

Functional Multilingual Extensions to European Keyboard Layouts

This is an initial draft for the CWA.

FOREWORD

The production of this document which addresses the considerations and guidance for functional multilingual extensions to European keyboard layouts, primarily for the users of the Latin script, was agreed by the CEN/ISSS Workshop Functional Multilingual Extensions to European Keyboard Layouts (WS/MEEK) in the Workshop's Kick-Off meeting on 2008-01-25.

The document has been developed through the collaboration of a number of contributing partners in WS/MEEK. WS/MEEK representation gathers a wide mix of interests, coming from academia, public administrations, IT-suppliers, and other interested experts. The present CWA (CEN Workshop Agreement) has received the support of representatives of each of these sectors. A list of experts who have supported the document's contents may be obtained from the CEN/ISSS Secretariat. The final WS/MEEK internal review/endorsement round of this CWA started on 2009-mm-dd and was successfully closed on 2009-mm-dd. In addition, the CWA has been the subject of a public comment period from 2009-mm-dd to 2009-mm-dd. The final text of this CWA was submitted to CEN for publication on 2009-mm-dd.

The CEN Workshop Agreement has only been made in English.

SCOPE

This CWA is aimed to assist in the preparation of functional multilingual extensions to European keyboard layouts. They are aimed to allow “ordinary users” and Public Authorities to input primarily Latin-script characters, especially in light of current and potential future legal and common educational requirements. The CWA builds on existing implementations of common official and de facto standards for national and regional computer keyboard layouts and related input methods. Based on this, recommendations and guidance are given on handling multilingual data entry requirements, taking account of existing international standards in the field.

Since the European single market allows for free movement of people and goods, one should be able to correctly enter the names of people, places, products, and companies and other legal entities in a consistent, easily comprehensible manner, which implies that the capability should exist in all kinds of applications, including those based on open source software that are traditionally based on freely available information.

The CWA does not define any specific, let alone a Pan-European keyboard layout. Liaison has been sought with ISO/IEC JTC1/SC35 to minimize the risk of further divergence between actual implementations and formal standards (ISO/IEC 9995 series and other relevant standards). Liaison has also been sought with the Unicode Consortium.

INTRODUCTION

Year 2008 has been proclaimed the International Year of Languages by the United Nations General Assembly. It has also been declared the European Year of Intercultural Dialogue by the European Parliament and the Council of the European Union. The MEEK Workshop can be seen as a small step in support of these themes.

The European Union has currently 23 official languages: Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, and Swedish. In addition, many more native and immigrant languages are spoken in Europe.

Of the regional languages, only Catalan/Valencian, Basque and Galician of Spain have a degree of official EU status. Luxembourgish, the national language of Luxembourg, and the other regional and minority languages have not been put forward for official EU language status by the governments of the respective Member States.

Of the official EU languages, Bulgarian is written with Cyrillic letters and Greek with Greek letters, whereas all the others are written with Latin letters. The Latin letter repertoire includes several letters beyond the “basic a-to-z” and a very large number of letters with diacritics, some of which are perceived as basic letters in one or more languages, and also collated separately (e.g., å, ä and ö after z in Swedish and Finnish).

BACKGROUND FOR THE CURRENT SITUATION

In the past, IT technology has imposed severe restrictions on the character repertoires to be processed at once by implementations of character encoding schemes that utilize 8 bits at most (which only allow 256 bit combinations, of which some are reserved for control characters). As a consequence, each country has historically concentrated on having its own official languages properly supported in ICT, more recently also its own regional and minority languages.

The expanding European single market, however, with free movement of goods and people, is drastically changing the support requirements. For both accuracy and politeness, all kinds of proper names should be communicated in their correct form. This change is driven not only by increasing business and legal reasons but also as a personal choice by many individuals in appreciation of cultural diversity.

FORMAL TREATIES

The more formal reasons for expanding the character repertoire stem from treaties such as:

1. The Lisbon treaty. *** To be expanded. ***

2. UNESCO Convention on the Protection and Promotion of the Diversity of Cultural Expressions (approved in October 2005, entered into force in March 2007), see <http://ec.europa.eu/culture/portal/action/diversity/pdf/st08668.EN06.pdf>.

