# Letter Frequency Analysis of Languages Using Latin Alphabet

Gintautas Grigas[1] & Anita Juškevičienė[1]

[1] Institute of Data Science and Digital Technologies, Vilnius University, Lithuania

Correspondence: Anita Juškevičienė, Institute of Data Science and Digital Technologies, Vilnius University, Akademijos str. 4, LT-08663, Vilnius, Lithuania. Tel: 370-5210-9314. E-mail: gintautas.grigas@mii.vu.lt, anita.juskeviciene@mii.vu.lt

Received: February 23, 2018; Accepted: March 8, 2018; Published: March 26, 2018

## Abstract

The evaluation of the peculiarities of alphabets, particularly the frequency of letters is essential when designing keyboards, analysing texts, designing alphabet-based games, and doing some text mining. Thus, it is important to determine what might be useful for designers of text input tools, and of other technologies related to sets of letters. Knowledge of common features among different languages gives an opportunity to take advantage of the experience of other languages. Nowadays an increasing amount of texts is published on the Internet. In order to adequately compare the frequencies of letters in different languages used in the online space, Wikipedia texts have been selected as a source material for investigation. This paper presents the Method of the Adjacent Letter Frequency Differences in the frequency line, which helps to evaluate frequency breakpoints. This is a uniform evaluation criterion for 25 main languages using Latin script in order to highlight the similarities and differences among them. Research focuses on the letter frequency analysis in the area of rarely used native letters and frequently used foreign letters in a particular language. The frequency of the letters is one of the factors that determines the location of the keys for the language specific letters on the keyboard.

**Keywords:** diacritics, keyboard layout, Latin script, letter frequency, language statistics, language similarity

## 1. Introduction

Latin script is used in many languages but alphabets of these languages differ. A certain uniformity is ensured by the Basic Latin alphabet, which is actually the English alphabet with 26 letters. It can be regarded as the backbone of the alphabets of all languages that use Latin script. However, some letters of the Basic Latin are not included in the alphabets of other languages (Dagienė et al. 2010). On the other hand, every language has some specific letters. Thus, the variety of Latin-based alphabets is high and far higher than of other script groups (e.g. Greek, Cyrillic).

It is important to evaluate peculiarities of alphabets, particularly the frequency of letters, when designing keyboards, analysing texts, designing alphabet-based games, and performing other text-based tasks. This is especially important in the case of on-screen keyboards. They are included into the software, and thus, it is possible to select the optimal number of keys for each language, depending on the frequency of letters and the screen size restrictions that are common for mobile devices.

Another issue is the layout of the characters on the keyboard. The selection of the optimal layout, or at least a layout close to the optimal one, requires a considerable amount of research. On the basis of such a study (Bi et al. 2012), the English on-screen keyboard was simultaneously optimized for five languages: English, German, French, Spanish, and Chinese (pinyin). Although the five languages studied in this work represent a large population of potential users, one would still want to ask whether the result of such optimization can be further extended towards a number of other languages and their similarity groups.

Typing for mobile devices has some peculiarities due to their small size and holding on hands while typing. A model of two-thumb text entry was developed (MacKenzie, 2002). Keyboard layout QWERTY was adopted to this method by splitting it in two parts: one for the left hand, another for the right hand. Solution was simple but typing was slow. A special layout for two-thumb text typing has been designed by (Oulasvirta et al. 2013) and named KALQ. However, for English language only.

Similar investigation using bigrams of letters was made by Chun (2015). In referring to KALQ app for a certain type of smart phones he has indicated that this layout allows a 34% increase in typing speed. As a result of research Chun has developed two-thumb text entry keyboard layout for Korean language.

Conduction of similar studies for each of the languages is a difficult task. Could the research results of any other language be used? For this purpose, the frequencies of alphabet letters of different languages will be compared and similarities as well as differences among languages will be disclosed.

Our goal is to analyse the alphabets of various languages and to compare the letter frequencies of different languages by focusing on language specific letters and frequently used foreign letters. A similar research has been conducted for Lithuanian alphabet letter frequency comparison with the alphabets of other European languages (Grigas & Juškevičienė, 2015). In this paper, the investigation is expanded to 25 languages.

The number of language specific letters is usually lesser than of those taken from the Basic Latin alphabet (Dagienė et al. 2010) and their frequencies are usually lower. In order to reduce the keyboard size of mobile phones, tablet computers and other small devices, some language specific letters are left without keys. Thus, their typing becomes complicated: usually requiring to press (touch) more than one key. Language specific letters or at least part of them are often left out. Nevertheless, the increasing globalisation requires foreign language letters that are not included in the alphabet. In order to specify a threshold below which the typing of letters becomes considerably slower, for example, when typing with a few keystrokes, the frequencies of rarely used letters should be known.

Frequencies of letters depend on the type of text (topic, purpose, etc.), as well as the style of its author. There are many sources of letter frequency statistics available, often several per language. However, their samples differ. Our analysis requires to look for larger sources covering the same type or at least similar types of texts.

Letter frequencies of Danish, English, Finnish, French, German, Icelandic, Polish, Russian, Spanish and Swedish languages are available on the Practical Cryptography (2015) website. The presented samples are large with no less than 90 million characters for each language. Unfortunately, only ten languages are presented on this website.

The character (including letters) frequency statistics of the texts of Wikipedia is provided by Denny Vrandečić on his website (Vrandečić 2012) and is additionally analysed in the paper Language Resources Extracted from Wikipedia (Vrandečić et al. 2011). The author provides numbers of unigrams (characters), bigrams, trigrams of 262 languages in Wikipedia. All letters, including all foreign letters (regardless of the language) are counted. These two unique features are essential for our research.

Wikipedia is developed by many authors. The topics of articles vary, but have a lot in common among all languages. Moreover, in Wikipedia global phenomena are described, thus foreign letters are frequently used in its texts revealing more evident similarities and differences between languages. In addition, the texts of Wikipedia are posted online. This feature corresponds to contemporary tendencies in text.

Thus, this source of data is well-suited for our research and for this reason we have selected it. In the article, we will call this source (Vrandečić 2012) a Base of Wikipedia characters or only a Base.

The paper is structured as follows. Section 2 presents the sets of letters. Section 3 covers the proposed Method of the adjacent letter frequency differences in the frequency line, which helps to evaluate frequency breakpoints in the subsequent section 4 for all 25 languages. The areas of large jumps in the letter frequency lines, in order to decide which letters are reasonably to include in the main plane of keyboard, are covered in Sections 5. Similarities (or differences) among alphabets of the languages using correlation method are described in Section 6. The paper ends up with conclusions.

## 2. The Sets of Letters

The European Union has 24 official languages and 22 of these languages use Latin script. All of them are included in the analysis. We have also included some additional languages that are commonly used in Europe: Icelandic, Norwegian and Turkish. Thus, a total of 25 languages have been analysed. Statistics of all these languages is presented in the Base.

