**General Approach to breaking a substitution cipher.**

**Letter Frequency.**

If we look at English text, there are certain letters that are more frequent than others. ETAOIN SHRDLU - Typesetting letter arrangement

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

English text frequency (most to least), but different kinds of sources may have different frequencies (News, Science, Religion, etc.)

**A message to break:**

TSSTAHNEMGIOCEESTRSRIAESA

**Record the frequencies**

|  |  |
| --- | --- |
| E | xxxx |
| T | xxx |
| A | xxx |
| O | x |
| I | xx |
| N | x |
| S | xxxxx |
| R | x |
| H | x |
| L |  |
| D |  |
| C | x |
| Other | MG |

This is consistent with English… Make a guess that it is a transposition of some kind.

If it is an anagram, then we can look for words that we expect are in the message to reduce the number of letter arrangements to try.

**Other frequencies we can use:**

**First letter of words**

t o a w b c d s f m r h i y e g l n p u j k

**Second letter of words**

h o e i a u n r t

**Third letter of words**

e s a r n i

**Last letter of words**

e s t d n r y f l o g h a k m p u w

**digraphs**

th he an in er on re ed nd ha at en es of nt ea ti to io le is ou ar as de rt ve

**Trigraphs**

the and tha ent ion tio for nde has nce tis oft men

**Double Letter**

ss ee tt ff ll mm oo

**Sample substitution cipher with word breaks preserved.**

**UIJT JT OPU B TFDSFU NFTTBHF**

|  |  |
| --- | --- |
| A |  |
| B | xx |
| C |  |
| D | x |
| E |  |
| **F** | **xxxx** |
| G |  |
| H | x |
| I | x |
| J | xx |
| K |  |
| L |  |
| M |  |
| N | x |
| O | x |
| P | x |
| Q |  |
| R |  |
| S | x |
| T | xxxxx |
| U | xxx |
| V |  |
| W |  |
| X |  |
| Y |  |
| Z |  |

**Guess: ETAOINSRHDLC** map to **TFUJB**

**Guess: TT** is **SS** or **EE**

**Guess: B** is **A** or **I**

**Guess: ESTDNR** map to **TUBF**

**Try: TS**

**UIJT JT OPU B TFDSFU NFTTBHF**

**...s .s ... . s..... ..ss...**

**Guess: J** is **A** or **I. Based on first word, more likely to be I.**

**Try: JI**

**UIJT JT OPU B TFDSFU NFTTBHF**

**..is is ... . s..... ..ss...**

**And then Try: JI**

**UIJT JT OPU B TFDSFU NFTTBHF**

**..is is ... a s..... ..ssa..**

**UF are the remaining high frequency letters. U both starts and ends words.**

**And then Try: UT**

**UIJT JT OPU B TFDSFU NFTTBHF**

**t.is is ..t a s....t ..ssa..**

**TH common digraph. Try: IH**

**UIJT JT OPU B TFDSFU NFTTBHF**

**this is ..t a s....t ..ssa..**

**Based on the last work, F is likely a vowel and e is common end.**

**UIJT JT OPU B TFDSFU NFTTBHF**

**this is ..t a se..et .essa.e**

**P is likely to be a vowel and o/u are left. We try letters and see if we can put together a message that makes sense.**

Table to record frequencies.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| A |  |  |  |  |  |  |  |
| B |  |  |  |  |  |  |  |
| C |  |  |  |  |  |  |  |
| D |  |  |  |  |  |  |  |
| E |  |  |  |  |  |  |  |
| **F** |  |  |  |  |  |  |  |
| G |  |  |  |  |  |  |  |
| H |  |  |  |  |  |  |  |
| I |  |  |  |  |  |  |  |
| J |  |  |  |  |  |  |  |
| K |  |  |  |  |  |  |  |
| L |  |  |  |  |  |  |  |
| M |  |  |  |  |  |  |  |
| N |  |  |  |  |  |  |  |
| O |  |  |  |  |  |  |  |
| P |  |  |  |  |  |  |  |
| Q |  |  |  |  |  |  |  |
| R |  |  |  |  |  |  |  |
| S |  |  |  |  |  |  |  |
| T |  |  |  |  |  |  |  |
| U |  |  |  |  |  |  |  |
| V |  |  |  |  |  |  |  |
| W |  |  |  |  |  |  |  |
| X |  |  |  |  |  |  |  |
| Y |  |  |  |  |  |  |  |
| Z |  |  |  |  |  |  |  |