Informing Flexible Abbreviation Expansion for Users with Motor Disabilities

Tim Willis[1], Helen Pain[1], Shari Trewin[2], Stephen Clark[1]

[1] Division of Informatics, University of Edinburgh, 2, Buccleuch Place, Edinburgh EH8 9LW, UK +44 6504415

{twillis, helen, stephenc}@cogsci.ed.ac.uk

www.cogsci.ed.ac.uk/~twillis

[2] IBM TJ Watson Research Center, P.O. Box 218, Yorktown Heights, N.Y. 10598, U.S.A. trewin@us.ibm.com

Abstract. In this paper we present results and conclusions of an initial study, intended to inform the design of a system to automatically reconstruct full sentences from abbreviated typed text. Such a system will reduce the number of keystrokes that cognitively-unimpaired users with motor disabilities need to input in order to communicate in full, rich language. A number of techniques for abbreviating messages have been identified, based on empirical studies of users. These include vowel deletion, phonetic replacement, and word truncation. How such techniques might be used to support users with motor disabilities is considered and further empirical work proposed.

1 Introduction

Communication for some people with motor disabilities can be tiring and time-consuming. For users of 'standard' QWERTY keyboards, it may be very useful to reduce the number of letters they must type in order to produce a message. Flexible Abbreviation Expansion is one approach. In order to inform the design of a system to automatically reconstruct full sentences from abbreviated typed text, experiments are being run to investigate the techniques which people use to shorten a message. Analysis of the empirical data, it is hoped, will reveal a number of high-level strategies people use to generate abbreviations, and also provide simple counts for the relative frequencies with which these are applied. These strategies will be implemented in an algorithm which can generate and rank candidate full form sentences.

The target user group is primarily users with adult-onset conditions and disabilities, including Motor Neurone Disease, spinal injury, and Muscular Dystrophy, who are literate, very familiar with typing and would prefer not to learn a symbol system, and who desire or need an alternative to speech recognition.

The next section considers some existing related systems, and the abbreviation strategies people use. Section 3 describes the preliminary experiments designed to

elicit such strategies, and section 4 gives some results and analysis. Section 5 offers some conclusions and describes how we plan to build upon these experiments.

2 Background

There are a number of Augmentative and Alternative Communication (AAC) systems which will expand an input of a few letters or symbols into anything from basic greetings and expressions of need[1], through to fully-formed essays on any topic.

2.1 Word Prediction

'Word Prediction' systems allow the user to input a few words or letters, and be offered best guesses as to the intended 'full form' of the utterance. At the letter level, *I like fo* might trigger an offer of the noun phrases '*food*', '*football*', '*foreign travel*' etc., ideally based upon a well-informed background knowledge of the user. At the word / morpheme level, *I go town PAST* might generate an offered full form of 'I went into the town yesterday' [2], [3].

This process is limited by its requirement for constant feedback, and at the start of a sentence there is little information from which to guess what is coming. Strategies include the use of syntactic, semantic, case, and selectional constraints, for instance used in the COMPANSION system [4],[5],[6],[7].

2.2 Abbreviation Expansion

One way of reducing the number of keypresses needed to generate a message is to take an abbreviation and expand it automatically. So an input of *Jn wk scl yrdy* might expand to *John walked to school yesterday*.

'Flexible' abbreviation expansion systems do not require the user to memorise a fixed set of abbreviations, and allow them to invent new forms spontaneously, as there is no longer a one-to-one correspondence between an abbreviation and a full form. The system itself interprets them, in the context of a sentence, and may offer alternative interpretations where it cannot resolve the ambiguity.

SMS (Short Message Service) text usage via mobile phones has exploded in the UK in the last two or three years. Some very basic word prediction is available, but the task is still quite laborious on most phones as a primarily numeric keypad means some letters take 3 or 4 repeated presses to obtain. This, combined with a limit of 160 characters per message, has led to a number of abbreviation conventions, many of them originally used on the internet, which has similarly exploded in use in less than a decade. Techniques include standardised abbreviations for words or phrases (e.g. ASAP – "as soon as possible"), omission of words – usually function ones, pho-

netic replacement (U for ‘you’), and a generally fragmentary nature to utterances. These techniques have also been observed in this study.