All characters, not only letters, are included in the Base. However, only letters of alphabets from the selected 25 languages are important for our analysis. Accordingly, the alphabets of these languages have been joined into one common set of letters. Everson (2004) has collected a lot of information about many alphabets. Alphabets of European languages are defined by ETSI in ETSI ES 202 130 v.2.1.2 (ETSI, 2007) standard and the annex of ISO 12199 (ISO 2000) standard. As the annex is of informative nature, the priority is given to the ETSI standard. Further on we mention only these two standards, so no ambiguity will arise if we call them only as ETSI and ISO.

In ETSI, letters of every language are divided into two groups: A and B. Group A includes all letters of the alphabet. They are compulsory. The letters that may be necessary for borrowed words or foreign personal names are assigned to group B. Nevertheless, this division has some doubtful points. For example, group A of the English alphabet includes a few letters that are not in the Basic Latin alphabet (e.g.: æ, ñ). However, such letters usually belong to some other alphabet (e.g.: æ – Danish, ñ – Spanish) and thus they will be included into the joint letter set through those languages.

The letter frequencies can reveal the real picture of the letter usage. Our preliminary study of the frequencies showed that two letters of the Latvian language are assigned to group A in ETSI, but their frequencies in Latvian are very low (ō – 0.0005% ŗ – 0.0003%). In addition, they are absent in all other languages. Therefore, these two letters have not been included into the joint letter set.

There are less problems with letters from group B. If some letters are unreasonably moved into group B, they will be included into joint set from some other language.

Thus, the set of 102 letters has been formed:

aáàâäăāãåąæbcćċčçdďđðeéèėêëěēęfgġğģhħiıíìîïīįjkķlĺľļłmnńňñņoóòôöõőøœpqrŕřsśšşșßtťţþuúùûūůųűüvwxyýÿzźżž

All of the other characters (i.e. not letters) have been removed from the Base. Accordingly, the calculation calculation samples of different languages letter frequencies have been developed (Table 1).

Table 1. The amount of the analysed texts

| Language code | Language name | Text amount (in millions of letters) |
|---|---|---|
| cs | Czech | 294 |
| da | Danish | 170 |
| de | German | 820 |
| en | English | 806 |
| es | Spanish | 787 |
| et | Estonian | 84 |
| fi | Finnish | 349 |
| fr | French | 763 |
| ga | Irish | 12 |
| hr | Croatian | 132 |
| hu | Hungarian | 358 |
| it | Italian | 784 |
| is | Icelandic | 24 |
| lt | Lithuanian | 103 |
| lv | Latvian | 43 |
| mt | Maltese | 53 |
| nl | Netherlands | 734 |
| no | Norwegian | 297 |
| pl | Polish | 714 |
| pt | Portuguese | 672 |
| ro | Romanian | 173 |
| se | Swedish | 388 |
| sl | Slovak | 109 |
| sk | Slovenian | 90 |
| tr | Turkish | 216 |

## 3. Ratios, Differences and Breakpoints of Letter Frequencies

It is common to express letter frequencies as percentages. Two additional attributes for the letter pair analysis will be considered: the ratio r of the frequency f and the relative difference d of the letter pair. The ratio $r(i, j) = f_j/f_i$ shows how many times the frequency of letters reduces in the descending order line of the frequencies, where $i<j$.

In the most cases we will be interested in the ratio of the adjacent letters, that is, when $j = i+1$. The relative difference of frequencies $d(i, j) = (f_i – f_j)/f_j = f_i/f_j – 1$ will be called the difference.

The difference shows how many times the frequency $f_i$ of the letter i is higher than the frequency $f_j$ of the letter j (or vice versa). The difference of the adjacent letters will be 0 if the frequencies of both letters are equal or greater than 0 if the letters are sorted by the descending order of their frequencies. In this case the difference is more convenient to use than the ratio.

The large differences of the adjacent letters will be indicated as the breakpoints of frequencies. The breakpoints of frequencies, which are greater than 1 (frequency of the next letter is reduced two or more times) will be identified.

The English language will be used as an example. Table 2 contains letter frequencies and differences (presented in four decimals until the frequency value has at least two significant digits). Only 20% of rarely used characters are not included.

Table 2. Letter frequencies and differences of the English language

| Letter | Frequency | Difference |
|---|---|---|
| e | 12.15476703 | 0.401597 |
| a | 8.672083734 | 0.008609 |
| t | 8.598064819 | 0.141554 |
| i | 7.53189498 | 0.020975 |
| o | 7.377156799 | 0.005413 |
| n | 7.337440171 | 0.106191 |
| s | 6.633069973 | 0.001225 |
| r | 6.624957169 | 0.396094 |
| h | 4.745351843 | 0.119989 |
| l | 4.236963167 | 0.085339 |
| d | 3.90381717 | 0.140023 |
| c | 3.424333016 | 0.272688 |
| u | 2.690630195 | 0.016587 |
| m | 2.646728388 | 0.149707 |
| f | 2.302089594 | 0.083845 |
| p | 2.124003257 | 0.09196 |
| g | 1.945129235 | 0.16274 |
| w | 1.672883441 | 0.078678 |
| y | 1.550863958 | 0.044312 |
| b | 1.48505829 | 0.399725 |
| v | 1.060964146 | 0.767008 |
| k | 0.600429726 | 1.974128 |
| x | 0.201884315 | 0.079297 |
| j | 0.187051741 | 0.412901 |
| z | 0.13238847 | 0.204144 |
| q | 0.109944054 | 8.833333 |
| é | 0.011180751 | 1.473379 |
| á | 0.004520435 | 0.421285 |
| ö | 0.003180528 | 0.155275 |
| ü | 0.002753048 | 0.016871 |
| í | 0.002707371 | 0.015173 |
| ó | 0.002666907 | 0.315818 |
| ā | 0.002026804 | 0.091511 |
| ä | 0.00185688 | 0.207523 |
| è | 0.00153776 | 0.28757 |
| ø | 0.001194311 | 0.168853 |
| ç | 0.001021781 | 0.098332 |
| ñ | 0.000930302 | 0.149893 |
| š | 0.000809034 | 0.025649 |

