CASE FOR SUPPORT
NUMGEN: Generating intelligent descriptions of numerical quantities for people with different levels of numeracy

Sandra Williams, 19th September 2007

1. Introduction

Natural Language Generation (NLG) technology enables computers to automatically generate documents from machine-readable data in a language such as English. Currently, NLG systems cannot describe numerical quantities such as proportions (e.g. '0.25') in a flexible way, with appropriate variations for different contexts (types of document, writing styles, or audiences). This limitation is serious because nearly all NLG systems produce language from underlying numerical data – for example, weather forecasts from numerical weather data, stock reports from stock market figures, or medical reports from medical data. In this project, we will investigate this important but neglected area.

To present numerical data appropriately, NLG systems require two types of knowledge:

- ***a varied range of linguistic descriptions*** for a given type of numerical quantity,
- ***rules about which descriptions fit different situations***.

We propose to investigate both of these in our study which will focus on *proportions* (e.g. 'a half'). We will incorporate both types of knowledge into an existing NLG system and evaluate it.

Linguistic variation in numeric descriptions. We know of no previous research on linguistic variation in proportions; in fact, a recent special issue on numerical expressions focussed mostly on cardinals, ordinals and quantifiers (Corver et al. 2007). Nevertheless they are extremely common: open any newspaper and you will see articles about *percentage* increases in interest rates, *decimal* exchange rate for Euros, or sales of goods at a *fraction* of their original cost. Mathematics helps us to identify some variations – 'one half' can be expressed not only as a fraction but also as a decimal or a percentage – but linguistic analysis is needed to identify which phrases actually occur in texts, e.g., 'half a teaspoon', '½ teaspoon' and '0.25g', but probably not 'fifty percent of a teaspoon' or '0.5 of a teaspoon'. Proportions are expressed in many ways in English as we found in a pilot study in which we analysed proportions in articles from journals, newspapers and science magazines about the discovery of an Earth-like planet. These varied widely in expressing exactly the same quantities: e.g., for size of the planet, 'a planetary radius of ~1.5$R_\oplus$', or '50% wider than the Earth', '1.5 times the Earth's radius', 'one-and-a-half times the diameter of Earth'.

We propose to gather examples of variations in proportions by analysing sets of English texts from the Internet where each set describes the same numerical data, as in our pilot study. The examples will be classified and enumerated. From them we will build a grammar which will be evaluated on its coverage of the examples.

Tailoring numerical expressions to fit different situations. A valuable area of research in NLG is devising algorithms that tailor output texts to different audiences. In previous work we built a system that tailored output for people with poor literacy (Williams 2004, Williams and Reiter 2005). A novel extension would be to tailor output for groups with different levels of numeracy. This would address a real social need since in 2003, the U.K. Government surveyed 8,041 adults and estimated that

47% of the population have 'limited mathematical skills' (Skills for Life Survey, HMSO, 2003).

We propose to adopt the National Curriculum for Schools as a convenient measure of numerical ability. It distinguishes skill levels by the concept of 'Key Stages'. We will develop rules that choose suitable numerical expressions for people at each stage (e.g., 'nearly three-quarters' for people at Key Stage 2 and '0.747' for people at Key Stage 3). The rules will be evaluated by asking maths tutors to judge the appropriateness of sentences generated for different Key Stages in mathematics.

1.1 Aims and objectives

The project aims to extend Natural Language Generation technology with techniques for generating numerical expressions; to contribute to linguistic knowledge about numerical expressions in English; and to investigate how to communicate numerical information to different audiences.

Since this is a feasibility study to demonstrate proof-of-concept and potential for subsequent research, it will focus on a subset of the wider problem by (a) restricting the numerical quantities to be investigated to proportions, and (b) restricting contextual factors that we consider to the mathematical ability of the intended audience. Our objectives are as follows:

1. To investigate how proportions are described in English, and develop a grammar that can generate such descriptions;
2. To develop rules to select appropriate descriptions for different levels of maths skills (represented by the Key Stages of the National Curriculum);
3. To implement these in an NLG system and evaluate them.

Objective (1) will contribute to linguistic knowledge and extend NLG technology to generate numerical expressions more flexibly, (2) will extend NLG technology by enabling it to tailor texts for different audiences, and (3) will evaluate the potential for future research.

