How did TJB encode B2?

Carl Hammer, Ph. D.
Director, Computer Sciences
Univac, Washington, D.C.

Invited Paper

Beale Cypher Symposium
15 April 1972
Washington, D.C.

In the context of this study, it is quite immaterial whether the Beale Cyphers are a hoax or not. The question of authorship and authenticity of the three cyphers is properly the subject of another investigation. Here we are concerned only with the question how B2 was encoded since both cleartext (B2C = "I HAVE DEPOSITED…") and keytext (DOI = Declaration of Independence, Beale Version) are known to us.

Having decided on this particular method of encoding the (B2C) cleartext by using first letters of numbered words in the keytext (DOI), how would a schooled cryptographer proceed? The answer depends on such factors as the desired coding efficiency, time available, and support materials at hand. Given much time and ample quantities of paper and pencils, an expert cryppie may want to begin by making a complete list of all available keytext letters so that he can develop a cypher with a maximum number of degrees of freedom. For example, B2C requires 42 A's but DOI can furnish 167 words starting with that letter. Therefore, our cryptographer could encode every letter A in B2C with a different number. In fact, he can make his selection in many permutative ways. On the other hand, the DOI has only two V's but to encode B2C we need eighteen of them; at best, we must use each V in DOI nine times. Finally, as is well known, there are no X's and Y's available in the DOI and our cryptographer must make some provisions to cope with this deficiency.

TJB, or whoever authored the three cyphers, has given us a clue what to do about the last question by encoding X = 994 ("Sexes") and Y = 822 ("Fundamentally"). We observe casually that a better pun for the letter choice would have been the word "by" which occurs sufficiently often at position numbers 90, 221, 577, 682, 706, 755, 852, 957, 1020, 1030, 1069, 1106, and 1198. Naturally, we would not want to use these elements for B but that would not cause any

problems; DOI provides us with 48 B-elements but B2C requires only twelve of them. Thus, we would still have an abundance of B's, even if we were to use up nine of them for the punned Y's.

Evidently, the author of B2 was not at all that efficient. Additionally, he chose most of his elements from the lower numbered elements and repeated them rather unnecessarily. For example, instead of using a differently numbered element for each occurrence of the letter A, he chose a subset of 15 from the whole set of 167 available elements, repeating some of them as often as five times. An analysis of the code elements used indicates that a majority of them come from the beginning of the DOI: 407 below 100; 574 below 200; 637 below 300; 673 below 400; 690 below 500; 718 below 600; 732 below 700, none used between 700 and 800; and 818, 882, 994 used for V, Y, and X, respectively.

A modern cryptographer - or even an experienced one of the early nineteenth century - desiring to encode B2C with the DOI as keytext in a most efficient manner would, therefore, have gone through the following steps:

(1) Develop a masterplan for this project.

(2) Prepare a listing of required letter frequencies for the B2C cleartext.

(3) Prepare a listing of available letter frequencies in the DOI keytext.

(4) Analyze the letter frequency ratios (see table I) for available to required letters and decide which course of action to follow in four cases:

(4.1) More elements available than required for certain type letters: Develop a strategy of random selection without duplication;

(4.2) The same number of elements is available as required for certain type letters: Develop a more tightly controlled strategy for random selection, still without duplication;

(4.3) Fewer elements available that required for certain type letters in the cleartext: Develop a strategy for randomly repeating the use of the available letters, with a uniform distribution for repeated usages;

(4.4) No elements available for some letters required (examples: X, Y): Develop a clever substitution strategy so as to minimize repeated usages (examples: X = Sexes, Y = By).

Obviously the author of B2 did very little in this direction. Rather, we may suspect that he "learned by doing".

TABLE I

Letter Frequency Distributions

λ	DOI n(λ)	DOI f(λ)	B2C n(λ)	B2C f(λ)	E f(λ)	$R\left\{\frac{[n(\lambda)_{DOI}]}{[n(\lambda)_{B2C}]}\right\}$
A	167	1267	43	.0564	.0781	3.9
B	48	363	11	144	.0128	4.4
C	53	401	17	223	.0293	3.1
D	37	280	49	642	.0411	0.8
E	36	273	105	1376	.1305	0.3
F	64	484	21	275	.0288	3.0
G	19	144	15	197	.0139	1.3
H	80	606	37	485	.0585	2.2
I	68	515	55	721	.0677	1.2
J	10	76	2	26	.0023	5.0
K	4	30	1	13	.0042	4.0
L	34	257	32	419	.0360	1.1
M	29	220	6	79	.0262	4.8
N	19	144	69	904	.0728	0.3
O	144	1075	63	826	.0821	2.3
P	63	477	12	157	.0215	5.2
Q	1	8	-	-	.0014	∞
R	40	303	38	498	.0664	1.1
S	64	484	48	629	.0646	1.3
T	254	1923	70	919	.0902	3.6
U	28	212	25	328	.0277	1.1
V	2	15	18	236	.0100	0.1
W	59	447	13	170	.0149	4.5
X	-	-	4	52	.0030	0.0
Y	-	-	9	118	.0151	0.0
Z	-	-	-	-	.0009	-
	1323	1.0001*	763	.9999*	1.0000	

* Due to roundoffs not exactly 1.0000

Notes: DOI Refers to the first letters of the 1323 words in the Beale version of the Declaration of Independence.

