

How to Make Numerical Information Accessible: Experimental Identification of Simplification Strategies

Susana Bautista¹, Raquel Hervás¹, Pablo Gervás²,
Richard Power³, and Sandra Williams³

¹ Universidad Complutense de Madrid, Spain

² Instituto de Tecnología del Conocimiento, Madrid, Spain

³ Department of Computing, The Open University, Milton Keynes MK76AA, UK
{subautis, raquelhb}@fdi.ucm.es, pgervas@sip.ucm.es,
{r.power, s.h.williams}@open.ac.uk

Abstract. Public information services and documents should be accessible to the widest possible readership. Information in newspapers often takes the form of numerical expressions which pose comprehension problems for people with limited education. A first possible approach to solve this important social problem is making numerical information accessible by rewriting difficult numerical expressions in a simpler way. To obtain guidelines for performing this task automatically, we have carried out a survey in which experts in numeracy were asked to simplify a range of proportion expressions, with three readerships in mind: (a) people who did not understand percentages; (b) people who did not understand decimals; (c) more generally, people with poor numeracy. Responses were consistent with our intuitions about how common values are considered simpler and how the value of the original expression influences the chosen simplification.

Keywords: numerical information, simplification strategies.

1 Introduction

A United Nations report [1] recommends that public information services and documents should be accessible to the widest possible readership. Information in newspapers often takes the form of numerical expressions (e.g., economic statistics, demographic data) which pose comprehension problems for people with limited education. A UK Government Survey in 2003 estimated that 6.8 million adults had insufficient numeracy skills to perform simple everyday tasks, and that 23.8 million adults would be unable to achieve grade C in the GCSE maths examination for 16-year-old school children [2].

A first possible approach to solve this important social problem of making numerical information accessible is to rewrite difficult numerical expressions more simply. Such an approach would require a set of rewriting strategies yielding expressions that are linguistically correct, easier to understand than the original, and as close as possible to the original meaning. For example, ‘25.9%’ could be rewritten as ‘just over a

quarter'. Simplification may in some cases entail loss of precision, but this is not necessarily a bad thing, for several reasons. Loss of precision can be signaled linguistically by numerical hedges such as 'around', 'more than' and 'a little under', so it need not be misleading. As Krifka has argued, competent writers and speakers frequently approximate numerical information and readers and hearers can readily recognize this, even when no hedge is present, especially when numbers are round [3]. For instance, in 'the distance from Oxford to Cambridge is 100 miles' it is clear that 100 miles is an approximation. Williams and Power [4] showed that writers tend to approximate numerical quantities early in a document, then give more precise versions of the same quantities later. As Krifka argues in the same paper [3], an inappropriately high level of precision would flout Grice's Maxim of Quantity [5] by giving too much information. There cannot be many situations in which we need to know that the distance from Oxford to Cambridge is 100.48 miles, for example.

This paper presents an exploratory survey in which experts in numeracy were asked to simplify numerical expressions (presented in context) for several kinds of readership, with the aim of collecting a repertoire of rewriting strategies that can be applied in an automatic text simplification system.

2 Background

Text simplification, a relative new task in Natural Language Processing, has been directed mainly at syntactic constructions and lexical choices that some readers find difficult, such as long sentences, passives, coordinate and subordinate clauses, abstract words, low frequency words, and abbreviations. Chandrasekar et al. [6] introduced a two-stage process, first transforming from sentence to syntactic tree, then from syntactic tree to new sentence; Siddharthan [7] instead proposed a three-stage process comprising analysis, transformation and generation. In 1998, the project PSET [8] employed lexical as well as syntactic simplifications. Other researchers have focused on the generation of readable texts for readers with low basic skills [9], and for teaching foreign languages [10]. However, to our knowledge, there have been no previous attempts to automatically simplify *numerical* information in texts.

