

Specification for a new Finnish Keyboard Layout

FINAL DRAFT 2006-06-20

This specification has been prepared by the Keyboard Working Group of the Finnish national “Kotoistus” Initiative under guidance from the Steering Group on Cultural Diversity Issues in ICT.

*

Introduction

This specification defines a new keyboard layout to replace the “de facto” layout that is presently in common use in both Finland and Sweden, in order to meet the growing national and European requirements for multilingual support. The earlier implementations have been severely restricted by the limitations imposed on the character repertoire and its size by the then prevalent character encoding schemes.

The implementation of the layout is done by software (“keyboard driver”). The physical keyboard doesn’t have to be affected, although additional engravings may preferably be added to identify certain additional capabilities.

This implementation-oriented specification is intended to be further developed into a more general specification for the same functionality (repertoire). The follow-up is planned to use a more generic, system independent terminology that is more in line with that of ISO/IEC 9995.

This specification is based on a proposal from November 2005. It is influenced by the feedback from the national public review on the “Kotoistus” web site and from the international IT industry, including the BOF discussion (on Design Principles for A Regional, Multilingual Keyboard) at the 29th Internationalization & Unicode Conference in March 2006. The end result has been further affected by the compromise solution that has been worked out together with the relevant Swedish standards committee, to facilitate in practice a choice between their new standard and ours. They both now work with the same engravings (defined as mandatory in the Swedish standard), in spite of their very different design principles. This has been a practical target partially due to the common cultural background, but also due to the fact that Finland is a relatively small market for keyboard hardware.

Highlights

The new general purpose keyboard layout extends the functionality of the presently common SWE/FIN keyboard layout considerably.

In spite of this, there are no disturbing changes to the users of the base functions of the regular (i.e., desktop) keyboards. In particular, the number and placement of keys is the same as in the full size QWERTY keyboards currently in use, and there is no change to the current allocation of keys.

The keyboard layout is designed to produce UCS/Unicode encoding. (The hexadecimal UCS code position is presented in the text as U+xxxx, and in Appendix 1 as xxxx, without the U+ prefix.)

The design is open-ended and based on intuitively recognizable positioning of additional letters. It can thus be learned with a minimum effort on an as-needed basis.

User Communities

The keyboard is intended for use by the speakers of the majority languages of Finland and Sweden. It can also be used by speakers of many other languages written using Latin letters, though less conveniently than using dedicated keyboards.

Basic Goals

The keyboard layout is designed to meet three important basic goals.

First, it provides for easy entering of text in the Fenno-Scandinavian majority languages, i.e. Finnish, Swedish, Norwegian, and Danish, using a familiar keyboard layout, with new capabilities, though, for the letters š, ž, æ and ø.

Second, it is designed to give an easy, intuitively recognizable way to enter text in all of the official Fenno-Scandinavian regional and minority languages, including those with the requirement for additional characters (i.e., for Finland, Sweden and Norway: Northern Sámi, Southern Sámi, Lule Sámi, Inari Sámi, Skolt Sámi, and Romany as spoken in Finland, and for Denmark: Faroese, Kalaallisut aka Greenlandic, and German).

As a third objective, it allows for relatively easy entering of particularly the names (of persons, places or products) with a variety of Latin letters used in other languages, such as the official languages of the European Union (though not, of course, including Greek) plus the remaining Nordic language, Icelandic.

The Latin characters in modern use of the MES-2 (Multilingual European Subset 2 of the UCS/Unicode) repertoire are the basis for this design. The needs of writing Vietnamese names have also been taken into account due to the size of that immigration population.

(The above mentioned characters plus others can presently be entered using various methods, such as selection from a table or entering the UCS code position. These methods, however, are generally more laborious than using the keyboard directly, and often they are also application dependent.)

It is also of particular importance to the end users that specific punctuation be provided to support the orthographies of the languages in question.

Presently, the basic ASCII quotation marks and hyphen-dash are often used by various applications (correctly or incorrectly) to produce localized punctuation. These characters are now augmented by new punctuation marks that produce the desired result independently of the application programs. With these additions, the keyboard supports much better the proper punctuation of particularly the Nordic languages and English or German.

In addition, the no-break space is provided as an easy to use tool for preventing unsuitable line breaks when texts are formatted.

Thus, there would be considerable justification to call the keyboard layout “Nordic International”.

The Means

To achieve the goals, room is needed for non-decomposable characters in intuitively recognizable

positions. For this purpose, almost all characters that are decomposable in UCS/Unicode terms are to be produced in the basic mode using a systematic approach utilizing dead keys to add diacritical marks. The letters å, ä, and ö are however kept in their currently allocated positions, for obvious reasons.

The new letters are positioned on the keys in such a way that their positions are intuitively recognizable even without new engravings. It is thus relatively easy to learn to use the additional capabilities that one needs. For diacritics, however, additional engravings are being proposed.

The keyboard layout in current use has two specific (dead) keys for diacritics. The new layout preserves these keys and provides for several additional diacritics. Because no new physical keys can be added, the new diacritics have to be added to present keys.

