Fall 2006
Chris Christensen
MAT/CSC 483

Appendix
Cryptograms

Here is a more detailed discussion of the history and techniques for solution of aristocrats and patristocrats (the generic term for them is *cryptograms*).

> *Codebreaking appears to be such a popular sport because it is literally fancy-free. If cryptography [codemaking] is a form of abstract algebra, then inventing a new cipher system is nothing more than building abstract castles in the air with material and design of one's own choosing. To make the system work is little more than to avoid self-contradiction, yet when the answer comes out right it always satisfies the inventor. Codemaking is much more popular than codebreaking because it is easier and more esthetic; it flings together shining theories however it pleases, whereas cryptanalysis forces the mind to concentrate upon the data, upon the coarse rubble of reality. But cryptanalysis is much more rewarding. For it subdues these hard and unyielding facts; it represents a victory of the mind over something, whereas codemaking represents a triumph over nothing. This mental mastery is the keen pleasure-pang of solution; it is what men of the intellectual caliber of Babbage and Wheatstone see in cryptanalysis, …*
>
> *Consequently it is not surprising to learn that those addicted to this mental enjoyment have banded together to assure themselves of it. The American Cryptogram Association was founded in 1932 by members of the National Puzzlers League who wanted to concentrate more on cryptology, taking as their motto "The cryptogram is the aristocrat of puzzles."* David Kahn, *The Codebreakers* by David Kahn, p. 768 f.

Here is information from the American Cryptogram Association website http://www.cryptogram.org/

> *The American Cryptogram Association is dedicated to promoting the hobby and art of cryptanalysis -- that is, learning to break ciphers.*

Membership benefits include six issues of The Cryptogram *each year, each of which has about 100 challenge ciphers of varying difficulty: typically 25 simple substitution with word divisions (called "Aristocrats"), 14 simple substitution without word divisions (called "Patristocrats"), 14 cryptograms (simple substitution and others) in languages other than English (called "Xenocrypts"), 26 English cryptograms of types other than simple substitution ("Cipher Exchange"), 16 Cryptarithms, and 5 or so difficult ciphers ("Analyst Corner"). The Cryptogram includes articles of general and specific interest to hobby cryptanalysts, a column devoted to computer users, another devoted to children, and general news about the ACA.*

New members also receive one copy of The ACA and YOU, *a 64-page manual that describes the procedures and ciphers used by the American Cryptogram Association.*

One of the founders of the ACA was Helen Fouché Gaines. The material that follows is adapted from her book *Cryptanalysis: A Study of Ciphers and their Solutions*. Her book (which is still available) is a standard for the study of classical ciphers. From Gaines … :

Simple Substitution

Simple substitution is ordinarily defined as a cipher in which each letter of the alphabet has one fixed substitute, and each cryptogram-symbol represents one fixed original. When this cipher is used for puzzle purposes, as we find it in our newspapers and popular magazines, the substitutes (which are invariably letters of the alphabet) may be chosen at random, and the cryptograms must follow certain arbitrary rulings which are designed to make them "fair": Word-divisions and punctuation must follow religiously those of the original text; a certain minimum of length must be provided; no letter may act as its own substitute [This is very important information; the German enigma machine had a similar property that the British codebreakers at Bletchley Park were able to exploit.]; foreign words are not permissible; and so on. Aside from the observance of such rules, however, no holds are barred; the constructor of such a cryptogram, totally unconcerned with the meaning of the plaintext (except that it must have one), sometimes gives his chief attention to distorting the normal language characteristics in an effort to baffle the analyst …

[Gaines follows with tips for solving aristocrats (cryptograms for which word length and punctuation is given) and other simple substitution cryptograms. Below, is a summary of the tips that she gives in Chapter IX. The words are mostly hers, but the sentences have been rearranged (a lot) and edited (somewhat).]

Entry

No matter how resistant the cryptogram, all that is really needed is an *entry*, the identification of one word, or of three or four letters. The experienced solver knows well that persistence will find this entry, and trusts largely to instinct and perseverance; the beginner, however, may feel at a loss for a "system," and, if so, may perhaps, be able to find suggestions for one in [what follows. Short words often provide an entry. The only two one-letter words in English are `a` and `I`. The most common two-letter words are: `an, at, as, he, be, in, is, it, on, or, to, of, do, go, no, so, my.` The most frequent word in English is `the`.]

Do a frequency count

First of all, in any substitution problem, there should be a counting of the letters of the cryptogram in order to find out their frequencies. [Frequency tables might disagree as to which letter is, say "the third most frequent," but] the same nine letters `E T A O N I R S H`, will constitute the *high-frequency group* of letters. These particular letters will make up about 70% of any English text, and it is almost impossible to prepare one, no matter how short, without using them in about that proportion, though in the shorter texts, `L` and `D` will sometimes creep up into the high-frequency class, taking the place of `H`. Aristocrats are arbitrarily confined to lengths which run between 75 and 100 letters. Even without manipulation, a text of this length will not always show `E` as a frequent letter, and may, for some reason, show `Z` or `X` with a fairly high frequency.