- to protect and promote the diversity of cultural expressions
- to create the conditions for cultures to flourish and to freely interact in a mutually beneficial manner
- to encourage dialogue among cultures with a view to ensuring wider and balanced cultural exchanges...
- to foster interculturality...
- to promote respect for the diversity of cultural expressions and raise awareness of its value at the local, national and international levels
- to reaffirm the importance of the link between culture and development for all countries...
- to give recognition to the distinctive nature of cultural activities, goods and services as vehicles of identity, values and meaning
- to reaffirm the sovereign rights of States to maintain, adopt and implement policies and measures that they deem appropriate for the protection and promotion of the diversity of cultural expressions on their territory
- to strengthen international cooperation and solidarity in a spirit of partnership...

3. European Charter for Regional or Minority Languages (signed in 1992, in force since 1996) at <http://conventions.coe.int/Treaty/en/Treaties/Html/148.htm>.

- Concentrates on languages traditionally used in a state by a minority and not an official language of that state.
- Ratification of this Council of Europe treaty (by the national parliaments) entails certain responsibilities, details of which are specified upon ratification.
- the facilitation and/or encouragement of the use of regional or minority languages, in speech and writing, in public and private life (§7.1.d)
- the provision of appropriate forms and means for the teaching and study of regional or minority languages at all appropriate stages (§7.1.f)
- to ensure that users of regional or minority languages may submit oral or written applications [to public authorities] and receive a reply in these languages (§10.1.a.iii)
- to ensure that users of regional or minority languages may validly submit a document in these languages (§10.1.a.v)
- to allow the administrative authorities to draft documents in a regional or minority language (§10.1.c)

4. On the recording of surnames and forenames in civil status registers, Convention relative à l'indication des noms et prénoms dans les registres de l'État Civil, CIEC 14 (signed in 1973 in the context of the International Commission on Civil Status, the Commission Internationale de l'État Civil; some member states have not signed it and some signatories have not ratified it, though). The official French text and the unofficial English translation are available at <http://www.ciec1.org/ListeConventions.htm>.

Spelling of names (Article 2):

- Where a record is to be made in a civil register by an authority of a Contracting State and there is produced for that purpose a copy of or extract from a civil status record or some other document that shows the surnames and forenames in the same characters as those used in the language in which the record is to be made, those surnames and forenames shall be reproduced literally without alteration or translation.
- Any diacritic marks forming part of such surnames and forenames shall also be reproduced, even if such marks do not exist in the language in which the record is to be made.
- Where a record is to be made in a civil status register by an authority of a Contracting State and there is produced for that purpose a copy of or extract from a civil status record or some other document that shows the surnames and forenames in characters other than those used in the language in which the record is to be made, those surnames and forenames shall be reproduced as far as possible by transliteration, without being translated.
- If there are standards recommended by the International Organisation for Standardisation (ISO), they shall be applied.

In practice, this would mean that letters such as the Danish (also Norwegian, Icelandic, Faroese, and Kalaallisut) letter AE (æ, which is an ae-ligature in some European languages), or the Icelandic THORN (þ->th) and (also Faroese) ETH (ð->d), and several other characters would not have to be supported in several other countries as such. It would appear from the timing of this agreement that a stroke should probably be treated as a diacritic.

5. German-Polish Treaty on good neighbourly relations and friendly cooperation (of 17 June 1991) at http://untreaty.un.org/unts/60001_120000/28/1/00054050.pdf.

Art. 20 § 3 ... shall in particular have the right, individually or in community with other members of their group:

- To use their mother tongue freely, both privately and in public, to disseminate and exchange information in it and to have access to such information; ...
- To express their given and family names in the mother tongue form; ...

§ 4 ... membership in the groups referred to in § 1 ... may not place him at any disadvantage.

CHANGING ENVIRONMENT

Most of the keyboard layouts that are presently in common use in Europe have been designed at a time when the number of different characters that could be

entered was severely limited by the then prevalent character encoding schemes. In addition, some limitations of the original mechanical typewriters have been carried over into their design. These keyboards have been designed for each local language environment, although often with some capability to enter “foreign” letters.

The advent of the Universal Character Set (ISO/IEC 10646 and Unicode) has fortunately removed the limitations imposed by the ICT systems on the number of different characters to be handled, i.e., processed and rendered, at once. Consequently, one could expect that any EU wide solution for interoperable public registers is likely to follow the lines of CIEC 14, which isn’t currently being enforced. Also, some countries are already committed to using the proper names in their formally recognized minority languages as a result of the European Charter on Regional or Minority Languages.

