limited because they suggest that the metric system might be suitable for scientists, but not for ordinary folks. But why? Why is *7 miles* interpreted as less precise than *11.265 kilometers*?

The phenomenon illustrated by these examples is well known. In Krifka (2002) I called this the Round Numbers Round Interpretation (RNRI) principle:

---

[1] I had the opportunity to present various precursors of this paper at a number of conferences, including *Sinn & Bedeutung* 2001 in Amsterdam, the annual meeting of the *Deutsche Gesellschaft für Sprachwissenschaft* in Munich 2002, the Workshop on *Cognitive Foundations of Interpretation* at the Royal Academy of Science of the Netherlands (KNAW) in 2004, and the *West Coast Conference in Formal Linguistics* in Berkeley in 2007. A version of this paper was published at the proceedings of the KNAW conference, see Krifka (2007). I would like to express my thanks to numerous suggestions and critiques that helped me to develop the points presented here, in particular by David Beaver, Anton Benz, Reinhard Blutner, Peter Bosch, Regine Eckardt, Gerhard Jäger, Jason Mattausch, Robert van Rooy, Philippe Schlenker, Uli Sauerland, Torgrim Solstad, Theo Vennemann, Henk Zeevat and three anonymous reviewer.

(1) RNRI principle:
Round number words tend to have a round interpretations in measuring contexts.

The RNRI principle is far too specific to be an irreducible axiom of language use. How can we derive it from more general principles?

In Krifka (2002) I tried to show that this is possible within the framework of Bidirectional Optimality Theory. In the present article, I will point out various problems with this account and propose a more convincing explanation. But first I will turn to my previous theory.

A General Preference for Approximate Interpretations?

The explanation of the RNRI phenomenon in Krifka (2002) runs as follows:

First, there is a well-known pragmatic principle of economy that prefers simple expression over complex ones. This principle has been identified by numerous researchers, and is most prominently expressed in the Principle of Least Effort in Zipf (1929). In the Neo-Gricean theories of Horn (1984) and Levinson (2000), it has been captured by the R-Principle and I-Principle, respectively. These principles express that the speaker should say only as much as necessary in order to be understood, as the hearer will fill in information that is not expressed. In the Swiss street sign example, this will lead to a preference of the simple number term *one hundred* over *one hundred and three*, provided that *one hundred* can be interpreted in an approximate way.

Secondly, I assumed a principle that prefers approximate interpretations over precise ones. Something like this principle has been proposed before occasionally, and can be motivated in various ways. For example, Duhem (1904) speaks of a balance between precision and certainty; if one wants to increase the latter, one has to decrease the former, and so it might be prudent to be imprecise. Ochs Keenan (1976) has argued, quite similarly, that speakers (in her case, the population of rural Madagascar) might prefer vagueness over precision in order to save face in case what they said turns out to be not true. We also can argue that a more coarse-grained representation of information might be cognitively less costly than a more fine-grained one. It is easier to remember that the speed of light is 300,000 kilometers per second than to remember that it is 299,792,458 meters per second. Also, digital watches met with less than total success, partly because they present too much information, compared to analog ones. In general, I argued, we have the following pragmatic preferences, one for linguistic forms and one for their interpretation.

(2) SIMPEXP: simple expression > complex expression

(3) APPRINT: approximate interpretation > precise interpretation

These two pragmatic principles interact in the way proposed in Bidirectional Optimality Theory, cf. Blutner (2000) and Jäger (2002). That is, pairs of expressions and interpretations, or forms and meanings ⟨F, M⟩, are compared, and among various candidates the optimal pairs are selected according to the following rule:

(4) A form-meaning pair ⟨F, M⟩ is optimal iff
a. there is no optimal pair ⟨F′, M⟩ such that ⟨F′, M⟩ > ⟨F, M⟩
b. there is no optimal pair ⟨F, M′⟩ such that ⟨F, M′⟩ > ⟨F, M⟩

This type of interaction has been invoked to explain so-called M(arkedness)-implicatures (cf. Levinson 2000), according to which a marked expression receives a marked interpretation.

The RNRI phenomenon can be explained in the following way. Consider the following four form-meaning pairs as candidates to be evaluated by the constraints SIMPEXP and APPRINT:

(5) ⟨*one hundred*, precise⟩
⟨*one hundred,* approximate⟩
⟨*one hundred and three*, precise⟩
⟨*one hundred and three*, approximate⟩

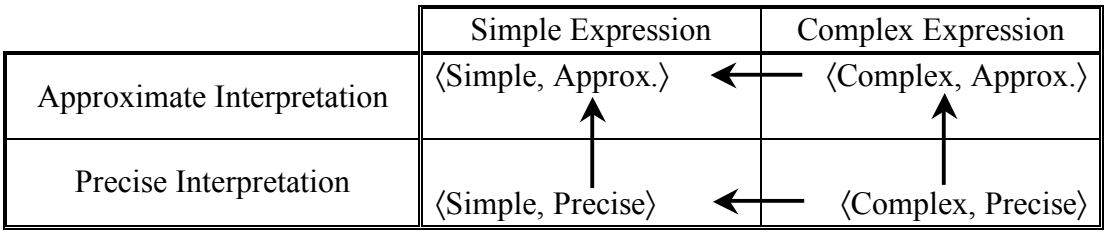
Clearly, ⟨*one hundred,* approximate⟩ is an optimal pair because there is no other pair that is better, hence there is no other **optimal** pair that is better:

(6) a. ⟨*one hundred*, approximate⟩ > ⟨*one hundred*, precise⟩, due to APPRINT
b. ⟨*one hundred*, approximate⟩ > ⟨*one hundred and three*, approximate⟩, due to SIMPEXP

From (6) it follows that ⟨*one hundred,* precise⟩ and ⟨*one hundred and three,* approximate⟩ are not optimal. But then ⟨*one hundred and three*, precise⟩ is an optimal pair, as it does not compete with any optimal pair: It does not compete with ⟨*one hundred,* approximate⟩, and the pairs it does compete with, ⟨*one hundred*, precise⟩ and ⟨*one hundred and three*, approximate⟩, are not optimal.

We can summarize the preference structure in the following diagram:

(7)



The pair ⟨Complex, Precise⟩ is an optimal pair because the two competing pairs ⟨Simple, Precise⟩ and ⟨Complex, Precise⟩ are preferred, but they are themselves not optimal.

A Conditional Preference for Simple Expressions?

In this section I will address two arguments against the theory developed in Krifka (2002), by showing that they can be countered by another theory that accounts for the RNRI phenomenon.

