Manfred Krifka*

Approximate interpretations of number words: A case for strategic communication

Abstract

A pragmatic theory of the approximate vs. precise interpretation of round number words such as *one hundred* vs. *ninety-seven* is developed that improves on the previous account of Krifka (2002) insofar as it does not postulate a general bias for approximate interpretations. It is shown that under a speaker preference for number expressions that are short and/or refer to cognitively salient values on scales, approximate interpretations are favored for these expressions even if there is no general bias for them. Also, it is shown that there is a general tendency for values on coarse-grained scales to be denoted by shorter expressions, which can be seen as evidence of optimization of language. Finally, evidence is discussed that speakers may preferentially select those messages for which an economic code is available.

1 Round number words and round interpretations

In Switzerland, one can find curious street signs like the one near to Zurich 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 very 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 completely unremarkable? Why is *100 meters* interpreted less precise than *103 meters*? This paper wants to give an answer to this question that is derived from more general pragmatic principles.

The phenomenon that number words like *one hundred* invite a less precise, more approximate interpretation than number words like *one hundred and three*, at least in a context in which they are used for measuring or counting, is well known and has often been observed before. In Krifka (2002) I called this the Round Numbers Round

*I had the opportunity to present various precursors of this paper at a number of conferences before the KNAW colloquium, including *Sinn & Bedeutung* 2001 in Amsterdam and the annual meeting of the *Deutsche Gesellschaft für Sprachwissenschaft* in Munich 2002. 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, Reinhard Blutner, Peter Bosch, Regine Eckardt, Gerhard Jäger, Jason Mattausch, Robert van Rooy, Philippe Schlenker, Uli Sauerland, Theo Vennemann, Henk Zeevat and an anonymous reviewer.

Interpretation (RNRI) principle:

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

Where *measuring* should henceforward include measuring the size of a set by specifying the number of elements in the set, which is normally done by counting.

The question is: Can the RNRI principle, which just states something about number words used for measuring or counting things, be derived from more general pragmatic principles? It is hard to imagine that it could be an irreducible axiom of language use.

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

2 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 (1949). In the neo-Gricean theories of Horn (1984) and Levinson (2000), it has been captured by the R-Principle and I-Principle, respectively. They express that the speaker should say only as much as he must in order to be understood, but not more, and in particular can rely on the addressee to fill in information that is not expressed, if necessary. In our example, this will lead to a preference of *one hundred* over *one hundred and three*, because the gain in precision that comes with the latter term is not necessary.

Secondly, I assumed a principle that prefers approximate interpretations over precise ones. This principle is not as well known, but it has been proposed before, 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 as it is often important to say something one is certain about, it is often necessary 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, witness for example the widespread preference for analog watches over digital ones. In general, then, we have the following pragmatic preferences, one for linguistic forms and one for their interpretation:

> SIMPEXP: simple expression $>$ complex expression, (2)

> APPRINT: approximate interpretation $>$ precise interpretation. (3)

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 $\langle F, M \rangle$, are compared, and among various candidates the optimal pairs are selected according to the following rule:

$$\begin{array}{l}\text{A form-meaning pair } \langle F, M \rangle \text{ is optimal iff} \\ \text{(a) there is no optimal pair } \langle F', M \rangle \text{ such that } F' > F \\ \text{(b) there is no optimal pair } \langle F, M' \rangle \text{ such that } M' > M.\end{array} \tag{4}$$

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:

$$\begin{array}{l}\langle \textit{one hundred}, \text{precise} \rangle, \\ \langle \textit{one hundred}, \text{approximate} \rangle, \\ \langle \textit{one hundred and three}, \text{precise} \rangle, \\ \langle \textit{one hundred and three}, \text{approximate} \rangle.\end{array} \tag{5}$$

Clearly, $\langle$*one hundred,* approximate$\rangle$ is an optimal pair because there is no other pair that is better, hence there is no other optimal pair that is better:

$$\langle \textit{one hundred}, \text{approximate} \rangle > \langle \textit{one hundred}, \text{precise} \rangle, \text{ due to IMPRINT} \tag{6a}$$

$$\langle \textit{one hundred}, \text{approximate} \rangle > \langle \textit{one hundred and three}, \text{approximate} \rangle, \text{ due to APPREXP} \tag{6b}$$

Notice that the pairs $\langle$*one hundred*, approximate$\rangle$ and $\langle$*one hundred and three*, precise$\rangle$ do not compete with each other according to the evaluation algorithm (4). From (6) it follows that $\langle$*one hundred*, precise$\rangle$ and $\langle$*one hundred and three,* approximate$\rangle$ aren't optimal pairs. In a second step, it can be shown that $\langle$*one hundred and three*, precise$\rangle$ is an optimal pair, as it does not compete with any optimal pair: It does not compete with $\langle$*one hundred,* approximate$\rangle$, and the pairs it does compete with, $\langle$*one hundred*, precise$\rangle$ and $\langle$*one hundred and three*, approximate$\rangle$, 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.