1.2 Research Questions

The research questions we will address are:

- How do descriptions of proportions vary in English?
- Can an NLG system describe proportions in terms that are appropriate for different levels of mathematical ability?

2. Previous Work

Communicating numerical information in NLG. This is an important problem in NLG because input data is wholly or partially numerical in *nearly every* NLG system, but the problem has received insufficient attention. Our own systems SkillSum and GIRL (Williams and Reiter, 2007) generated feedback on basic-skills tests, but variations in the presentation of numerical data were limited to choosing digits or number words. The CLEF system for describing medical results for patients (Williams et al., 2007) has a similar limitation. Some recent NLG systems summarise numerical time-series data – e.g., SumTime (Reiter et al., 2005) summarises data from weather prediction systems for oil rig personnel, and BabyTalk-Doc (Portet et al., 2007) summarises data from medical monitors (such as blood-pressure monitors) for clinicians – but both of these describe numerical data in the formulaic language of professionals, e.g. '1.0-1.5 mainly SW swell falling 1.0 or less mainly SSW swell by afternoon' (SumTime) and 'toe/core temperature gap rises for 7 minutes to 2.4' (BabyTalk-Doc); they would require much greater flexibility of expression to generate comprehensible numerical descriptions for non-professionals. Our proposed project will produce results that will be invaluable to NLG engineers who build systems for a wide range of people.

Previous NLG research on communication of numerical quantities has been limited to specialised contexts. In a system designed to provide intelligent answers to numerical questions, Moriceau (2006) considered how expressions of numerical information depend on contextual features such as time and place: e.g., for the average ages of women marrying in 2004, it generated an answer with whole-number ages associated with different places. Our proposal differs in that it addresses the wider and more fundamental problem of how to express the numerical information itself without (for now) taking into account other associated contextual features. The approaches are, however, complementary and should be combined in the future.

The difficulty of communicating numerical information. This has been highlighted in a large volume of previous educational and psychological research. Hansen et al. (2005), in their book on children's misconceptions in learning primary school maths, provide ample evidence of the confusions that many children and adults have about for instance, decimal places, in believing that '68.95%' is larger than '70.1%'. Resnick et al.'s paper (1989) is just one example of studies on the misconceptions afflicting older children when learning to calculate with decimal fractions – misconceptions that often persist into adulthood. Studies in psychology and medical informatics show that even professionals misunderstand the slightly more complex mathematics of risk. Gingerenzer and Edwards (2003) found that doctors could calculate more accurately with reference sets rather than proportions, and recommended using frequency expressions in place of percentages.

Previous linguistic research does not address proportions. Research on numerical expressions has been almost entirely limited to cardinals (number names such as 'one', 'two', 'three'), ordinals ('first', 'second', 'third', etc.) and quantifiers (such as 'some', 'few', 'many') – see for example the recent special issue of Lingua entitled 'Linguistic perspectives on numerical expressions' (Corver et al., 2007); we are not aware of any linguistic research on expressing fractions, decimals and percentages. Our proposed linguistic analyses will therefore be entirely novel.

Style guides do not help. There is little help from style guides, such as the 'Chicago Manual of Style' and the 'Plain English Campaign'. Both have little to say about alternative ways to express numbers; instead, they merely state rules for writing different types e.g., percentages:

> 'In humanistic copy the word *percent* is used; in scientific and statistical copy, or in humanistic copy that includes numerous percentage figures, the symbol % is more appropriate.' (Chicago Manual of Style, p. 384).

There is no advice on whether a writer might be wiser to use an alternative such as '¾' or 'three in four' in place of '75%' for certain audiences. This omission is rather puzzling because the guide suggests alternatives for expressing many other kinds of content, e.g. for labels on medicines, they advise changing 'This adult nasal spray is for local application in the nose' to 'Spray directly into the nose' (www.plainenglish.co.uk). Our proposed research will enable us to publish advice for authors on how to vary numerical phrases for different audiences.

3. Methodology

Our investigation will consist of three activities: linguistic analysis and grammar construction, construction of rules to guide generation for audiences with different levels of maths ability and evaluation. These are described in the following sections.