2.3 Abbreviation strategies

We hypothesise that if ambiguity can be sufficiently resolved by automatic means, the physical and cognitive effort required from the user will be less than for both word prediction and fixed abbreviation schemes. A clearer understanding of how people abbreviate may reduce such ambiguity. The following example demonstrates some ways in which an abbreviated string might be expanded.

		I	*wt*	*2*	*c*	*th*	*nw*	*fm*	
=>	[1]	I	want	to	collect	the	new	film	
	[2]	I	went	to	see	the	new	film	
	[3]	I	want	to	contact	the	new	firm	…

The intention of this research is to identify the techniques which users use to abbreviate, and to reverse-engineer the processes which have produced the input sequence of reduced forms. By obtaining the relative probabilities for each technique, it is hoped that candidate original forms for each abbreviated word might be accurately ranked.

Some of the questions addressed in this study were: Are some letters, or classes of letters, deleted more frequently than others? What are the contexts in which such deletions take place? What are common types of substitutions? What kind of information is used in choosing abbreviations, e.g. phonetic information? Syntactic information? Other types? What role does the initial length of the word play?

3 Experiments

An initial experiment was run on 10 able-bodied, frequent computer users, around 18-25 years old, educated to at least university entrance level. A small number of users with physical disabilities in the target user group were available, but not drawn upon, so as not to over-use them. It was assumed that the target users had at a previous time been very similar to the subjects, in their computer usage.

3.1 Materials

Four texts were created, each of 500 letters. There were no numbers (digits). Two unrelated topics were written about: learning to ski (texts S1 and S2) and Loughborough, a university town in the English Midlands (texts L1 and L2). Subjects were given one text to abbreviate, and another to reconstruct. As far as possible the text to

reconstruct was from another subject, and a different topic to the text to be abbreviated, so that they would not become acclimatised to the vocabulary.

For each topic, both a grammatically 'simple' (L2 and S2) and 'complex' (S1 and L1) form was written, so it could be determined whether syntactic complexity had an effect on how it was abbreviated. The 'complex' forms were unrestricted in use of subordinate clauses, co-ordination, and the location of adverbial phrases.

The 'simple' texts as far as possible contained only active, declarative, single-clause sentences. All adjuncts occurred after the verb. The intention was to convey the same information in both versions, which necessitated an occasional infinitive clause or gerund.

All four texts were fairly formal, and were written by the first author, who was also the experimenter. No external confirmation of complexity was carried out.

3.2 Procedure

Subjects were randomly allocated one of the four 500-letter texts, and asked to perform four abbreviation tasks on it, followed by a reconstruction task.

The subjects were given the following imaginary scenario: that they needed to send their text down an expensive telegraph line, such that the person at the other end would be able to reconstruct it accurately. Each letter they omit would save money, whilst spaces and punctuation would be free. There would be a heavy charge, equivalent to ten letters' saving, however, for any words that were either missing, or incorrectly reconstructed at the other end.

Subjects were asked to progressively delete letters, as follows: In their first task, down to 300; in their second, to 260; in the third, to the absolute minimum number of letters, which they thought would be intelligible to another person; and in the fourth to the minimum number of letters which, the subject judged, would allow him- or herself to reconstruct the text at some distant time.

Their final task was to try and recreate either a different text that had been abbreviated by another subject, or their own[1]. Either way, the shortened text was from the third task. For each of the first four tasks, the subject could see a full, untouched version of the text for reference. Subjects were given the following rules: They may delete individual letters, and replace one or more letters with others (upper or lower case) or with numbers. Non alpha-numeric characters were not allowed. They could delete an entire word, if they were sure it could be reinserted by a reader, and were asked not just to paraphrase. They were given the example shown in Fig. 1:

---

[1] It was intended that subjects would reconstruct each other's texts, and to explore whether subjects had more difficulty with unrelated texts, but this proved difficult to arrange, so that half used their own task three texts, with a delay of at least an hour. Of the three subjects whose Task 5 texts were significantly shorter than the originals (missing over 5% of letters), 2 had worked on unrelated text, and the third had worked on their own abbreviations.

