# STRATHCLYDE

**DISCUSSION PAPERS IN ECONOMICS**

# SEPARATING MYTH FROM PROBABILITY IN THE ORIGINS AND EVOLUTION OF QWERTY


**BY**

**NEIL KAY**



**NO. 11-27**



**DEPARTMENT OF ECONOMICS**
**UNIVERSITY OF STRATHCLYDE**
**GLASGOW**

# Separating Myth from Probability: the Origins and Evolution of QWERTY

Neil Kay

Department of Economics,
Sir William Duncan Building
University of Strathclyde
Rottenrow
Glasgow,
UK, G4 OLN

n.m.kay@strath.ac.uk

5th November 2011

## Abstract

We use basic probability theory and simple replicable electronic search experiments to evaluate some reported "myths" surrounding the origins and evolution of the QWERTY standard. The resulting evidence is strongly supportive of arguments put forward by Paul A. David (1985) and W. Brian Arthur (1989) that QWERTY was path dependent with its course of development strongly influenced by specific historical circumstances. The results also include the unexpected finding that QWERTY was as close to an optimal solution to a serious but transient problem as could be expected with the resources at the disposal of its designers in 1873.



Keywords: path dependency: QWERTY: technology: technological standards: innovation

# Separating Myth from Probability in the Origins and Evolution of QWERTY

## 1. Introduction

Paul David's 1985 AER article and W. Brian Arthur's 1989 EJ article (David, 1985; Arthur, 1989 respectively) are two of the most influential and cited articles in modern social science. Kim et al., (2006) put them in the top hundred most cited articles in economics published in major refereed economic journals between 1970 to 2005.[1] In addition, Harzing's Publish and Perish[2] records 4206 citations for David (1985) and 4248 citations for Arthur (1989)[3]. The articles have had a powerful impact on many areas of academic research including innovation studies, history of technology, economics, organization science, and strategic management. Although the articles were authored separately and made significant original contributions in their own right, in terms of their contribution for our purposes here they may be considered as complementary assets.

We shall not review the enormous literature that these two articles have spawned, Arthur (1994) gives some flavour of the breadth and depth of the applications both real and potential of what David labelled QWERTY-nomics. Instead our concern here will be what was the seminal example chosen by Arthur and David to explore the implications of increasing returns and lock-in in a path dependent world; the evolution of the QWERTY keyboard standard.

As Arthur explains (1994, pp.xvii-xviii), the creation of the QWERTY standard helped provide much of the basis for the development of his and David's theory of technological evolution in which increasing returns, path dependence and historical accidents and circumstances play powerful and influential roles. But what is noteworthy about the QWERTY story is the extent to which the historical accidents and circumstances that have been cited as crucial factors in the evolution of QWERTY have been questioned and dismissed as speculation, myth and apocrypha. In this paper we draw upon standard examples in probability theory based on urn selection and round table seating problems and use some simple electronic search experiments to assess the likelihoods that core elements in the QWERTY story are true.

We introduce two central alleged "myths" in David's (1985) account in Section 2, and review characteristics of the early technology that will be important to our analysis in Section 3. We explore the probabilities that the "myths" are in fact true in Sections 4 and 5, look at the role of contiguity and non-letter buffers in Section 6 and the search for infrequent letter combinations in Section 7. We finish with a short concluding section.

## 2. The QWERTY "myths"

David's (1985) account of the evolution of QWERTY cites technical interrelatedness, economies of scale, quasi-irreversibility of investment and path dependent processes as the influences which led to QWERTY becoming locked in as the dominant keyboard standard. The technical interrelatedness arose from the need for system compatibility between the keyboard "hardware" and the "software" of the touch typist's human capital stored in the memory of a specific keyboard format. Decreasing cost conditions (system scale economies) obtained from positive externalities for mobile labour trained in one generally available keyboard format, and in the market for instruction in touch typing. Quasi-irreversibility obtained from the difficulties that touch typists would face in "unlearning" QWERTY-based skills once acquired if they were subsequently required to move to another format, even those such as Dvorak which were alleged to be more efficient than QWERTY in terms of ergonomics and/or typing speed.

An integral part of both David and Arthurs' arguments is that historical accidents or transient factors in the early stages can exert significant and permanent influence as to which evolutionary path the technology takes, and which standards it adopts. David (1985) cites Arthur's (1983) more formal modeling of QWERTY-type scenarios to show how specific formats and technologies can become dominant standards through earlier marginal success (which can be through historical accidents or circumstances specific to place and time) becoming reinforced and amplified through positive feedback processes. In turn, Arthur (1983) was an early version of Arthur (1989) and the latter paper cross-cites David (1985) as providing a historical account of the mechanisms by which lock-in for QWERTY was obtained

QWERTY-nomics or the economics of QWERTY has been subject to much debate and some criticism (notably Leibowitz and Margolis, 1990), but perhaps the simplest and clearest defense of the basic argument is the general absence of credible alternative explanations. For example, Leibowitz and Margolis (1990) cite three studies of the relative merits of keyboard formats from the ergonomics literature which found that average typing speeds with Dvorak were 2.3%, 5% and 6.2% respectively faster than with the QWERTY format (1990, pp.15-16) and conclude "the consistent finding in the ergonomic studies is that the results imply no clear advantage for Dvorak" (1990, p.17).

However, competition in a neoclassical world is about marginal advantages, and a key performance parameter which showed an advantage of 2.3%, 5% or 6.2% over a rival standard or technology would normally be expected in such a world to help that standard or technology outcompete its rivals, in the absence of significant blocks or barriers to such adoption. In the economics of QWERTY these blocks or barriers are technical interrelatedness, economies of scale, and quasi-irreversibility of investment, and we shall assume these are present here in the absence of evidence-based alternative explanations. This will allow us to concentrate on the role of early historical accidents or circumstances in determining how QWERTY eventually became the dominant standard.

The QWERTY format was developed in the early 1870s by Christopher Latham Sholes, a printer and newspaper editor, with financial support provided by James Densmore and marketing support from George Yost, a petroleum salesman. It was Densmore and Yost

that secured the sale of the typewriter in 1873 to E. Remington and Sons (Wershler-Henry, 2007, p.70).

What was first labeled "the Sholes and Glidden Type Writer" was sold commercially by Remington in 1874, was renamed the No. 1 Remington in 1876, and after case redesign became the No. 2 Remington in 1878 (Wershler-Henry, 2007, pp.70-71). As Wershler-Henry notes, "there are a wide variety of competing explanations for how the configuration of the QWERTY keyboard came about" (2007, p.153).