| | | |
|---|---|---|
| æ | 0.000788802 | 0.017451 |
| ú | 0.000775272 | 0.066234 |
| ł | 0.000727113 | 0.028802 |
| ū | 0.000706756 | 0.01047 |
| å | 0.000699433 | 0.134488 |
| ć | 0.000616519 | 0.016994 |
| â | 0.000606217 | 0.013909 |
| ë | 0.0005979 | 0.114272 |
| à | 0.000536584 | 0.032728 |
| ã | 0.000519579 | 0.020229 |
| ī | 0.000509277 | 0.049898 |
| č | 0.000485073 | 0.025722 |
| ô | 0.000472909 | 0.184333 |
| ð | 0.000399304 | 0.218099 |
| ă | 0.000327809 | 0.040173 |
| ş | 0.000315148 | 0.011151 |
| ê | 0.000311673 | 0.050188 |
| ē | 0.000296778 | 0.022669 |
| ś | 0.0002902 | 0.037267 |
| ž | 0.000279773 | 0.101662 |
| ß | 0.000253956 | 0.039106 |
| ı | 0.000244398 | 0.049014 |
| î | 0.000232979 | 0.005895 |
| ń | 0.000231614 | 0.019115 |
| đ | 0.000227269 | 0.077059 |
| ï | 0.000211009 | 0.370968 |
| ò | 0.000153913 | 0.036789 |
| ì | 0.000148451 | 0.065004 |
| þ | 0.00013939 | 0.158927 |
| ý | 0.000120275 | 0.004145 |
| û | 0.000119779 | 0.072222 |
| ğ | 0.000111711 | 0.026226 |
| ř | 0.000108856 | 0.013873 |
| ę | 0.000107366 | 0.075871 |
| ș | 0.0000998 | 0.166909 |
| ż | 0.0000855 | 0.004373 |
| ą | 0.0000851 | 0.007342 |
| ù | 0.0000845 | 0.022523 |
| ě | 0.0000827 | 0.045526 |
| ő | 0.0000791 | 0.00315 |
| œ | 0.0000788 | 0.480186 |
| õ | 0.0000532 | 0.043796 |
| ė | 0.0000510 | 0.374582 |

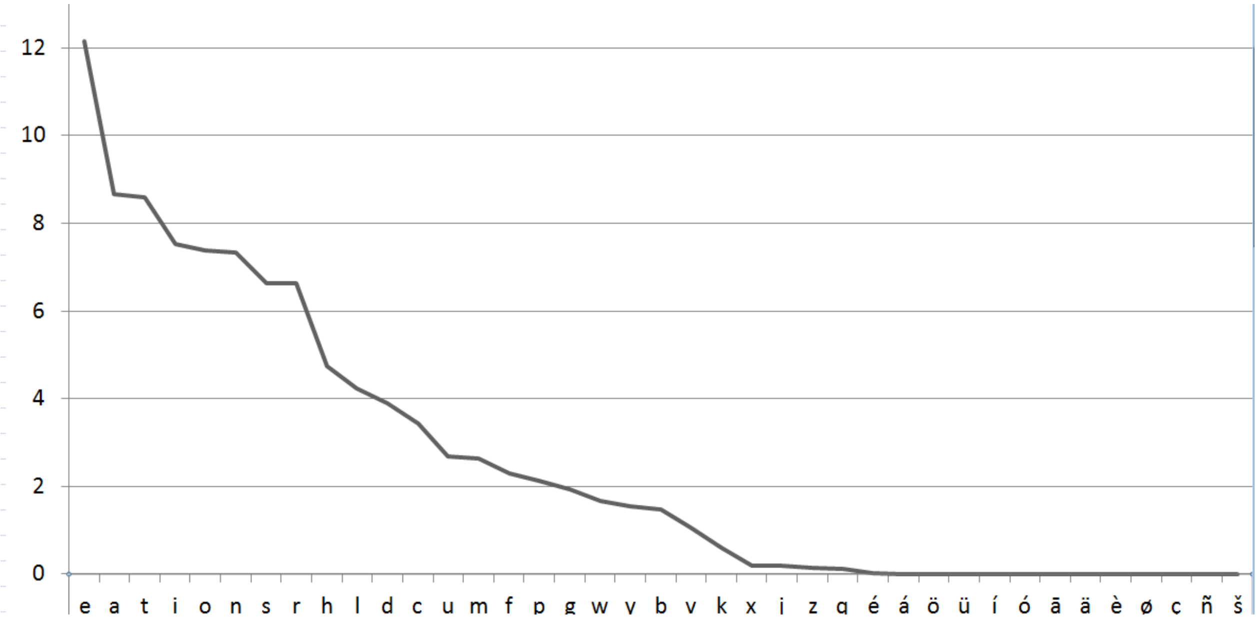

Figure 1. Letter frequencies of the English language in percent

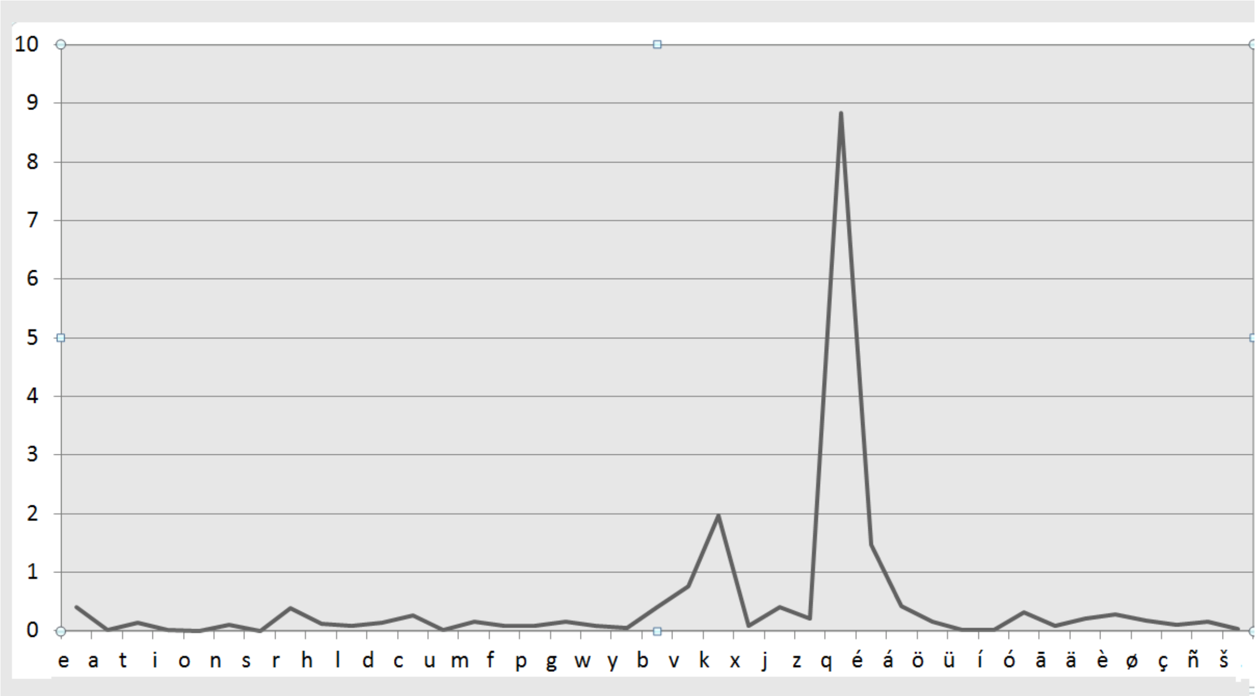

Figure 2. Letter frequency differences of the English language

The letter frequencies are presented in Figure 1, whereas the differences are introduced in Figure 2.