What remains to be done is to provide for a way to enter the correct data. However, there is no real possibility to start from scratch, or switch the base completely to formal international standards, such as the current ISO/IEC 9995-3 for multilingual keyboard layouts, which at present does not appear to provide an evolutionary, user friendly base for these extensions. This assertion is supported by the formally recognized need to both revise this standard and also functionally replace it by a new part of ISO/IEC 9995. The new development project in ISO/IEC JTC1/SC35 has the long term potential to become the basis for new national standards for and implementations of multilingual keyboard layouts. In the meantime, however, there is an urgent need to satisfy the current needs.

A PRAGMATIC APPROACH

Most users are familiar with one keyboard layout only, and they wouldn’t welcome any disturbing changes to it. Thus, any expansion should preferably be intuitive and, as such, easily comprehensible and carefully based on current layouts.

One approach, which has made into a national standard and several experimental implementations, stems from the fact that many European users of other than UK keyboards are familiar with the dead key method for keying “foreign” letters with

diacritics. In this method, the key for the required diacritical mark is pressed first, followed by the base letter with which the diacritic is to be combined. This method has been in wide use since early mechanical typewriters, and there could be considerable benefit in continuing its use and expanding it to cover essentially all letters with single diacritics. It should be noted, though, that most users consider as base characters those letters of their own alphabet that have diacritics, and thus cannot be asked to utilize any other method than a single keystroke to enter them on their national keyboards. Such letters, e.g., å, ä and ö in Swedish and Finnish, ñ in Spanish, etc. will thus have to continue to appear directly on their respective keyboards, although they would be created stepwise, as combinations, on most other keyboards.

Although one could envisage the dead letter key method to be extended to entering letters with multiple diacritics, this is not generally supported by the current table driven methods that the industry employs to create keyboard drivers (e.g., MSKLC by Microsoft). This, however, could well be the method of choice for many among the few users who need to enter characters with multiple diacritics. Several of the letters with multiple diacritics (e.g., many of the Lithuanian accented letters) cannot be encoded as pre-composed characters, and they'll thus have to be entered as decomposed characters. They could also be entered using such an extended dead letter key method as characters that could not be explicitly listed as being supported.

Another method to enter composite characters as decomposed characters would be to enter the base letter followed by the combining diacritical marks. This method could be more suitable for those users who are not familiar with the dead letter key method. It could also be easier and possibly faster to implement than the support for multiple diacritics using the dead letter key method. It would not, however, produce pre-composed characters.

If both of these modes of operation are to be supported, it would make sense to implement them as technically two separate keyboard drivers utilizing the same keyboard layout.

The underlying principle for the design approaches described herein is not to impose any disturbing changes to those users who are content with current

implementations. Those who need additional capabilities, could easily choose between the two keyboard drivers, as appropriate. It should be noted that for ease of processing, a facility to normalize the decomposed input characters will be needed, but this is probably best to implement outside the keyboard driver itself.

The dead key method could, and probably should, also be extended to letters with diacritic-like additions, specifically to characters with a stroke (although they are not decomposable). This would help avoid having to assign key positions to each of the various letters, such as the Polish L WITH STROKE [ł] or the Sámi T WITH STROKE [t̥]. In addition to keeping key positions available for other use, this would also meet the criteria that the method and the positioning of the characters should be intuitive.

For ease of use, the diacritical marks should be engraved on the key tops, since they can hardly be placed following any intuitively recognizable scheme. If properly placed, the other additional characters don't necessarily need to be engraved.

It should be noted that the intuitive position for a character is not absolute. A user who knows the pronunciation of a character will likely think that its intuitive position is on a key which corresponds to its pronunciation. However, a user who is not familiar with the character is likely to think that the intuitive position is on a key which matches its visual appearance. A good example is the Kalaallisut or Greenlandic letter KRA (κ) with an apparent similarity to letter K but with a pronunciation similar to letter Q.

USAGE SCENARIOS

There can be found three basic scenarios in which users have a need to enter characters which are not directly provided for by the present keyboards.

(1) The first scenario involves a user who needs to enter characters from one or more foreign languages, using his or her own national keyboard. An example would be a clerk entering a foreign name.