It is in this context that Paul David wrote that;

> "the tendency of the typebars to clash and jam if struck in rapid succession was a particularly serious defect….. Sholes struggled for the next six years to perfect "the machine". From the inventor's trial-and-error rearrangements of the original model's alphabetical key ordering, in an effort to reduce the frequency of typebar clashes, there emerged a four-row, upper case keyboard approaching the modern QWERTY standard … Thus were assembled into one row all the letters which a salesman would need to impress customers, by rapidly pecking out the brand name: TYPE WRITER" (1985, p.333)[4]

David does not give any evidence for these assertions while Stephen Gould describes the alleged deliberate placing of the letters needed to type "typewriter" (or "type writer") in the top line as apocryphal (1991, p.68) and Lundmark (2002, p.19) also noted this feature of the keyboard exists "whether by intent or accident". Beeching added (also without evidence) that this was apparently "helpful to typewriter salesmen who could not type" (1974, p.x). We shall label this supposed myth "Myth 1" (while retaining quotations marks).

The second alleged design aspect of QWERTY identified by David was said to be a response to an early problem with the tendency for typebars to collide, jam and stick together. The problem was in fact well documented by a number of sources (Wershler-Henry, 2007, p.156), and indeed Richards notes that there was little that was original in Sholes's invention except that the key system and linkage to the typebars had been designed to enable rapid typing without "fouling" of the typebars (1964, p.25). The standard version of the solution to this problem is that trial and error was augmented and aided by Sholes's partner Densmore asking his son in law who was superintendent of schools in Western Pennsylvania to prepare a list of the most common two-letter sequences in the English language (Wershler-Henry, 2007, p.156; Richards, 1964, p.24)). Sholes and Densmore are then said to have used this list to "split up as many of these pairs as they could" (Wershler-Henry, 2007, p.156).

If such a list existed, then Sholes and colleagues would likely have jealousy guarded the contents of what would have been a valuable (but non-patentable) piece of intellectual property and even (at least to begin with) kept secret the existence of such a list[5]. Gould (1949, p.29), Richards (1964, p.24) and Lundmark (2002, p.17) cite this solution without supporting evidence, and Wershler-Henry concludes it is just "speculation", stating

"beyond the fact that they requested a table of letter frequencies … no one seems to have a copy of the list or any evidence of how Sholes and Densmore implemented it" (2007, p.156) . We shall label this "Myth 2"

In view of the central importance that the QWERTY standard has assumed in the economics and history of technologies, such lack of concrete evidence for such key elements in the story might seem surprising at first sight. However, at a perhaps obvious self-referential level, the primitive nature and restricted distribution of writing machines limited opportunities to record and debate these aspects of the development of writing machine technology. Also, the emergence of the QWERTY format was only one of numerous innovations that characterised the early days of the typewriter. Much of the early analysis was on the mechanical and engineering aspects of this competition, with little if any consideration given to keyboard formats in this context. For example, an early work, analyses differing machines in great technical detail but keyboard format (and QWERTY as standard) is only mentioned in passing (Mares, 1909, pp.47-49) and with no mention of the "myths". The focus instead in such accounts tends to be on the large variety of technical hardware problems and solutions to the core problem of replicating the characters of a keyboard on a paper page. It is perhaps not surprising that what could have appeared to have been the more tractable problem of alternative keyboard formats was not well documented.

## 3. The technology

The basic QWERTY format as developed in 1873 was based around four rows with eleven characters in each row. The 26 letters of the alphabet occupy parts of three rows and are flanked by numbers and punctuation marks in the top row and ends of some rows. QWERTY takes its name from the first six letters of the second line in Figure 1.

The 1873 keyboard is very close to the modern QWERTY format, though the numbers 1 and 0 are missing since they could be approximated with the letters I and O. Later the "C" and "X" were transposed and the M dropped to the end of the bottom row so creating what was to become the QWERTY standard that persists to the present day. .

However, there is a second aspect of the technology here which is also of relevance to consideration of "Myth 2". A typebar is the long metal strip inside a typewriter with a character to be printed on the page at the end of it. It was the typebars which would often jam when keys were depressed together in early typewriters, the propensity to jamming being greatest for adjacent typebars. Gould gives this description of the typebar arrangement in the Sholes Glidden Type Writer[6]:

> "Forty-four type-bars, each bearing one character at the inner end, are pivoted into short bearings arranged tangentially to a common circle. When at rest, these bars hang downwards in a "type-basket" …. forming an inverted truncated cone".(Gould, 1949, p.27).

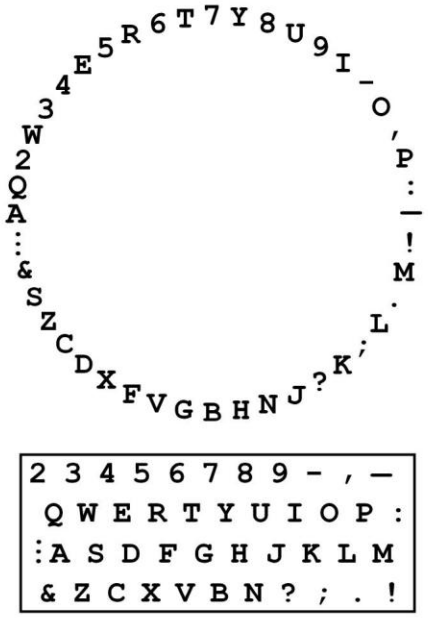


**Figure 1: Sholes QWERTY keyboard and typebasket 1873**
Source: Koichi Yasuoka and Motoko Yasuoka (2009)

The circular array of forty-four typebars relative to a "common circle" is shown at the top of Figure 1. The 44 keyboard characters in rows and the 44 typebars in a circle in the typebasket together form the basic physical characteristics of the QWERTY technology of relevance to our analysis. In the next section we shall consider whether probability theory can assist in terms of assessing the likelihood of either myth being true in the light of these technological characteristics and constraints.

We shall see that the likelihood of either QWERTY "myth" being true is dependent on hardware configurations in the respective cases. We shall also see that problems can in principle be made more tractable by treating them as analogous to standard problems in probability theory – urn selection problems in the case of "Myth 1" and round table seating problems in the case of "Myth 2".

## 4. "Myth" 1: Putting the letters for "typewriter" on one line

The first scenario we consider is the probability of the letters that make up "typewriter" finishing up on the top letter row of the keyboard (the second row in Figure 1) by chance, under the original Sholes design.

