Precision and mathematical form in first and subsequent mentions of numerical facts and their relation to document structure

Sandra Williams and Richard Power
The Open University
Walton Hall, Milton Keynes MK7 6AA, U.K.
s.h.williams@open.ac.uk; r.power@open.ac.uk

Abstract

In a corpus study we found that authors vary *both* mathematical form and precision[1] when expressing numerical quantities. Indeed, within the same document, a quantity is often described vaguely in some places and more accurately in others. Vague descriptions tend to occur early in a document and to be expressed in simpler mathematical forms (e.g., fractions or ratios), whereas more accurate descriptions of the same proportions tend to occur later, often expressed in more complex forms (e.g., decimal percentages). Our results can be used in Natural Language Generation (1) to generate repeat descriptions within the same document, and (2) to generate descriptions of numerical quantities for different audiences according to mathematical ability.

1 Introduction

This study is part of the NUMGEN project[2], which aims (a) to investigate how numerical quantity descriptions vary in English, (b) to specify a grammar that covers these variations, and (c) to develop an algorithm that selects appropriate descriptions for people with different levels of mathematical ability. We collected, from newspapers, popular science magazines and scientific journals, examples of numerical facts that were mentioned more than once, so that first mentions could be compared with subsequent mentions. For example in the following text, two mentions of the same numerical fact – the proportion of A grades in UK A-level examinations in 2008 – are underlined:

> ***A-level results show record number of A grades***
> ***Record numbers of teenagers have received top A-level grades***
> *By Graeme Paton, Education Editor*
> *More than a quarter of papers were marked A as results in the so-called gold standard examination reach a new high.*
> ...
> *According to figures released today by the Joint Council for Qualifications, 25.9 per cent of A-level papers were awarded an A grade this summer ...*
> (Daily Telegraph, 14 August 2008)

Comparing the two, (a) the first (*More than a quarter*) is less precise than the second (*25.9 per cent*), (b) its mathematical form, a common fraction, is less complex than the decimal percentage form of the second, and (c) its string has more characters (i.e., it is *not* shorter in length as might be expected if it were a summary). Also, the two mentions occur in different parts of the document – the first paragraph, and the fifth paragraph.

1.1 What do we mean by precision?

To compare the **precision** of numerical expressions we needed a more exact definition of the concept. We derived the following rules to determine precision:

- Precision increases with the number of significant figures
- Round numbers imply vagueness (implicit approximation)
- Modifiers increase the precision of round numbers when they indicate the direction of approximation ($>$ or $<$)
- Common proportional quantities imply vagueness (implicit approximation similar to round numbers)

[1] Our use of the term *precision* has nothing to do with precision in information retrieval (i.e., the percentage of documents retrieved that are relevant).

[2] http://mcs.open.ac.uk/sw6629/numgen

Proceedings of the 12th European Workshop on Natural Language Generation, pages 118–121,
Athens, Greece, 30 – 31 March 2009.

Our first rule concerns arithmetical precision — i.e., the number of significant figures. Thus 344 with three significant figures is more precise than 340 with only two and *56%* with two significant figures is more precise than *50%* with one.

Second, we adhere to Krifka's RNRI (round number round interpretation) theory that when speakers or writers mention a round figure such as *sixty*, they mean that the actual figure is slightly less than or more than the round number unless they explicitly modify it with (say) *exactly*, and similarly, hearers or readers interpret it as rounded (Krifka, 2007). As a consequence, *sixty* and *around sixty* have the same level of precision, while *exactly sixty* is more precise than *sixty*.

Third, we take into account modifiers (or numerical hedges) such as *under*, *over*, *more than*, and verbs such as *topped*. So we say that *over sixty* and *topped sixty* are more precise than *sixty* since they give more information.

Finally, we extend Krifka's ideas (2007) to cover common proportional quantities. Krifka confined his ideas to scalar and numerical quantities, but we propose that they can also be applied to common proportions such as *half*, *two thirds* and *three quarters* and their ratio, decimal, percentage and multiple equivalents. We hypothesise that when speakers or writers use a common proportion, they implicitly round up or down just the same as with round whole numbers, so we would argue that *around a half* is the same level of precision as *a half*, whereas *more than half* is more precise than *half*. When comparing different types, we take the implied vagueness of common proportions into account, so that we consider *25%* to be more precise than *one quarter*.