In addition to the original dead keys, several keys provide for diacritic functionality when used in conjunction with the AltGr key (and possibly a Shift key). For example, the apostrophe key is used for the caron when used together with the AltGr key. In the sequel, the term “diacritic key” refers to both the two diacritic keys proper and the diacritic key functionality of other keys.

N.B. For compatibility with the mandatory engravings of the Swedish standard, the dot above diacritic is assigned to the AltGr + Shift position in B10 (hyphen key) in addition to its more natural placement in B09 (dot key).

Compositions

The diacritic keys produce logically a combining diacritical mark as defined in UCS/Unicode.

The keyboard layout has two modes of operation, basic (default) mode and decomposed mode. The method of switching between these two modes is to be determined, and the decomposed mode does not necessarily have to be implemented in the first phase. Switching between these modes is expected to be rare, since most users will only need the capabilities provided by the basic mode.

In the basic mode, the diacritic combines with the base character that is typed next, i.e. after the diacritic, in a manner that is familiar to the users, and has been since the early days of mechanical typewriters. Thus, e.g., ´ and e will produce é. This dead key method produces a pre-composed character, like é (U+00E9), for the repertoire defined in Appendix 3. This implies no change to current functionality for Latin 1 letters, but it extends them in a well-defined, universal way.

Ultimately, there is a need to attach also multiple, different diacritical marks to the same base letter. Thus, e.g., u together with ¨ diaeresis and ˇ caron would result in ů (which exists in UCS/Unicode also as a pre-composed character U+01DA). Such functionality is needed for correct writing of e.g. Lithuanian and Vietnamese. From an implementation point of view, this particular character ů used as the example is produced by the specified keyboard driver only in the decomposed mode, i.e., as <U+0075,U+0308,U+030C>. All composite characters that cannot be encoded as pre-composed characters are encoded in the decomposed mode, which can also be used to produce the same characters that can be produced in the basic mode, though encoded differently. In the decomposed mode, the diacritical marks are keyed in after the base letter. The keyboard driver does not produce decomposed characters in a normalized form (here the NFC form would be <U+00FC,U+030C>); if necessary, some other software will be responsible for the normalization.

The stroke modifier key

The composition principle is extended to some characters that are not composite in the UCS/Unicode sense but can intuitively be seen as consisting of a basic Latin letter and a stroke,

such as đ. The stroke in such characters is not a diacritic, but in the proposed keyboard layout, a special modifier key is used to create a character with a stroke. Although the stroke is not a decomposable element of a character, the keyboard driver can handle it in a manner comparable to the use of dead keys for diacritical marks. This makes typing such letters easy and natural, as soon as the user has learned the simple principle.

This approach makes it possible to type several Sámi letters using the modifier key (with AltGr) and a basic letter key. The characters involved are đ - d with stroke (U+0111, in Northern, Inari and Skolt Sámi), ġ - g with stroke (U+01E5, in Skolt Sámi), and ƞ - t with stroke (U+0167, in Northern Sámi). – The stroke is also used to create e.g. ł - l with stroke (U+0142, in Polish) or ħ - h with stroke (U+0127, in Maltese), and can be used to write ø - o with stroke (U+00F8, also available directly on the ö key).

Since the stroke is not a decomposable element, all combinations of the modifier key and a letter that produce a character are to be specified separately. When used in any other than the intended combinations (defined in Appendix 3), the modifier key would have no effect or could produce some stroke-like character. The function of the stroke modifier key is identical independently of the mode, basic or decomposed.

In addition to the primary position for the stroke modifier key, a fallback position has been provided for use on those keyboards (mostly in some laptops) that don't have the primary position available.

The Changes

The specification changes the meaning of some key sequences in the basic mode, but this will probably not cause much difficulties. At present, pressing e.g. the circumflex dead key and the c key produces ^c, whereas in the new scheme, it produces ĉ. The sequence ^c would need to be produced by pressing the space key between the circumflex key and the c key. This is how users generally behave anyway even now. The sequence ~c, on the other hand, will be produced by pressing the tilde dead key followed by the c key with or without a space in between (because the combination has not been specified in Appendix 3). If the diacritic tilde is to be combined with the letter c, this can be done in the decomposed mode.

The additional letters and diacritical marks are shown on the layout.

Open-endedness

The layout is open ended. Only a minimum repertoire is defined.

In the basic mode, each of the characters defined in Appendices 1 and 3 is being encoded using one code position. In the decomposed mode, the keyboard driver cannot check the applicability of any characters with combining diacritical marks.

The responsibility for checking the acceptability of any resulting character belongs first to the user and ultimately to the application (which may set its own limitations to the repertoire – as always, since e.g. names in a population registry may not contain otherwise perfectly valid characters such as digits).

For the sake of consistent application of the design principle, no other decomposable letters than å, ä and ö have their own key position.

In Appendix 1, the specific new function (when compared with the common keyboard layout in current use) of each key is discussed.

