Texting, Texting, One Two Three (20 marks)

The respected espionage-supply company Z Enterprises is about to release a new version of their Z1200 model wristwatch, popular among spies (and also high-school students) for its ability to discreetly send text messages. Although the Z1200 had only four buttons in total, the user could input characters by pressing three-button sequences. For example, if we call the buttons 1, 2, 3, and 4, a was 112, A was 113, b was 114, SPACE was 111, the END sequence that finished the message was 444, etc.

The Z1300 has the same button layout, and it was planned that it would use the same text-input method. In the design stage, however, a new engineer proposed that he could significantly reduce the number of button presses needed for each message. Unfortunately, the manual had already been printed and the new Z1300 shipped without any information regarding how to use this new input method. But if you were a spy, you could work out how it works just from a few examples – couldn't you? Here are the examples:

Testing testing
332221432241423411222143224142341331

Does anyone copy
3323332214313142343324221124232342343331

be vewy vewy qwiet im hunting wabbits
23412112342213443431234221344343123442344412122141243123124142224142341134443123412341412243331

Mission failed Tango not eliminated
332434143434132421244314123221233133223142341321423222121232412434142312221233331

my boss Z is a pain in the
24334312341324343133234441414313113423141421414212223121331

uh oh no backspace on this thing
24123113223114232123413124223434334231242211324212223141431222314142341331

just kiddin boss
2344324143221234341233233414212341324343331

[The questions are on the next page.]

Questions:

7.1 (10 marks)

What are the codes for each of the lowercase letters?

7.2 (5 marks)

What message does the following sequence of button presses encode?

231212322323214143131423432341322333431232414322214241423413 31

7.3 (5 marks)

With what sequences of button presses would you input the following messages?

a. help

b. Mayday mayday SOS

Answer sheet

7.1	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	
7.2									
7.3a									
b									

Answers

7.1	a	b	c	d	e	f	g	h	i
	31	**2341**	**242**	**233**	**21**	**244**	**341**	**231**	**41**
	j	k	l	m	n	o	p	q	r
	23443	**2343**	**232**	**243**	**42**	**32**	**342**	**23442**	**[44]**
	s	t	u	v	w	x	y	z	
	43	**22**	**241**	**2342**	**344**	**[23441]**	**343**	**23444**	
7.2	**hello is anybody listening**								
7.3a	**23121232342331**								
b	**332433134323331343124331343233313431334333323343331**								

Comments by Patrick Littell

From examining repeated elements and letters, we can work out most, but not all, of the character codes for the letters, along with SPACE being 1, the SHIFT sequence that creates a capital letter being 33, and the END MESSAGE sequence being 331 (SHIFT + SPACE, a sequence that otherwise wouldn't be used).

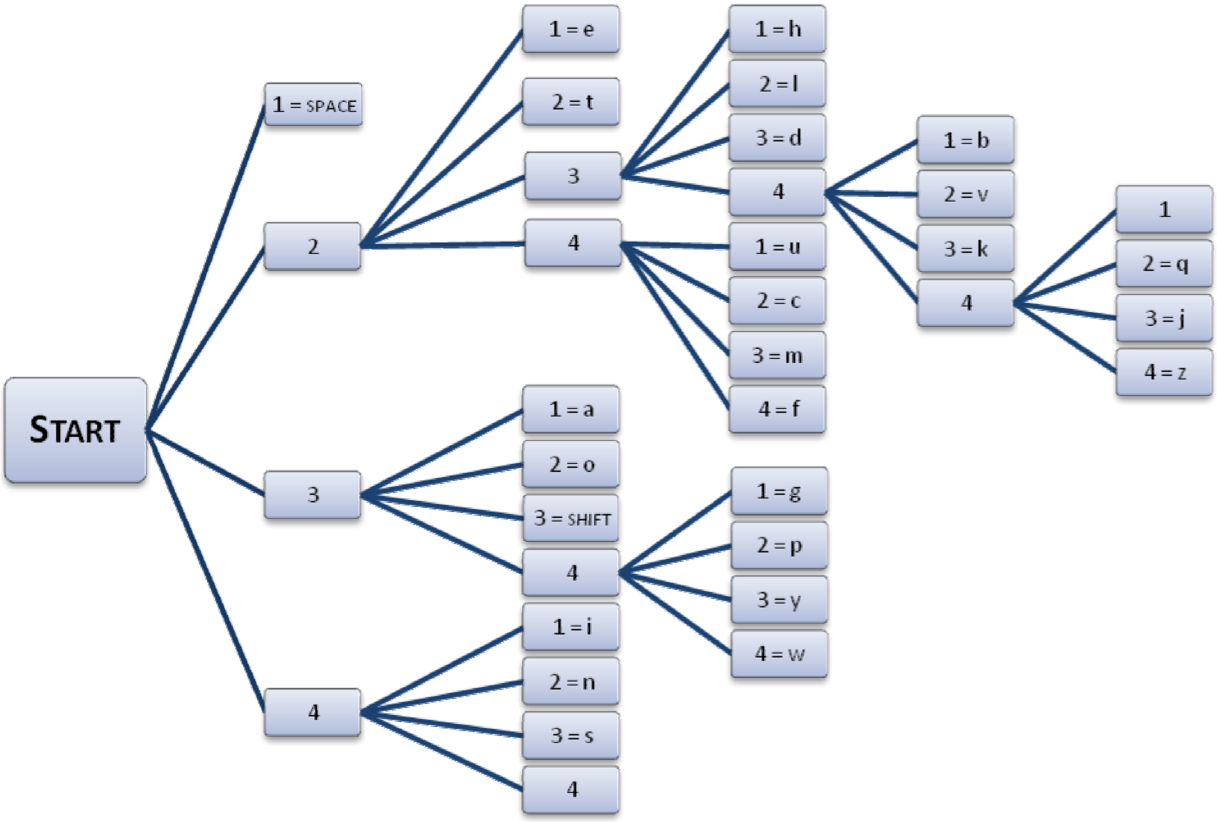
Lowercase 'z' doesn't appear in the plaintext, but knowing that uppercase 'Z' is 3323444 and "shift" is 33 we can conclude that lowercase 'z' is 23444.

