MODERN GREEK LANGUAGE FREQUENCY COUNTS FOR TEXT ENTRY DEVICES

Ilias Sarafis[1], Anastasios Markoulidis[2]

[1] *Technological Educational Institute of Kavala Agios Loukas, GR65404, Kavala, GREECE*
isarafis@teikav.edu.gr

[2] *Technological Educational Institute of Kavala Agios Loukas, GR65404, Kavala, GREECE.*

Abstract: In this paper we present and analyze tabulated case-sensitive single letter, digram and trigram frequency counts from a 450 Kword modern Greek corpus. Special characters, such as space, comma, paragraph change and full stop, used for text input, are also included. Data analysis revealed moderate correlation between English and Greek frequency counts, justifying a different approach when studying Greek keyboards. A keystroke-level model using the digram frequency data and applying Fitts' law was employed to predict entry speed rates for Qwerty, Fitaly, and Telephone virtual keyboards, both in English and Greek layouts. The results of this paper could be used when designing new software or hardware keyboard layouts and in order to improve their efficiency and text entry speed.

Keywords: digram frequencies; linguistic data; modern Greek; text input devices; virtual keyboard

1. Introduction

Moving towards mobile computing and portable devices has brought to the fore, the need for smaller, more versatile, and space efficient text entry methods. The use of T9 (http://www.t9.com, Tegic Communications 1998, James and Reischel 2001) on mobile phones for SMS editing is a characteristic example of a smart and effective way to enter text on a keyboard initially designed for numeric input. Only 9 alphanumeric keys (12 including space and special characters), and a disambiguation algorithm and lexicon small enough to fit into mobile phones memory, provide millions of users the ability to communicate quickly using text messages.

Apart from hardware mini-keyboards (Green et al. 2004, Clarkson et al. 2005) there are also many virtual (soft-type) keyboards, which have been studied and applied on mobile devices (Masui 1998, Goldstein et al. 1999, Zhai et al. 2005). Research for better keyboard layouts and disambiguation techniques is an on-going process, emphasizing both in better modelling algorithms and in designing smaller and more user-friendly layouts with higher text entry rates.

Much effort was made to develop theoretical models, in order to evaluate the performance of such virtual keyboards, in terms of maximum words-per-minute. These models mainly use the well-known Fitts' law (Fitts 1954) and keystroke-level timing together with linguistic data (digram frequencies) (Mackenzie et al. 1999, Soukoreff 2002).

Concerning research on Greek language keyboards, extended literature review has revealed two matters: first, no published research exists about optimized and efficient Greek language keyboard layouts, not even actual performance measurements of existing Greek keyboards. Second, there are no linguistic data published for Greek digram and trigram frequencies, but only for letter frequencies, despite the fact that there is a large (47 Mwords) corpus available.

Thus, our research effort was focused, on one hand to compute and publish Greek digram and trigram frequencies and on the other hand to apply a keystroke-level model using Fitts' law to a number of Greek keyboards. One of the modelled keyboards is a novel design, based on FITALY English layout (http://fitaly.com/, Textware Solutions 1996) and applying the philosophy of minimizing key distances between the most frequent digrams.

2. Modern greek linguistic data

A large (47 Mwords) corpus is available for Greek language, by the Institute of Language and Speech Processing (ILSP - website http://hnc.ilsp.gr). This corpus is known as "Hellenic National Corpus"™ (HNC) and it is not publicly available as a whole, rather through a web interface for searching words and lemmas.

Mikros et al. (2005) have published quantitative characteristics based on HNC corpus, mainly for linguistic purposes. Thus, their data do not include frequencies for special characters, such as space, comma, paragraph change and full stop, used for text input. Moreover, they do not provide data for digram and trigram frequencies.

In our study, we decided to use corpora from two nationwide Greek newspapers, "Ta Nea" and "Macedonia". The corpora are available through the Center for Greek Language (http://www.greek-language.gr/). We have selected the "articles" section and created a combined corpus having 453.407 words, 2.574.047 characters without spaces, 3.018.705 characters with spaces and 8.807 paragraphs.