1.2 Maths form and conceptual complexity

Numerical proportions may be expressed by different **mathematical forms**, e.g., fractions, ratios, percentages. Complexity of mathematical form denotes the amount of effort and numerical skill required by readers to interpret a numerical quantity; as complexity of mathematical concepts increases, the amount of effort required for comprehension also increases.

As a convenient measure of the complexity of mathematical forms, we employ a scale corresponding to the levels at which they are introduced in the Mathematics Curriculum for Schools (1999); that is, we assume that simple concepts are taught before difficult ones, so that a child learns whole numbers up to ten at Level 1, then much later learns standard index form (e.g., $4.12 \text{x} 10^6$) at Level 8 (table 1).

Maths Form	**Level or Complexity**
Whole numbers 1–10	Level 1
Whole numbers 1–100	Level 2
Whole numbers 1–1000	Level 3
1-place decimals	Level 3
Common fractions	Level 3
Money and temperature	Level 3
Whole numbers $>$ 1000	Level 4
3-place decimals	Level 4
Multiples	Level 4
Percentages	Level 4
Fractions	Level 5
Ratios	Level 5
Decimal Percentages	Level 6
Standard index form	Level 8

Table 1: Scale of Level/Complexity extracted from the Maths Curriculum for Schools (1999)

2 Hypotheses

Our hypotheses about repeated mentions of numerical facts are as follows:

- Precision will increase from first to subsequent mentions.
- Level of complexity of mathematical forms will increase from first to subsequent mentions.
- Changes in precision and mathematical form are related to document structure.

3 Empirical Study

3.1 The NUMGEN Corpus

The corpus has 97 articles on ten topics, where each topic describes the same underlying numerical quantities, e.g., 19 articles on the discovery of a new planet all published in the first week of May 2007 (from Astronomy and Astrophysics, Nature, Scientific American, New Scientist, Science, 11 newspapers and three Internet news sites). In total, the corpus has 2,648 sentences and 54,684 words.

3.2 Corpus analysis and annotation

The articles were split into sentences automatically, then checked and corrected manually. We annotated 1,887 numerical quantity expressions (788 integers, 319 dates, 140 decimals, 87 fractions, 107 multiples, 66 ordinals, 336 percentages and 44 ratios).

In this study, we looked for coreferring phrases containing numerical quantities, such as the sentences *... of papers were marked A* and *... of A-level papers were awarded an A grade* in the above text, and compared the numerical expressions associated with them.[3] Then, for each fact, we noted the linguistic form of first and subsequent mentions in each text and their document positions.

3.3 Judgements on precision and mathematical level

Two readers (the authors) judged whether precision had changed from first to subsequent mentions of a numerical fact in a text, and if so, whether it had increased or decreased, according to the rules set out in the list in section 1.1. We also judged the conceptual complexity of mathematical forms, ranging from 1 to 8 (as defined in table 1). For precision, the judges agreed on 94% of cases (Cohen's kappa is 0.88). Differences were resolved by discussion.

3.4 Results

Table 2 shows results for binomial tests on 88 cases of repeated numerical facts. They show a clear trend towards *unequal precision* between first and subsequent mentions and, in the 62 cases where it is unequal, an overwhelming trend for precision to *increase*. Regarding mathematical level (i.e., the complexity scale for mathematical form), the trend is for subsequent mentions to have a level *equal* to that of first mentions, but in the 31 cases where it is unequal, they show a significant trend towards an *increase in level* — i.e., subsequent mentions are conceptually more difficult.

Our first hypothesis (precision increases from first to subsequent mentions) is thus clearly supported. Our second hypothesis (level of conceptual complexity increases from first to subsequent mentions) is supported by significant increases in level only where the level changed. Note that by our definition, complexity of mathematical concepts is distinct from precision: for example, 59 is more precise than 60 but equally complex (both are taught at Level 2 – whole numbers up to 100). Further investigation revealed that mathematical level tended to remain the same where both mentions were at the beginning of a document (n=14, $p < 0.005$, in a 2-tailed binomial test, as above).

Observation	n	Prop.	Sig.
Precision: Equal	26	.30	.0002
Unequal	62	.70	
Precision: Increase	56	.90	
Decrease	6	.10	.00001
Maths Level: Equal	57	.65	
Unequal	31	.35	.007
Maths Level: Increase	25	.81	
Decrease	6	.19	.0009

Table 2: Binomial tests on repeated mentions, based on .5 probability, 2-tailed, Z approximation.

Hypothesis three (changes in precision and mathematical form are related to document structure) is partially validated in that precision and mathematical level both increase from early to later positions in the document structure.

