COMPRESSION OF ENGLISH TEXT

A. ROSS ECKLER
Morristown, New Jersey

Information theorists maintain that typical English-language text is approximately 75 per cent predictable; for example, in a text-reconstruction experiment carried out by Claude Shannon, a subject presented with various incomplete sentences was able to guess the next letter correctly in 79 out of 102 attempts. However, if one attempts to retain the sense of a sentence after weeding out letters according to some predetermined rule (such as all the vowels), one cannot achieve such an extreme limit; in fact, the removal of even half the text is likely to introduce severe difficulties. Although Shannon cites an experiment in which six subjects restored an average 93 per cent of the 50-per-cent-deleted FOOTSTRINGTHEFT, surely the triteness of the phrase made it more easily recognizable than unfamiliar text would have been. It is the purpose of this article to shed light on the question of just how much can be trimmed from text without losing the meaning.

Of course this has been attempted before, most frequently by speedwriting systems designed to aid the part-time stenographer who does not want to take the time to master a non-literal shorthand system such as Pitman or Gregg. Most of these systems are tailored to a "business English" vocabulary and assume that the stenographer is the only reader of the compressed text; they are often rather cryptic in appearance to the untrained reader. The rules of compression are elaborate, involving both literal and phonetic properties; furthermore, 100 or more high-frequency words are encoded by single letters, making guessing especially hard for the outsider if the word is long (1 for 'letter', s for 'sincerely', p for 'price'). In short, if the reader of compressed text is not the writer of it as well, he has a better chance of recovering the original if he knows the rules used for compression, so that he can if necessary reconstruct the set of words corresponding to a particular shortened form.

In this article, I propose three sets of literal (not phonetic) rules for compressing words -- low, medium and high. These rules have been kept reasonably simple, but try to leave the most informative letters of each word in view. Of course, there is a great deal of ambiguity present in individual shortened forms (the letter t can stand for 'the', 'to', 'it' and 'at' in high-compression text, for example), but often this can be overcome by considering the sentence as a whole. Following the rules and a discussion of their characteristics, I give various examples of compressed text for the reader to try deciphering.
Low-compression Rule

1. Create a reduced word by replacing doubled letters with single ones (as 'bookkeeper' to 'bokeper')
2. Compress the reduced word as follows:
   - put down the first letter (b)
   - put down consonants in order of appearance (bk, bkp, bkpr)
   - insert vowels in order of appearance (bokpr, bokepr, bokeper)
3. Write down reduced words of five letters or less without compression; compress reduced words of six letters or more to five

For instance, 'we should elect only able representatives' becomes 'we shld elc onl abl rprsn', a reduction of 36 letters to 25.

Medium-compression Rule

1. Same as above
2. Same as above
3. If the reduced word is i letters long, compress it to n letters

The above-mentioned sentence becomes 'we shld elc onl abl rprsn', a reduction of 36 letters to 20.

High-compression Rule

1. Create a reduced word by replacing doubled letters with single ones and omitting h after t (as 'three' to 'tre')
2. Same as before
3. If the reduced word is i letters long, compress it to n letters

The above-mentioned sentence becomes 'w shl elc on ab rprsn', a reduction of 36 letters to 16.

How much compression of text does each rule achieve on the average? Using the Kucera and Francis corpus of a million words taken from English-language writings of 1961, one can easily calculate the average reduction due to steps 2 and 3: low, medium, and high compression reduces text to 0.79, 0.64 and 0.52 of its original length. It is somewhat harder to estimate the additional compression achieved by eliminating doubled letters and h following t, but the effect is relatively small. For example, there are about 84,000 occurrences of h following t in 3-letter, 5-letter, 8-letter and 10-letter words occurring 100 or more times in Kucera and Francis. Since words occurring more than 100 times use up nearly 70 per cent of the total corpus of 4.7 million alphabetic letters, one can use a simple extrapolation to predict a total of (10/7)(84000) = 120,000 occurrences of h following t in the entire corpus -- a leap for doubled letters per cent. The idea was to shorten the word, 69,971 better distinct 'them', 'then' common words.

Obviously, The most baffling, 8: letter, and I say:

a: a
b: be, bee, b
- do, odd
f: of, if, off,
g: go, egg

Among two-letter culos, represe: true, three, the turn -- all occur highly-ambiguous fire, farm, fa, step, still, stat sit, stuff, sto...