(2) In the second scenario, a user needs to enter foreign characters on a foreign keyboard. An example would be a clerk in a multilingual work environment, who

needs to enter foreign names, and must also be able to move from one workstation (and keyboard) to another.

(3) The third scenario applies to users who visit Internet cafes, or who must work in a foreign office, or use a foreign keyboard for some other reason. The scenario also applies to multilingual users. The common factors are that these users need to produce text for some language which is not supported by the keyboard they are using, and they need to enter the characters frequently.

Scenarios one and two should be solved with the same mechanism. In both scenarios, the user must enter characters from a vast array of possible characters. The characters can be alien to the user.

(The user will need visual aids for locating and selecting the correct characters. Visually similar characters should be arranged next to each other, so the user can pick the correct one. There should also be textual aids to help in the selection process, so the user can make the correct decision, e.g., information on the languages and contexts where the character is used. These visual aids and documentation are outside the scope of this document.)

In scenario 3, users wish to enter text in a single language, other than the one the keyboard was designed for. Typically, users will frequently need to enter certain letters of their own alphabet and that small set of special characters that are used in e-mail addresses and website URLs (e.g. ~). The repertoire of the needed characters is probably limited to 1-6 characters (e.g. a Finnish user on a French keyboard should get by with finding the characters ä, ö, @ and ~).

Users should ideally be able to enter the characters using the same method, regardless of which foreign keyboard is being used. This means that once users learn how to enter their own language's characters on one foreign keyboard, they can use the skill on all foreign keyboards. It is important that as few key strokes as possible are required for entering text, as the user will be needing them often.

The solution should not rely on the availability of an IM for the targeted language always being available on the system. The IMs that are used should be designed for people who are not touch typists, and therefore rely on keyboard engravings. After all, touch typists could just switch to their national layout on any keyboard. Since

they don't need the engravings, they can use any keyboard the same, regardless of the engravings.

It should also be generally noted that there is a difference between users who are touch typists, and those who need to look at the engravings on the keyboard when typing. As an example, from a touch typist's perspective, a letter positioned under the letter "y" on a qwerty keyboard should be under the letter "z" on a qwertz keyboard. From a non-touch typist's perspective, the letter should be under "y," even though "y" is positioned elsewhere on the keyboard. This document will give priority to those users who rely on keyboard engravings. As a related matter, it should be noted that it is impossible to create a mechanism through which it would be totally transparent for users to switch between keyboards. To use the qwerty/qwertz issue as an example, because the position of the letter "y" on qwerty and qwertz keyboards is different, even though uniform positions for keys were to be specified, if such positions involved the "y" key, users switching between qwerty and qwertz keyboards would still find themselves looking for the "y" key.

Sample design and implementation

The details pertaining to the "pragmatic" design principles for what is described above are the result of the discussions at the BOF session on the Design Principles for a Regional, Multilingual Keyboard at the 29th IUC (Internationalization and Unicode Conference) in San Francisco on 6 March 2006 (see: http://www.unicodeconference.org/iuc29/program-d.htm#BOF_3).

A sample specification (for particularly Finnish and Swedish and the recognized minority languages in the Nordic countries plus the official EU languages using the Latin script) was published in Finland in November 2008 as SFS 5966, Keyboard layout: Finnish-Swedish multilingual keyboard setting. In addition, this layout also provides for specific punctuation used in a number of European languages. Since

2006 there had been available at least three different evaluation or trial versions implementing a near final basic (pre-composed) mode of operation, two for the Windows environment (of which one was a formal evaluation version by Microsoft) and one for the Linux environment. To illustrate its use, Appendix 1 contains an informal translation of the text of a quick users' guide produced by SFS.

For comparison, a keyboard layout intended for governmental use with a similar multilingual repertoire had been published in Sweden as SS 662241/T1:2006. The approach taken there is to distribute a large number of characters with diacritics or stroke over practically all available key positions.

It should be noted that the requirement to cover the characters of the local minority languages, which the Finnish standard responds to, is of course dependent on the region.

A proposed design for Scenario 3 relies on online access to IM configuration files that are not already installed on the system. Each of these IM files relates to the user's requirements for their own needed characters and the specific keyboard design that is being used.