Although our corpus is much smaller than HNC, it is certainly larger than the limited 20 Kwords corpus (Mayzner and Tresselt 1965), used by keystroke-level models for English language (Mackenzie et al. 1999), thus the results can be considered reliable enough. Further on, we computed Pearson product-moment correlation coefficients in order to find out if there is a significant statistical difference between HNC and our corpus. We compared single letter frequencies and ranking from the two corpora and we found to be highly correlated ($p = 0.9996$ & $p=0.9991$). Therefore, one could assume that the digram frequencies and other linguistic data we present here would be reliable enough, at least for keystroke-level modelling.

2.1. Word-length and single letter frequencies

Word length distribution for the studied corpus is presented in Figure 1. The average word length of 5.551 characters is very close to that reported for HNC (5.33 total and 5.56 for "miscellaneous" media type). We have also tested three, randomly selected subsets of the 450k corpus, one with 20 Kwords, one with 70 Kwords, and one with 180 Kwords. The word length for these subsets showed no significant differences (5,529, 5,552 and 5,557 characters respectively). The word length distribution shown in Figure 1, is also consistent with the one reported for HNC.

Single letter frequencies for lower-case, upper-case, accented and non-accented letters, are shown in Table 1. Notice that in Greek language there are two types of accent: (a) the "tonos" (΄) that can go over all the seven vowels and (b) the "dialitika" (¨) that go over "ι" or "υ" in certain cases. In order to type an accented letter on a keyboard, one should first press the accent key, prior to the letter key. There are also cases where "tonos" and "dialitika" are combined, e.g. ΐ and ΰ. In these cases also, one should first press the proper accent key, prior to the letter key.

Upper case letters relative frequency is 2.85%, the seven vowels count for 54.83% of the corpus and accented vowels are the 14.83% of the total letters. Correlation of the single letter frequencies list between the 450k corpus and the subsets of 20k, 70k and 180k corpora, was high (0.9995 0.9997 and 0.9999 respectively).

By comparing upper and lower-case counts we found a Pearson correlation of 0.7697 which is moderate but higher than the reported for English language (0.6337) (Jones and Mewhort 2004). Correlation between upper-and lower-case vowels is 0.7953 (accented 0.4969 & non-accented 0.8249), between upper- and lower-case consonants 0.7350 and between accented and non-accented vowels 0.8564.

In order to compare the Greek with the English frequency counts, we used the standard QWERTY character mapping, as in Table 2. We have used English linguistic data published by Jones and Mewhort (2004). We can observe that there are 13 totally identical letters and 7 letters that are similar (Delta, Fi, Gamma, Lamda, Pi, Ro, Sigma).

Using this mapping, correlation between the two lists of single letter frequency counts gives a $p=0.8157$. This result reveals the need for a different approach when designing and evaluating Greek keyboard layouts. Any proposal for optimized Greek layouts could not be based on work done for English ones.

After reordering the mapping to achieve a higher correlation ($p=0.9703$), we concluded to the results shown in Table 3, referred as "Greek reordered" layout. It is obvious that such a mapping would be totally confusing for the user of hardware keyboards, where two letters (English and Greek) are printed on the same key

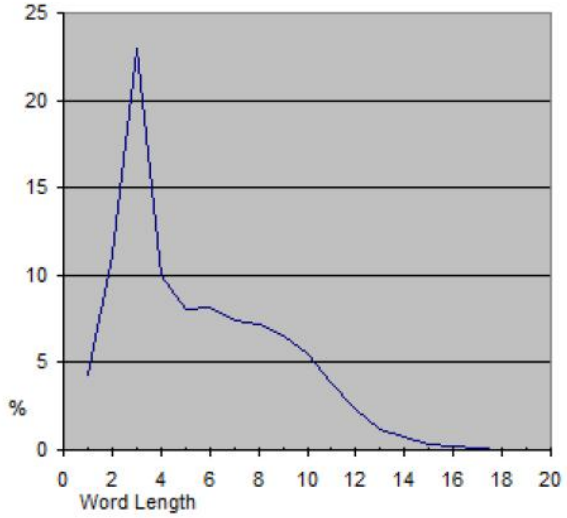

Figure 1. Word length distribution.