Two jumps can be seen in the differences chart: a small one (1) between the letters k and x, and another one, a very large one (8), between the letters q and é that clearly draws the line between the letters belonging to the English alphabet and those that do not belong to it or, in other words, between the commonly (often) used and rarely used letters. This threshold is important when designing data input devices (keyboards).

The use of the letters of the alphabet is balanced and the alphabet corresponds to the actual usage of the letters in the online space if all of the alphabet letters of the used language are listed before the first larger jump and all foreign language alphabet letters take the position after the jump.

There are some borrowed words in English texts that contain accented characters, such as à la carte, abbé, Ægean, archæology, belovèd, café. According to Everson (2004) the English alphabet contains 16 extra letters (à, æ, ç, ð, é, è, ë, ê, ï, ñ, ö, ô, œ, ȝ, þ, ƿ). However, as it can be seen in Figure 2, these letters are positioned after the high jump (8). Therefore, their position in the frequency line shows that they must be left outside the English alphabet.

## 4. Frequency Differences and Jumps Between Adjacent Letters in the European Languages

The frequency differences are introduced in a simplified form, as the presentation of a separate chart for every language (Table 3) would occupy a lot of space. The letters are divided into 4 groups of frequencies and are set out in the descending order. The most common letter is first in the line. It can be seen that the most common letters are: e (in 12 languages) and a (in 10 languages).

Table 3. The letters of 25 languages presented in their frequency descending order with inserted jump indicators

| | f≥0.1 | 0.1>f≥0.01 | 0.01>f≥0.001 | 0.001>f≥0.0001 |
|---|---|---|---|---|
| cs | oeantsilvrkdump**í**chz**á**yjb**ěéřýčžš**g**ů**f❶**ú***w* | x**ň** | **óť**❶**ď**q | üöä❶ćèłėľ |
| da | erntaisdlogmkfvubhp**åyøæ**cj❷w | z | xéqöü | ä❶áóíèðāšßłç |
| de | enirstadhulgocmbfkwzpv**üäö**jy**ß** | ❶x | ❶q❶é | ❶áóíèšçčāłøâćúåūô |
| en | eationsrhldcumfpgwybvk❶xjzq | ❽**é** | ❶á**ö**üíóāä**èøç** | **ñ**š**æ**…**ëà**…**ê**…**ï**…**œõ** |
| es | eaonsrildtcumpbgvfy**ó**hq**í**jz**áé**x**ñ**k**ú** | w | ■ **ü**❶öèçàäã | āâūëšćôò |
| et | aiestlunokrdmvgpjh**ä**b**õü** ❶ fc**ö**❶y | wz | šx❶žqé | áāóøíå |
| fi | ainteslouk**ä**mrvjhpyd❶**ö**gcbfw | z | x❶šéq | ❶áüåóží❶èøć |
| fr | easnitrluodcpm**é**vgfbhq**à**x**è**yjk**ê**z | w**çôâîùœûï** | ❷**ë**áüöíóäāúñ | šćūã |
| ga | ainhrestclodgmub**íáé**f**óú**p❹vyk | wjzxq | ❶ä**ö**ü❶**àłëèçñ**òšāč | ãð**ô**…**æ**…**ê**…ą**ï**…**œ** |
| hr | aioenrjstuklvdmpgzbc**č**h**šžć**f**đ**yw | ❷x❶qé | áüöíóäúãâçñāè | |
| hu | eatlsnkrioz**áé**gmbyvdhupj**ö**fc**óőíüúű** | ❶wx | ❸q | ❷šäõčćăèșçžłø |
| is | arniestulgm**ð**kfovhd**íá**jb**óþ**yp**öæúé**c**ý** | ❶wx | z❷q | üøäåāèšçłàīâūã |
| it | eiaonltrscdupmgvfbzh❶q**èà**ky**ò** | w**ù**jx**ìé** | ❽á**ó**üöíāčšäçşëăúīñâ | ūô |
| lt | iasoretnukmlpvdjg**ė**by**ųšž**c**ąįčū**fzh**ę** | ❶x | w | ❺qéáóíöüāłäīñ |
| lv | aisetrun**ā**oklmdpvjz**ī**gb**ē**c**š**❶**ņ**f**ļū**h**žķģč**\| | ❶ywx | ❷q❶éäüö | óı…**ō**…**ŗ** |
| mt | ailtenrumsojkdfb**ħ**pgwhz**ġż**xqv**ċ**c❷y | à❹éù | òèáüçíóöćìščâ | |
| nl | enairtodslghvmukcpbwjzf❷y | ❶x**ë**é | ❶qè**ïüö** | óá**ä**íłçšôčâåßã |
| no | erntasioldgkmvfpubh**å**yj**ø**c❶**æ**w | zxéqöä | üáóíèāðàúãšłçīßūćô | ñë |
| pl | aioenrzwsctkydpmulj**ł**gbh**ąęóżś**f❶**ńć**v | **ź**x❷éq | áüö | íäšèčø |
| pt | aeosirdntmuclp❶gvbfh**ã**q**éçá**z**í**jx**ó**ky**ê**w**õú** | **àâô** | ❹**ü**❶öèäñłş | āăšë |
| ro | eiarntulocsdpm**ă**fv**î**gb**șț**zh**â**jxkyw | ❹éq | áüöèóíäôãşçšłúćč | ßāàőë |
| sk | oaenirvtslkdmpuchjbz**á**y**ýíčéšžú**gf**ľťó**ôwx | **ňďä** | ❹**ĺ**q | řěö**ŕ**èü❶łůî |
| sl | eaionrsltjvkdpmuzbgh**č**c**šž**f❷y | w❷xéqćá | üöóíèäúđçâôőøł | êàåîë |
| sv | eanrtsildomkgvhfup**ä**cb**öå**yj❸xw | zé❶q | ❶üáøèóíæôâçšćëčú | ðñłã |
| tr | aeinrl**ı**kdtsmyuob**üş**vgzhcp**çğö**f❹j | w❷xâîq | é❶ûáóíäāèš | ñćúêãč |

Language specific letters are bolded in group A of the ETSI standard (Table 3). The letters of Basic Latin alphabet, which ETSI assigned to group B, are underlined. The integer value of the differences is written in white on a black background and it is inserted in the breakpoint place.

The biggest jump (11), marked by a square, belongs to the Spanish language. English and Italian languages also have quite big jumps (8). Next is the Lithuanian language (5). The bigger the jump, the higher the disjuncture between often and rarely used letters (letter groups A and B according to ETSI).

All languages sorted in descending order by their biggest jumps (J) and the distances from the biggest jump to the nearest smaller jump (D) are presented in Table 4.

If a different gradation of jumps was chosen (e.g., 0.5, 1.5, 2), a slightly different language distribution into the groups would be achieved. However, languages would not move more than two positions in the table (after the appointment of the position for each of the two criteria).