The following, medium: 100 or more times in Kucera and Francis. Since words occurring 100 or more times use up nearly 70 per cent of the total corpus of 4.7 million alphabetic letters, one can use a simple extrapolation to predict a total of (10/7)\(\times\)84000 = 120,000 occurrences of h following t in the entire...
corpus -- a letter-reduction of only 2.5 per cent. A similar argument for doubled letters predicts an even smaller reduction -- perhaps 0.5 per cent. The principal reason for introducing the h-following-t rule was to shorten the word 'the' (occurring more often than any other word, 69,971 times in the corpus) to a single letter, and to permit better distinction among the two-letter versions of 'this', 'that', 'them', 'then', 'than' and 'they' -- all occurring among the 71 most common words.

Obviously, there is much ambiguity in the high-compression rule. The most baffling words are those that have been reduced to a single letter, and I summarize the common alternatives below:

```
a: a  h: he
b: be, bee, by, ebb  i: I
b: do, odd  l: all, ill
f: of, if, off, fee  m: me, my, am
r: or
r: dr, odd  n: an, in, on, no
```

Among two-letter compressed words, tr is probably the most ambiguous, representing there, their, tree, trees, tare, trap, trip, try, true, three, third, truth, trim, torn, tour, tire, threw and turn -- all occurring 20 or more times in Kucera and Francis. Other highly-ambiguous combinations include fr (for, from, far, fear, firm, fire, farm, fair, four, forth, form, ford, free, Fred), st (stem, step, still, staff, south, star, stay, stiff, steel, set, sat, seat, site, sit, stuff, stop) and ts (this, these, those).

The following paragraph is written according to the high-compression, medium-compression and low-compression rules; try and decipher each representation before moving on to the next. Most people will agree, I think, that full recovery of a message with half the letters missing is extremely difficult, but with only one quarter missing it is quite easy. A final thought: compressed text might be used as the basis of a party game, with individuals or teams competing against each other to determine the sense most quickly. (It might be combined with a treasure hunt or dictionary rally, in which a series of decoded messages give instructions to do something.)

```
A final thought: compressed text might be used as the basis of a party game, with individuals or teams competing against each other to determine the sense most quickly.
```

---

The following paragraph is written according to the high-compression, medium-compression and low-compression rules; try and decipher each representation before moving on to the next. Most people will agree, I think, that full recovery of a message with half the letters missing is extremely difficult, but with only one quarter missing it is quite easy. A final thought: compressed text might be used as the basis of a party game, with individuals or teams competing against each other to determine the sense most quickly. (It might be combined with a treasure hunt or dictionary rally, in which a series of decoded messages give instructions to do something.)

```
A final thought: compressed text might be used as the basis of a party game, with individuals or teams competing against each other to determine the sense most quickly.
```

---
observe how they do their work, do they make laws for which there is a real need? do they carry out the laws and see that the people obey them? these and many other matters should be of special concern to every citizen who wants good and honest government.

As a matter of interest, here is another paragraph, first written in Carter Brie hand (a form of speedwriting developed in 1957 by Theodore H. Carter) and then according to the high-compression rule.

Which is more understandable?

n a rprnsn dmrc, e ppl gvrn tmsslvs b ntrstng e ntr admstrs e e stat t tr rprnsnvs, hom y cos b blt. e fwrns dsdd t r gvrnn nd b a r mblic r rprnsn dmrc. y rcnznzd t a pur dmcr c i ntr pctltr n ndur. n a pur dmcr, o crs, mgn o plbl afs rmns n e hns o e ppl tmsslvs, s t y klws, lv txs, a dtrm l mtr

in a representative democracy, the people govern themselves by entrusting the entire administration of the state to their representatives, whom they choose by ballot. the founders decided that our government should be that of a republic or representative democracy. they recognized that a pure democracy is neither practical nor enduring. in a pure democracy, of course, management of public affairs remains in the hands of the people themselves, so that they make laws, levy taxes, and determine all matters.

(420 letters to 222 letters)

250 CRYPTOGRAMS WITH CLUES

David Fellman of Rochester, Minnesota writes: "I received a comfortable little booklet of 250 Cryptograms in the mail. It is cleanly printed, has ample space between lines to make solving pleasant, and totally lacks the usual confusing typos. This is not a book for the beginner. There are no solutions given, nor any instructions on how to solve. Clues given in the back of the book are for the first and last letters of each puzzle. Puzzles range in length from a few words to three lines; all are original with the author. The fun is in solving the puzzles, but occasional ones reveal a sly wit. There are no gimmicks. It is sold by mail order and can be obtained by sending five dollars to Margot Pyrc, Box 702, Woodburn, OR 97071."