3.1 Linguistic analysis and grammar construction

We will undertake a small-scale, narrowly-focussed analysis of the range of ways in which proportions can be expressed. A large-scale, widely-focussed corpus analysis is inappropriate for our purposes, since it would yield only a wide range of expressions, not a set of alternative expressions for the same content. We expect the analysis to identify which numerical phrases actually occur in texts and the types of entity that they quantify, e.g., for part of a teaspoon we might find 'half a teaspoon', '½ teaspoon', but probably not 'fifty percent of a teaspoon' or '0.5 of a teaspoon'.

Obtaining such data requires the analysis of sets of texts that describe the same numerical quantities, such as the set describing the discovery of an Earth-like planet that was mentioned in the introduction. No such data set is currently available; as pointed out above, we cannot rely on existing corpora from the corpus linguistics or Machine Translation communities, since these do not provide text samples that describe exactly the same content.

In our pilot study of Earth-like planet articles, we used a search engine (Google) to find the original article from a science journal (Astronomy and Astrophysics) and others from popular science magazines and newspapers, downloading a total of fourteen articles all published in the same week. The articles mentioned five numerical quantities, each expressed in the texts in 5-10 different ways, so yielding some thirty ways of describing (in this case) very large numbers; e.g.,'20.5 light years'. We plan to collect at least forty sets of texts on a variety of topics, covering around 200 proportions with a total of over 1000 expressions (excluding repetitions).

Proportion expressions will be identified by hand, since current text analysis tools could not recognise, for example, that '20.5 light years' and '120 trillion miles' express the same distance. We will classify phrases syntactically (e.g., 'noun phrase', 'adjective'), and also semantically according to entity type (e.g. teaspoon, inter-planetary distance), mathematical form (e.g., fractions, decimals or percentages), and conceptual difficulty (the level at which they first appear in the Mathematics Curriculum for Schools). Syntactic classification will be assisted by automatic tools (part-of-speech taggers, parsers). From the annotated corpus we will develop a grammar for generating numerical expressions, extending the coverage of existing grammars developed for information extraction (e.g., Mikheev et al. 1998).

3.2 Construction of rules to fit the expression to the reader

Since mathematical concepts build on one another, their difficulty can be measured by the order in which they are introduced in mathematics classes, as specified in the National Mathematics Curriculum for Schools – e.g., large numbers such as ten thousand are introduced at Key Stage 1 (ages 5 to 7), simple decimals and fractions at Key Stage 2 (ages 7 to 11), more complex decimals and fractions at Key Stage 3 (ages 11 to 14), and standard index forms such as 9.9999×10^3 at Key Stage 4 (ages 14-16). We intend to use this curriculum as a model for choosing numerical expressions for people with different levels of ability.

Our technique will be similar to that of one of our past NLG systems, SkillSum, in which we derived preference rules that scored candidate expressions for their suitability in documents for people with limited literacy (Williams and Reiter, in press). The parameters used in SkillSum were ones generally considered to aid readability, such as short sentences and short common words both for objects and relations (including discourse connectives such as 'but' and 'even so' that provide rhetorical links between statements in a text). Proportion expressions will be graded according to the Key Stages at which their underlying mathematical concepts first appear in the Mathematics Curriculum, and a set of preference rules will select expressions that fit both the Key Stage and the context in the text to be generated.

The grammar and the selection algorithm will be incorporated into an NLG system (from a previous project) which generates summaries of electronic health records for cancer patients (Williams et al. 2007). These health records contain many numerical quantities which are currently rendered only as raw digits –i.e., the form in which they appear in the health record database.

3.3 Evaluation

We will evaluate the grammar by measuring the extent to which it covers the examples enumerated in our linguistic analysis. The result will be expressed as a raw percentage and also assessed qualitatively (i.e., in trading off comprehensive coverage against reliability and efficiency).

To evaluate how well the generated expressions conform to levels of mathematical ability, as represented by the Key Stages in the Mathematics Curriculum, we will ask domain experts (i.e., maths tutors) to judge the appropriateness of examples generated by our system for readers at different levels. We have contacts with maths tutors from an earlier project (SkillSum, see Williams' C. V.) and in the Open University.