Figure 2. Greek QWERTY layout.

Table 1. Single letter frequency counts.

Letter	Upper case Count	Lower case Count	Letter	Upper case Count	Lower case Count	Letter	Upper case Count	Lower case Count	Letter	Upper case Count	Lower case Count
A	7,977	224,437	Λ	1,261	65,746	Υ	755	86,169	Ύ	35	24,891
B	1,804	15,129	M	4,144	77,192	Φ	852	18,792	Ώ	12	18,946
Γ	2,305	38,238	N	2,250	156,083	Χ	1,390	27,875	Ϊ	6	1,047
Δ	3,022	38,996	Ξ	123	9,620	Ψ	66	3,381	Ϋ	0	168
E	6,751	160,718	O	4,602	194,841	Ω	342	36,597	ΐ		268
Z	260	7,536	Π	5,401	94,512	Ά	595	46,237	ΰ		1
H	3,502	91,607	P	928	107,199	Έ	1054	43,384			
Θ	1,000	27,751	Σ	5,940	104,505	Ή	215	32,879			
I	2,587	168,352	ς		86,686	Ί	147	58,928			
K	5,177	94,606	T	5,146	188,610	Ό	894	50,339			

Table 2. Typical QWERTY mapping EN-GR.

EN	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
GR	A	B	Ψ	Δ	E	Φ	Γ	H	I	Ξ	K	Λ	M	N	O	Π	-	P	Σ	T	Θ	Ω	ς	X	Y	Z

Table 3. EN-GR mapping for maximized correlation.

EN	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
GR	I	Θ	K	Π	A	Ω	Δ	Σ	T	Ξ	B	P	ς	N	E	Λ	-	Y	H	O	M	Φ	X	Z	Γ	Ψ

Table 4. Frequency counts of special characters.

Symbol	Space	,	.	"	Paragraph	-	)	(	:	'	;	/
Count	444,658	27,382	22,716	11,162	8,969	3,019	1,999	1,922	1,192	1,015	629	546
Symbol	%	!	*	&	=	+	[	>	<	]	`	
Count	460	315	212	10	6	6	5	3	3	3	1	
Symbol	1	2	3	4	5	6	7	8	9	0		
Count	2,710	2,334	1,270	1,139	1,210	721	799	703	1,101	4,043		

2.2 Special Characters

Special characters are of significant importance for keyboard design and efficiency. As shown in Table 4, special characters, count for a total of 17.98% of the corpus. Space only, counts for 14.74% of the total characters, in comparison to 18.65% reported for English language (Mackenzie et al. 1999). Other significant special characters are the full-stop (0.75%), comma (0.91%) and paragraph change (0.29%).

2.3 Digram and Trigram Frequencies

Digram and trigram frequencies are useful in many cases. From anagram solutions (Dominowski and Duncan 1964) and word recognition to keyboard modeling and even for developing disambiguation algorithms (Maragoudakis et al. 2002). An Appendix that contains case sensitive bigram frequencies, including space, full stop and comma, together with a list of the 500 most common trigrams (case independent), is available by the authors, upon request.

We have tested the correlation between the following lists: (a) single letter counts, (b) letter as predecessor counts and (c) letter as successor counts. We found high consistency between them that was over 0.9997 in the worst case. There is moderate agreement between single letter counts and the frequency counts of a letter at the beginning or the end of the word (0.6992 & 0.6838 respectively).

High inconsistencies are observed between the lists of (a) the first letter and (b) the last letter of a word (p=0.2377), even after ignoring the final ς, which can only be found at the end of the word. This could affect the positioning of the "space" key, since space is the dominant character, both before (97.9%) and after a word (89.8%). Other characters following the end of a word are paragraph change (i.e. the "Enter" key), full stop and comma.