3 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 discussed in Section 2, in fact, 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 cannot be truly applicable in the same situation. It would be easy enough to respond to this argument by removing the pair ⟨*one hundred*, precise⟩ from competition. 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 know whether the two forms *one hundred* and *one hundred and three* do compete or not.

A perhaps more important objection against the theory in Section 2 is that the preference for approximate interpretations is not warranted. It might well be that in many situations approximate interpretations are what the speaker intends for the addressee, but one may doubt that there is a general overriding tendency for such interpretations. There are certainly 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 and 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. In the situations considered here, this holds only under the condition of approximate interpretation. In case the interpretation is precise, the speaker does not have any other choice than to select a particular expression, regardless whether it is simple or complex.

Following this reasoning, we don't have to assume a general bias for a precise or an approximate interpretation, such as the one that we did assume with APPRINT. This view is especially attractive in the context of research results concerning the concept of number in human cognition. In this body of research, evidence is forthcoming that our sense of numbers rests on two distinct cognitive systems: First, a nonsymbolic system of quantity estimation, addition and subtraction that can also be found with animals and infants, and secondly, a symbolic system resting on counting that appears to be genuinely human (cf. Dehaene, 1997, and Lemer *et al.*, 2003, for case studies involving acalculia patients). The second system is precise, and it is enforced with small numbers. The numbers 3 and 4 are considered different because they represent different steps in the counting sequence. The first system is inherently approximative, and it mostly shows up with larger numbers. The numbers 44 and 47 may be so close to each other as to be indistinguishable for this system. The two systems are related

to each other, but they fulfill separate roles, and we cannot say that one – say, the approximative one of quantity estimation – is to be preferred over the other.

If precise and approximate interpretation are not ordered with respect to each others, then they cannot be used to evaluate candidates of forms and interpretations. Rather, precise interpretation and approximate interpretation should be candidates themselves among which one or the other can be selected, according to pragmatic principles.

Let us assume a principle INRANGE, a consequence of the Gricean maxim of Quality, which says that assertions must be truthful:

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

Let us consider a simple but clear example. Assume that an integer in the interval $[1, 2, \ldots, 100]$ is to be reported as the result of a counting or a measurement. This can be reported 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 $[i-2, \ldots, i+2]$. For example, reporting the value by *forty* stands for the range $[38, \ldots, 42]$ under the approximate interpretation, and for [40] under the precise interpretation. We now can construct tables like the following, in which pairs of form and interpretation and the true value constitute the input:

	Form/Interpretation pairs	Value	INRANGE	SIMPEXP
✓	⟨*forty*, $[38, \ldots, 42]$⟩	39		
	⟨*forty*, $[40]$⟩	39	*	
	⟨*thirty-nine*, $[37, \ldots, 41]$⟩	39		*
✓	⟨*thirty-nine*, $[39]$⟩	39		
✓	⟨*forty*, $[38, \ldots, 42]$⟩	40		
✓	⟨*forty*, $[40]$⟩	40		
	⟨*thirty-nine*, $[37, \ldots, 41]$⟩	40		*
	⟨*thirty-nine*, $[39]$⟩	40	*	

(9)

If the true 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. 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. It is unclear whether this second result is also desired. One might argue that even in this case the approximate interpretation of *forty* is preferred, and if the precise interpretation is intended, *exactly forty* is preferred.

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.

4 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 and 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 investigates the coding of information in a setting in which the probability and/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 a danger, expressions like *biological mother* can be used (cf. Horn, 1993).