A corpus of numerical expressions was collected for the NUMGEN project [4]. The corpus contains 10 sets of newspaper articles and scientific papers (110 texts in total). Each set is a collection of articles on the same topic - e.g., the increased risk of breast cancer in red meat eaters, and the decline in the puffin population on the Isle of May. Within each set, identical numerical facts are presented in a variety of linguistic and mathematical forms.

3 Experiment

Candidate rewriting strategies may be obtained in two ways: one is to collect them directly from human authors, another is to validate strategies mined from a large corpus. Our experiment employs the first option.

3.1 Underlying Assumptions

In this paper we consider a ‘numerical expression’ (NE) to be a phrase that presents a quantity, optionally modified by a numerical hedge as in ‘more than a quarter’ or ‘around 97 %’. To date, we have restricted coverage to proportions - i.e., fractions, ratios and percentages. We have two working hypothesis:

- H1:** When experienced writers choose numerical expressions for readers with low numeracy, they tend to prefer round or common values to precise values. For example, halves, thirds and quarters are usually preferred to eightieths or forty-ninths, and expressions like *N in 10* or *N in 100* are chosen instead of *N in 365* or *N in 29*.
- H2:** The choice between different simplification strategies (fractions, ratios, percentages) is influenced by the value of the proportion, with values in the central range (say 0.2 to 0.8) and values at the extreme ranges (say 0.0-0.2 and 0.8-1.0) favouring different strategies.

3.2 Materials

We focused on simplification strategies at two levels: decimal percentages, and whole-number percentages. Three sets of candidate sentences were chosen from the NUMGEN corpus for presentation to participants: eight sentences containing only decimal percentages, and two sets of eight sentences containing mixed whole-number and decimal percentages. Although the number of sentences in each set was eight, the number of numerical expressions was larger as some sentences contained more than one proportion expression.

A wide spread of proportion values was present in each set, including the two end points at nearly 0.0 and almost 1.0. We also included some numerical expressions with hedges and sentences from different topics in the corpus. In short, we included as many variations in context, precision and different wordings as possible.

3.3 Participants

Our experimental evaluation involved 34 participants, considering only the ones that answered at least one question. They were primary or secondary school mathematics teachers or adult basic numeracy tutors, all native English speakers. The task of simplifying numerical expressions is difficult, but it is a task that this group seemed well qualified to tackle since they are highly numerate and accustomed to talking to people who do not understand mathematical concepts very well. We found participants through personal contacts and posts to Internet forums for mathematics teachers and numeracy tutors.

3.4 Survey Design and Implementation

Our survey took the form of a questionnaire in which participants were shown a sentence containing one or more numerical expressions which they were asked to simplify. The survey was divided into three parts as follows:

1. Simplification of numerical expressions for a person who can not understand percentages. We will refer to this part as ‘No Percentages’.
2. Simplification of numerical expressions for a person who can not understand decimals. We will refer to this part as ‘No Decimals’.
3. Free simplification of numerical expressions for a person with poor numeracy. We will refer to this part as ‘Free Simplification’.

For part (2), the set of sentences containing only decimal percentages was used. One of the two mixed sets of sentences with whole-number and decimal percentages was used for part (1) and the other for part (3). The experiment was presented on SurveyMonkey¹, a commonly-used provider of web surveys.

We asked participants to provide simplifications for numerical expressions that were marked in each sentence. Below the sentence, each numerical expression was shown beside a text box in which the participant was asked to type the simplified version. Our instructions said that numerical expressions could be simplified using any format: number words, digits, fractions, ratios, etc. and that approximators such as ‘more than’, ‘almost’ and so on could be introduced if necessary. Participants were also told that the meaning of the simplified expression should be as close to the original expression as possible and that, if necessary, they could rewrite part of the original sentence.