The first objection is that one of the four form/interpretation pairs in (5) should be out of consideration. There is no situation in which the pairs ⟨*one hundred*, precise⟩ and ⟨*one hundred and three,* precise⟩ can compete with each other from the perspective of the speaker, as the two pairs are not applicable in the same situation. If the actual distance is 103 meters, we would have to remove the pair ⟨*one hundred,* precise⟩ from the optimization process. The algorithm in (4) still would identify ⟨*one hundred,* approximate⟩ and ⟨*one hundred and three*, precise⟩ as the optimal pairings of forms and interpretations. But the argument points to a more general problem: We did not distinguish between the perspective of the speaker and the perspective of the hearer. The speaker might know that the two mentioned pairs do not compete with each other, but the addressee does not.

A more important objection is concerned with the preference for approximate interpretations. Why should there be such a general preference?. There are many situations in which the speaker wants to be interpreted in a precise way. For example, if someone offers to sell a car for *one thousand euros*, then he would not be satisfied if the buyer offers him less than that, with the excuse that approximate interpretations are preferred.

The basic idea for an improved explanation of the RNRI phenomenon is that the two constraints SIMPEXP und APPRINT in (2) and (3) are not independent of each other. The constraint that prefers simple expressions over complex ones, SIMPEXP, can be operative only if there is a choice between simpler and more complex expressions. Such a choice exists only under the approximate interpretation; under the precise interpretation, we are bound to one value only.

If precise and approximate interpretation are not ordered with respect to each other, then they cannot be used to evaluate candidates of forms and interpretations. Rather, precise interpretation

and approximate interpretation should be candidates themselves from which one or the other can be selected, according to pragmatic principles.

This can be made clear as follows. Let us assume a principle INRANGE, a consequence of the Gricean maxim of Quality, which says that assertions must be truthful:

(8) INRANGE:
The true value of a measure must be in the range of interpretation of the measure term.

Let us consider a simple example. Assume that an integer in the interval [1, 2, … 100] is to be reported as the result of a measurement. This can be done in a precise way, or in an approximate way, where the latter means that if the value i is reported, it may stand for the range [i-2…i+2]. For example, reporting the value by *forty* stands for the range [38…42] under the approximate interpretation, and for [40] under the precise interpretation. We now can construct tableaus like the following, in which pairs of form and interpretation and the actual value constitute the input:

(9)

	Form / Interpretation Pairs	Actual Value	INRANGE	SIMPEXP
☞	⟨*forty*, [38…42]⟩	39		
	⟨*forty*, [40]⟩	39	*	
	⟨*thirty-nine*, [37…41]⟩	39		*
☞	⟨*thirty-nine*, [39]⟩	39		
☞	⟨*forty*, [38…42]⟩	40		
☞	⟨*forty*, [40]⟩	40		
	⟨*thirty-nine*, [37…41]⟩	40		*
	⟨*thirty-nine*, [39]⟩	40	*	

If the actual value is 39, then two winners emerge: *forty* under an approximate interpretation, and *thirty-nine* under a precise interpretation. This is certainly a desired result, as it corresponds to the RNRI principle. In particular, it shows that the approximate interpretation of *thirty-nine* is ruled out, even if it would result in a true statement, because it violates SIMPEXP.

If the true value is 40, then again two winners emerge: *forty* under an approximate interpretation, and *forty* under a precise interpretation. This is not quite the desired result, as in this case the approximate interpretation of *forty* is preferred. It seems that we still need a general preference for approximate interpretations to model preference for *forty* here.

A more general problem of the tableau in (9) is that it assumes that the actual value of the reported measurement is known, as the candidates consist of an expression and an interpretation. But this is usually not the case for the addressee (except perhaps in answers to exam questions), and it is often not even the case for the speaker either, who might be uncertain about the precise actual value. I will show in the following section how these intrinsic problems can be solved within a framework of strategic communication.

Conditional Preferences in Strategic Communication

Let us assume a game-theoretic setting of strategic communication, as developed by Parikh (2001). This is not alien to the bidirectional approach to pragmatic tendencies of interpretation; in fact, Dekker & van Rooy (2000) have given a game-theoretic formulation in terms of Nash equilibria for cases like the one depicted in diagram (7). In this setting, the preference for approximate interpretations of simple measure terms can be derived under the assumption that addressees

hypothesize about the coding strategies of speakers, and speakers make use of this hypothesizing in their coding.

Parikh investigated the coding of information in a setting in which the probability or utility of a message is taken into consideration, an idea already put forward by Shannon (1948). For example, if an expression F is ambiguous between two meanings M, M′, where M is, in the given context, much more likely than M′, then a speaker can safely encode the meaning M by F. If the meaning M is less likely, then the speaker better refers to it by a more complex expression F* that denotes M but not M′. For example, *mother* is usually applied to biological mothers, but also to step mothers, foster mothers, etc. In many cases, there is no need for further specification, but if there is, expressions like *biological mother* can be used (cf. Horn 1993). We see a similar phenomenon at work when a round number, which is typically interpreted as vague

In the case at hand, the idea of economical encoding allows us to explain the RNRI phenomenon without a general bias towards approximate or precise interpretation. The only bias we have to assume is the uncontroversial one towards simple expressions.

Assume as before that measurements may be reported in precise or approximate ways, where reporting in an approximate way means that a reported value stands for range of possible values. For example, if *thirty-nine* is reported in an approximate way, it would optimally represent 39, less optimally 38 and 40, still less optimally 37 and 41, and so on. This could be captured by using normal distributions as measures of fit, but here I will just use intervals like [39 ±2], that is, [37…41] for the range of admissible interpretations of a number word like *thirty-nine*. I will talk of the precision level as a real number r that determines the level of precision at which a number word is interpreted. If the number word strictly denotes i, and the level of precision is r, then the number word under this level of precision denotes the interval [i ±ir]. For example, for the precision level r = 1/15 the numeral *thirty-nine* is interpreted as [39 ±2.6], that is, [36.4…41.6]. At a precision level of 0, numbers are interpreted in a precise way. Cf. Dehaene (1997) for further discussion of approximation levels and their relation to the Weber-Fechner law of discriminability of stimuli.

Two expressions that are interpreted in an approximate way may be indistinguishable with respect to each other for a given value. For example, the expressions *thirty-nine* and *forty* are indistinguishable for values like 38, 39, 40 and 41 if reported under a precision level of 1/15, as they would report the intervals [39 ±2.6] and [40 ±2.7], respectively, which overlap for these values. In everyday conversation, the information carried by *thirty-nine* and *forty* are equivalent if the given values are to be reported. This does not apply for reporting measurement values with a margin of error in physics, where 39 ±2.6 and 40 ±2.7 may mean something different. But then reporting values with margin of errors is not an indication of a coarse-grained representation, but rather a hallmark of high precision.