Suppose we were to assign individual letters at random to rows in the keyboard. To do this, imagine we have an urn with 26 balls identical in all respects expect that 10 balls are marked *T*, 10 balls are marked *M* and 6 balls are marked *B*, the letters corresponding to each of the three letter rows to which they will be assigned, *T* (top) *M* (middle) and *B* (bottom). We then assign the 7 letters that make up the word "typewriter" (t, y, p, e., w, r, i) by drawing balls blindfolded from the urn and pairing the 7 letters considered in

succession with a corresponding ball also drawn in succession from the urn. Each of the seven letters are then assigned a row associated with the letter on the ball with which they are paired.

Asking what the probability is of the 7 letters that make up "typewriter" finishing up on the top row is equivalent to determining the probability of drawing 7 *T* balls in succession from the urn.[7]

We can set up the problem as follows.

For the first ball the probability of drawing a *T* ball is:

$$P(T_1) = \frac{X}{Z}$$

Where $P(T_1)$ is the probability of the first ball drawn being a *T* ball, *X* is the number of *T* balls and *Z* is the total number of balls in the urn.

For the second ball, the probability of a *T* ball being drawn, assuming that the first ball was also a *T* ball, is:

$$P(T_1T_2) = \frac{X(X-1)}{Z(Z-1)}$$

Where $P(T_1T_2)$ is the probability of both of the first draws being a *T* ball. The probability of the second ball being a *T* ball is affected by there being one less *T* ball and one less ball in total available for selection after the first draw.

By extension, the probability of all *T* in a series of *n* draws is:

$$P(T_1T_2 \ldots T_n) = \frac{X\ (X-1)\ldots(X-n+1)}{Z(Z-1)\ \ldots(Z-n+1)}$$

$$= \frac{X\ !(Z-n)!}{(X-n)!Z!}$$

In this scenario X = 10, Z = 26 and n = 7. Consequently:

$$P(T_1T_2 \ldots T_n) = 0.00018$$

The probability that seven *T* balls would be drawn from the urn in succession in random choosing is therefore about 1 chance in 5,000. It is therefore highly improbable that this occurred by chance. By analogous reasoning, the probability that the seven letters that make up "typewriter" could finish up on the top row by chance is about 0.0002, or about one chance in 5,000. This would certainly be sufficient to reject any hypothesis that such

an allocation would be the product of a random process in standard tests of statistical significance.

Many of the popular accounts of "Myth 1" note that arraying the letters of "typewriter" along the top line could make it easier to type the word. However, David's account above has "assembled into one row all the letters which a salesman would need to impress customers" (1985, p.333). It could be argued that these sales tricks could be achieved by locating the "typewriter" letters in any single row, not just the top one. If that case is accepted, then the relevant question is: what is the probability of the 7 letters that make up "typewriter" finishing up on any one of the three rows by chance?

This may be said to analogous to determining the probability of drawing a series of *n* balls with any one specific set of markings (*T, M or B*) in succession from the urn in our previous example.

That is:

$$P(A) = P(T_1T_2 \ldots T_n) + P(M_1M_2 \ldots M_n) + P(B_1B_2 \ldots B_n)$$

Where $P(A)$ is the probability that only balls with one specific set of markings would be drawn in the *n* draws, $P(M_1M_2 \ldots M_n)$ is the probability that only *M* balls would be drawn in the *n* draws, and $P(B_1B_2 \ldots B_n)$ is the corresponding probability that only *B* balls would be drawn in the *n* draws.

Since the number of T and M balls are the same at 10 and the number of B balls are 6, then:

$$P(T_1T_2 \ldots T_n) = P(M_1M_2 \ldots M_n) = 0.00018$$

$$\text{And } P(B_1B_2 \ldots B_n) = 0, \text{ since } P(B_n) = 0$$

$$\text{So } P(A) = P(T_1T_2 \ldots T_n) + P(M_1M_2 \ldots M_n) + P(B_1B_2 \ldots B_n) = 0.00036$$

Since there are equal numbers of *T* and *M* balls there is an equal chance of all the *n* balls drawn bearing just the *T* brand or just the *M* brand. Since there are only six *B* balls and there are 7 iterations, if all the six *B* balls have been drawn in the first six iterations the probability of the seventh ball also being a *B* is zero. The probability that the letters that make up "typewriter" would fall on any one line by chance is 0.00036, twice as likely than in the top row case (here about one chance in 2,500) but still highly improbable.

In short we may dismiss any hypothesis that the letters making up TYPEWRITER fell on the first letter row (or indeed any other letter row) by chance.[8] Can this be seen as lending support to an argument that "Myth 1" is in fact true?

That would seem reasonable in the absence of competing hypotheses. In fact there is one consideration that could lead to a possible minor qualification of "Myth 1". In the early days of typewriter development, "typewriter" could refer to the machine, the typist, or could also be a verb meaning "to type" (Gitelman, 1999, p.208). We should also bear in

mind that in the early days typewriters were extremely expensive[9], objects of curiosity, with patchy and limited distribution. It is entirely possible that the Remington salesmen were not only trying to impress the potential customer with the ease of using this new technology, but to flatter their ego by encouraging them to type the name of the small elite band of practitioners they were being invited to join – the "typewriters".

## 5. "Myth" 2: use of a frequent letter list to design QWERTY

The second so-called "myth" relates to whether or not Sholes did enlist the help of a third party to construct a list of the most frequently encountered letter pairings in the English language, and if they did construct such a list whether or not this list was used to physically separate typebars associated with such letter pairs and "split up as many of these pairs as they could" (Wershler-Henry, 2007, p.156).

In this connection it is important to note first that the physical layout of the QWERTY keyboard with its three rows of letters should not itself be taken as a direct guide as to whether or not such deliberate sorting took place. What matters is the physical arrangement of the typebars themselves and "the tendency of the typebars to clash and jam if struck in rapid succession" (David, 1985, p.333). As Wershler-Henry notes "the matter on which all sources seem to agree is that whatever configuration Sholes started with, the type bars began to collide with each other and stick when a typist of even moderate speed began to type" (2007, p.156). It is contiguity of typebars in this circle that matters, not the position of letter keys on the QWERTY keyboard, with neighbouring or adjacent typebars facing the greatest risk of clashing.

A crucial aspect of the design for our purposes is the use of an alternating or zig-zag protocol to transpose letters from the keyboard to their corresponding position on the circular typebasket. This can be seen in Figure 1 where the 22 characters from the top two rows of the keyboard alternate in the top half of the typebasket, while the 22 characters from the bottom two rows of the keyboard alternate in the bottom half of the typebasket.