"Class distinctions" among the letters are always, to some extent, dependable. High-frequency letters, moderate-frequency letters, and low-frequency letters, all tend to be very exclusive. They will exchange frequencies with letters of their own class, but all three classes are disinclined to welcome outsiders. The vowels, also have their fraternity; if

the frequency of E is lowered, some other vowel, even U and Y, will insist upon making up the difference, rather than yield this privilege to a consonant.

The high-frequency group includes the nine letters E T A O N I R S H. Even in this exclusive club, there are cliques – not ironclad, but clearly noticeable:

> *Class I*. The letters T O S appear frequently *both as initial letters and final letters* in their own words, with terminal O confined largely to short words. All three of these are freely doubled.
>
> *Class II*. The letters A I H appear frequently as initial letters, but far less frequently as finals, especially A I. Not one of these is readily doubled.
>
> *Class III*. The letters E N R appear frequently as final letters, but far less frequently as initials. The letter E is very freely doubled; the other two not so often.

When one of these letters changes its class, the least likely exchange is one occurring between classes II and III.

The following lists of frequent initial and final letters are taken from Abraham Sinkov's *Elementary Cryptanalysis: A Mathematical Approach* .

Frequencies of initial letters

```
t, a, s, o, i, c, w, p, b, f, h, m, r, d, e, n,
l, g, u, y, v, j, k, q, x, z
```

Frequencies of final letters

```
e, s, d, n, t, r, y, o, f, l, a, g, h, m, w, k,
c, p, i, x, u, b, v, j, z, q
```

We return to Gaines … .

Other points

In words of three and five letters, the central one is nearly always a vowel, taking it for granted that the words `the` and `and` will never present any difficult cryptogram. In the longer words, the favorite positions of the vowels are the two positions which follow the initial letter and the two positions which precede the final letter. The favorite position of `I`, in fact, is well-known as the third-to-last. About half the words used in any written text are of the type called *negative*, or *empty*; that is, the pronouns and auxiliary verbs, and particularly the various kinds of connectives *without which no sentence can be put together*. If your cryptogram is an "aristocrat," you will probably find that most of the prepositions begin with `A`. Every sentence contains a verb, and these are more or less limited in their possible terminations. Any letter used only two or three times, and always followed by the same letter, is good material for `Q`.

Contact letters

[A more sophisticated analysis involves determining contact between letters. When dealing with a patristocrat (word length is not given), a list is made showing every letter with the two which have flanked it right and left each time it was used. When word length is known, contacts between letters from different words are not listed.]

1. The vowels `A E I O` are normally found in the high-frequency section of the frequency count; the vowel `U` in the section of moderate frequencies, and the vowel `Y` in the low-frequency section.

2. Letters contacting low-frequency letters are usually vowels.

3. Letters showing a wide variety in their contact-letters are usually vowels.

4. In repeated digrams, one letter is usually a vowel.

5. In reversed digrams, one letter is usually a vowel.

6. Doubled consonants are usually flanked by vowels, and vice versa.

7. It is unusual to find more than five consonants in succession.

8. Vowels do not often contact one another. If the letter of highest frequency can be assumed as E, any other high-frequency letter which never touches E at all is practically sure to be another vowel, and one which contacts it very often cannot be a vowel.

Establishing the vowels

The most frequent vowel is E. The one which never touches it is most likely to be O. Both of these are very freely doubled, and for that reason are often confused with each other, but seldom with any other vowel. They rarely touch each other.

The vowel which follows E and almost never precedes it, is A.

The vowel which reverses with it is I.

The same observations will apply to the vowel O; but a distinction occurs when the vowel U can be found; this vowel precedes E and follows O.

The only vowel-vowel digrams of any real frequency are OU, EA, IO.

Three vowels found in succession may represent IOU, EOU, UOU, EAU.

Identifying the consonants

Those letters remaining in the high-frequency section of the frequency count will usually include T N R S H. Of these, the most easily recognized is H, which precedes all vowels and seldom follows one; it may be identified often as part of the repeated sequences TH, HE, HA.

Next to H, the most recognizable of the consonants, aside from frequency, is probably R, which reverses freely and indiscriminately with all vowels, and has a strong affinity for other high-frequency letters.

The consonant T can usually be identified by its frequency, by its tendency to precede vowels rather than follow them, and by its almost inevitable combination with H on more than one occasion. If is also notably difficult to distinguish from the vowels.

The letter N has characteristics which are to some extent the opposite of those mentioned for H; it prefers to follow vowels and precede consonants, and, to a lesser extent, the same is true of S, according to some charts. However, N, S, and T are all readily reversible with vowels, and are sometimes hard to tell apart.

The only frequent reversals of two consonants are ST-TS and RT-TR.

The doubles TT and SS are among the most frequent in the language.