4 Results

The results of the survey were carefully analyzed as follows. First, within each block of questions, a set of simplification strategies was identified for each specific numerical expression. These strategies were then grouped together according to the mathematical forms and/or linguistic expressions employed (fractions, ratios, percentages). Where necessary, they were subdivided further according to choices of numerical values for the constituents of the simplified expressions (denominators in fractions, or reference value in ratios, for example). Not all simplification strategies occur with enough frequency to merit detailed analysis; the approach followed here has been to group together (under a generic label of *Others*) all simplification strategies with a low frequency of use with respect to the total (for example, in the case of fractions, a total of ten different kinds of fractions were used (hundredths, sixths, tenths, etc.), but we only represent in labeled sub-columns the ones with significant usage; the rest are summed in the *Others* sub-columns). The non-numeric column represents simplified expressions where no numbers were used like ‘almost all’ or ‘around none’. Remaining simplifications (*Rem.* column) are rewritings of the whole sentence or parts of it, coinciding with comments expressed by the participants that sometimes the whole sentence would be better understood if the non-numerical part was also simplified, and some deletions. The observed frequencies (represented in percentages) of the different simplification strategies are given in Table 1. Rows do not add up 100% as not all participants gave an answer for all numerical expressions.

¹ www.surveymonkey.com

Table 1. Frequencies of simplification strategies for 34 participants: (1) No Percentages: intended for people who do not understand *percentages*, (2) No Decimals: intended for people who do not understand *decimals*, and (3) Free Simplification: intended for people with poor numeracy. Frequencies are represented in percentages

NO PERCENTAGES (%)												
Numerical Expression	Fractions				Total	Ratios			Total	Non-numeric	Percent-ages	Rem.
	Halves	Thirds	Quarters	Others		N in 10	N in 100	Others				
more than 1%	3			15	18		6		6	15	18	24
2%				6	6		12	6	18	3	12	38
16.8%			3	24	26		15	50	65		9	
27%		9	71	3	82			12	12		6	
at least 30%		21	9	12	41	29		6	35		3	9
40%	21	6		26	53	29			29		6	3
56%	82				82						6	3
63%	24	41		9	74	9		15	24		3	
75%			32		32			29	29	3		24
97.2%				3	3	3	29	6	38	21	18	12
98%				6	6		12		12	65	3	9
Mean	12%	7%	10%	9%	39%	6%	7%	11%	24%	10%	7%	11%
NO DECIMALS (%)												
Numerical Expression	Fractions				Total	Ratios			Total	Non-numeric	Percent-ages	Rem.
	Halves	Thirds	Quarters	Others		N in 10	N in 100	Others				
0.6%	3			3	6	3	6		9	6	47	3
2.8%				3	3	24			24		47	9
6.1%						15		3	18		50	3
7.5%				12	12	3	6	3	12		50	6
15.5%				15	15	3	6	3	12		44	9
25.9%			15		15		3	9	12		38	3
29.1%				3	3	9	3	3	15		50	3
35.4%		9		3	12	9	3	3	15		41	3
50.8%	44				44		3		3		21	3
73.9%			44		44		3	3	6		18	3
87.8%				3	3	9	3	3	15		47	3
96.9%				3	3	6	3	3	12		29	12
96.9%				6	6	9	6	3	18		21	6
97.2%				3	3	3	6	6	18	3	41	6
97.2%				3	3	12	3	3	18	3	32	6
98.2%				3	3	9	3	3	15	6	44	3
Mean	3%	1%	4%	4%	11%	7%	3%	3%	14%	1%	39%	5%
FREE SIMPLIFICATION (%)												
Numerical Expression	Fractions				Total	Ratios			Total	Non-numeric	Percent-ages	Rem.
	Halves	Thirds	Quarters	Others		N in 10	N in 100	Others				
0.7%							6		6	18	9	26
12%				6	6	12	3	6	21		21	3
26%			41		41			12	12			3
36%		41			41	3		6	9			3
53%	41				41						6	6
65%	6	15			21	3	9	6	18		3	12
75%			15		15			9	9	6	3	15
91%						21	9		29	6	6	12
above 97%						3	29		32	12	6	
Mean	5%	6%	6%	1%	18%	5%	6%	4%	15%	5%	6%	9%

In order to analyse the results we performed a one-way analysis of variance (ANOVA), which results are represented in Table 2. When considering the whole survey (*Whole* column), there is no significant difference in the use of fractions, ratios and percentages. Only the use of non-numeric expressions is significant, but this is due to their low usage. However, when analysing the survey by parts we find interesting results.