I	*will*	*see*	*you*	*later*	might become
I	*wil*	*C*	*U*	*l8r*	

- *losing an 'l' from 'will'*
- replacing 'see' with 'C', and 'you' with 'U'
- replacing the middle three letters of 'later' with an '8' because it sounds similar.

The number of letters has been reduced from 16 to 9.

Fig. 1. The example abbreviation given to subjects

Subjects were asked to fill in a 'notes' document as they progressed, in which they would outline the techniques they used, and their order, with an indication of the time taken per task. They were asked for an indication of their use of SMS texting on a scale 1-10.

4 Results and Analysis

Most of the subjects deviated from their instructions to some extent, but useful information could still be extracted. Despite being asked to abbreviate in such a way that it would be possible for another person to reconstruct the text 100% faithfully, several subjects paraphrased whole phrases they felt were comparatively superfluous, which made reconstruction very difficult.

Each original word in the text was considered along with its phonetic form and the forms generated or untouched by that subject in each of the four tasks. Each form was compared with the input to see what letters had been deleted, added and substituted, and the possible intended phonetic representations used in each case were noted. Also any other hypothesis in relation to the particular form was noted (e.g. use of a digit to replace a phoneme). Context information was also noted (e.g. word initial v. word final replacements; vowels v. consonants; word length, function word v. content word).

A set of rules were developed, and where two or more could be said to apply, the one judged to be more specific was noted. Techniques were considered both for each participant across all words, and for each word across all participants. Table 1 gives a number of the rules we elicited from the data.

Table 1. Some of the rules elicited

Deletion rule	**Context**	**Example**
e → ø	where it is 'magic 'e''	make → mak
e → ø	final morpheme, either directly preceding or directly following 'r'	centre → centr
double letter → single	anywhere	tall → tal
g → ø	after 'in', i.e. a gerund 'ing' becoming 'in'.	skiing → skiin
ing → ø	deletion of whole gerund morpheme	skiing →ski
final x of y letters → ø	word final	distance → dist
any letter → ø	where it is phonetically silent	cupboard → cuboard
vowel → ø	no other rule already applying	tip → tp
[ə] → ø (vowel deletion)	unstressed vowel (Phon.)	simile → simle
Substitution rule	**Context**	**Example**
ch / qu → k	anywhere it is phonetically a [k]	technique → teknike
er / re → a	word-final	scorcher → scorcha
ce → s	anywhere it is phonetically an [s]	face → fas
ph → f	anywhere it is phonetically [f], i.e. not overlapping syllable boundaries in e.g. 'haphazard'	phone → fone
letters → digit	anywhere phonetically similar	won → 1

Non-standard deletions and replacements, which could not easily be assigned to a rule were also recorded as such. Particularly useful examples were highlighted where only one of two possible rules was used, and the other was not, despite being still an option. One such was the word 'shopping', which was treated in a variety of ways, listed in Table 2.

Table 2. Abbreviation forms produced by six subjects for the word 'shopping'

Subject	Task 1	Task 2	Task 3	Task 4
D	shpng	shpn	shpn	shpn
E	shoping	shop	shop	shop
P	shop	shp	shp	shp
A	shopping	shop	shppng	[MISSING]
F	shpin	shpn	shpn	spn
M	shop	shop	shop	shop

1. Subject D first used double-to-single-letter, plus deletion of all vowels. This was followed by deletion of 'g' of 'ing', even though the full 'ing' was no longer there.

2. Subject E first used double-to-single-letter, then deleted 'ing'. This can be seen either as end-of-word-deletion, or hypothesised as a more specialised rule whereby the gerund morpheme is deleted. 'ing-deletion' was used by a number of subjects for different words.

3. Subject P used double-to-single-letter, and ing-deletion, followed by vowel-deletion.

4. Subject A left the word intact for task 1, then used double-to-single-letter, and ing-deletion for task 2. In the only instance of reinsertion with this word, the subject added 'png', for task 3. This last form is lacking only the vowels.

5. Subject F used double-to-single-letter , deleted the first vowel, and used ing-to-in. This was followed with deletion of the remaining vowel. In task 4 the 'h' is deleted, which of the four letters is arguably the least important phonetically, though it does modify the 's'.