4. Project management and workplan

The project will be managed within the Natural Language Generation Group at the Open University which was formed in 2005 from key members of the Information Technology Research Institute (ITRI) at the University of Brighton who established a strong research record in NLG. These members, Dr. Richard Power, Prof. Donia Scott and Dr. Paul Piwek will form an advisory panel to provide excellent support for the project.

Dr Williams is PI and research fellow for this proposal. She became a Postdoctoral Researcher in the NLG group at the Open University in February 2006 after spending two years as a Research Fellow in the NLG group at the University of Aberdeen where she also completed her PhD. She is ideally suited to carry out the research of this project since she has experience in successfully managing research projects at BT, Macquarie University and CTAD; also, she has developed a number of NLG systems. For example, systems that generated reports for people with limited literacy and numeracy for the GIRL and SkillSum projects in which experiments showed that document content and structure as well as choice of words contributed to improving the readability of NLG output. At the Open University, she has developed a system for presenting medical information to patients either as written reports, or as scripted dialogues between animated agents (Williams et al. 2007). Before Williams' PhD, her background was in industrial NLP research at BT Labs. See also her C.V.

Our workplan shown in the chart has numbered activities described beneath it.

Activity	**F**	**M**	**J**	**M**	**J**	**J**	**A**	**S**	**O**	**N**	**D**	**J**
1. Data analysis, grammar	■	■	■	■	■	■						
2. Rule construction, interface					■	■	■	■	■	■		
3. Evaluation							■				■	
4. Dissemination								■	■	■	■	■

1. **Months 1–6.** Data analysis, classification, and grammar construction.
2. **Months 5–10.** Construction of rules to tailor generation for different levels of maths ability. Interface the grammar and rules to an existing NLG system
3. **Months 7 and 11.** Evaluation of grammar and system. **Deliverable:** evaluation results.

4. **Months 8–12.** Dissemination of results. **Deliverables:** conference papers, journal article, magazine article.

The plan follows a style successful in our previous NLG project, SkillSum, except that since this is only a one-year feasibility study, there is only one cycle of development and evaluation (in SkillSum, we completed six cycles in two years).

5. Outcomes

- The generation grammar and method that we develop for choosing appropriate numerical expressions will both contribute to NLG research.
- The enumeration and classification of proportion expressions will provide a novel contribution to linguistics research.
- The evaluation will provide validation of our approach for subsequent research.
- Our algorithms could form the basis of a set of guidelines for writers.

We expect a number of future applications to stem from this research. One is a system that could help writers to express quantities, i.e. a numerical-term 'grammar checker' and help tool. Another would be generation of mathematical exercises and feedback for a maths e-learning system.

6. Timeliness and need

Theoretically, we would argue that research into expressing numerical quantities is clearly needed by the NLG community since most NLG systems generate from numerical data. Our linguistic analysis of proportion expressions also meets a clear need, since linguistic research on numerical expressions has so far been limited to consideration of cardinals, ordinals and quantifiers.

Practically, there is a strong case for the relevance of research on communicating numerical information to people with limited numeracy. Following the Moser report (1999), the U.K. Government's policy has been to increase the level of numeracy in the adult population by providing basic skills education. However, adult enrolments on basic skills courses have been low, and in 2003 another survey reported that 49% of adults were still handicapped (e.g., on the job market) by poor mathematical skills (Williams et al., 2003). Large amounts of research funding have recently been spent on improving accessibility (e.g., the PACCIT-LINK programme), but the needs of the large proportion of the population that is innumerate have been mostly overlooked. Furthermore, nobody has thought of tackling this problem from the angle of providing better explanations of numerical quantities.

In short, for numerate people the benefit will be a gain in *precision*: numerical quantities will be expressed in ways that convey the maximum amount of useful information. For innumerate people, the benefit will be a gain in *accessibility*, allowing them some understanding of figures that would have been previously incomprehensible – perhaps accompanied by some educational side-effects.

7. Benefits

This study is important because it illuminates two issues of practical as well as theoretical interest: how numerical expressions vary in different contexts, and how to express them when communicating with different audiences. It should benefit human authors as well as researchers who build automatic document generation systems and linguists who study the meaning and structure of mathematical language, i.e.:

- The NLG and Linguistics research communities.
- Authors who communicate numerical quantities through writing.
- People who read communications containing numerical quantities.