As with "Myth 1", we can draw on familiar examples in probability theory to explore the possibilities as to whether the present disposition of QWERTY happened by accident or design in this scenario. The coupling of typebars in this circle can be treated as analogous to the seating of couples around a round table[10], with the chances of a couple sitting together if seats are allocated randomly being a similar problem to the chances of two particular typebars being next to each other in the circle.

By analogy with a standard round table seating problem, we could first of all ask what is the probability of a particular letter pair (say $X$ and $Y$) being adjacent to each other on the typebar circle if the typebars have been allocated slots at random? We shall look at how we could answer the problem in principle before looking at the possibility of doing so in practice.

First, if there *n* typebars, then the total number of possible permutations of typebar arrangements is:

$$(n\text{-}1)!$$

If we treat *X* and *Y* as a single typebar in the case where they are located together, this means that we can then the total number of typebar permutations around the circle with *X* and *Y* treated as a single typebar is:

$$(n\text{-}2)!$$

However, *X* and *Y* can be located next to each other in two different ways, *XY* or *YX*.

This means that the total number of possible permutations of *X* and *Y* being located next to each other is:

$$(n\text{-}2)!2$$

Which means the probability P(N) of X and Y being located next to each other if typebars as allocated slots randomly is:

$$P(N) = \frac{(n-2)!2}{(n-1)!}$$

And the probability of the two typebars <u>not</u> being located next to each other if slots are allocated randomly is:

$$\{1 - P(N)\} = \left(1 - \frac{(n-2)!2}{(n-1)!}\right)$$

We can now introduce a second pair of typebars and ask what are the probabilities of both pairs being separated if the typebars are allocated spaces randomly, At this point we part company with the standard round table seating problem which typically assumes that each couple is (implicitly at least) in a monogamous relationship, and that once they are paired with their partner they cannot pair with anyone else. However, if we are considering the problems of frequently paired letters, there is no <u>a priori</u> reason to expect such fidelity on the part of individual letters, and indeed there could be grounds for suspecting a propensity to alphabetic promiscuity, especially on the part of some freewheeling vowels. What this means is that just because, say, E has coupled with R frequently in the past, that has not precluded it from striking up a similarly close relationship with S. Another way of expressing this is that just because one pair of letters have established a frequently close relationship on the written page, this does not preclude one or other of the partners displaying the same degree of intimacy with another partner on the same page. For the purposes of our typebar location problem, what this means is that it is reasonable to assume, at least for a first approach to the problem, that

whether or not a particular letter already has a frequent partner does not affect its chances of being a frequent partner to another letter.

On that basis, we can also treat the probabilities of each pair of frequent partners (from any such list of frequent pairings or partners in the English language) being located together on the typebar as being independent of the probability of any other pair of frequent partners being located together. So if we have two frequent partners, the probability of both pairings being split up if slots are allocated randomly is:

$$\{1 - P(N)\}^2 = \left(1 - \frac{(n-2)!2}{(n-1)!}\right)^2$$

And the probability of <u>all</u> pairs being separated if there is a list of *m* letter pairs is:

$$\{1 - P(N)\}^{\mathrm{m}} = \left(1 - \frac{(n-2)!2}{(n-1)!}\right)^m$$

We can turn now to the question of whether or not the physical arrangement of the typebars reflected design considerations with frequent partners in mind. Assume to begin with that the physical design of QWERTY did result in frequently typed pairs being separated. What would be the probability of this happening by chance?

The two variables that matter here are the number of typebars (*n*) and the number of pairs on the frequent partner list (*m*). There were 44 typebars in the Sholes typebasket, so we may take *n* to be 44. What is less certain is the length of the frequent partner list, but if we take m = 12 as a first estimate (as we note below, this may seem fairly reasonable in the light of actual frequent partner lists) then the probability of a single letter pair being separated in the typebar circle assuming the letters have been allocated randomly is:

$$\left(1 - \frac{(n-2)!2}{(n-1)!}\right)^m = 1 - \frac{2}{43} = 0.9535$$

Where n = 44 and m = 1

While the corresponding probability of all pairs on a given list of 12 letter pairs being separated in the typebar circle if the letters have been allocated randomly is:

$$\left(1 - \frac{2}{43}\right)^{12} = 0.52$$

Where n = 44 and m = 12

So even if the list is as long as a dozen pairs there would still be a better than 50/50 probability that random allocations would do the comprehensive job of pair separation with no need for human intervention on that front from the typewriter designer. A list of 20 would probably encounter no more than one or two pairs that would need separation. In fact, there is a further consideration of relevance if such a list was to be drawn on for

keyboard design purposes. Table 1 shows three lists of frequent letter pairings in the English language drawn from three independent sources.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| pair | Cornell frequency | Cornell rank | Dickens Rank | Zim Rank | pair | Cornell frequency | Cornell Rank | Dickens Rank | Zim Rank |
| th | 5632 | 1 | 1 | 1 | is | 1660 | 21 | | |
| he | 4657 | 2 | 2 | 2 | or | 1556 | 22 | | 16 |
| in | 3429 | 3 | 5 | 6 | ti | 1231 | 23 | | 17 |
| er | 3420 | 4 | 3 | 5 | as | 1211 | 24 | 18 | 19 |
| an | 3005 | 5 | 4 | 3 | te | 985 | 25 | | 14 |
| re | 2465 | 6 | 7 | 4 | et | 704 | 26 | | |
| nd | 2281 | 7 | 6 | 9 | ng | 688 | 27 | 14 | |
| at | 2155 | 8 | | 8 | of | 569 | 28 | 13 | 13 |
| on | 2086 | 9 | 12 | 7 | al | 341 | 29 | | |
| nt | 2058 | 10 | 19 | | de | 332 | 30 | | |
| ha | 2040 | 11 | 20 | | se | 300 | 31 | | |
| es | 2033 | 12 | 8 | 11 | le | 298 | 32 | | |
| st | 2009 | 13 | 9 | 10 | sa | 215 | 33 | | |
| en | 2005 | 14 | 15 | 12 | si | 186 | 34 | | |
| ed | 1942 | 15 | 11 | 15 | ar | 157 | 35 | | |
| to | 1904 | 16 | | 20 | ve | 148 | 36 | | |
| it | 1822 | 17 | 16 | | ra | 137 | 37 | | |
| ou | 1820 | 18 | | | id | 64 | 38 | | |
| ea | 1720 | 19 | 10 | | ur | 60 | 39 | | |
| hi | 1690 | 20 | | 18 | ro | n/a | n/a | 17 | |

**Table 1; Frequent letter pairings in English**

Sources: (1) rows 2, 3, 7 and 8 from Cornell University[11]: (2) rows 4 and 9 from t Dorit (2009)[12] (3) rows 5 and 10 from Zim (1962)

The table cites any pairing that appears in any one of the three lists. As can be seen there is a close correspondence between the results of the three lists, especially for the top eight on the Cornell list. Further, all 20 pairings from the Zim list and all but one of the 20 pairings from the Dickens list also appear on first 28 items in the Cornell list. The other feature of the list shown by the Cornell results is the tendency towards diminishing returns from list extension. The most frequent letter pairing on the Cornell list is about ten times more frequent than its 28th ranked pairing, and nearly 100 times more frequent than its 39th ranked pairing.