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

Assume as before that measurements may be reported in precise or approximate ways, where reporting in an approximate way should mean that a reported value stands for range of possible values. The proper way of capturing this idea is to represent the range by a normal distribution; for example, if *thirty-nine* is reported in an approximate way, then it would optimally represent 39, less optimally 38 and 40, still less optimally 37 and 41, and so on. To keep things simple, I will continue representing approximate interpretations of i by an interval $[i - s, \ldots, i + s]$, which might be interpreted by a normal distribution with mean i and standard deviation s. The magnitude of s indicates the coarseness of the interpretation; with smaller s, the interpretation gets more precise. With $s = 0$, the interpretation is maximally precise. When we define levels of approximate interpretations, we have to allow higher standard deviations for higher means; the interpretation $[998, \ldots, 1002]$ is much more precise than the interpretation $[8, \ldots, 12]$. A level of approximation can be defined by the quotient of standard deviation s and the mean i. For example, at a level of approximation of $1/10$, a reported value *ten* would stand for $[9, \ldots, 11]$, whereas a reported value of *one hundred* would stand for $[90, \ldots, 110]$. At a level of approximation of 0, numbers are interpreted in a precise way. Cf. Dehaene (1997) for further discussion of such relationships, which relate to the Weber–Fechner law of discriminability of stimuli.

Two values that are interpreted in an approximate way may be indistinguishable with respect to each other. For example, the values *thirty-nine* and *forty* may be indistinguishable if reported under an approximate interpretation with a standard deviation of 2. They would report the normal distributions $[37, \ldots, 41]$ and $[38, \ldots, 42]$, respectively, which largely overlap. In everyday conversation, the information carried by *thirty-nine* and *forty* are equivalent. This does not apply, of course, for reporting measurement values with a margin of error in physics, where 39 ± 2 and 40 ± 2 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. To be specific, we can assume that two normal distributions are not distinguishable if their means are within their standard deviations. Hence, the normal distributions $[37, \ldots, 41]$, with mean 39, and $[38, \ldots, 42]$, with mean 40, would be indistinguishable.

Consider now the same task as before: A result of a measuring has to be reported that is an integer in the interval $[1, \ldots, 100]$. 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 $1/10$, and a precise one with level of approximation 0. Let us furthermore assume that both levels of approximation are equally likely. As always, the speaker knows about this, and takes it into account when encoding the report.

Let us first consider the case the speaker reports *thirty-eight*. There are two possible interpretations. We first consider the approximate interpretation, under which the expression reports the normal distribution $[38 - 3.8, \ldots, 38 + 3.8]$. This interpretation is indistinguishable from a number of alternative utterances, in particular *thirty-five*, *thirty-six*, *thirty-seven*, *thirty-nine*, *forty* and *forty-one*. 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 them, as it is the shortest. The speaker has not done so; he uttered *thirty-eight*. 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-eight* stands for 38, 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 normal distribution $[40 - 4, \ldots, 40 + 4]$. This is indistinguishable from the interpretations of the alternative utterances *thirty-six*, ..., *forty-four*. But among these utterances, *forty* is the shortest and would have been chosen. Hence the utterance is consistent with the assumption of the approximate interpretation.

Of course, assuming the precise interpretation is also consistent with the utterance, just as before. Reasoning about alternative utterances with indistinguishable interpretations does not resolve this case. But the addressee has an a-priori assumption about the reported value: namely, that all the values have the same probability. The addressee now evaluates the two interpretation possibilities under this a-priori assumption. Under

the precise interpretation, *forty* would report the value 40, which has an a-priori probability of 0.01. Under the approximate interpretation, *forty* would report the normal distribution $[40-4, \ldots, 40+4]$, which has an a-priori probability of 0.08, if we count for that the likelihood that the value is within one standard deviation. Hence the more conservative assumption is that the speaker applied the approximate interpretation.

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 toward 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 $[38, \ldots, 42]$ 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.

5 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:

I wrote this article in twenty-four hours. (approximate) (10a)

I wrote this article in twenty-three hours. (precise) (10b)

The house was built in twelve months. (approximate) (11a)

The house was built in eleven months. (precise) (11b)

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.

Even worse, sometimes the idea that simple expressions lead to approximate interpretations is plain wrong:

Mary waited for forty-five minutes. (approximate) (12a)

Mary waited for forty minutes. (precise) (12b)

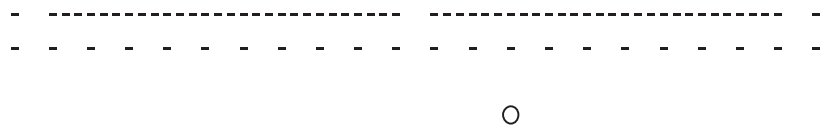
The wheel turned one hundred and eighty degrees. (approximate) (13a)

The wheel turned two hundred degrees. (precise) (13b)

Her child is eighteen months.	(approximate)	(14a)
Her child is twenty months.	(precise)	(14b)
John owns one hundred sheep.	(approximate)	(15a)
John owns ninety sheep.	(precise)	(15b)

Such examples point to the fact that speakers do not just prefer simplicity of expressions, but also simplicity of 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: 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 or in addition to a bias for short expressions, a bias for simple representations. This can be seen as a bias for coarse-grained representations, in the sense of Curtin (1995). The basic idea is that 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). 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 (16).