B2C Refers to the cleartext obtained by decoding Beale Cypher 2 with the Declaration of Independence as keytext.

E Refers to standard English letter frequency.

We propose, herewith, the following hypothesis: B2 was encoded in a grossly suboptimal manner, because the author did not care to produce a most efficient code and was probably not even familiar with such a concept. Rather, he set up a limited list of available elements for the first several hundred words of the DOI and added to this list, in spurts, as he proceeded with his encoding task. He also noted quickly (i) the paucity of V's in the DOI and he decided to use consistently the "distant" element 818, (ii) the absence of X's and he decided to "pun" it by using "distant" element 994, and (iii) the absence of Y's and he decided to use "distant" element 822. All these "distant elements" are far beyond the limited range of available elements he had "planned" to use for the balance of his encoding job.

What evidence do we have in support of this hypothesis? First, B2 contains 179 different numerical elements (after removing repetitions), 80 of which are below 100, another 36 between 100 and 200, and only 63 are above 200. As we pointed out earlier, most of the repeated elements (75 percent, to be exact) are below 200. Details are shown in Table II which gives the order in which the author picked "new" elements (obviously, the first elements in every row is always "new") to encode his cleartext. For example, to encode "IHAVEDEPOSITED…" he chose I =115 (rather than I =2) for his very first element; but he gets around to using I =2, 8, 140, 154, 158, 657 eventually. Yet he fails to employ I = 139, 151, 165, 167, and many others which occur "early" in his numbered keytext.

TABLE II

Order of Elements by First Occurrence in B2

	In Order of Appearance																	Sorted															Counts		
																																		E	O	T
A	24	36	28	177	45	81	98	51	284	150	27	229	83	25	152			24	25	27	28	36	45	51	81	83	98	147	150	152	229	284			7	8	15
B	308	9	77	18	134	495	193											9	18	77	134	193	308	495											3	4	7
C	84	65	92	4	94	199	21											4	21	65	84	92	94	199											3	4	7
D	52	15	210	118	63	252	135	246	320	406	591							15	52	63	118	135	210	246	252	320	406	591							7	4	11
E	37	49	7	79	85	138	190	629	496	520	557	612	584	33				7	33	37	49	79	85	138	190	496	520	557	584	612	629				6	7	14
F	195	159	122	273	131	360	676	11										11	122	131	159	195	273	360	676										3	5	8
G	270	48	113	133														48	113	133	270														2	2	4
H	73	107	394	6	20	301	204	466										6	20	73	107	204	301	394	466										5	3	8
I	115	657	140	2	8	154	314	158	67									2	8	67	115	140	154	158	314	657									6	3	9
J	120	590																120	590																2	-	2
K	305																	305																	-	1	1
L	42	101	102	233	400	157	196	420	176	405								42	101	102	157	176	196	233	400	405	420								6	4	10
M	58	82	117	207														58	82	117	207														2	2	4
N	47	10	287	353	616	549	44	566										10	44	47	287	353	549	566	616										4	4	8
O	31	56	5	136	46	106	12	43	57	125	143	302						5	12	31	43	46	56	57	106	125	136	143	302						6	6	12
P	17	105	30	121														17	30	105	121														1	3	4
Q																																			-	-	
R	59	53	96	219	248	344	112											53	59	96	112	219	248	344											4	3	7
S	62	35	71	78	110	38	216	515	609	297	242	285	275					35	38	62	71	78	110	216	242	275	285	297	515	609					6	7	13
T	22	29	26	543	3	41	16	34	60	61	14	50	32	64	39	653	288	3	14	16	22	26	29	32	34	39	41	50	60	61	64	288	543	653	10	7	17
U	238	316	95	250	371	388	409	440										95	238	250	316	371	388	409	440										5	3	8
V	818																	818																	1	-	1
W	72	290	19	66	40	1	459											1	19	40	66	72	290	459											4	3	7
X	994																	994																	1	-	1
Y	882																	882																	1	-	1
Z																																					
																																			95	84	179

NOTE: E = Even Counts, O = Odd Counts, T = Total Counts

Table III analyses the pattern he employs to introduce new numbers against repetition of elements already used. The first four "I's" are encoded with all different, new numbers 115, 657, 140, 2; but 657 is a surprise! The next I = 140 is a repetition; it is followed by three new numbers 8, 154, 314. The resultant pattern (details not shown here) indicated that halfway down the road the author decided that he had enough different numbers and no longer had to replenish his stockpile. Typically, the last 33 I's encoded are all repetitions of numbered elements used earlier. A surprise element in this pattern, in addition to an occasional use of a very high number, is the introduction of a new element toward the very end of his encoding process: the last three B's, the last T, the last U needed to encode B2C were seemingly picked out of thin air and for no apparent reason. On the other hand, the last 76 E's are encoded by re-using earlier elements! A weighted density plot would clearly indicate his preference for a seemingly lazy and very inefficient process.