Appendix 2 shows the proposed keyboard layout.

Appendix 3 defines the repertoire required for pre-composed encoding with dead letter key input.

Appendix 1:

The functions of the various data keys are as follows:

<i>Pos</i>	<i>Base function</i>	<i>+ Shift</i>	<i>Alt Gr function</i>	<i>Alt Gr + Shift</i>	<i>Comment</i>
E00	§ (00A7)	½ (00BD)	Modifier: stroke		Added stroke modifier. If E00 not available, see C09!
E01	1 (0031)	! (0021)		¡ (00A1)	Added inverted exclamation mark.
E02	2 (0032)	" (0022)	@ (0040)	” (201D)	Added quotation mark.
E03	3 (0033)	# (0023)	£ (00A3)	» (00BB)	Added quotation mark.
E04	4 (0034)	¤ (00A4)	\$ (0024)	« (00AB)	Added quotation mark.
E05	5 (0035)	% (0025)	‰ (2030)	“ (201C)	Added per mille sign + quotation mark.
E06	6 (0036)	& (0026)	, (201A)	„ (201E)	Added low quotation marks.
E07	7 (0037)	/ (002F)	{ (007B)		
E08	8 (0038)	((0028)	[(005B)		
E09	9 (0039)) (0029)] (005D)		
E10	0 (0030)	= (003D)	} (007D9)	° (00B0)	Added degree sign.
E11	+ (002B)	? (003F)	\ (005C)	¿ (00BF)	Added inverted question mark.
E12	Cd: ´ (0301) acute	Cd: ` (0300) grave	Cd: ¸ (0327) cedilla	Cd: ˙ (0328) ogonek	Added diacritics.
D01	q (0071)	Q (0051)			
D02	w (0077)	W (0057)			
D03	e (0065)	E (0045)	€ (20AC)		
D04	r (0072)	R (0052)			
D05	t (0074)	T (0054)	þ (00FE)	Þ (00DE)	Added letter thorn.
D06	y (0079)	Y (0059)			
D07	u (0075)	U (0055)			
D08	i (0069)	I (0049)	ı (0131)		Added letter dotless i.
D09	o (006F)	O (004F)	œ (0153)	Œ (0152)	Added oe ligature.
D10	p (0070)	P (0050)	Cd: ˆ (031B) horn	Cd: ˆ (0309) hook above	Added diacritics (esp. for Vietnamese).
D11	å (00E5)	Å (00C5)	Cd: ˆ (030B) double acute	Cd: ˆ (030A) ring above	Added diacritics.
D12	Cd: ¨ (0308) diaeresis	Cd: ^ (0302) circumflex	Cd: ~ (0303) tilde	Cd: ¯ (0304) macron	Added diacritic.
C01	a (0061)	A (0041)	ə (0259)	Ə (018F)	Added letter schwa.

C02	s (0073)	S (0053)	ß (00DF)		Added letter sharp s.
C03	d (0064)	D (0044)	ð (00F0)	Ð (00D0)	Added letter eth.
C04	f (0066)	F (0046)			
C05	g (0067)	G (0047)			
C06	h (0068)	H (0048)			
C07	j (006A)	J (004A)			
C08	k (006B)	K (004B)	ƙ (0138)		Added letter kra.
C09	l (006C)	L (004C)	Modifier: stroke		Added fallback position for stroke modifier.
C10	ö (00F6)	Ö (00D6)	ø (00F8)	Ø (00D8)	Added letter o with stroke.
C11	ä (00E4)	Ä (00C4)	æ (00E6)	Æ (00C6)	Added letter ae.
C12	' (0027)	* (002A)	Cd: ˇ (030C) caron	Cd: ˘ (0306) breve	Added diacritics.
B00	< (003C)	> (003E)	(007C)		
B01	z (007A)	Z (005A)	Ʒ (0292)	Ʒ (01B7)	Added letter ezh.
B02	x (0078)	X (0058)	× (00D7)	• (00B7)	Added multiplication sign and middle dot.
B03	c (0063)	C (0043)			
B04	v (0076)	V (0056)			
B05	b (0062)	B (0042)			
B06	n (006E)	N (004E)	ŋ (014B)	ŋ (014A)	Added letter eng.
B07	m (006D)	M (004D)	μ (00B5)	— (2014)	Added micro sign and em dash.
B08	, (002C)	; (003B)	' (2019)	‘ (2018)	Added single quotation marks.
B09	. (002E)	: (003A)	Cd: (0323) dot below	Cd: ˙ (0307) dot above	Added diacritics.
B10	- (002D)	_ (005F)	– (2013)	Cd: ˙ (0307) dot above	Added en dash + diacritic.
A	Space (0020)		NBSP (00A0)		Added no-break space.

N.B. The layout intentionally leaves many positions unassigned, even though rather natural allocations could be presented. The need for additional assignments varies by the user, by the nature of the text, and by the program. Thus, they are best left unassigned in the basic layout, so that there is more room for varying useful customizations.