Due to the fact that there is no language that would not have any breakpoints, and values of these breakpoints vary in a wide range (from 1 to 11), it can be assumed that the given gradation by integer 1 is appropriate for our purposes.

Table 4. Jumps of the letter frequencies

| J | D | Languages |
|---|---|---|
| 11 | 10 | es |
| 8 | 7 | en, it |
| 5 | 4 | lt |
| 4 | 4 | ro |
| 4 | 3 | ga, pt, sk |
| 4 | 2 | mt, tr |
| 3 | 2 | sv |
| 3 | 1 | hu |
| 2 | 2 | fr |
| 2 | 1 | da, hr, is, lv, nl, pl |
| 2 | 0 | sl |
| 1 | 1 | no |
| 1 | 0 | cs, de, et, fi |

Adding the language specific letters into the ETSI standard raised objections inside it. The letters of the main group A should be treated and typed as the other letters of this group and the letters of the group B are moved further away, after numerals, and thus require more keystrokes to type them. Unfortunately, this rule applies only to the English language, which alphabet has only 26 letters of the Basic Latin. Other language specific letters within group A are typed in the same way as the letters of group B.

It can be assumed that a contradiction appeared in the ETSI standard due to the layout of the frequency unjustified letters (layout does not meet the distinction of the letters between the groups A and B) and it resulted in the decreased SMS typing speed in languages other than English.

Jumps of the frequencies can be used when designing onscreen keyboards. The number of keys is not mechanically restricted. Therefore, an optimal number of keys can be selected for each language by splitting the list of letters at the most appropriate breakpoint.

Table 5. The ratio of frequencies

| Language | Language specific letters | | | | Last in the en | | Ratio of freq. | |
|---|---|---|---|---|---|---|---|---|
| | List | $f_{s1}$ | $f_{sn}$ | $f_{sall}$ | Letter | $f_{bn}$ | $f_{sn}/f_{bn}$ | $f_{s1}/f_{bn}$ |
| cs | íáěéřýčžšůúňóťď | 2.7952 | 0.0170 | 12.5090 | q | 0.0120 | 1 | 233 |
| da | åøæ | 0.7926 | 0.7320 | 2.2866 | q | 0.0154 | 48 | 51 |
| de | üäöß | 0.5672 | 0.1567 | 1.5488 | q | 0.0344 | 5 | 16 |
| es | óíáéñú(ü) | 0.7998 | 0.1252 | 2.2796 | w | 0.0831 | 2 | 10 |
| et | äõüö(šž) | 1.0694 | 0.2330 | 2.9515 | q | 0.0117 | 20 | 91 |
| fi | äö (å) | 3.3433 | 0.4363 | 3.7796 | q | 0.0136 | 32 | 246 |
| fr | éàèêçôâîùœûï(ë) | 2.4438 | 0.0151 | 3.6890 | w | 0.0881 | 0 | 28 |
| ga | íáéóú | 1.8851 | 0.8751 | 7.1006 | q | 0.0124 | 71 | 152 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| hr | čšžćđ | 0.8837 | 0.1981 | 2.7016 | q | 0.0154 | 13 | 57 |
| hu | áéöóőíüúű | 3.5430 | 0.2253 | 11.3774 | q | 0.0122 | 18 | 290 |
| is | ðíáóþöæúéý | 3.6600 | 0.2497 | 12.0487 | q | 0.0109 | 23 | 336 |
| it | èàòùìé(ó) | 0.2381 | 0.0411 | 0.7084 | x | 0.0585 | 1 | 4 |
| lt | ėųšžąįčūę | 1.6643 | 0.1721 | 6.8919 | q | 0.0054 | 32 | 308 |
| lv | āīēšņļūžķģč(ōŗ) | 4.0671 | 0.1169 | 10.2346 | q | 0.0064 | 18 | 635 |
| mt | ħġżċ | 2.0288 | 0.5814 | 4.1295 | q | 0.6844 | 1 | 3 |
| pl | łąęóżśńćź | 1.7848 | 0.0660 | 6.0278 | q | 0.0121 | 5 | 148 |
| pt | ãéçáíóêõúàâô(ü) | 0.6669 | 0.0412 | 3.2523 | w | 0.1235 | 0 | 5 |
| ro | ăîșțâ | 2.2527 | 0.4485 | 5.7356 | q | 0.0156 | 29 | 144 |
| sk | áýíčéšžúľťóňďäĺŕ | 1.7484 | 0.0067 | 9.0332 | q | 0.0126 | 1 | 139 |
| sl | čšž | 1.1240 | 0.5336 | 2.5372 | q | 0.0132 | 40 | 85 |
| sv | äöå | 1.6661 | 1.2724 | 4.2160 | q | 0.0194 | 66 | 86 |
| no | åøæ(é) | 1.0216 | 0.1678 | 1.9692 | q | 0.0176 | 10 | 58 |
| tr | ıüşçğö | 8.5689 | 0.8869 | 10.5980 | q | 0.0124 | 72 | 691 |
| Average | | 2.1136 | 0.3303 | 5.5394 | | 0.5609 | 22 | 166 |

Note. English and Dutch languages which do not have the language specific letters are not included in the Table.

## 5. Frequency Ratio of Remote Letters

The ratio of the frequencies and their differences can be analysed not only between adjacent letters but also among remote letters. Such differences may be useful when considering the letter layout options on the keyboard, for example, what letter or group of letters should be allocated in the more convenient positions for typing..

Frequency differences of language specific and lesser-used Basic Latin letters are presented in Table 5. The following notations are used:

$f_{s1}$ – the frequency of the most frequent (at the front of the list) language specific letter,

$f_{sn}$ – the frequency of the rarest (at the end of the list, not in parenthesis) language specific letter,

$f_{sall}$ – the sum of frequencies of all language specific letters,

$f_{bn}$ – the frequency of the rarest Basic Latin alphabet letter,

$f_{s1}/f_{bn}$ – the ratio of frequencies of the most frequent language specific letter and the rarest letter of the Basic Latin alphabet,

$f_{sn}/f_{bn}$ – the ratio of frequencies of the rarest language specific letter and the rarest Basic Latin alphabet letter.

The letters in parenthesis are language specific letters that are included in group A of the ETSI standard. Nevertheless, they should not belong to this group due to their frequencies (see Table 5). Language specific letters are mostly used (more than 10%) by the Czech, Hungarian, Icelandic, Latvian and Turkish.

The ratio of the frequencies indicates the importance of the language specific letters and numeric evaluation of this importance with regard to rarely used ($f_{sn}/f_{bn}$) and commonly used ($f_{s1}/f_{bn}$) language specific letters. These ratios can be used to determine which letters should be given the priority when the data input devices (keyboards) have a limited number of keys.

These two columns of values can be considered as *from ... to* values. *From* – the number of times the frequency of the rarest language specific letter exceeds the frequency of the rarest Basic Latin letter. *To* – the number of times the frequency of the most frequent language specific letter exceeds the frequency of the rarest Basic Latin letter. Thus, if a language specific letter will be brought to the background of the keyboard, and a rarely used Basic Latin letter will remain in the foreground, this decision will be far away from the optimal option by *from* times.