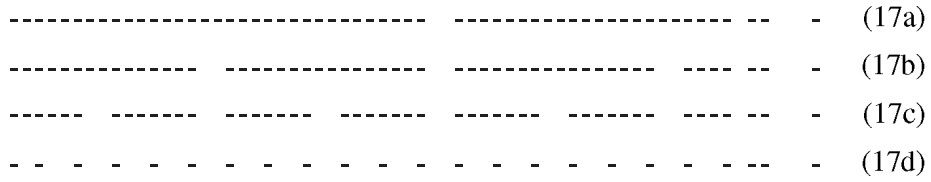


(16)

The more coarse-grained scale (16a) has fewer values for representing measurements. For example, any measurement between 35 and 45 is represented by a single value on the scale, 40. A result of counting or measuring has to be reported using the value that is closest, at the chosen granularity level. In the indicated example, the small circle in (16) will be represented by 43 on the fine-grained level (16b), and by 40 on the coarse-grained level (16a).

In examples (10) to (15), 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 (12). The relevant scales are the scale that counts the hours; then a scale that counts half-hours; then a scale that counts quarter hours; then a scale that counts in 5 minute intervals; and finally a scale that counts in

minutes, not represented here.



(17a)
(17b)
(17c)
(17d)

We can now explain why (12a), *forty-five minutes*, is interpreted in a less precise way than (12b), *forty minutes*, following the reasoning of Section 4.

Let us first consider the case *forty-five minutes*; we have to explain why this invokes the more coarse-grained scale (17c) rather than the fine-grained scale (17d). With respect to level (17c), the scale point 45 represents the times in $[38, \ldots, 52]$. With respect to level (17d), the scale point 45 represents the times in $[43, \ldots, 47]$. Let the a-priori probability for a report of each time in the range to be considered here be the same, say r. Then the probability that the measured time is in $[43, \ldots, 47]$ is 5_r, and the probability that the measured time is in $[38, \ldots, 42]$ is 10_r. Let the a-priori probability on hearing *forty-five minutes* that one of the scales (17c) or (17d) be used be the same, say s. Then on hearing *forty-five minutes* the probability that the more fine-grained scale (17d) is used is 5_{rs}, and the probability that the more coarse-grained scale (17c) is used is double the value of that, 10_{rs}. Hence the hearer will assume the more coarse-grained scale.

Let us now consider the case *forty* minutes. This does not denote any scale point on the more coarse-grained scale (17c), so it cannot be interpreted at this level. The first scale it can be interpreted at is (17d), where it captures time in $[37, \ldots, 42]$. By a similar reasoning as before, this is the level at which *forty* will be interpreted, and not the even more fine-grained level of single minutes with scale points like 38, 39, 40, 41, etc.

In this way, we can deal with apparent exceptions to the preference for short expressions by assuming a preference for simple representations, where the logic of motivating the RNRI phenomenon in strategic communication essentially stays the same. The obvious question, now, is: What is relevant, simplicity of expressions or simplicity of representations? I would like to argue in the next section that both are relevant, that they are in unison to a large extent, and that there are evolutionary pressures to realign them in case they diverge too much.

6 From simplicity of representation to simplicity of expression

Simplicity of expression may not count in the general case as a motivation why certain measurement expressions are interpreted in a less precise way than others, as we have seen in the last section. However, it certainly is not an accident that very often, simplicity of expressions correlates with coarse-grainedness of scales, that is, simplicity of representations. For example, the following scale hierarchy, while satisfying the

requirement of equidistance of scale points, would be decidedly odd:

- (18a)

- ------- ------- ------- ------- ------- ------- ------- (18b)

The scale (18b) 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 is also not motivated as a translation of another system into the decimal one. We can measure simplicity of expressions on a specific scale in various ways, for example by counting syllables. We have the following average numbers of syllables for the number words at the three following scales, from one to one hundred:

one, two, three, four, ..., one hundred: $273/100 = \mathbf{2.73}$ syllables per word, (19a)

one, five, ten, fifteen, ..., one hundred: $46/20 = \mathbf{2.3}$ syllables per word, (19b)

one, ten, twenty, thirty, ..., one hundred: $21/10 = \mathbf{2.1}$ syllables per word. (19c)