Consider now the same task as before, in the following setting. A result of a measuring has to be reported that is an integer in the interval [1 … 100]. The addressee has no initial hypothesis about the value of the measurement, so he assumes an a-priori likelihood of p = 0.01 for each of the integers. Let us assume two possible interpretations, an approximate one with level of approximation of 1/15, and a precise one with level of approximation 0. Let us furthermore assume that both precision levels are equally likely.

Let us first consider the case the speaker reports *thirty-nine.* There are two possible interpretations. We first consider the approximate interpretation, under which the expression reports the interval [39 ±2.6]. This interpretation is indistinguishable from a number of alternative utterances, in particular *forty.* The speaker could have uttered any of these alternative utterances to convey the message. But if he had this choice, he would have uttered *forty*, which is the preferred utterance among the alternatives, as it is the shortest. The speaker has not done so; he uttered *thirty-nine*. Consequently, the premise that the speaker intended the approximate interpretation must be false. The speaker must have intended the precise interpretation.

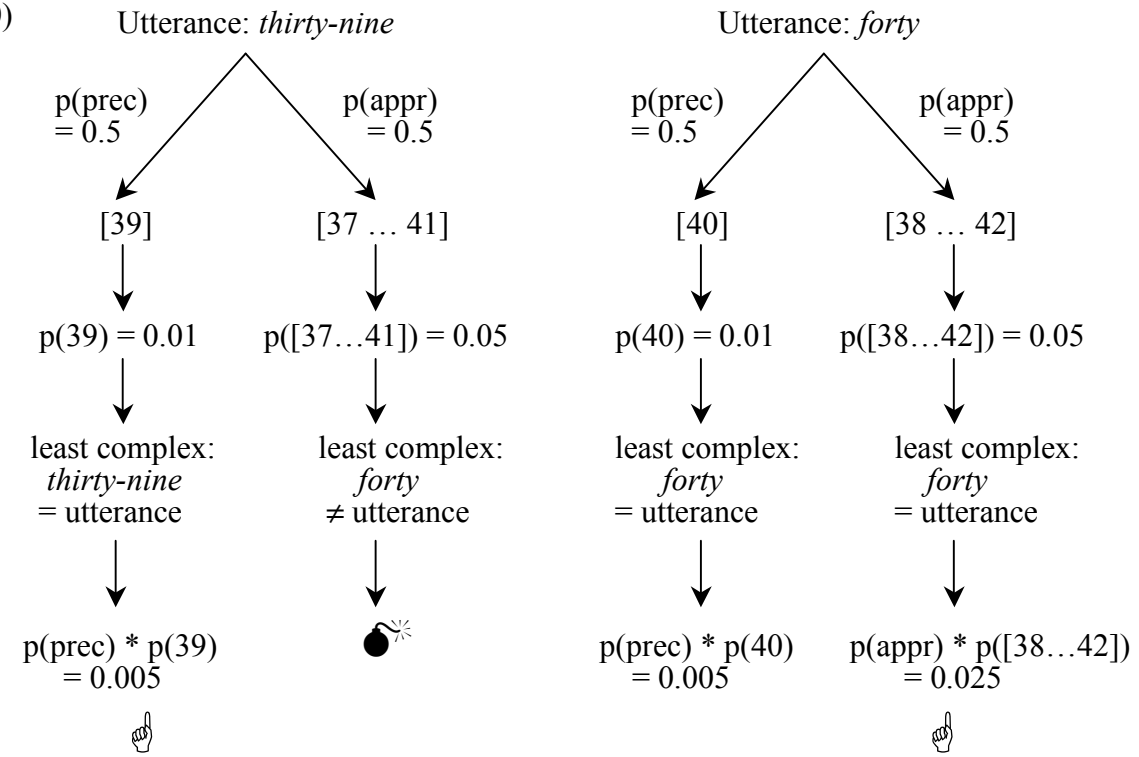
Under the precise interpretation, no problem arises. *Thirty-nine* is interpreted as 39, there are no indistinguishable alternative utterances. The utterance is consistent with the assumption that the speaker intended a precise interpretation.

Consider now the case that the speaker utters *forty*. Again, there are two possible interpretations. Under the approximate interpretation, the utterance reports the interval [40 ±2.7]. This is undistinguishable from the interpretations of the alternative utterances like *thirty-nine*. But among the alternative utterances, *forty* is the shortest and would have been chosen. Hence the utterance is consistent with the assumption of the approximate interpretation.

Of course, the utterance *forty* is also consistent with the precise interpretation, as before. But now we can make an argument that overall the approximate interpretation is preferred, following Parikh's reasoning about the a priori likelihood of the message. Under the precise interpretation, *forty* would report the value 40, which has an a-priori likelihood of 0.01. Under the approximate interpretation, *forty* would report the integer values in [40 ±2.7], that is, [38, 39, 40, 41, 42], which together have a likelihood of 0.05. Hence the more conservative assumption is that the speaker had in mind the approximate interpretation, as it would report a value of a greater a-priori likelihood. If the speaker wants to block this inference, some indication like the adverb *exactly* would have to be applied.

We can summarize the computation of overall probabilities in these two cases in the following diagrams, where we assume that the a-priori probability of each value is 0.01, and that the a-priori probability of precision level 0 and precision level 0.15 is 0.5 in either case. We see that for *thirty-nine*, only a precise interpretation is possible, whereas for *forty*, the approximate interpretation is selected because the reported values have a greater a-priori likelihood.

(10)



I have stressed that speaker and addressee know about each other's knowledge concerning the communicative situation. This means that by using a round number word, like *forty*, a speaker can signal a round interpretation, provided that the context is not skewed towards a precise one (as e.g. in arithmetic class, when the sum of 13 + 26 is requested).

In the *gedankenexperiment* above, we have assumed that all measurement values have the same a-priori likelihood. This is a simplifying assumption which turns out to be unnecessary. Even if e.g.

40 is more likely to be reported than 38, 39, 41 and 42, the cumulative likelihood of [38..42] is still greater than the likelihood of 40, and this is all that is necessary for the argument to get through. Furthermore, if the context is such that a precise interpretation is a-priori more likely (as e.g. when in arithmetic class the sum of 13 + 26 is requested), then this can override any other interpretation tendency when *forty* is uttered.

Simplicity of Expressions vs. Simplicity of Representations

In Krifka (2002) I pointed out that simplicity of expression is not always the decisive factor for an approximate interpretation. A bias for simple expressions cannot explain all interpretation preferences, as in the following examples:

(11) a. I wrote this article in twenty-four hours. (approximate)
b. I wrote this article in twenty-three hours. (precise)

(12) a. The house was built in twelve months. (approximate)
b. The house was built in eleven months. (precise)

Even in our original example concerning Swiss street signs, it might be difficult to argue that complexity of expression is the reason if we consider the fact that the distance is given in Arabic numbers; after all, *100* and *103* consist of the same number of digits.