Table 2. Results of ANOVA test. Strategies which do not share a letter are significantly different.

Strategy	No Percentage			No Decimals			Free Simplif.		Whole	
Fractions	A			A			A		A	
Ratios		B		A			A		A	
Percentages			C		B			B	A	
Non-Numeric			C			C		B		B

Overall, fractions are the preferred simplification for *people who do not understand percentages*. Although ten different types of fractions were used by the participants, the most commonly used were halves, thirds and quarters. The second preferred type of expression is ratios. From the nine different types of ratios employed (ranging from N in 10 to N in 1000), the most common were N in 10 and N in 100. It is surprising that 7.5% of the expressions chosen were percentages, even though participants were asked to simplify for people who do not understand percentages. We are unsure whether they ignored the instructions, did not agree with them, or just did not find another way of simplifying the expression. However, the use of percentages is not significant with respect to the use of non-numeric expressions.

Whole number (cardinal) percentages are the preferred simplification for *people who do not understand decimals*. This reinforces the idea that they are easier to understand than the original number, while at the same time being the closest to the original value and mathematical form. Frequencies of use of fractions and ratios are very similar and are not significantly different. Non-numeric simplifications were seldom used, in contrast to the first part of the survey; in fact, they occurred only for the peripheral points on the proportion scale, e.g., *almost everyone* or *a little*.

Fractions and ratios are similarly used when simplifying for *people with poor numeracy*. The frequencies of non-numeric expressions and percentages are similar to the ones in the first part of the survey.

In order to test hypothesis H1 (round or common values are preferred to precise ones), we carried out a series of two sample t -tests on common and uncommon fractions and ratios. The results showed that there was significant difference between the use of common and uncommon fractions in the three parts of the survey and the whole survey (no percentages: $p < .001$, no decimals: $p = .07$, free simplification: $p < .0001$, whole: $p < .0001$). However, in the case of ratios there was no significant difference except in the case of free simplification (no percentages: $p = .48$, no decimals: $p = .36$, free simplification: $p = .006$, whole: $p = .14$).

As can be seen in the results, the use of different types of fractions seems to depend on the value being simplified, with quarters, thirds and halves (common fractions) preferred in the central range from 20% to 80%, and greater variety (and rarer use of fractions) at the peripheral. These phenomena can also be observed in non-numeric expressions. This was our hypothesis H2, and in order to test it we performed a series of two sample t -tests on the use of fractions, ratios, percentages and non-numeric in central and peripheral values. The results showed that the use of the four strategies was significantly different for central and peripheral values of the proportions (fractions: $p < .0001$, ratios: $p = .03$, percentages: $p < .0001$, non-numeric: $p < .0001$). The only exception was the use of ratios in the first part of the survey (simplification for people who do not understand percentages), with a p -value of 0.14.

5 Discussion

When asked to simplify for people who do not understand percentages, or for people with poor numeracy, the participants preferred fractions, followed by ratios; when asked to simplify for people who do not understand decimals, they preferred whole-number percentages. Responses show that fractions are considered as the simplest mathematical form, followed by ratios, but this did not mean that fractions were preferred to ratios in every case: the value of the original proportion also influenced choices, with fractions heavily preferred for central values (roughly in the range 0.2 to 0.8), and ratios or non-numeric preferred for peripheral values (below 0.2 or above 0.8), always depending on the kind of simplification being performed.