TABLE III

Alternate Appearance of New and Used Elements in B2

	N	U	N	U	N	U	N	U	N	U	N	U	N	U	N	U	N	U	SUMS
A	4	1	2	2	2	1	4	1	1	4	1	3	1	16					43
B	3	2	1	2	3														11
C	3	2	3	1	1	7													17
D	2	1	3	7	1	4	1	3	1	1	1	4	1	6	1	12			49
E	5	2	2	3	1	5	2	4	2	1	2	76							105
F	3	4	1	2	1	1	1	1	1	1	1	4							21
G	3	2	1	9															15
H	2	1	4	5	1	18	1	5											37
I	4	1	3	3	1	9	1	33											55
J	2																		2
K	1																		1
L	8	8	1	11															32
M	3	1	1	1															6
N	4	5	1	1	2	6	1	49											69
O	6	2	3	6	1	7	1	7	1	29									63
P	3	1	1	7															12
Q	-																		-
R	2	1	1	2	1	1	1	1	1	4	1	22							38
S	4	1	1	4	2	1	1	4	1	1	1	13	1	1	1	7	1	3	48
T	9	2	1	1	1	3	1	10	1	1	2	13	1	23	1				70
U	4	1	2	13	1	3	1												25
V	1	17																	18
W	6	5	1	1															13
X	1	3																	4
Y	1	8																	9
Z	-																		-

NOTE: N = New, U = Used

763

Summarizing our findings, we observe that our author proceeded substantially in the following way:

(1) First, he wrote out his cleartext B2C; we may surmise that he edited it at least once before he began with the encoding process.

(2) He then numbered (at least) the first 1000 words in the DOI, noting (in passing?) what elements to use conveniently for X, V, and Y. We state here specifically that the author did not make a complete list of all available letter elements which would have been more efficient and also speedier. We also assert that he affixed position numbers to each of the circa 1000 first words of the DOI. As shown in Table II, of 179 different numbers used, 95 are even and 84 are odd. A statistical test reveals that the ratio 95/84 does not differ significantly from 90/89 or 89/90! Thus there is no difference of evens over odds, nor any other bias.

(3) He then encoded his cleartext, jumping randomly from one section of the keytext to another. These jumps were noted in our earlier paper (1) when we observed significant periodicities of lengths 3 and 5, as well as significant deviations between runs up and down. A casual inspection of B2 is quite convincing in this matter; if we single out elements belonging to specific hundreds, B2 assumes the following appearance: 115 - (73, 24) - 818 - (37, 52, 49, 17, 31, 62) - 657 - (22, 7, 15) - 140 - (47, 29) - 107 - (79, 84, 56) - 238 - (10, 26) - etc. This simplistic notation enhances our ability to see the pattern which was basic to the author's encoding job.

Finally, we observe that TJB was certainly not beyond making many clerical errors in his encoding process. We have several communications giving details about the forty-four specific code elements (numbers) which were edited by George Hart against the earlier Hiram Herbert papers in an effort to "force" the solution. Specifically, George Hart applied corrections of plus-one to two elements, 53 and 84; of minus-one to twenty-one elements ranging from 96 to 241; of plus-ten to four elements ranging from 449 to 505; of plus-nine to eleven elements ranging from 511 to 620; of plus-ten to three elements 643, 647, 666; of plus-eleven to two elements 807, 811; and of minus-eleven to the largest element 1005.

Anyone who has ever tried to number words by hand in a given keytext will at once recognize these corrections as being of a very common type. Incidentally, in the ranges shown above are also some few numbers which required no corrections suggesting that "Uncle TJB" went about his task rather sloppily, to say the least.

Unfortunately, we have no Wellsian Time Machine to help us "see" the author of these cyphers at work. But we have the tools of analysis and simulation both of which point strongly in the direction of our original assertion: TJB was not a professional cryptographer! However, he must have had some exposure to the coding techniques employed in his time; in his hour of need he resorted to this type of multiple substitution cypher. His choice has the advantage of great relative security which lasts exactly as long as the keytext is unknown. After that, even a high school student can decypher any such codes. TJB botched his job rather badly making numerous mistakes in the numbering of the words and in the selection of clever substitutes for missing letters. he also did not guard himself too well against attempts to break the cypher by probable word or letter substitutions, probably never even thought of it. Otherwise he would have used a

greater diversity of available keytext elements to obtain a maximum in degrees of freedom. May we infer that B1 and B3 also contain such errors?

The latter two have withstood the rather disorganized attack by many individuals and groups successfully for well over a hundred years. How much longer will they be able to protect "the exact locality of the vault" with its deposits of gold, silver, and jewelry? Judging by the many rumblings about "solutions" we are inclined to believe that B1 and B3 will shortly yield to the massive force exerted by organized and determined experts employing the latest tools of modern technology. It will be interesting to see what epitaph we can write about them as the decoded messages B1C and B3C filter into the public domain and the press. It will be equally interesting to look at the faces of our golddiggers as they enter the empty "excavation or vault six feet below the surface of ground"….