In contrast, the scale suggested in (18b) does not show any such decrease of expression complexity when compared to the basic scale (19b):

three, six, nine, twelve, ..., ninety-nine: $92/33 = \mathbf{2.79}$ syllables per word. (20)

Hence we can assume the following general principle that I call SER, which relating Simplicity of Expressions and Simplicity of Representations:

SER:
For any two related scales $S1$, $S2$: $S1$ is more fine-grained than $S2$ iff
the average complexity of expressions of the values of $S1$ is greater than
the average complexity of expressions of the values of $S2$
(measured over a reasonably large interval of expressions). (21)

SER is not exception-less. For example, consider the following scales of reporting the ages of young kids in months:

- (22a)

--- ----- ----- ------- ------- ------- ------- ------- (22b)

The average complexity of English number words at the scale (22a) 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 certainly quite rare.

If simplicity of expressions and simplicity of representations diverge, then one can sometimes detect evolutionary pressure to realign them. One example of this is the expression of amounts of money, e.g., in the US: *quarter – two quarters – three quarters* instead of the more verbose *twenty-five cents – fifty cents – seventy-five cents*. For similar reasons, Dutch had expressions like *kwartje* for 25 cents and *rijksdaalder* 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 $\frac{1}{2}$, as illustrated:

---- ---- ---- ---- ---- ---- ---- ---- ---- (23a)

- - - - - - - - - - - - - - - - - - (23b)

Refining the granularity by $\frac{1}{2}$, as in the step from (23a) to (23b), has the following advantages: First, values that are located right in between, say, 40 and 50 cannot be clearly reported by either '40' or '50'. The granularity of representation distorts such values more than others. Refining the granularity of representation by $\frac{1}{2}$ is the optimal way to overcome this problem. Secondly, if granular representations are in general evidence of a magnitude estimation system for numbers, then the simplest way to refine the granularity level of a scale $S1$ is to divide the distances between the values of $S1$ in half. Halving is a particularly simple operation of dividing: the two resulting parts must be equal in size, a comparison which can be handled by estimation of magnitude.

Other from that, the number 5 has no special property in decimal counting systems. Nevertheless, we find that the expression of this number is sometimes special. In English and in colloquial German, we have the following phonological irregularities:

fifteen [fɪfti:n], instead of regular **fiveteen* [fajfti:n], (24a)

fifty [fɪfti], instead of regular **fivety* [fajfti]; (24b)

colloquial form *fuffzehn* [fuftˢe:n], standard form *fünfzehn* [fynftˢe:n], (25a)

colloquial form *fuffzig* [fuftˢɪg], standard form [fynftˢɪg]; (25b)

Notice that the irregular forms are phonologically simpler: The stem *fif-* consists of two morae, compared with the three morae of *five*, and it has a monophthong in place of a diphthong. The stem *fuff-* also has two morae instead of three, a loss of rounding, and a loss of the nasal. We can see this as evidence for the simplification of expressions that invite an approximate interpretation. In the case of English, the Old English stem *fif* [fi:f] was reduced to *fif* [fɪf] in these environments, while undergoing regular development to *five* [fajf] during the Great Vowel Shift in others. A likely reason that facilitated such simplifications is by the higher frequency of the phoneme *five* in combinations like *five* + *teen* or *five* + *ty*, which is a typical cause for 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.

4-5-6:	IV-V-VI,	(26a)
40-50-60:	XL-L-LX,	(26b)
400-500-600:	CD-D-DC.	(26c)

Other phonological simplifications in number words can also be explained as the result of an increase in frequency that comes with the possibility of an approximate use of these number words. The most prominent case in English is *twelve* (Germanic **twa-libi* 'two + remnant'), which translates the basic unit of an older duodecimal system based on the number 12, for which we also have evidence in the term *dozen.*

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 and 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 and 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 a quantity. It might be a factor behind the frequent vigesimal systems, if they are not just motivated by counting fingers and toes, and it might be the motivation for the special role of 40 in number systems such as the ones in Danish or Albanian, which are simpler than expected. The special role of 5 is motivated by the prominent operation of dividing a prominent quantity, 10, in half, as mentioned above.