The system we find is a "variable-length", rather than "fixed-length", code system. Although some of the codes are much longer than three digits, overall most codes are much shorter, because very common characters (like e, t, "space", etc.) are given very short codes whereas only fairly rare letters are given the longer codes.

a	31	**n**	42
b	2341	**o**	32
c	242	**p**	342
d	233	**q**	23442
e	21	**r**	44
f	244	**s**	43
g	341	**t**	22
h	231	**u**	241
i	41	**v**	2342
j	23443	**w**	344
k	2343	**x**	23441
l	232	**y**	343
m	243	**z**	23444

Two letters remain, however, 'r' and 'x', neither of which appear in the plaintext. To determine their values, we have to work out the overall logic of the system.
Looking at the numerical codes, we notice that they aren't random: there are frequently repeated initial sub-codes, and a lot of gaps. For example, many codes begin in 23-, 234-, and 34-, but none begin in, for example, 1-.
But why shouldn't a code begin with 1? If you consider the use of such a device, what would happen if a letter code began with one? What would happen is that, since 1 is "space", the device wouldn't know whether that 1 was intended as a space or as the first number of a longer code.
Looking further, we can see that *none* of the codes begins with another letter's code. That is, since 'a' is 31, no other letters' codes have 31- as their first two numbers, since 'b' is 2341, no other codes have these as their first four numbers, etc.
"Fixed-length" code systems, like the original three-number code system, always know when the user has keyed in a complete code. But since this system has "variable-length" codes, it needs some system to tell it whether some sequence, of whatever length, is a complete code or just the first part of a longer one. In this case, it knows when a code is complete because no beginning part of a valid code is a valid code.

It's especially clear if we draw a "tree" of the codes: only those nodes that don't have further "branches" are assigned characters. Assigning "31" to "a" is fine, because there aren't any "311", "312", etc. to confuse the system. On the other hand, we can't assign "34" to anything because then it would prevent "341", "342", etc. from being entered.



Looking carefully at our tree, there are exactly two "free" nodes – that is, ones that don't already have a character assigned and that don't have any "branches": "44" and "23441". These are where "r" and "x" have to go – if they go anywhere else, the internal logic of the system is compromised.
Since frequent letters (like "e","t","a","o","i","n","s") get short codes, and rare letters (like "q","j","z") get long codes, "r" must be "44" while "x" is "23441".
Now we have all 26 letters, SPACE, SHIFT, and the END sequence, and can encode and decode any message for this device.

Comments by Adam Hesterberg

The codes for almost all of the letters can be filled in just by guessing the division between letters in the examples. For example, in "Testing testing", there's a repeated block of "222143224142341," presumably "testing", plus the codes 33 (capitalization), 1 (space), and 331 (end). (Choosing a smaller repeated block quickly leads to contradictions). There are 15 digits for the 7 letters in "testing", so most of its letters have 2-digit codes. From there, all of the letters except r and x appear easily. R and x, however, don't appear in the given data.

The other letters' codes, sorted by code, are:

1: SPACE
21: e
22: t
231: h
232: l
233: d
2341: b
2342: v
2343: k
23442: q
23443: j
23444: z
241: u
242: c
243: m
244: f
31: a
32: o
33: SHIFT (END is SHIFT+SPACE)
341: g
342: p
343: y
344: w
41: i
42: n
43: s

In this form, it's easy to see that the codes have the property that if you start writing any sequence of 1s, 2s, 3s, and 4s, you get to one and only one code: for instance, in "23134342," "23" isn't a code, "231" is 'h', and no other code begins with "2313." There are two codes missing, though: "44" starts no code, and "23441" starts no code. Hence these are 'r' and 'x' in some order. Also, 'r' is a very common letter and 'q' is a very rare one, so writing a typical message would require fewer digits if 'r' had the shorter code than if 'x' did. Hence 'r' is 44 and 'x' is 23441.

Google "Huffman code" for more on the type of code used in this problem.