For example, if the Latvian tablet computer keyboard foreground has only the Basic Latin letters, including rarely used q, and all the language specific letters are typed by few keystrokes, they are typed several times more slowly, even though that the letters č and ā are 18 and 635 times, respectively, more common than letter q. By the way, the frequency of letter q is the lowest among Basic Latin letters in 19 languages (from the 23 listed in the Table).

## 6. Character Frequency Correlation

The linguistic similarity among the characters in terms of the frequency can be judged from their alphabets' frequency correlations. Unigram frequency correlations of all the analysed language pairs are provided in Table 6.

In order to make Table 6 more compact, correlation coefficients (CC) are expressed as integer percentage that has been obtained by fractional CC multiplied by 100 and rounded to the nearest integer.

These percentage numbers range from 74 (da/ga) to 99: da/no (Danish/Norwegian) and hr/sl (Croatian/Slovenian). These two cases are depicted in frequency charts (Figure 3 and Figure 4). The letters are arranged in the decreasing order according to their frequency in the first language. As a result, the first language frequency polyline is permanently decreasing, whereas the second one is a zigzag curve. As it can be seen in the case of maximum correlation, the curves almost coincide, while in the case of minimum correlation the curves differ, although their correlation coefficient is not small.

Table 6. Unigram frequency correlation of languages as percentage

| | cs | da | de | en | es | et | fi | fr | ga | hr | hu | is | it | lt | lv | mt | nl | no | pl | pt | ro | sk | sl | sv | tr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cs | | 85 | 82 | 88 | 90 | 87 | 87 | 87 | 76 | 92 | 88 | 81 | 90 | 86 | 83 | 82 | 86 | **86** | 86 | 88 | 85 | **98** | 94 | 89 | 82 |
| da | 85 | | **96** | 94 | 91 | 85 | 80 | 94 | 74 | 84 | 87 | 85 | 89 | 79 | 79 | 81 | **97** | **99** | 79 | 87 | 87 | 84 | 88 | **96** | 84 |
| de | 82 | 96 | | 94 | 90 | 85 | 81 | **95** | 79 | 82 | 82 | 83 | 89 | 77 | 77 | 80 | **97** | 95 | 79 | 86 | 90 | 80 | 85 | 93 | 83 |
| en | 88 | 94 | 94 | | **95** | 91 | 89 | **95** | 86 | 88 | 87 | 86 | **95** | 86 | 84 | 87 | **94** | 94 | 87 | 93 | 93 | 89 | 90 | **96** | 85 |
| es | 90 | 90 | 90 | 95 | | 91 | 86 | **97** | 85 | 90 | 85 | 85 | **97** | 87 | 86 | 86 | 93 | 90 | 86 | **98** | **95** | 91 | 91 | 93 | 87 |
| et | 87 | 85 | 85 | 91 | 91 | | **96** | 91 | 86 | 93 | 87 | 88 | 92 | 94 | 94 | 94 | 86 | 89 | 83 | 90 | 91 | 88 | 82 | 91 | 89 |
| fi | 87 | 80 | 81 | 89 | 87 | 96 | | 86 | 83 | 91 | 85 | 86 | 90 | 93 | 90 | 92 | 81 | 85 | 83 | 84 | 88 | 88 | 91 | 90 | 87 |
| fr | 87 | 94 | 95 | 95 | 97 | 91 | 86 | | 90 | 86 | 87 | 85 | 94 | 85 | 84 | 85 | 94 | 94 | 81 | 93 | 94 | 86 | 89 | 94 | 84 |
| ga | 78 | 74 | 79 | 86 | 85 | 86 | 83 | 80 | | 83 | 77 | 85 | 87 | 82 | 82 | 86 | 78 | 76 | 77 | 84 | 87 | 80 | 81 | 84 | 83 |
| hr | 92 | 84 | 82 | 88 | 90 | 93 | 91 | 86 | 83 | | 83 | 85 | 93 | 94 | 91 | 92 | 87 | 86 | 89 | 91 | 90 | **95** | **99** | 89 | 87 |
| hu | 88 | 87 | 82 | 87 | 85 | 87 | 85 | 87 | 77 | 83 | | 84 | 85 | 79 | 83 | 83 | 86 | 90 | 80 | 82 | 81 | 87 | 87 | 90 | 84 |
| is | 81 | 85 | 83 | 86 | 85 | 88 | 86 | 85 | 85 | 85 | 84 | | 85 | 86 | 86 | 88 | 80 | 86 | 76 | 82 | 86 | 82 | 85 | 90 | 86 |
| it | 90 | 89 | 89 | 95 | 97 | 92 | 90 | 94 | 87 | 93 | 85 | 85 | | 91 | 87 | 92 | 91 | 89 | 88 | **95** | **97** | 91 | 94 | 92 | 88 |
| lt | 86 | 79 | 77 | 86 | 87 | 94 | 93 | 85 | 82 | 94 | 79 | 86 | 91 | | 94 | 92 | 78 | 81 | 85 | **88** | **89** | 89 | 93 | 86 | 86 |
| lv | 83 | 79 | 77 | 84 | 86 | 94 | 90 | 84 | 82 | 91 | 83 | 86 | 87 | 94 | | 90 | 78 | 82 | 81 | 86 | 88 | 86 | 90 | 87 | 85 |
| mt | 82 | 81 | 80 | 87 | 86 | 94 | 92 | 85 | 86 | 92 | 83 | 88 | 92 | 92 | 90 | | 81 | 83 | 82 | 84 | 91 | 83 | 91 | 88 | 90 |
| nl | 86 | 97 | 97 | 94 | 93 | 86 | 81 | 94 | 78 | 87 | 86 | 80 | 91 | 78 | 78 | 81 | | 97 | 82 | 90 | 90 | 86 | 90 | 93 | 84 |
| no | 86 | 99 | 95 | 94 | 90 | 89 | 85 | 94 | 76 | 86 | 90 | 86 | 89 | 81 | 82 | 83 | 97 | | 80 | 87 | 87 | 86 | 90 | 96 | 85 |
| pl | 86 | 79 | 79 | 87 | 86 | 83 | 83 | 81 | 77 | 89 | 80 | 76 | 88 | 85 | 81 | 82 | 82 | 80 | | 86 | 85 | 88 | 89 | 83 | 82 |
| pt | 88 | 87 | 86 | 93 | 98 | 90 | 84 | 93 | 84 | 91 | 82 | 82 | 95 | 88 | 86 | 84 | 90 | 87 | 86 | | 93 | 91 | 91 | 91 | 83 |
| ro | 85 | 87 | 90 | 93 | 95 | 91 | 88 | 94 | 87 | 90 | 81 | 86 | 97 | 89 | 88 | 91 | 90 | 87 | 85 | 93 | | 86 | 90 | 91 | 88 |
| sk | 98 | 84 | 80 | 89 | 91 | 88 | 88 | 86 | 80 | 95 | 87 | 82 | 91 | 89 | 86 | 83 | 86 | 86 | 88 | 91 | 86 | | **96** | 89 | 83 |
| sl | 94 | 88 | 85 | 90 | 91 | 92 | 91 | 89 | 81 | 99 | 87 | 85 | 94 | 93 | 89 | 91 | 90 | 90 | 89 | 91 | 90 | 96 | | 92 | 87 |
| sv | 89 | 96 | 93 | 96 | 93 | 91 | 90 | 94 | 84 | 89 | 90 | 90 | 92 | 86 | 87 | 88 | 93 | 96 | 83 | 91 | 91 | 89 | 92 | | 88 |
| tr | 82 | 84 | 83 | 85 | 87 | 89 | 87 | 84 | 83 | 87 | 84 | 86 | 88 | 86 | 85 | 90 | 84 | 85 | 82 | 83 | 88 | 83 | 87 | 88 | |