We can draw some tentative conclusions from Table 1. First, the strong correspondence it shows between independently produced lists varying in time and source suggests that if Sholes did obtain a reasonably accurate list that it would show strong family resemblance to those in Table 1, especially for the more frequently cited parings. Second, if we add in the issue of diminishing returns, any such list would have been unlikely to be as long (or much longer) than a dozen of the most frequent letter pairs to provide effective coverage

in relation to the resources that could have been reasonably devoted to this task. But if the list was restricted to that length, there would have been about an even chance that these frequent pairs would have already been separated by chance anyway.

Even if Sholes did use such a list, the combination of diminishing returns and chance separations suggests that it would (probably) have been of little assistance at best, and if it did lead to variations in the design these would (again probably) have been marginal tinkering. So when Wershler-Henry reports the "speculation" that Sholes and his partners used such a list to "split up as many of these pairs as they could" (2007, p.156), it is inferences that they would have a serious number of separations to undertake and that they may have ran into some technical barriers to completing this task to a satisfactory level that are questionable. This does not necessarily mean that such a list was not used, only that it would not have to have been a long list to identify the most frequent pairings, and in turn a short list would probably have not led to any need to tinker with the design. It also means that if any further splitting suggested by that list was indeed needed, it would probably have been minor and manageable with considerable freedom and scope for switching and swapping of any typebar locations if and when this should prove necessary or desirable. However there is a further element to take into account here, and that is the role of numbers and punctuation marks in buffering letters from contagion with each other. We look at that in the next section

## 6. QWERTY and the role of contiguity and buffers

The last section suggests that random allocation of letters would have been a fairly effective way of separating frequently occurring letter pairs and that any reallocations of letters to avoid frequently occurring letter pairs would have delivered marginal gains at best. However there is a further feature of QWERTY and typebasket design which reduces even further the probability that such a list would be useful for separating any specific cases of frequently occurring and adjacent letter pairs on the typebasket. QWERTY did not start from a random allocation of letters, numbers and punctuation marks but instead with a keyboard structure which clustered letters contiguously along three rows flanked by numbers and punctuation marks[13] as shown in Figure 1. When this was transposed using the zig–zag alternating protocol onto the circular typebasket, this meant that a dozen letters (here WERTYUIOPMLK) were automatically buffered from contiguity with any other letter on the typebasket through separation by numbers and punctuation marks.

These separations can provide a systemic basis for immunizing a letter from coupling with any other letter, let alone a frequent partner. Taken to the limit, a 52 typebar typebasket would have had the potential to separate all possible letter pairings for the 26 letters if dummy typebars were used as buffers in some cases.

Why was this not done? If this had occurred to Sholes and partners it might have appeared a technologically inelegant solution and in any case it would have added further engineering and expense to a product which was already very expensive at $125. Adding redundant typebars might also have created additional engineering problems. But even if

these problems could have been dealt with it would potentially have suffered problems of visibility and imitability in that any solution based around such buffers would have been obvious to any rival with the danger of it being designed around with alternative configurations.

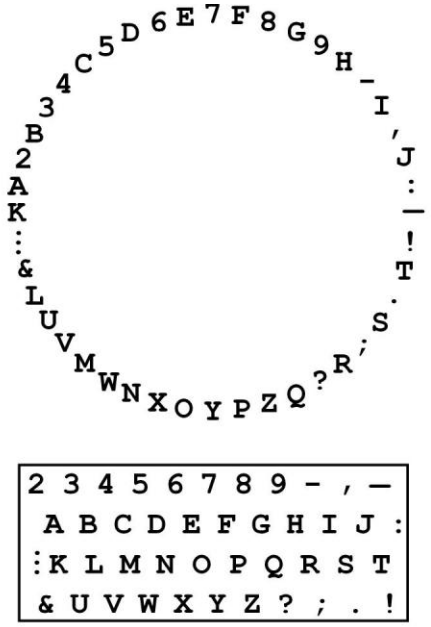


**Figure 2: Sholes keyboard and typebasket with ABCDEF format**

In any case, the partial immunization effects that QWERTY provided with the systemic buffering of letters with numbers/punctuation already added much more value than a purely random allocation solution would have provided, and would have almost certainly have ensured that a frequent letter pairing list would be of no direct use in the design of QWERTY. This can be easily demonstrated by experimenting with alternative letter configurations for the keyboard using the same protocol for keyboard/typebasket transposition that QWERTY used.

Figure 2 shows what is possibly the simplest and most obvious configuration with an alphabetic ordering of ABCDEF replacing the QWERTY ordering for all the letter keys reading left to right on all three rows. "A" takes the place of the "Q" on the QWERTY keyboard in Figure 2 and also takes the corresponding place that had been occupied by "Q" on the typebasket in that figure, and so on. It was also the default arrangement which sources suggest that Sholes started off with (Wershler-Henry, 2007, pp.153-155) for easy reference, and indeed the consonant string DFGHJKL on the middle letter row may be residual evidence of this.

An alphabetic ABCDEF keyboard is in fact just as effective as QWERTY in dealing with all the three frequent letter pairing lists in Table 1 in that no letter pair that appears on any of these lists is contiguous on the ABCDEF typebasket. While experiments with other keyboard configurations can produce pairings when transposed on to the typebasket, these are typically occasional and limited events and reinforce the general point that any

frequent letter pairing list would most likely be of no direct use, and at best only very limited use, in dealing with the potential problem of contiguous letter pairs on the QWERTY typebasket.

However there is one more feature of the design of QWERTY which is obscured by reference to frequent partner lists, and we turn to this in the next section.