4 Discussion

Are these results surprising? We believe they show that appropriate presentation of numerical information requires surprising sophistication. It is usual to *summarise* information early in an article, but with numerical facts, summarisation cannot be equated with lower precision or with simpler mathematical form. If summarisation means identifying important facts and presenting them in a condensed form, then why are early mentions of numerical facts *not* condensed? A surprisingly large proportion of first mentions (45%) had longer (or equally long) strings than subsequent mentions (see the text in the introduction, where *More than a quarter* is longer than *25.9 per cent*). Also, why change the mathematical form? It is not obvious that 25.9% should be converted to a common fraction. Intuitively we might reason that *25.9%* is close to *25%* which can be expressed by the simpler mathematical form *a quarter*, but it is far from obvious how this reasoning should be generalised so that it applies to all cases.

A side-effect of our analysis is that it provides some empirical evidence in support of

[3] Note that the numerical facts themselves do not corefer, since they are merely properties of coreferring sets or scales (Deemter and Kibble, 2000).

Krifka's RNRI theory (2007); however, the data is sparse. Ten repeated mentions of numerical facts had round, whole number first mentions and subsequent mentions that were more precise, e.g., *200,000...207,000*. Thus demonstrating that authors do indeed write round numbers which they intend readers to interpret as being approximate. There is similar evidence from 22 examples demonstrating that RNRI can be extended to common proportions.

5 Related work

Communicating numerical information is important in Natural Language Generation (NLG) because input data is wholly or partially numerical in *nearly every* NLG system, but the problem has received little attention. For example, SUMTIME summarises weather prediction data for oil rig personnel e.g., *1.0-1.5 mainly SW swell falling 1.0 or less mainly SSW swell by afternoon* (Reiter et al., 2005) but would require much greater flexibility to present the same numerical facts to non-professionals.

The difficulty of communicating numerical information has been highlighted in educational and psychological research. Hansen *et al.*'s book (2005) provides ample evidence of confusions that many children have about e.g., decimal places; indeed, they demonstrate that many believe *68.95%* is larger than *70.1%* -- misconceptions that often persist into adulthood. Even professionals misunderstand the mathematics of risk. Gingerenzer and Edwards (2003) found doctors calculate more reliably with reference sets than with proportions.

We are not aware of any research on linguistic variation in proportions; in fact, a recent special issue on numerical expressions contained *no* papers on proportions (Corver et al., 2007).

6 Conclusions and Future Work

In this paper we presented:

- A set of rules for determining precision in numerical quantities that is sufficient to cover the examples in our corpus
- A scale for conceptual complexity in numerical expressions derived from the Mathematics Curriculum for Schools.
- A corpus of sets of articles whose main message is to present numerical facts
- Empirical results demonstrating trends towards increasing precision and complexity in repeat mentions of numerical facts with position in document structure.

Our results identify an interesting and well-defined problem that will be addressed in the final stage of NUMGEN: how to derive appropriate simplified expressions (less precise, simpler mathematical form) for use in contexts like the openings of articles, or communications intended for readers with lower levels of mathematical ability.

Acknowledgements

Our thanks to members of The Open University NLG Group. NUMGEN is supported by ESRC[4] Small Grant RES-000-22-2760.

References

N. Corver, J. Doetjes, and J. Zwarts. 2007. Linguistic perspectives on numerical expressions: Introduction. *Lingua, Special issue on Linguistic perspectives on numerical expressions*, 117(5):751–775.

K. Van Deemter and R. Kibble. 2000. On Coreferring: coreference in MUC and related annotation schemes. *Computational Linguistics*, 26:629–637.

G. Gigerenza and A. Edwards. 2003. Simple tools for understanding risks: from innumeracy to insight. *British Medical Journal*, 327:714–744.

A. Hansen, D. Drews, J. Dudgeon, F. Lawton, and L. Surtees. 2005. *Children's Errors in Maths: Understanding Common Misconceptions in Primary Schools*. Learning Matters Ltd, Exeter, UK.

M. Krifka. 2007. Approximate interpretation of number words: A case for strategic communication. In G. Bouma, I. Kraer, and J. Zwarts, editors, *Cognitive foundations of interpretation*, pages 111–126, Amsterdam. Koninklijke Nederlandse Akademie van Wetenschapen.

Qualification and Curriculum Authority. 1999. *Mathematics: the National Curriculum for England*. Department for Education and Employment, London.

E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169.

[4]Economic and Social Research Council