We have tested the correlation between case sensitive bigram counts and compared them with the values reported for English language (Jones and Mewhort 2004). Correlation between upper/upper and upper/lower digrams exhibits poor consistency at 0,4316 (English 0,372). Correlation between upper/upper and lower/lower is also low at 0,3526 (English 0,665) and there is moderate agreement between upper/lower and lower/lower digrams at 0,5945 (English 0,587). The total number of instances for each bigram type is:

upper/upper: 9.919 occurrences upper/lower: 52.152 occurrences lower/lower: 1.962.339 occurrences lower/upper: 28 occurrences.

3. EVALUATION OF GREEK SOFT KEYBOARDS

We have used a keystroke-level model proposed by Mackenzie et al. (1999), to predict the performance of three different Greek keyboard layouts and compare them to their English equivalents. The model is valid for one-hand typing on soft (virtual) keyboards.

At first, a spatial layout of each keyboard is drawn and the x-y coordinates of the keys are derived. Hence, the distance between every two keys in the bigram table can be calculated:

$$A_{ij} = \sqrt{|x_j - x_i|^2 + |y_j - y_i|^2} \quad (1)$$

where i is the predecessor and j is the successor in a bigram.

By applying Fitts' law (Fitts 1954), which models the time a human needs to move from one key to the other, we could calculate the movement time for each possible bigram

$$MT_{ij} = RT + \frac{1}{BW}\log_2\left(\frac{A_{ij}}{W_j} + 1\right) \quad (2)$$

where RT is the reaction time a human needs to locate a specific key on the keyboard, BW is the human motor bandwidth as defined by Fitts, Aij is the distance between the centers of the two keys, as in (1) and Wj is the width or the size of the target key.

Based upon the above prediction, we can then summate the movement time of all bigram combinations, after weighting them with the probability Pij of occurrence of a digram with first letter i and second letter j (Mackenzie et al. 1999):

$$\overline{MT} = \sum_i \sum_j (P_{ij} \times MT_{ij}) + RT \quad (3)$$

Equation (3) gives the mean movement time in seconds, between two characters. To convert this to words-per-minute (WPM) we use:

$$WPM = \left(\frac{1}{\overline{MT}}\right) \times \frac{60}{CPW}. \quad (4)$$

where CPW is the mean characters per word for a specific language. In order to have comparative results, we used CPW=5 as proposed by Mackenzie et al. (1999), both for English and Greek language.

In our calculations we have used the dimensions of a common hardware keyboard with key width W = 18mm but with zero clearance between keys. We also focused on maximum entry speed prediction, thus we used RT=0 (expert user) and the value BW=4,9 proposed by Mackenzie et al. (1999). For the key repeat time (bigrams with same letters), we used the 0,153 seconds value, also proposed by Mackenzie et al. (1999). English bigram data were

downloaded from http://www.dynamicnetservices.com/~will/academic/bit95.tables.html.

All the above model equations and bigram tables were entered into a spreadsheet and produced the final wpm results for each of the following keyboard layouts.

3.1 QWERTY layouts

We used typical QWERTY layouts with English-to-Greek mapping as in Table 2. For English language we used case independent bigram tables (a 27X27 matrix including the space character). For the Greek layout, we applied in addition, an analytical model with case-sensitive bigrams with accented letters, including full-stop and comma, as well as the space character. The analytical bigram table is given in the Appendix and leads to a 66X66 matrix with bigram probabilities.

In the analytical model, we assumed that if the second bigram letter is in upper case (e.g. at the begin of a sentence), a shift key should be pressed before the letter key, thus increasing the travel time. In the same manner, the accent key should be pressed before any accented letter that is successor in the bigram.

Figure 3. FITALY keyboard – English layout.

Figure 4. Proposed FITALY-like Greek layout – Detailed and reordered.

The Greek QWERTY layout for the analytical model is shown in Figure 2. In order to calculate the distance between the space key and the letter keys, we divided the space key length to three equal parts and used the distance from the part that was closer to each key.

We have also tested the EN-GR mapping in Table 3, which results from a reordering of letter mapping that gives a maximum correlation between English and Greek single letter frequency counts.