Worse yet, the idea that simple expressions lead to approximate interpretations sometimes is plain wrong. Consider the following minimal pairs:

(13) a. Mary waited for forty-five minutes. (approximate)
b. Mary waited for forty minutes. (precise)

(14) a. The wheel turned one hundred and eighty degrees. (approximate)
b. The wheel turned two hundred degrees. (precise)

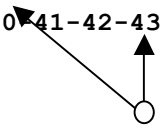
(15) a. Her child is eighteen months. (approximate)
b. Her child is twenty months. (precise)

(16) a. John owns one hundred sheep. (approximate)
b. John owns ninety sheep. (precise)

I would like to argue that such examples show that speakers do not (just) prefer simple expressions, but rather simple representations. When speaking of time intervals, 24 hours is simpler than 23 hours because it denotes the length of a day, a prominent conceptual unit. The same holds for the other examples mentioned above: 12 months correspond to a year, 45 minutes to three quarters of an hour, 180 degrees to half of a complete turn, 18 months to one and a half years, and 100 is not just a multiple of ten, but a power of 10. The expressions that have a more approximate interpretation all refer to more salient units on their scales.

It is easy to replace the argumentation of the previous sections by postulating, instead of a bias for short expressions, a bias for simple representations. This bias for simple representations is, more specifically, one for coarse-grained representations, in the sense of Curtin (1995). The basic idea is this: The results of measuring can be reported with respect to various levels of granularity, which differ from each other in the density of representation points. For example, a distance can be specified on scales listing hundreds of kilometers, tens of kilometers, kilometers, etc. In a rather transparent way, scales are optimal if the points are distributed in an equidistant fashion (or sometimes according in other regular ways, e.g. logarithmically, cf. Hobbs 2000). Furthermore, scales of different granularity levels should align, which simplifies conversion from one granularity level to the other. The most frequent type in our culture is the one based on the powers of ten, as illustrated in (17):

(17) a. …-30----------40----------50-…
b. …-30-31-32-33-34-35-36-37-38-39-40-41-42-43-44-45-46-47-48-49-50-…



The more coarse-grained scale (17.a) has fewer values for representing measurements than the more fine-grained scale (b). For example, any measurement between 35 and 45 is represented by a single value on the scale, 40. Results of counting or measuring have to be reported using the value that is closest, at the chosen granularity level. In the indicated example, the small circle in (17) will be represented by 43 on the fine-grained level (b), and by 40 on the coarse-grained level (a).

The scale hierarchy in (17) can be refined by introducing an intermediate scale. The optimal choice appears to be the one that adds the numbers 25, 35, 45 etc., thus creating scale points with a distance of five. This refinement is optimal insofar as it allows for the best overall representation of random measure values. On this scale, the circle in (17) would be represented by the number 45, as this is closer than 40.

In examples (11) to (16), the scales of different granularity are not based on the powers of ten, but on some other principle that is merely translated into the decimal system. As an example, take the minute scale (13). The relevant scales are the scale that counts the hours; then the scale that counts half-hours; then the scale that counts quarter hours; then the scale that counts ten minutes, then the scale that counts in 5 minute intervals. I will not represent even more fine-grained scales here, like the scale that counts single minutes, half minutes, quarter minutes, etc.

(18) a. 0----------60----------…--120-…
b. 0----------30----------60----------90----…--120-…
c. 0------15-------30-------45-------60-------75-------90----…--120-…
d. 0---10----20----30----40----50----60----70----80----90----…--120-…
e. 0-5-10-15-20-25-30-35-40-45-50-55-60-65-70-75-80-85-90-95-…--120-…

We can now explain why (13.a), *forty-five minutes*, is interpreted in a less precise way than (13.b), *forty minutes*. It follows from the following principle:

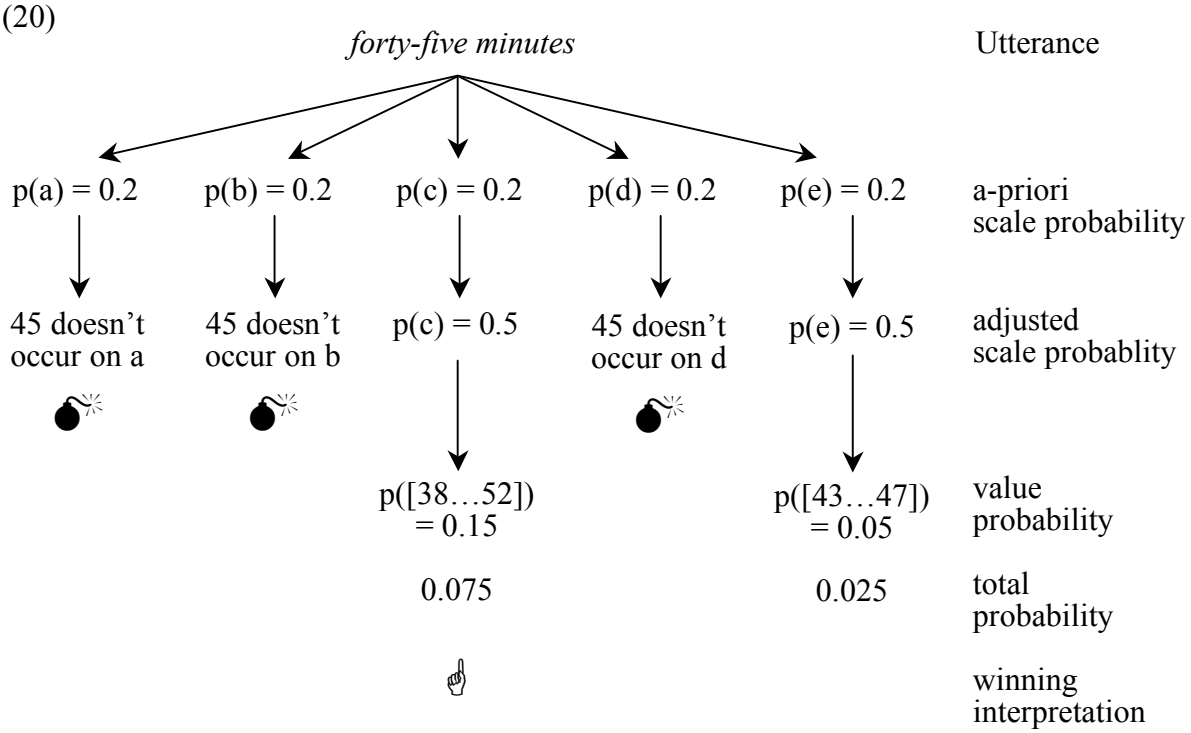
(19) The Coarsest Scale Principle:
If a measure expression α occurs on scales that differ in granularity,
then uttering α implicates that the most coarse-grained scale on which α occurs is used.