## 7. QWERTY and the search for infrequency

In the Sholes Glidden QWERTY typebasket there are 26 letters and only 18 places for numbers and punctuation. This means that some letters are forced to be adjacent to each other. There are two sequences with contiguous letters: AQ and SZCDXFVGBHNJ. The remaining letters can be seen as following a non-contiguous order in the WERTYUIOPMLK sequence, this sequence separated by numbers and punctuation.

| **Contiguous series** | | | | |
|---|---|---|---|---|
| **AQ sequence** | | | | |
| **Pairing** | **Frequency** | | | |
| AQ/QA | 2 | | | |
| | | **Non-contiguous series** | | |
| **SZCDXFVGBHNJ sequence** | | **WERTYUIOPMLK sequence** | | |
| **Pairing** | **Frequency** | **Pairing** | **Frequency** | **Cornell list?** |
| SZ/ZS | Zero | WE/EW | 2,969 | No |
| ZC/CZ | Zero | ER/RE | 17,275 | Yes |
| CD/DC | 9 | RT/TR | 3,046 | No |
| DX/XD | Zero | TY/YT | 1,086 | No |
| XF/FX | Zero | YU/UY | 23 | No |
| FV/VF | Zero | UI/IU | 531 | No |
| VG/GV | Zero | IO/OI | 525 | No |
| GB/BG | 2 | OP/PO | 1,942 | No |
| BH/HB | 22 | PM/MP | 737 | No |
| HN/NH | 66 | ML/LM | 93 | No |
| NJ/JN | 45 | LK/KL | 264 | No |
| **Total** | 146 | **Total** | 28,491 | |

**TABLE 2: Frequencies for letter pairs in "Life on the Mississippi" PDF**

If the objective is separation of letter pairings in words that are also contiguous on the typebasket, the WERTYUIOPMLK sequence already achieves that with the help of number and punctuation buffers. The sequences that matter in this context are the two contiguous sequences.

The first clues to what QWERTY achieved can be pursued informally by considering the two contiguous sequences AQ and SZCDXFVGBHNJ. It can be extremely difficult to find any words that contain any contiguous letter pair read in either direction (e.g. XF or FX). This can be explored more systematically by measuring letter pair frequency for

texts that were contemporaneous with the design of QWERTY, and as a first experiment we shall use Mark Twain's "Life on the Mississippi" (Hence "LotM").

"LotM" was allegedly the first text actually typed using a typewriter, reportedly after Twain, a lover of gadgets, had first seen a Remington no.1 in 1873 (Wershler-Henry, 2007, pp.225-26) and we can see how efficient QWERTY would have been in dealing with contiguity in this case in Table 2 above which shows frequencies for different letter pairs from an electronic PDF version of this book (Manis, 1999)

The outcomes that matter for this purpose are the letter pair frequencies for the two contiguous sequences on the left of the table, but for comparison and contrast purposes we also show the letter pair frequencies on the right of the table for each letter pair contained within the non-contiguous WERTYUIOPMLK sequence.

The results are notable in that Twain's typist would have been able to type the whole of this 145,000 word manuscript with no more than 146 occurrences of letter pairs[14] that were contiguous in the AQ and SZCDXFVGBHNJ sequences in the typebasket. Further analysis shows that of the 146 occurrences in the contiguous sequences, half (73) were either in proper names (such as "John" or "Hamlet" or in two words "neighbor" and "enjoy" (or their derivations such as "neighborhood" or "enjoyment").

Suppose, as an experiment and to obtain a sample comparator the SZCDXFVGBHNJ and the WERTYUIOPMLK sequences were to swap places in the typebasket so the letters in the WERTYUIOPMLK sequence are now contiguous. In that case the number of letter pairings that would be contiguous on the typebasket in "LotM" would rise from 146 to 28,493 (including AQ/QA). While Twain's typist would have encountered letter pairings on the text that were also contiguous on the typebasket only about once every 1,000 words, switching these two sequences on the typebasket would have raised that frequency to about once every five words

We have already noted that it was very unlikely that a list of frequent letter pairings would have been of any direct help in the design of QWERTY but QWERTY goes much further in that it avoids even reasonably frequent letter pairs that would not have appeared on any such list. This is strongly reinforced by the pairings associated with the non-contiguous WERTYUIOPMLK sequence. Only one set (ER/RE) appears in the Table 1 frequent letter pairing lists, yet the typical frequency in the non-contiguous series is of an order of magnitude many times greater than any of those associated with the contiguous series.

QWERTY reduces the number of occurrences of letter pairings involving contiguous letters on the 44 character typebasket to as close to zero as it might be reasonable to expect with the resources at the disposal of Sholes at the time of development. Further experiments with alternative keyboard arrangements reinforce QWERTY's achievement in those respects. For example as Figure 2 shows, the obvious default alphabetic or ABCDE keyboard has two contiguous sequences AK and LUVMWNXOYPZQ if these are translated into the typebasket using the same alternating or zig-zag protocol used for

QWERTY. A search run on the "LotM" PDF using this format resulted in 5,599 letter pairings on the text that were also contiguous on the typebasket, with four couples showing extremely high incidence compared to even the QWERTY aggregate incidence: i.e. AK/KA (740); LU/UL (2,214); WN/NW (888); and OY/YO (1,389). None of the individual letter pairings even appear on any of the frequent letter pairing lists of Table 1. Yet had Twain's typist set out to type "LotM" with such an alphabetic keyboard, she or he would have encountered letter pairings on the text that were also contiguous on the typebasket on average about once every 26 words, compared to QWERTY's once every 1,000 words.

The conclusion that this leads to is that the design of QWERTY is not consistent with the deliberate separation of letter pairings using a frequent letter pairing list. Rather it is consistent with ensuring that when contiguity took place on the QWERTY typebasket, it was for letter pairings that are highly infrequent in the English language.

The QWERTY solution was to substitute one difficult, costly, and clumsy approach with a much easier, tractable, and effective approach. It can be difficult to know a priori whether or not a given letter pairing is likely to appear on any frequent letter pairing list. But if we restrict contiguity to consonants, then it can be relatively easy to find letter pairings which are highly infrequent or even effectively non-existent in the English language. For example, in the case of "B", searching for all incidences of letter pairings with B* and *B (where * indicates any consonant) confirmed that there was no occurrences for about one-third of all possible consonant pairings for "B" in the whole of "LotM".