As some participants commented, not only are percentages mathematically sophisticated forms, but they may be used in sophisticated ways in a text, often for example describing rising and falling values, for which increases or decreases can themselves be described in percentage terms. Such complex relationships are likely to pose problems for people with poor numeracy even if a suitable strategy can be found for simplifying the individual percentages. Another danger is that simplifying several related percentages might obscure the relationship between them. One obvious case would be to render two different values identical, e.g. by simplifying both 48% and 52% to one-half. Another would be to replace two percentages by apparently simpler raw data, for two values with different totals, thus making it harder to see which of the two proportions is larger. In some of the examples with more than one numerical expression being compared, some of the evaluators reported a tendency to phrase them both according to a comparable base - e.g., both in terms of tenths, rather than one as a fifth and one as a third. Thus we should consider the role of context (the set of numerical expressions in a given sentence as a whole, and the meaning of the text) in establishing what simplifications must be used.

6 Conclusions

Through a survey administered to experts on numeracy, we have collected a wide range of examples of appropriate simplifications of percentage expressions. Our aim is to use this data to guide the development of a system for automatically simplifying percentages in texts. With the knowledge acquired from our study we will improve our algorithm to simplify numerical expressions. Our initial hypothesis was that in choosing suitable simplifications, our experts would favor certain mathematical forms - those corresponding to simpler mathematical concepts taught earlier in the curriculum. As expected, the results supported a ranking in which fractions were the simplest form, followed by ratios, whole-number percentages and decimal percentages. However, it did not follow that for *any* proportion value a fraction was the most appropriate simplification, because other forms (e.g., non-numeric expressions) were preferred for peripheral values (near to 0% or 100%) that would have required unfamiliar fractions such as one-hundredth. The value of the original proportion also influenced choices, depending on its correspondence with central or peripheral values. Our results also show that the experts use different options for each simplification strategy.

We have also collected a parallel corpus of numerical expressions (original and simplified version). This corpus will be shared with other researches so it can be used to different applications to improve the readability text. This could be a very useful resource because simplification of percentages remains an interesting and non-trivial problem.

References

1. Nations, U., Standard Rules on the Equalization of Opportunities for Persons with Disabilities. Technical report (1994)
2. Williams, J., Clemens, S., Oleinikova, K., Tarvin, K.: The Skills for Life survey: A national needs and impact survey of literacy, numeracy and ICT skills. Technical Report Research Report 490, Department for Education and Skills (2003)
3. Krifka, M.: Be brief and vague! And how bidirectional optimality theory allows for Verbosity and Precision. In: *Sounds and Systems: Studies in Structure and Change: A Festschrift for Theo Vennemann*. Trends in Linguistics, vol. 141, pp. 439–458. Mouton de Gruyter, Berlin (2002)
4. Williams, S., Power, R.: Precision and mathematical form in first and subsequent mentions of numerical facts and their relation to document structure. In: *Proc. of the 12th European Workshop on Natural Language Generation*, Athens (2009)
5. Grice, H.P.: Logic and Conversation. In: Cole, P., Morgan, J.L. (eds.) *Syntax and Semantics*. *Speech Acts*, vol. 3, pp. 41–58. Academic Press, San Diego (1975)
6. Chandrasekar, R., Doran, C., Srinivas, B.: Motivations and Methods for Text Simplification. In: *COLING*, pp. 1041–1044 (1996)
7. Siddharthan, A.: Resolving Attachment and Clause Boundary Ambiguities for Simplifying Relative Clause Constructs. In: *Proceedings of the Student Research Workshop, 40th Meeting of the Association for Computational Linguistics* (2002)
8. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of English newspaper text to assist aphasic readers. In: *AAAI 1998 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Wisconsin (1998)
9. Williams, S., Reiter, E.: Generating readable texts for readers with low basic skills. In: *Proceeding of the 10th European Workshop on Natural Language Generation*, Aberdeen, Scotland, pp. 140–147 (2005)
10. Petersen, S.E., Ostendorf, M.: Text Simplification for Language Learners: A Corpus Analysis. In: *Speech and Language Technology for Education (SLaTE)* (2007)