For example, the most coarse-grained scale of *forty-five minutes* is (18.c), and for *forty minutes*, this is scale (18.d). As scale (c) is coarser than scale (e), *forty-five minutes* is interpreted in a more approximate way than *forty minutes*.

Why are measure expressions interpreted at the coarsest scale? Consider *forty-five minutes*. This term is represented only on scales (c) and (e), hence we can disregard the other scales. With respect to scale (c), the scale point 45 represents the times in [38…52]; with respect to scale (e), the scale point 45 represents the times in [43…47]. Let the a-priori probability that the measured time is any particular minute within the range to be considered be r. Then the probability that the measured time is in [43…47] is 5r, and the probability that the measured time is in [38…42] is 10r. Let the a-priori probability on hearing *forty-five minutes* that one of the scales (c) or (e) be used be the same, say s. Then on hearing *forty-five minutes* the probability that the more fine-grained scale (e) is used is 5rs, and the probability that the more coarse-grained scale (c) is used is twice that value, 10rs. Hence the hearer will assume the more coarse-grained scale.

Let me illustrate this with the help of a diagram. Assume that for each scale the a-priori probability for using the scale is 0.2. Assume also that the a-priori likelihood that the measured event is n minutes to be 0.01, for each n in the range under consideration. We then have the following situation:

(20)



We see that the most coarse-grained scale on which the measure expression occurs is the winner. This is the case if the a-priori likelihood of each scale is the same. As before, we can assume that the context changes this probability, e.g. by increasing the a-priori likelihood of more fine-grained scales, which might tip the balance towards the other solution.

We might also factor in a context-dependent utility function that penalizes coarse-grained representations to greater or lesser degree – see Jäger (2007) for a suggestion that builds on the proposal discussed here. One can integrate this by assuming a function on the total probability in the last line of (20) that multiplies this value with a factor 1 if the real value is represented by the precise interpretation of the number word, and by factors that decrease to 0 the farther removed the real value is from the precise interpretation. The best representation, then, is one that maximizes the resulting value. To illustrate, consider the utterance of *thirty minutes* with respect to the scales in (18). We get the following

(21) Utterance *thirty minutes*

Scale (a): excluded
Scale (b): Scale probability 0.25, value probability p([15…45]) = 0.30, total: 0.075
Scale (c): Scale probability 0.25, value probability p([23…37]) = 0.14, total: 0.035
Scale (d): Scale probability 0.25, value probability p([35…45]) = 0.10, total: 0.025
Scale (e): Scale probability 0.25, value probability p([37…42]) = 0.05, total: 0.0125

As before, the most coarse-grain scale (b) is selected. Now assume a factor that gives a penalty in case the actual number differs from the precise interpretation of the measure expressions. For concreteness, we can assume the following function:

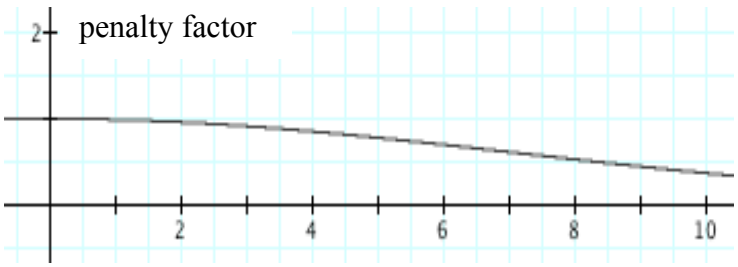
(22)

penalty for misrepresentation: $\pi_r(v, u) = (1+r)^{-(v-u)^2}$

where v: actual value,
u: precise value of utterance,
r: value ≥1 specifying severeness of penalty

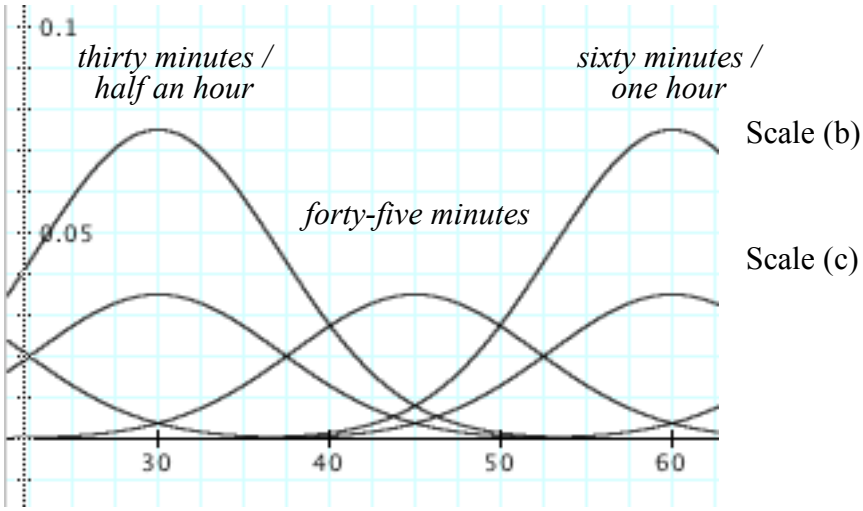
In case r = 0 we have a constant penalty factor of 1, that is, no penalty ensues. In case r = 0.1 we have the following penalty factor, where the x axis indicates the difference between v and u.

(23)



For example, if the utterance value differs from the actual value by 8, the probability that the scale is used decreases by 0.5. When we apply the penalty factor $\pi_{0.1}$ to the probabilities of the scales (18.b) and (c), we get the following picture:

(24)



We see that for most of the values in the given interval (between 20 and 60), scale (b) is preferred over scale (c), in line with what we have argued so far. However, for reporting values between 40 and 50, the more fine-grained scale (c) is better, and we expect *forty-five minutes* for reported values within this range. The net effect of this competition between scales is that for values in the middle between the scale points of coarse-grained scale that otherwise is preferred, a more fine-grained scale will be used.

This example also shows an important principle that governs scale refinement. If there is a scale of equidistant points [n, 2n, 3n, 4n, …], then the optimal refinement of this scale will be one with equidistant points that contains intermediary points, [0.5n, n, 1.5n, 2n, 2.5 n, …]. The reason is that at exactly these intermediary points, the penalty for expressing values by coarse-grained scales is highest. We will return to this principle of scale refinement below.

The penalty function in (23) is symmetric, that is, values that are higher or lower than the expressed value are penalized in the same way. Penalty functions don't have to be that way. If I tell someone that the train leaves in half an hour (or 30 minutes), then this is a bad choice of wording if the train actually leaves in 28 minutes, because the addressee might blame me if he missed the train. It would be more appropriate here to say that it leaves in 25 minutes, or even in 20 minutes. This can be modeled with a penalty function that is asymmetric, penalizing higher values more than lower values.