Those last points relate to how consonants and vowels played differing roles in how QWERTY was designed. We have argued that a frequent letter pairing list would have been of little if any direct use. That is almost certainly true, but if such a list was consulted by Sholes and his colleagues and the aim was to separate frequency from contiguity, the role that vowels play in these respects would quickly have become evident. This is illustrated in Table 1 where 35 out of the 40 most frequently occurring letter pairs compiled from the three lists involve a vowel. This implies a simple meta-rule: to reduce probability of contiguous typebasket letters also appearing together in text to be typed, ensure that vowels are not contiguous with any other letter in the typebasket

That rule was generally followed in Sholes-Glidden with one exception which helps test the rule: AQ/QA. If you wanted to neutralize the effect of any vowel (other than U), then it is difficult to find any better solution than to put it in an isolated two-letter contiguous sequence with Q. This is shown in the case of "LotM" which only records two incidences of this coupling.

This also helps to explain why both "E" and "I" are missing from what is the otherwise alphabetic consonant string DFGHJKL on the QWERTY keyboard. If "Q"s position is taken as given, then removing these two vowels to the top letter line eliminates any possibility of contiguity with any other letter in the typebasket given the protocol used by QWERTY to transpose letters from keyboard to typebasket.

QWERTY was not designed to separate known frequent letter combinations on the typebasket, it was designed for a task that was much more ambitious and indeed was to be far more effective, the task of minimizing any chances of a letter pairing on the text corresponding to contiguous letters on a 44 character typebasket. Considered purely in terms of that narrow and historically-dependent objective, the results of tests such as those run on "LotM" and other texts[15] suggests it represents a near-optimal engineering design solution.

On that last point, could QWERTY have been improved any further to minimize the chances that letter pairings in any text to be typed would correspond to contiguous letters in the typebasket? While the answer in principle might well be in the affirmative, the more appropriate way to phrase that question would be to ask if Sholes had any reason to believe it could be improved further and, as importantly, could his rivals have reasonably been expected to find any way to improve on it, given no-one had the advantage of electronic searches to check frequencies associated with alternative configurations? "Enjoy" and "neighbor" were the most frequent words in the search above of "LotM"to incorporate any letter pairings from the contiguous sequences in the QWERTY typebasket, but even these words are highly infrequent in the English language with a Lancaster University search of a 100,000,000 word electronic databank ranking "enjoy" at 1,617$^{th}$ and the UK English variant "neighbours" at 3,239$^{th}$ in terms of frequency in present-day English (Leech et al., 2001)[16].

In fact, the contiguous sequences are highly sensitive to perturbations. Further experiments with "LotM" and other texts confirm that rearranging letters associated with these two sequences typically runs the danger of creating high-tariff frequent couplings. In terms of the specific engineering problem of contiguous letters in the typebasket and jamming problems, QWERTY is arguably as close to optimal as could reasonably be expected at the time of its design.

## 8. Conclusions

We can come to some conclusions regarding the development of QWERTY and some associated "myths" that have grown up around it.

First, on "Myth 1", it is indeed highly probable that QWERTY was designed to enable Remington salesmen to impress their potential customers by being able to easily type out TYPE WRITER just using a single row, though since the term "typewriter" at the time also commonly referred to the typist, it is less clear to what or whom the word would have been referring to.

"Myth 2" is that QWERTY was allegedly designed with the help of the most frequently encountered letter pairings in the English langauge, and that this list was (again allegedly) used to physically separate typebars associated with such letter pairs to "split up as many of these pairs as they could" (Wershler-Henry, 2007, p.156)?

The conclusion here is that it is quite likely that such a list was compiled but it would have been of little direct help in designing QWERTY. The one really helpful clue that such a list could have provided would be to demonstrate the need to isolate vowels on the typebasket from contagion with each other and (most) consonants. Whether by accident or design, QWERTY reflects that meta-rule.

The evidence does not support QWERTY being designed around a list of most frequently occurring letter pairs. What it does support is rather more interesting and powerful in terms of any design objectives reflecting contiguity of letter pairs. Rather than focus on high frequency, QWERTY approaches the problem from the other end of the spectrum by focussing on low frequency. QWERTY ensures that where contiguity could not be avoided, that it involved letter pairs that were amongst the most infrequent couplings in the English language.

It is also conceivable that the "myth" of using a frequent letter pairing list to design QWERTY would have been actively encouraged (or at least not discouraged) by Sholes in order to conceal the real design principles underlying QWERTY. The focus on low frequency of letter pairings as opposed to high frequency is non-obvious as is evidenced by the strategy remaining unarticulated for nearly 140 years. However, once revealed and articulated, such a solution would have been easily replicated by rivals. While any original intention to design around a list of frequent letter may have been genuine, its eventual status, if any, would have been no better than red herring.

If, as might seem likely from our arguments and the balance of probabilities, Sholes did deliberately apply a rule based around conjunction of infrequency and contiguity, there remains the question of why Sholes apparently took the secret to the grave with him (he died in 1890). It is obvious the rule would have been concealed in 1873 and soon after to protect intellectual property, but why did he at least not eventually confide in friends or family and for the sake of posterity and reputation?

An answer may be found in Wershler-Henry who describes Sholes as worn down by the various trials and tribulations associated with the development of QWERTY. He was described by sources as a modest and retiring man who found it difficult to persevere in the face of these difficulties to the point that by the late 1880s Sholes had totally disowned his own invention, refusing to own one, use one, or even recommend it (Wershler-Henry, 2007, p.67). If you do not value your own creation, then you would not expect other people to.

QWERTY is often dismissed as just an exercise in marketing consistent with "Myth 1", along with any claims that a list of frequent letter pairings were used to design QWERTY being treated as dubious. Indeed, our analysis supports both these aspects of the history of QWERTY as far as they go. With this background it is perhaps unsurprising that Wershler-Henry (2007, p.153) quotes with approval that any idea that the arrangement of keys in QWERTY was "scientific" was "probably one of the biggest confidence tricks of all time" (Beeching, 1974, p.40).

But in fact our evidence here suggests rather the opposite conclusion. In terms of the narrow design objective of reducing jamming problems involving letter pairs on the text that were also contiguous on the typebasket, QWERTY reflects a genuine and elegant engineering solution of the highest quality that would have been difficult to better with the tools that were at the disposal of Sholes and his rivals in the 1870s.