6. Subject M (the only non-native speaker) used double-to-single-letter, and ing-deletion.

Four of the six subjects kept the same form for tasks 2 to 4. All subjects preserved the letters 'spn' throughout, and 'h' was only dropped once. The most popular early technique was double-to-single-letter (5 out of 6). 4 out of 6 keep some vestige of the 'ing' in task 1, and a similar number the first vowel.

Overall, in Task 3, both versions of the skiing text, and the informal version of the other text (L2) were reduced to an average of 270 characters (55% of the original text). However, the formal text L2 was reduced to an average of about 310 words (62%), which at least somewhat supported the hypothesis that text with more syntactic complexity would prove harder to abbreviate.

5 Conclusions and Further Experiments

We have drawn the following conclusions:

1. Vowels are deleted far more readily than consonants.
2. Letters at the beginning of words are much more keenly preserved than in the middle and end. The first letter is always preserved unless there is a phonetically influential 2nd letter (e.g. 'eNjoy', 'eXist'), when it is very likely to be deleted.
3. People will happily truncate words, reducing a word by up to half whilst preserving the front.
4. Phonetic shorthands are used readily. For instance, 4 out of 6 subjects replaced 'ch' in 'technical' with a k.
5. Users are idiosyncratic and generally consistent in their preferences for rule weightings (what order they are likely to choose rules in) so there is clearly great potential for adapting an expansion system to individuals.

Our pilot study established the presence of several commonly used abbreviation techniques. For a particular word, several of these techniques may apply, and the application of one may affect the applicability of others.

At present, there is only patchy indication (from some participants' notes files) of which letters were easily discarded within the first 40% pass (etc.), and which only reluctantly and in the very last few letters of the chunk. In order to assess whether one word is a more likely expansion than another for a given abbreviation, more detailed information about the priorities given to the observed techniques is needed. Future experiments will investigate the relative importance of the observed 'rules' by requiring participants to make three passes over a set of words. At each pass, the task is to make abbreviations that are:

1) Natural: the abbreviation is considered to be obvious
2) Short: the abbreviation is a significant reduction in the word size
3) Minimal: the shortest possible abbreviation considered comprehensible

Context will play a role in participants' decisions, both at the letter-level, and the word level (syntactic and semantic). The latter will be investigated by presenting the candidate words both with and without a full sentence context.

We hope to obtain relative frequencies of rule applications in the next experiments, which can be used as part of a probabilistic component in an automatic abbreviation expansion system.

References

1. Alm, N., Newell, A., Arnott, J. A Communication Aid which Models Conversational Patterns. In R. Steele & W. Gerrey (Eds.), Proceedings of the Tenth Annual Conference on Rehabilitation Technology Washington, DC: RESNA (1987) 127-129

2. Nisbett, P., Millar, S. CALL Centre, Edinburgh Accelerated Writing for People with Disabilities. (1993)
3. Klund, J., Novak, M., If Word Prediction Can Help, Which Program Do You Choose?, available from web server, trace.wisc.edu
4. Stum, G. M., Demasco, P. W., McCoy, K. F Automatic Abbreviation Generation. In J. J. Presperin (Ed.), Proceedings of the Fourteenth Annual RESNA Conference, Washington, D.C.:RESNA Press (1991) 97-99
5. Stum, G. M., Demasco, P. Flexible Abbreviation Expansion. In J. J. Presperin (Ed.), Proceedings of the RESNA International '92 Conference Washington, D.C.: RESNA Press. (1992) 371-373
6. McCoy, K. F., Demasco, P. W., Jones, M. A., Pennington, C. A., Vanderheyden, P. B., Zickus, W. M. A Communication Tool for People with Disabilities: Lexical Semantics for Filling in the Pieces. In Proceedings of the ASSETS '94 Marina del Ray, CA. (1994)
7. Demasco, P. W., McCoy, K. F. Generating Text from Compressed Input: An Intelligent Interface for People with Severe Motor Impairments. Communications of the ACM, 35(5), (1997) 68-78.

The IPA (International Phonetic Alphabet) font used in this paper was obtained from the following archive at Oxford University;- http://info.ox.ac.uk/departments/humanities/chcdown.html

Acknowledgements

The first author is funded by The Engineering and Physical Sciences Research Council (**EPSRC**), grant number 99303114.