The predicted time for QWERTY layouts, for expert users (RT=0), was as follows:

English layout simplified:	31,23 wpm
Greek simplified model:	29,91 wpm
Greek reordered-simplified:	30,75 wpm
Greek analytical model:	27,73 wpm

As we can see, reordering the Greek keys results in a typing rate close to the English QWERTY. The lower speed rate for the analytical model is mainly due to accented and upper case letters that require an extra key-press before the letter key.

3.2 FITALY layouts

FITALY layout shown in Figure 3, is a commercially available software keyboard layout (http://fitaly.com/, Textware Solutions 1996), that was designed for minimizing travel distance between the most frequent letters. According to TextwareT, a 56,7% of total keystrokes occurs in the central area of the keyboard (letters T-A-N-E-O-R and the two space bars)

We have computed speed rates for the following FITALY-like Greek keyboards: (a) one using the typical QWERTY mapping as in Table 2 and (b) one with reordered keys as in Table 3 (Figure 4). We have tested layouts using the simplified bigram tables (case independent without comma, full stop and accent keys) and a Greek reordered layout using the analytical bigram tables of the Appendix.

After modifying the spreadsheet to calculate optimum distances from the two space keys, we obtained the following results for expert users:

English layout simplified: 40,57 wpm

Greek simplified model: 37,22 wpm
Greek reordered-simplified: 3 9,67 wpm
Greek reordered-analytical: 34,81 wpm

As expected, FITALY layout is much more efficient than QWERTY, both in English and in Greek languages. Entry speed rate improvement is 29,9% for English layout, 24,43% for Greek simplified model, 29% for Greek reordered-simplified model and 25,53% for the analytical Greek model.

3.3 Telephone pad layouts

The telephone keyboard is used extensively for exchanging short text messages via mobile phones. As there are three or more letters assigned to each key, text entry is done (a) either in a multi-tap mode, or (b) using a disambiguation algorithm.
In multi-tap mode the user has to press a key one or more times until the desired letter appears. I.e. one should press three times the number 2 key, in order to enter the letter "C". In word disambiguation mode, each letter key is pressed once and the algorithm selects and suggests the possible word combinations from a built-in lexicon. In case of more than one ambiguous words, the user can select the desired word from a list of suggestions.

As an example, to enter the word "HELLO" the key sequence is "4-3-5-5-6" in word disambiguation mode. In multi-tap mode this would be "4-4-3-3-5-5-5-1 second pause-5-5-5-6-6-6". Notice that in multi-tap mode a delay of one second is needed if the subsequent keys are the same.

It is obvious that multi-tap mode is highly inefficient in terms of text entry speed. Furthermore, modern mobile phones and devices have enough memory and power to implement a disambiguation algorithm. Thus we focused on telephone keypads incorporating such an algorithm. In addition, we assumed an "ideal" algorithm and we have not taken into account the time a user needs to select the desired word from a list of suggestions.

With the assumptions above, modelling the telephone keyboard (Figure 5) for maximum entry speed (the case of an expert user), produced the following results:
English layout simplified: 40,86 wpm Greek simplified model: 41,30 wpm
Accent is automatically entered when typing in lower case with disambiguation. Thus, an analytical model for telephone keypads was not computed. Further on, since the telephone keypad follows a simple logic of alphabetically ordered keys, a reordered Greek layout for this type of keyboard would be meaningless and thus was not tested.

4. CONCLUSION

One main difference between English and Greek language is the presence of accented letters. The high frequency (~15%) of these characters should be taken into account when designing, modeling and evaluating keyboard layouts, since accent is an extra key that should be pressed prior to the letter key. An exception is telephone pads, where there is no provision for special accent key

Our predictions using an analytical model that takes accent and letter case into account, showed a 7,3% (QWERTY) to 12,2% (FITALY-like) decrease in entry speed, comparing to a simplified model.

Another key finding is the moderate agreement between English and Greek single letter counts, when using the typical QWERTY mapping. After reordering the keys so that the two frequency lists correlate, we predicted a slight 2,8% increase in typing speed for QWERTY-like keyboard, but a noticeable 6,6% increase in the FITALY-like layout.

