numgen GENERATING INTELLIGENT DESCRIPTIONS OF NUMERICAL QUANTITIES FOR PEOPLE WITH DIFFERENT LEVELS OF NUMERACY

The Open University

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, Milton Keynes, 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 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.

See Figure 1 for an example.

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)

A typical observation

	First mention	Subsequent mention
String	'More than a quarter'	'25.9 per cent'
Complexity	Level 3	Level 6
Significant figures	1	3
Modifier (hedge)	'more than'	-
Rounding	yes	no
Precision	less precise	more precise

Figure 1

Corpus

- 97 articles on ten topics
- Each topic describes the same underlying numerical quantities, e.g., 19 articles about a new planet published in May 2007 in Astronomy and Astrophysics, Nature, Scientific American, New Scientist, Science, 11 newspapers and 3 Internet news sites.
- 2,648 sentences, 54,584 words

Corpus Annotation

- 1,887 numerical quantity expressions (788 integers, 319 dates, 140 decimals, 87 fractions, 107 multiples, 66 ordinals, 336 percentages and 44 ratios)
- 390 inter-text co-referring phrases containing numerical quantities
- 88 intra-text co-referring phrases containing numerical quantities

Hypotheses about repeated mentions of numerical facts

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

Method

- Two readers (the authors) judged whether precision had changed from first to subsequent mentions of a numerical fact, and if so, whether it had increased or decreased.
- The same two readers judged conceptual complexity from 1 to 8 (see table 1).
- For precision, the judges agreed on 94% of cases (Cohen's kappa is 0.88).

A scale of conceptual complexity

Table 1 shows 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.

Maths Form	Level or complexity
Whole numbers 1-10	Level 1
Whole numbers 1- 100	Level 2
Whole numbers 1-1000	Level 2
1-place decimals	Level 3
Common Fractions	Level 3
Money and temperature	Level 3
Whole numbers > 1000	Level 3
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

Rules for determining precision

To compare the precision of numerical expressions we derived the following rules:

- Precision increases with the number of significant figures
- Round numbers imply vagueness – implicit approximation (Krifka, 2007)
- Modifiers increase the precision of round numbers when they indicate direction (> or <)
- Common proportional quantities imply vagueness – implicit approximation similar to round numbers

Results

Table 2 shows results for binomial test (significance is based on 0.5 probability, 2-tailed, Z approximation) on 88 cases of repeated numerical facts.

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

Table 2

They show:

- A clear trend towards unequal precision between first and subsequent mentions – supports hypothesis 1
- An overwhelming trend for precision to increase where it is unequal – supports hypothesis 1
- A trend for equal mathematical precision
- A significant trend for mathematical level to increase where it is unequal (i.e., subsequent mentions are conceptually more difficult) – supports hypothesis 2 – (futher investigation revealed that mathematical level remains the same when both mentions are at the beginning of an article)
- Hypothesis 3 is partially validated in that precision and mathematical form both increase from early to later positions in the document structure

Discussion

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 numerical form. If summarisation means identifying important facts and presenting them in a condensed form, then why are early mentions of numerical facts *not* condensed? 45% of first mentions had longer (or equally long) strings than subsequent mentions (e.g., *More than a quarter* is longer than *25.9 per cent*).

Why change the mathematical form? Intuitively, *25.9%* is close to *25%* which can be expressed by the simpler form *a quarter*, but it is far from obvious how this reasoning should be generalised so that it applies to all cases.

NumGen (Generating intelligent descriptions of numerical quantities for people with different levels of numeracy) is funded by the Economic and Social Research Council under Small Grant Ref. RES-000-22-2760