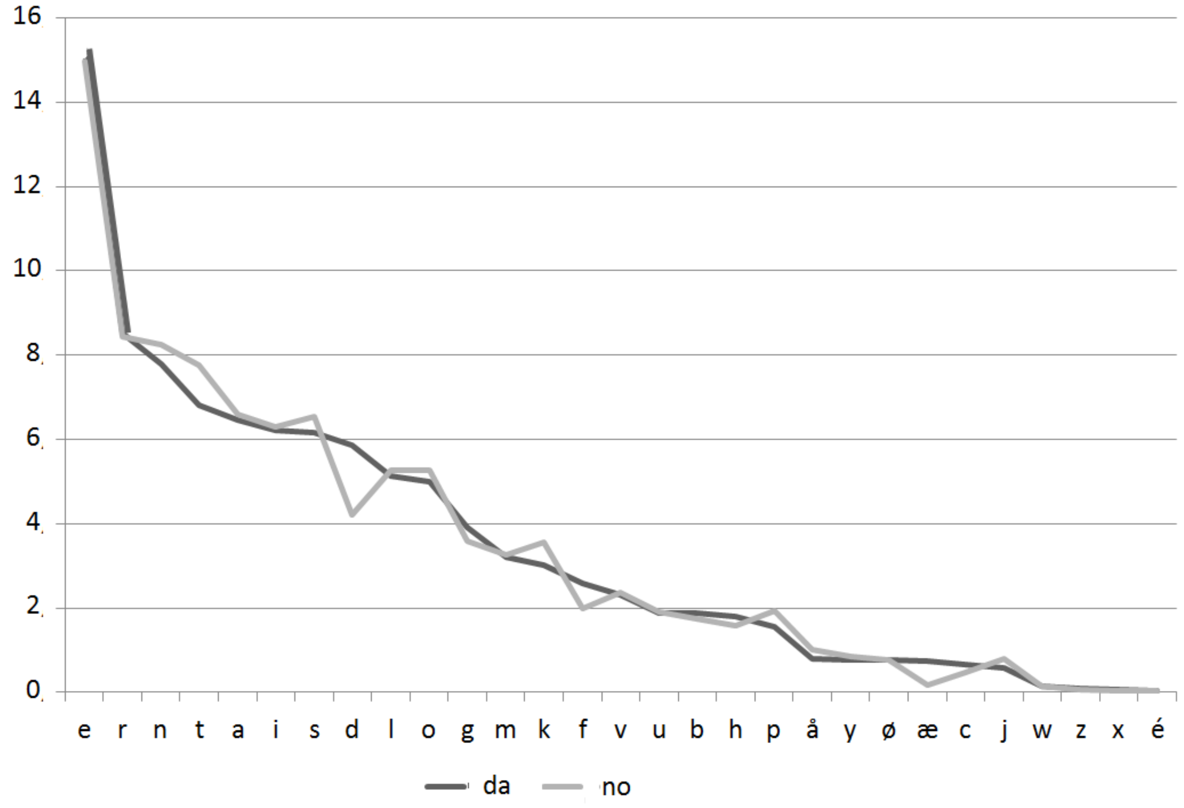

Figure 3. Comparison of letter frequencies in Danish and Norwegian languages

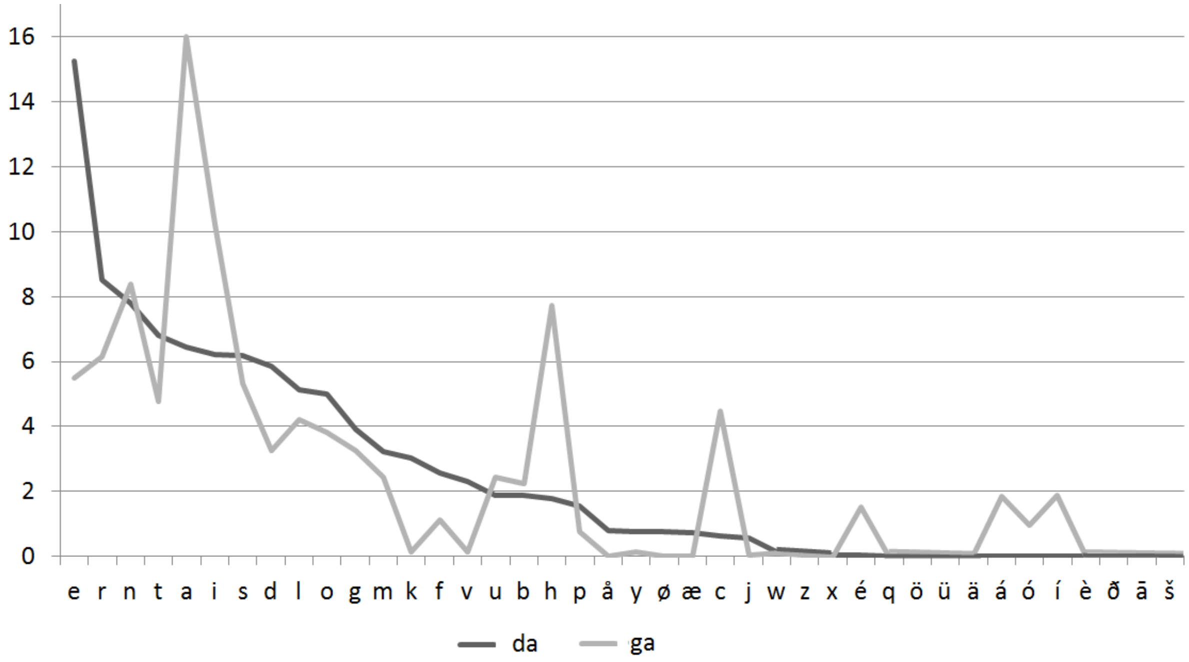

Figure 4. Comparison of Danish and Irish language unigrams

Strength of the correlation is usually described as follows:

CC = 0–0.2 very weak,

CC = 0.2–0.5 weak,

CC = 0.5–0.7 average,

CC = 0.7–1 strong.