More broadly, it is hoped that the use here of basic probability theory and simple easily replicable experiments is seen as supporting and reinforcing the arguments of David (1985 and1986) and Arthur (1983 and 1989). David argued that the early part of the evolution of a technology "is likely to be governed by 'historical accidents', which is to say, by the particular sequencing of choices made close to the beginning of the process. It is there that essentially random, transient factors are most likely to exert great leverage, as has been shown neatly by Arthur's (1983) model of the dynamics of technological competition under increasing returns"(1985, 335)

The present paper shows that these "historical accidents" and "transient factors" did indeed influence the initial development and eventual dominance of QWERTY. It also helps confirm the nature of crucial missing or vaguely defined "historical accidents" or "transient factors" that influenced this process. QWERTY was indeed in part a marketing solution reflecting temporary historical circumstances, but much more than that it was also a superb engineering solution by Christopher Latham Sholes[17] to a problem that no longer exists. In short, QWERTY was indeed a highly path-dependent outcome as David and Arthur have argued, though in some ways neither obvious nor expected.

**Acknowledgments**

I am grateful to Paul A. David; W. Brian Arthur; Darryl Holden, Rod Cross, Steve Thompson and Sam MacAulay for helpful comments on earlier drafts. Any errors of omission or commission are my responsibility

## References


Arthur, W. B., 1983. On Competing Technologies and Historical Small Events: The Dynamics of Choice Under Increasing Returns. Technological Innovation Program Workshop Paper, Department of Economics, Stanford University,

Arthur, W. B., 1989. Competing technologies, increasing returns, and lock-in by historical events. Economic Journal 99, 116-31.

Arthur, W. B., 1994. Increasing returns and path dependence in the economy. University of Michigan Press, Ann Arbour.

Beeching, W. A., 1990. Century of the Typewriter. British Typewriter Museum Publishing, Bournemouth.

David, P. A., 1985. Clio and the economics of QWERTY. American Economic Review 75, 332-37.

David. P. A. (1986). Understanding the Economics of QWERTY: The Necessity of History. in: Parker W. N., (Ed), Economic History and the Modern Economist. Basil Blackwell, New York. pp.30-49.

Dorit R. L., 2009. Keyboards, Codes and the Search for Optimality. American Scientist, http://www.americanscientist.org/issues/num2/keyboards-codes-and-the-search-for-optimality/1 (accessed June 20th, 2011).

Gitelman, L., 1999. Scripts, Grooves and Writing Machines. Stanford University Press, Stanford.

Gould, R. T., 1949. The Story of the Typewriter: from the Eighteenth to the Twentieth Centuries: Representing Technology in the Edison Era. Office Control and Management, London.

Gould , S. J., 1991. The Panda's Thumb of Technology. in: Gould, S. J., Bully for Brontosaurus. Penguin, London, pp.59-75.

Jaynes, E. T., 2003. Probability Theory: the Logic of Science. Cambridge University Press, Cambridge.

Kim, E. H., Morse, A. and Zingales, L., 2006. What has mattered to economics since 1970. Journal of Economic Perspectives 20, 189-202.

Leech, G., Rayson, P. and Wilson, A., 2001. Word Frequencies in Written and Spoken English: based on the British National Corpus. Longman, London.

Leibowitz, S. J., and Margolis, S. E., 1990. The fable of the keys. Journal of Law and Economics 33, 1-25.

Lundmark, T., 2002. Quirky QWERTY. Penguin, London.

Manis, J. (Ed), 1999. Life on the Mississippi by Mark Twain (Samuel L. Clemens) Pennsylvania State University, Electronic Portable Document File format http://www2.hn.psu.edu/faculty/jmanis/twain/lifeonmr.pdf (accessed June 20th, 2011).

Mares, G. C., 1909. The History of the Typewriter. Pitman, London.

Martin, G. E., 2001. Counting: the Art of Enumerative Combinatorics. Springer Verlag, New York.

Maths Explorers Club., 2011. Cornell Department of Mathematics. http://www.math.cornell.edu/~mec/ (accessed June 22nd 2011).

Richards, G. T., 1964. The History and Development of Typewriters. Her Majesty's Stationery Office, London.

Wershler-Henry, D., 2007. The Iron Whim: a Fragmented History of Typewriting. Cornell University Press, Ithaca.

Yasuoka, K., and Yasuoka, M., 2009. On the prehistory of QWERTY. ZINBUN 42, 161-74.

Zim, H. S., 1962. Codes and secret writing. Scholastic Book Services, New York.

**Endnotes**

[1] David (1985) was number 89 in the rankings and Arthur (1989) number 93.
[2] Search of author impact analysis. Harzing home page is at http://www.harzing.com/pop.htm?source=pop_3.2.4150 (accessed June 22nd, 2011).
[3] To 22nd June 2011.
[4] See also David (1986) for further discussion of this.
[5] We discuss below how and why Sholes and/or Remington could at some point have found it to their advantage to publicise the existence of such a list.
[6] Gould describes the Sholes Glidden as the Sholes–Densmore typewriter
[7] This interpretation is based on an urn selection example in Jaynes (2003, pp.52-53).
[8] These conclusions hold even more strongly for the modern QWERTY standard with 9 and 7 letters respectively in the bottom two rows. The same basic probability approach used here shows that the probabilities of the requisite letters falling on any one line are barely greater than the probabilities of this happening by chance for the top letter row alone.
[9] David (1985, p. 333) notes the price if these early Remingtons was $125
[10] For example see Martin (2001) for examples of this type.
[11] This was produced by a Cornell University NSF-funded project to develop materials and activities to give middle school and high school students experience of more advanced topics in mathematics. See Maths Explorers Club (2011) at http://www.math.cornell.edu/~mec/ and for the table see: http://www.math.cornell.edu/~mec/2003-2004/cryptography/subs/digraphs.html (both accessed June 20th, 2011).
[12] Dorit's analysis was based on extracts taken from one of Dicken's novels.
[13] This basic structure was by now quite stable, though there was subsequent adjustment on the margins such as with "M" subsequently being dropped to the end of the bottom row
[14] For easy of replication we have not dropped PDF-specific results such as web addresses in the PDF acknowledgments (there are two such results for HN in the longer contiguous sequence)
[15] Other texts we used for these purposes included a PDF version of "Wealth of Nations".
[16] The rankings of these two words were obtained with the help of companion software provided with this text, further details on the project and the software are available at: http://ucrel.lancs.ac.uk/bncfreq/flists.html and http://ucrel.lancs.ac.uk/bncfreq/ (accessed June 20th, 2011).
[17] There are various accounts for how the 1873 design was produced but Sholes does appear to have been the major guiding hand here.