From Simplicity of Representation to Simplicity of Expression

Optimal scale selection, as discussed in the last section, can explain everything that we have tried to explain by selection of shortest expressions, and more. We can apply exactly the same reasoning for the preference of *forty* over *thirty-nine* by assuming that the coarse-grained scale […, 30, 40, 50,…]

is preferred over the fine-grained scale […, 38, 39, 40, 41, …]. An additional advantage of optimal scale selection seems to be that we do not have to apply the concept of indistinguishable reported values (cf. (12)). The attentive reader might have questioned this concept – at least, an anonymous referee did. For example, the relation of being indistinguishable is not transitive: if expressions α and β are indistinguishable, and expressions β and γ, then α and β need not be indistinguishable).

Yet the explanation of the RNRI phenomenon by a preference for simple expressions works in the majority of cases as well, and this hardly can be an accident. For the most part, the two explanations yield the same result. This is so because the number words of more coarse-grained scales are, in general, shorter than the number words of less coarse-grained scales. For example, the following scale hierarchy, while satisfying the requirement of equidistance of scale points, would be decidedly odd:

(25) a. …**-30-31-32-33-34-35-36-37-38-39-40-41-42-43-44-45-46-47-48-49-50-**…
b. …**-30-------33-------36-------39-------42-------45-------48-------**…

The scale (25.b) is odd because the expressions that denote the scale points do not make use of the scale points for which the decimal system offers simple expressions. It neither is motivated as a translation of another system into the decimal one.

We can measure simplicity of expressions on a specific scale in various ways. One is by counting syllables. I have computed the following average numbers of syllables for the number words at the three following scales, from one to one hundred:

(26) a. *one, two, three, four, ... one hundred:* 273/100 = **2.73** syllables per word
b. *one, five, ten, fifteen, ... one hundred:* 46/20 = **2.3** syllables per word
c. *one, ten, twenty, thirty, ... one hundred*: 21/10 = **2.1** syllables per word

In contrast, the scale suggested in (25.b) does not show any such decrease of expression complexity when compared to the basic scale (26.b) – quite to the contrary:

(27) *three, six, nine, twelve, … ninety-nine*: 92/33 = **2.79** syllables per word

Hence we can assume the following general principle which relates Simplicity of Expressions and Simplicity of Representations.

(28) SER:
If S_1 is a more fine-grained scale than S_2, then
the average complexity of expressions of the values of S_1
tends to be greater than
the average complexity of expressions of the values of S_2
(measured over a reasonably large interval).

SER is not without exceptions. For example, consider the following scales of reporting the ages of young kids in months:

(29) a. **1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20-21-22-23-24**
b. **1---3-----6-----9-------12-------15-------18-------21-------24**

The average complexity of English number words at the scale (29.a) is 44/24 = 1.83, the average complexity of the more coarse-grained scale is 25/9 = 2.78, a clear violation of SER. However, granularity hierarchies like this one are quite rare. Furthermore, the age of children is more often given by expressions like *one year*, *one and half years*, *two years*, etc.

One indication that SER is an important principle of language is that in case simplicity of expressions and simplicity of representations diverge, there is evolutionary pressure to realign them. One example of this is the expression of amounts of money, e. g. in the US: *quarter* instead of *twenty-five cents.* For similar reasons, Dutch had expressions like *kvartje* for 25 cents and *rijksdaler* for 2,50 guilders in pre-Euro times. A particular curious case of translation from one system to another happened in German when the duodecimal system of *12 Pfennig = 1 Groschen* got replaced

by a decimal system; now *Groschen* was used for 10 Pfennig, and *Sechser* (literally, 'sixer') for half that value, 5 Pfennig.

A more subtle influence of this evolutionary force can be seen in the way how the number 5 is expressed in certain combinations. The number 5 is special in the decimal system because it allows for an optimal refinement of granularity by the factor ½, as illustrated:

(30) a. `10----20----30----40----50----60----70----80----90----100`
b. `10-15-20-25-30-35-40-45-50-55-60-65-70-75-80-85-90-95-100`

This is the reason why the expression of this number is sometimes special. In English and in colloquial German, we have phonological irregularities for the number words of 15 and 50. We also find a special, shortened form for 1½ in German.

(31) a. *fifteen* [fɪftiːn], instead of regular **fiveteen* [faɪftiːn]
b. *fifty* [fɪfti], instead of regular **fivety* [faɪfti]

(32) a. colloquial form *fuffzehn* [fʊftˢeːn], standard form *fünfzehn* [fʏnftˢeːn]
b. colloquial form *fuffzig* [fʊftˢɪg], standard form [fʏnftˢɪg]

(33) *eineinhalb* (long form), *anderthalb* (short form) for 1½.

Notice that the irregular forms are phonologically simpler: The stem *fif-* has a monophthong in place of a diphthong. The stem *fuff-* exhibits a loss of rounding, and a loss of the nasal. The form *eineinhalb* has three heavy syllables, resulting in an unusual molossus foot; the form *anderthalb* has a light second syllable, resulting in a more usual cretic foot (heavy-light-heavy).

We can see this as evidence for the simplification of expressions that invite an approximate interpretation for conceptual reasons. In the case of English, the Old English stem *fīf* [fiːf] was reduced to *fif* [fɪf] in these environments, while undergoing regular development to *five* [faɪf] during the Great Vowel Shift in others. Notice that *five / fifteen / fifty* developed differently from the phonologically similar *nine / nineteen / ninety*. The most likely reason for the distinct development of *fifteen* and *fifty* is that these numbers occurred more frequently due to the fact that they occur on coarse-grained scales, and that increased use resulted in phonological simplification.

In this connection, it is also interesting to note the special encoding of the numbers 5, 50 and 500 in Roman numerals, which always are shorter than their immediate neighbors on the same granularity scale. This is, of course, not a result of evolution but of design, also motivated by the iconic representation of the hand. Nevertheless, it gives evidence of the special function of 5.

(34) a. 4–5–6: IV–V–VI
b. 40–50–60: XL–L–LX
c. 400-500-600: CD–D–DC

Certain irregularities of numbers bear witness of the shaping effect of scales that are not operative anymore. One case in point is English *twelve* (Germanic **twa-libi* 'two+remnant'), which translates the basic unit of an older duodecimal system based on the number 12. Another is the suppletive form of the number word for 40 and 90 in Russian and other East-Slavonic languages:

(35)

'10'	*desjat'*	'60'	*šestdesjat'*
'20'	*dvadcat'*	'70'	*sem'desjat'*
'30'	*tridtcat'*	'80'	*vosem'desjat'*
'40'	*sorok*	'90'	*devyanosto*
'50'	*pjatdesjat'*	'100'	*sto*

According to Comrie (1992), the form *sorok* has replaced a regular Slavonic form; it is either a Greek loan or a classifier. The form *devyanosto* has been analyzed as a subtractive numeral or as evidence of an older nonal system. There is a way to connect these two oddballs by assuming that they represent remnants of an ancient octal subsystem, where 90 is the first multiple of 10 beyond 80, and 40 specifies ½ times 80. There are various other indications that have been adduced to argue for an octal, or quaternary system (cf. Winter 1992).