According to this scale the correlations of all language pairs are strong. This is due to the similarity of the usage of letters in different languages. Therefore, in our case, it is reasonable to describe correlations ad hoc, with epithets only for limited range of correlations, for example:

CC = 0.74-0.80 very weak,

CC = 0.81-0.85 weak,

CC = 0.86-0.90 average,

CC = 0.91-0.95 strong,

CC = 0.96-0.99 very strong.

Let us list the language pairs with very strong mutual correlation:

CC = 99: da/no (Danish/Norwegian), hr/sl (Croatian/Slovenian);

CC = 98: cs/sk (Czech/Slovak), es/pt (Spanish/Portuguese);

CC = 97: es/fr (Spanish/French), es/it (Spanish/Italian);

CC = 96: da/de (Danish/German), et/fi (Estonian/Finnish), en/sv (English/Swedish), Slovak/Slovenian.

Spanish language belongs to three pairs of the strongest correlating languages, whereas Slovak, and Slovenian belong to two. By joining the Spanish language with all of its strongest correlation languages mentioned here, we can get a group of four languages: Spanish, Italian, Portuguese, French. An analogous group can be constituted with the Slovenian language: Slovenian, Croatian, Slovak.

This is a comparison of the alphabet letters by giving their frequencies corresponding weights. Therefore, the results should be more realistic and the optimisation of the letter layout on the keyboard may be based on them. The search for an optimal letter variant for every language by experiments is costly and time consuming. If the letter frequencies of the two languages are well correlated (CC is strong), it is likely that they will have similarities in their layout on the keyboard.

Most researches and practical works are conducted to establish a rational layout of the characters in the English language keyboard. The tablet’s thumbnail keyboards are becoming relevant. KALQ layout for the English language is noteworthy (Bi et al. 2012). The English language correlates the best with the Swedish language (0.96). Therefore, Swedish has more opportunities to take advantage of what is done for English. If analogous works were done for Spanish, then three languages (Italian, Portuguese and French) could use their results. Similarly, two languages (Croatian and Slovak) could use the results obtained for Slovenian language.

## 7. Discussion and Conclusions

1. The letter frequencies are taken from a source common to all languages (Wikipedia), which ensures possibility for an adequate comparison of the letter frequency characteristics of all the official languages of the European Union and three additional languages (Icelandic, Norwegian and Turkish) employed in the online space that use Latin script. The common alphabet of all the analysed languages has been comprised of 102 letters and it has been included in the calculations of every language.

2. The original Method of the Adjacent Letter Frequency Differences is presented, which helps to highlight similarities and differences among languages. The results of this work can benefit screen-keyboard designers with the aim the number of keys to be minimal, which is important for mobile devices with small screens.

3. For all the analysed languages frequency jumps range from 1 to 11 that correspond to the frequency differences from 2 and 12 times between the two adjacent letters sorted by frequency. The biggest jumps are established in the area of rarely used letters. They divide the letters into two distinct groups: 1) commonly used and 2) rarely used. It is reasonable that the clear difference between the groups should be clearly reflected in the typing convenience and speed: the letters of the first group must be in the foreground of the keyboard and typed directly with one keystroke, whereas the second group may be in the background and typed by a few keystrokes.

4. There are some devices, especially mobile, which do not comply with the previous recommendations of these conclusions: some letters, usually the language specific ones, are located in the background layout and rarely used letters are left in the foreground. Such irrational solution is revealed by the ratio of the frequencies of those letters. This ratio depends on the particular letters and can be very high, it can reach even a few hundred times. Thus, it deviates from the optimal solution by so many times.

5. The maximum frequency jumps of particular language define the stability of the alphabet of that language. The largest is for Spanish (11), while the following languages are: English (8), Italian (8) and Lithuanian (5).

6. The correlation between the frequencies of all the analysed languages is calculated. The maximum correlation coefficient 0.99 is obtained from two pairs of languages: Danish – Norwegian and Croatian – Slovenian. The frequency of the letters is one of the factors that determines the location of the keys for the language specific letters on the keyboard. A high correlation coefficient indicates that the layout of the letters on those keyboards can be close to each other, and if the optimal layout of one language is established, then similar layout may be considered for the other language of the pair or at least may be used as a starting point for its layout optimisation.

7. Using the results of this work it is recommended to conduct researches with similar groups of languages identified or individual languages of those groups in order to optimize the layout of the keyboard for all group languages and the typing convenience (speed) for each language separately.

We believe that it would be useful to use the statistics of the bigrams.

## References

Bi, X., Smith, B. A., & Zhai, S. (2012). Multilingual touchscreen keyboard design and optimization. *Human–Computer Interaction, 27*(4), 352-382.

Chun, C. (2015). A systematic technique of smart phone keyboard layout design for optimal text-messaging. *2015 IEEE International Conference on Consumer Electronics* (ICCE), Las Vegas, 9-12 Jan. 2015. https://doi.org/10.1109/ICCE.2015.7066545

Dagienė, V., Grigas, G., Jevsikova, T. (2010). *Programinės įrangos lokalizavimas* [Software Localisation. Monograph] // MII, 328 p.

ETSI ES 202 130 v.2.1.2 (2007). Character repertoires, orderings and assignments to the 12-key telephone keypad (for European languages and other languages used in Europe).

Everson, M. (2004). The Alphabets of Europe. Evertype. Retrieved April 20, 2017, from http://www.evertype.com/alphabets/

Grigas, G., & Juškevičienė, A. (2015). Raidžių dažnių lietuvių ir kitose kalbose, vartojančiose lotyniškus rašmenis, analizė [Letter Frequency Analysis of Lithuanian and Other Languages Using the Latin Alphabet]. *Santalka: Filologija, Edukologija, 23*(2), 81-91.

ISO 12199:2000 (2000). Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet.

MacKenzie, I. S., & Soukoreff, R. W. (2002). A model of two-thumb text entry. *Proceedings of Graphics Interface, 2002*, 117-124.

Oulasvirta, A., Reichel, A., Li, W., Zhang, Y., Bachynskyi, M., Vertanen, K., Kristensson, P. O. Improving two-thumb text entry on touchscreen devices. *CHI '13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France - April 27 - May 02, 2013, p. 2765-2774. https://doi.org/10.1145/2470654.2481383

Practical Cryptography (2017). Letter frequencies for various languages. Retrieved April 20, 2017, from http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/

Vrandešić, D. (2012). Letter frequency. Retrieved April 20, 2017, from http://simia.net/letters/unigrams.zip

Vrandešić, D.; Sorg, Philipp; Studier, Rudi (2011). Language Resources Extracted from Wikipedia. *Proceedings of the sixth international conference on Knowledge capture*, June 26-29, Baniff, Alberta, Canada, p. 153-160. https://doi.org/10.1145/1999676.1999703

## Copyrights