A particularly interesting kind of evidence for the issue whether approximate interpretations are correlated to complexity of expression or to prominent ways of changing the granularity of scales are languages with number systems built on different bases, such as the vigesimal number systems based on twenty in Basque, Georgian, the Celtic languages and Danish, in which the number words for 40 ("two score") and 60 ("three score") are shorter than the number words for 50 ("two score and ten"). Hammarström (2004) could indeed show this effect for a number of languages. In the following, I report my own counts for the number words for multiples of 10 between 20 and 90 in the closely related Scandinavian languages Norwegian (decimal system) and Danish (vigesimal system) using the web sites that Google identified as Norwegian and Danish,

respectively (as of March 4, 2005):

Number	**Norwegian**	Occurrences	**Danish**	Occurrences
20	*tjue*	61,300	*tyve*	121,000
30	*tretti*	43,700	*tredive*	25,400
40	*førti*	**39,200**	*fyrre*	**26,800**
50	*femti*	**81,200**	*halvtreds*	**15,500**
60	*seksti*	**19,400**	*tres*	**36,400**
70	*sytti*	10,200	*halvfjerds*	581
80	*åtti*	13,100	*firs*	3,740
90	*nitti*	13,500	*halvfems*	540

(27)

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 does not change when we consider the complete form *halvtredsind-s-tyve* '50', literally 'half-third-times-of-twenty', which has only 1180 occurrences. Notice that the shortening of this complex form to *halvtreds*, and similarly for *halvfjerds* and *halvfems*, nevertheless can be seen as a development leading to simplifications of the expressions of prominent values. There also exists a form *femti* '50', mostly for monetary purposes, which occurred only 988 times on Danish web pages. The table also shows that we have 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. There is some evidence that written Danish resorts to the Arabic number notation more often when number words are complex, something that is not possible in spoken Danish, of course.

If we assume that for Danish and Norwegian speakers occasions to refer to particular numbers occur with similar likelihood, then one way to explain these striking differences is that for Danish speakers the number words for 40 and 60 indeed invite more easily an approximate interpretation than the number word for 50, whereas for Norwegian speakers, the number word for 50 invites an approximate interpretation more easily than the number words for 40 or 60.

This would be reminiscent of how the monetary system used influences the prices of goods. At flea markets in pre-Euro Netherlands, I am told, prices often came in multiples of *kwartjes* (25 cents) and *rijksdaalders* (2,5 guilders), for no obvious reason but to reduce the expense of the transaction itself. There is evidence for a similar principle of reduction of transaction costs in communication.

7 Conclusion

In this paper I have argued that the fact that simple numbers invite an approximate interpretation can be explained by general pragmatic principles. We found that even if there is no particular preference for approximate interpretations and precise interpretations, a weak preference for simple expressions that is checked by a strict preference for truthful interpretations predicts our findings. I have also discussed cases in which

it was not simplicity of expressions that mattered, but rather saliency of representations on more or less fine-grained scales. I have argued that, in the grand scheme of things, simplicity of expressions does matter after all, as salient representations tend to be shorter, and tend to be shortened in language change. We also found evidence, in the comparison of decimal and vigesimal number systems, that speakers might abide by the structure of their language, disregarding cognitive simplicity.

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 the reverse can be true as well: Speakers sometimes do most what grammars do best. 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 the coding in selecting what is said in the first place.

References

Blutner, Reinhard. 2000. Some aspects of optimality in natural language interpretation. J. Semantics **17**: 189–216.

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

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

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

Dekker, Paul and Robert van Rooy. 2000. Bidirectional 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.

Horn, Laurence. 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*, pp. 11–89. Georgetown University Press, Washington, DC

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

Jäger, Gerhard. 2002. Some notes on the formal properties of bidirectional Optimality Theory. J. Logic, Language and Interpretation **11**(4): 427–451.

Jansen, C.J.M. and M.M.W. Pollmann. 2001. On round numbers: Pragmatic aspects of numerical expressions. J. Quantitative Linguistics **8**: 187–201.

Krifka, Manfred. 2002. Be brief and vague! And how bidirectional optimality theory allows for Verbosity and Precision. In D. Restle and D. Zaefferer (eds), *Sounds and Systems. Studies in Structure and Change. A Festschrift for Theo Vennemann*, pp. 439–458. de Gruyter, Berlin.

Lemer, Cathy, et al. 2003. Approximate quantities and exact number words: dissociable systems. Neuropsychologica **41**: 1942–1958.

Levinson, Stephen. 2000. *Presumptive Meanings*. MA.

Ochs Keenan, Eleanor. 1976. The universality of conversational postulates. 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.
Zipf, George K. 1949. *Human Behavior and the Principle of the Least Effort*. Addison-Wesley, Cambridge, MA.