Though reordering the keys on hardware QWERTY keyboards might be confusing for expert users, novel software keyboards could benefit from it.

Finally, a very promising keyboard layout is the telephone pad, being very efficient, both in terms of space and in terms of entry speed rates, when disambiguation algorithms are used. A vast majority of the population is very familiar with it and especially young people, which in many cases make use of a mobile phone before they try any other mobile device.

REFERENCES

[1]. Clarkson, E., Clawson, J., Lyons, K., and Starner, T., 2005. An Empirical Study of Typing Rates on mini-QWERTY
[2]. Keyboards. *In: ACM CHI 2005 Conference on Human Factors in Computing Systems*, April 2-7 2005, Portland Oregon
[3]. USA. ACM Press, 1288 - 1291 Dominowski, R.L., and Duncan, C.P., 1964. Anagram

solving as a function of bigram frequency. *Verbal Learning & Verbal*

[4]. *Behavior,* 3, 321-325. Fitts, P.M., 1954. The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement.

[5]. *Experimental Psychology,* 47, 381-391. Goldstein, M., Book, R., Alsio, G. and Tessa, S., 1999. Non-keyboard QWERTY Touch Typing: A Portable Input Intereface

[6]. For The Mobile User. *In: ACM CHI 1999 Conference on Human Factors in Computing Systems,* 15-20 May 1999

[7]. Pittsburgh PA USA. ACM Press, 32-39. Green, N., Kruger, J., Faldu, C. and Amant, R.S., 2004. A Reduced QWERTY Keyboard for Mobile Text Entry. *In: ACM*

[8]. *CHI 2004 Conference on Human Factors in Computing,* 24-29 April 2004 Vienna Austria. ACM Press, 1429 - 1432. James, C.L., and Reischel, K.M., 2001. Text Input for Mobile Devices: Comparing Model Prediction to Actual Performance.

[9]. *In: SIGCHI 2001 Conference on Human factors in Computing,* 31 March - 4 April 2001 Seattle WA USA. ACM Press,

[10]. 365-371. Jones, M.N., and Mewhort, D.J.K., 2004. Case-sensitive letter and bigram frequency counts from large-scale English corpora.

[11]. *Behavior Research Methods, Instruments, & Computers,* 36 (3), 388-396. Mackenzie, I.S., Zhang, S. X., and Soukoreff, R. W., 1999. Text entry using soft keyboards. *Behavior & Information*

[12]. *Technology,* 18 (4), 235-244. Maragoudakis, M., et al., 2002. Improving SMS Usability Using Bayesian Networks. *In:* I.P. Vlahavas and CD.

[13]. Spyropoulos, eds. *Lecture notes in artificial intelligence (LNAI, Vol. 2308.* Springer-Verlag, 179-190. Masui, T., 1998. An Efficient Text Input Method for Pen-based Computers. *In: ACM CHI 1998 Conference on Human*

[14]. *Factors in Computing,* 18-23 April 1998 Los Angeles California USA. ACM Press, 328-335. Mayzner, M.S., and Tresselt, M.E., 1965. Tables of single-letter and digram frequency counts for various word-length and

[15]. letter-position combinations. *Psychonomic Monograph Supplements,* 1 (2), 13-32. Mikros, G., Hatzigeorgiu, N., and Carayannis, G., 2005. Basic Quantitative Characteristics of the Modern Greek Language

[16]. Using the Hellenic National Corpus. *Quantitative Linguistics,* 12 (2), 167-184. Soukoreff, R.W., 2002. *Text entry for mobile systems: Models, measures, and analyses for text entry research.* Thesis (MSc).

[17]. York University. Tegic Communications, 1998. *Reduced keyboard disambiguating computer.* US Patent no. 5,818,437. Textware Solutions, 1996. *Method for designing an ergonomic one-finger keyboard and apparatus thereof.* US Patent no.

[18]. 5,487,616. Zhai, S., Kristensson, P.O., and Smith, B.A., 2005. In search of effective text input interfaces for off the desktop computing.

[19]. *Interacting with Computers,* 17, 229-250