Another irregularity of Russian numerals can be found in their pronunciation. While forms up to '60' are pronounced shorter (e.g. *šestdesjat'* [šys'at']), higher forms are pronounced longer (e.g. *sem'desjat'* [s'emis'ət']). This is reminiscent of the different complexity of multiples of 10 up to 60 and beyond 60 in older Germanic languages, e.g. Anglo-Saxon *sixtig* '60' vs. *hund-seofontig* '70', which again might point to the role that 60 played since Babylonian times (cf. Menninger 1962).

Returning from such speculative excursions, it should be mentioned that there is evidence form linguistic corpora that the number words for 12 and 15 are indeed more frequent than other number words denoting 11 - 19 (cf. Dehaene & Mehler 1992 for English, French, Japanese, Kannada, Dutch, Catalan and Spanish). Also, there is evidence that between 10 and 100, the number words denoting the powers of ten are far more frequent than other numbers (cf. Sigurd 1988). Jansen & Pollmann (2001) define a notion of roundness in which multiples of 10, of 2 and of 5 play a special role. The special role of 10 is, of course, due to the accidental fact that humans have ten fingers, thus providing the basis for a popular type of number system. The special role of 2 is motivated by the prominent operation of doubling or a quantity, or dividing it in half.

Avoidance of Complexity?

Not all number systems are based on ten, and we might ask what will happen in systems with different base. In particular, vigesimal systems are of interest here, that is, systems based on twenty, as there are a number of European languages that have vestiges of it, even if sometimes only in part, as in Standard French. European vigesimal systems are mixed, at least in the sense that 100 is a simple number word (this does not hold for all vigesimal systems, e.g. for Nahuatl). This creates a tension between optimal scale geometry and complexity of expressions: If we want to refine the scale based on multiples of 100 minimally (cf. (36.a), we arrive at one that also adds the scale item 50 (cf. b). But in vigesimal systems, this is denoted by a relatively complex number word ("two scores and ten"), in any case more complex than the number words for 40 ("two scores") and 60 ("three scores"). Also, the number words for 30, 70 and 90 are more complex than their decimal counterparts.

(36) a. `0----------100`
b. `0----------50----------100`
c. `0------20------40------60------80------100` vigesimal
d. `0--10--20--30--40--50--60--70--80--90--100` decimal

I have looked into the frequency of number words in three languages: First, Norwegian, a decimal system, and Danish, a partially vigesimal system (partial insofar as a number like 51 will be represented as *en og halvtreds* 'one and fifty', not as "eleven and fifty"). I selected Norwegian and Danish as a pair of languages that are spoken by culturally similar communities, in an attempt to exclude cultural biases. Second, I also looked at Basque, which has a full vigesimal system (e.g., the number 51 is *berrogei ta hamaika* 'forty (= twice twenty) and eleven'). For Norwegian and Danish the Google web sites restricted to those languages could be used; the search was done on March 4, 2005. For Basque, restricted searches like that are not possible, all that one can do is to restrict searches to Spanish web sites. But sample inspections revealed that the words nearly exclusively occurred on Basque web sites. This search was done on November 20, 2007. I made sure that the number words did not have homonyms (for this reason, Basque *hogei* '20' had to be excluded, as this sequence occurs in other languages as well). One should be aware of the fact that the number words also occur as part of complex numbers, as e.g. the number 41 in the Norwegian spelling *førti en*, in addition to *førtien*, or in Danish *en og fyrre*, in addition to *enogfyrre*. For Basque, I made sure that the occurrences of, e.g., *berrogei* excluded complex numbers like *berrogei ta bat* '41' by subtracting the number of occurrences of the string *berrogei ta*. It should be stressed that these are still preliminary results that need to be done more carefully by language experts; the results could be influenced, e.g., by a big Danish company with many web sites that has the word *tres* in its title. Also, the reader should be warned that Hammarström (2004), who looked at (presumably much smaller) Danish and Welsh corpora did not find a difference between languages with decimal number system and languages with vigesimal number system.

(37)

Num-ber	**Norwegian**		**Danish**		**Basque**	
	Numeral	#	Numeral	#	Numeral	#
20	*tjue*	61300	*tyve*	121000		
30	*tretti*	43700	*tredive*	25400	*hogei ta hamar*	892
40	*førti*	**39200**	*fyrre*	**26800**	*berrogei*	**85000**
50	*femti*	**81200**	*halvtreds*	**15500**	*berrogei ta hamar*	**213**
60	*seksti*	**19400**	*tres*	**36400**	*hirurogei*	**34000**
70	*sytti*	10200	*halvfjerds*	581	*hirurogei ta hamar*	69
80	*åtti*	13100	*firs*	3740	*larogei*	9000
90	*nitti*	13500	*halvfems*	540	*larogei ta hamar*	7

Let us first consider Norwegian and Danish. We find for Norwegian the predicted relative (and even absolute) maximum for *femti* ‘50’; for Danish, we find the opposite, a relative minimum for *halvtreds* in comparison to *fyrre* and *tres*. This picture does not change when we consider the complete form *halvtredsind-s-tyve* ‘50’, literally ‘half-third-times-of-twenty’, which has only 1180 occurrences. There also exists a form *femti* ‘50’, mostly for monetary purposes, which occurred only 988 times on Danish web pages. The table also shows sharp local minima in Danish for the complex forms *halvfjerds* ‘70’ and *halvfems* ‘90’, which are absent in Norwegian. Interestingly, *tres* ‘60’ even occurs considerably more often than the longer number words *tredive* ‘30’ and *fyrre* ‘40’, which might be seen as further evidence of the influence of simplicity on use.

In the case of Basque, the numbers are even more extreme. In particular, it is stunning how rare the number word *berrogei ta hamar* ‘50’ occurs, compared to *berrogei* ‘40’ and *hirurogei* ‘60’. Yet within its immediate neighbors, it forms a local maximum, with 213 occurrences, as shown in (38). It is unclear whether this reflects the cognitive status of 50 or is a result of translations from a language with a decimal number system like Spanish. There are also local maxima at ‘45’ and ‘55’. The very low number of occurrences above ‘50’ is remarkable; in fact, most of these occurrences just stem from web sites that list all Basque number words.

(38)

Numeral	‘41’	‘42’	‘43’	‘44’	‘45’	‘46’	‘47’	‘48’	‘49’
Occurr.	37	151	40	118	147	72	70	109	54
‘50’	‘51’	‘52’	‘53’	‘54’	‘55’	‘56’	‘57’	‘58’	‘59’
213	29	10	7	6	50	5	3	3	7

There are, I think, two possible explanations for these differences between languages with decimal and with vigesimal number systems. One is that the scale hierarchy of vigesimal languages is the same as the scale hierarchy with decimal languages, but that number words for numbers meaning ‘50’, ‘70’ and ‘90’ are underused due to their complexity, as complex expressions are avoided. In written language, writers may resort to Arabic numbers, something that could be checked with more careful corpus research. The second explanation is that the scale hierarchy of vigesimal languages itself is different, insofar as these languages also have a prominent scale based on multiples of 20, which – in the case of Basque – appears to be more prominent than the scale based on multiples of 10. Of course, these two explanations do not exclude each other. In particular, the complexity of expressions can obviously be an obstacle against switching from one scale system to another, in the case at hand, from switching from a state with a prominent scale based on multiples of 20 to a state with a prominent scale based on multiples of 10.

Conclusion

In this paper I have argued that the fact that round numbers are interpreted in an approximate way can be explained by general pragmatic principles, even in the absence of an a-priori preference for approximate interpretations. The crucial argument was that the approximate interpretation is more likely because it is compatible with a wider range of possible meanings. This reasoning is blocked with non-round numbers as under the approximate interpretations, round numbers would be selected, as short expressions are preferred. We have discussed limitations of this motivation, and have argued that it should be generalized to a principle that predicts that the most coarse-grained scale compatible with the selected expression should be chosen. Simplicity of expressions then appears as a secondary phenomenon, as coarse-grained scales tend to have simpler expressions to denote their scale points. We have finally seen that there are interesting differences between languages based on different number systems. In particular, languages with an vigesimal number systems seem to make use of a scale based on multiples of 20, which is reflected in the frequency of use of number words.

This point is perhaps the most interesting one: According to a well-known phrase by DuBois (1987) that has been used to motivate frequency-based explanations in linguistics, grammars do best what speakers do most. It might very well be the case that sometimes the reverse is true: Speakers do most what grammars do best. A grammar that offers simple coding of multiples of 20 will motivate speakers to use multiples of 20 more often than other numbers. So far, pragmatics is mostly concerned with the role of simplicity or complexity of coding a particular message; we should perhaps also pay attention to the role of simplicity and complexity of coding in selecting what is said in the first place.

References

Blutner, Reinhard (2000), "Some aspects of optimality in natural language interpretation", *Journal of Semantics* 17, 189-216.

Comrie, Bernard (1992), "Balto-Slavonic", in Jadranka Gvozdanovic, *Indo-European numerals*, Berlin, Mouton de Gruyter, 717-834.

Curtin, Paul (1995). *Prolegomena to a theory of granularity*. MA Thesis, University of Texas at Austin.

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

Dehaene, Stanislas (1997), *The Number Sense: How the Mind Creates Mathematics*. Oxford: Oxford University Press.

Dekker, Paul & Robert Van Rooy (2000), "Bi-directional Optimality Theory: An application of game theory", *Journal of Semantics* 17, 217-242.

DuBois, John (1987), "The discourse basis of ergativity". *Language* 63, 805-855,

Duhem, Paul (1904), *La théorie physique, son objét et sa structure.* Paris.

Hammarström, Harald (2004), "Number bases, frequencies and lengths cross-linguistically." Abstract accepted at the conference *Linguistic perspectives on numerical expressions*, 2004, Utrecht, Netherlands. Article (draft) see http://www.cs.chalmers.se/~harald2/

Hobbs, Jerry R. (2000), "Half orders of magnitude.", in L. Obrst & I. Mani, *Workshop on semantic approximation, granularity, and vagueness,* 28-38.

Horn, Laurence R (1984), "Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature." In D. Schiffrin (ed.), *Meaning, form, and use in context: Linguistic applications.* Washington D.C.: Georgetown University Press, 11-89.

Horn, Laurence R. (1993), "Economy and redundancy in a dualistic model of natural language." In S. Shore & M. Vilkuna, eds., *SKY 1993: 1993 Yearbook of the Linguistic Association of Finland*, 33-72.

Jäger, Gerhard (2002), "Some notes on the formal properties of bidirectional Optimality Theory", *Journal of Logic, Language and Information* 11, 427-451.

Jäger, Gerhard (2007), “Communication about similarity spaces”. Talk given in Amsterdam, KNAW Academy colloquium, *New perspectives on games and interaction*, see http://wwwhomes.uni-bielefeld.de/gjaeger/talks/slidesKnawGames.pdf
Jansen, C. J. M. and M. M. W. Pollmann (2001), “On round numbers: Pragmatic aspects of numerical expressions.” *Journal of Quantitative Linguistics* 8:187–201.
Krifka, Manfred (2007), “Approximate interpretations of number words: A case for strategic communication.”, in Gerlof Bouma, Irene Krämer & Joost Zwarts, *Cognitive foundations of interpretation*, Amsterdam, Royal Netherlands Academy of Arts and Sciences, 111-126.
Krifka, Manfred (2002), “Be brief and vague! And how bidirectional optimality theory allows for Verbosity and Precision”. In David Restle & Dietmar Zaefferer (eds), *Sounds and Systems. Studies in Structure and Change. A Festschrift for Theo Vennemann*. Mouton de Gruyter, Berlin, 439-458.
Lemer, Cathy e.a. (2003), “Approximate quantities and exact number words: dissociable systems.” *Neuropsychologica* 41, 1942-1958.
Levinson, Stephen. 2000. *Presumptive Meanings.* Cambridge, Mass.: MIT Press.
Menninger, Karl (1962), *Number words and number symbols***,** Dover,
Ochs Keenan, Elinor (1976), “The universality of conversational implicature”, *Language in Society* 5, 67-80.
Parikh, Prashant (2001), “Communication, meaning and interpretation.” *Linguistics and Philosophy* 23, 141-183.
Shannon, Claude (1948), “A mathematical theory of communication.” *Bell Systems Technical Journal* 27, 379-432, 623-656.
Sigurd, Bengt (1988), “Round numbers.” *Language in Society* 17, 243–252.
Winter, Werner (1992), “Some thoughts about Indo-European numerals”, in Jadranka Gvozdanovic, *Indo-European numerals*, Berlin, Mouton de Gruyter, 11-28.
Zipf, George K. (1929), “Relative frequency as a determinant of phonetic change”, *Harvard Studies in Classical Philology* 40,