A library of several of these IM configuration files could be accessible, with many "standard" versions that provide a mapping of the characters that are available on the national keyboard (e.g. the characters that the CLDR lists as applying to that locale) to the keyboard being used. E.g. one could provide a mapping of the keys needed by a typical Finnish user to the keys of a French keyboard.

In addition to a library of such default common cross-keyboard IM (configuration) files, the proposed solution describes a very simple mechanism that would permit people to create their own personalized IM files using a simple user interface that could be installed on all PCs for public use. These IM files could be created in advance for later use or could be created on-the-fly when first encountering a new keyboard and stored online for subsequent re-use. These personalized IM files could contain mappings for characters that are of particular interest to that specific user and could be located in a position that the user prefers, rather than in the default position that would exist in the "standard" IM files that were available.

Summary notes on the solutions

The SFS 5966 keyboard is a good sample solution for user case 1, for Finland. User case 1 can be solved for other countries using similar methods.

User case 2 could have an “ideal” solution only if all countries create an international keyboard layout solution which is based on a common model. If a common set of rules is followed, only those characters which are specific to the national keyboard will be different on each keyboard. All other keys should be positioned using a uniform layout. In the end, we would have a common way for producing some finite set of characters. That set could be discovered only after every layout has been designed. Once the layouts would have been designed, we would know the shared set. Every keyboard layout should then be provided with extra documentation for producing the set of characters which fall outside the shared set. Once this is done, any user anywhere in Europe could sit at any keyboard following the shared layout, and just look up those keys which differ between his/her national keyboard layout and the one she/he is sitting in front of at the moment.

In other words, every country in Europe would have to follow a common set of guidelines for designing their international keyboard layout which can be used to generate those characters which fall outside their national set of characters. This means that user case 2 could not be solved immediately using the SFS 5966 solution. However, if countries would adopt the solution, user case 2 would gradually get solved.

User case 3 cannot be solved satisfactorily with the SFS 5966 solution, as frequent use of certain characters requires that those characters be produced with a single key press (with at most one modifier key in addition to the shift key). Multiple key presses to generate one character, except in rare cases of characters with multiple diacritics, are not an acceptable solution for a user who needs to enter large volumes of text.

DESCRIPTION OF ALL THREE SOLUTIONS:

SFS solution: All base characters (without diacritics) can be found in levels 1-4 (with stroke-modified letters considered having a diacritic). All letters with a diacritic are entered by first entering a diacritic, which is behind one key in level 1-4, and then a base letter. In addition, nationally defined characters with a diacritic can have their own position so that they can be produced with a single key in levels 1-4. Covers

most symbols commonly used in Europe. Additional work is required to cover all needed symbols, even if the work ends up concluding that all symbols are covered.

Global IM solution: Covers all characters, and even non-character strings. No implementation is currently available, but due to simplicity of implementation, one could be achieved in a short amount of time. The advantage is that all characters can be entered with a single keystroke in levels 1-4, even characters with a diacritic (even a limited number of characters with multiple diacritics). The disadvantage is that users must always select a specific keyboard layout before they start typing (for sometimes a very small amount of text). This burden can be reduced by using a simple interface for selecting the layout.

ISO/IEC-9995-3 type solution: Covers all Latin characters, not just those in use in Europe. Includes support for Cyrillic, Greek and other scripts, and symbols. A single universal means for entering all Latin characters. The big advantage is that once a user learns the method, it will work the same on all keyboards. Disadvantages are: (a) Each "international" character (characters outside "a-zA-Z0-9,-") requires multiple keystrokes. (b) Each character's sequence must be memorized, which can be overly difficult for ordinary, casual users. No implementation is currently available. Implementation work cannot probably start before the method has been finalized.

Further considerations

One should not ignore the ramifications of the requirements by the Greeks (users of the Greek script) and Bulgarians (users of the Cyrillic script) to enter Latin characters and vice-versa, although the details thereof are clearly not within the scope of this workshop. It should also be noted that new member states may use the Cyrillic script in a more complex form.

Whenever a keyboard layout is being worked on, a requirement is often brought up to revamp the layout and base it on the Dvorak keyboard. Any such activity would clearly not be within the scope of this Workshop. Additionally, although the Dvorak keyboard is based specifically on the English language, it hasn't gained wide acceptance among the English speaking users in spite of it having been available

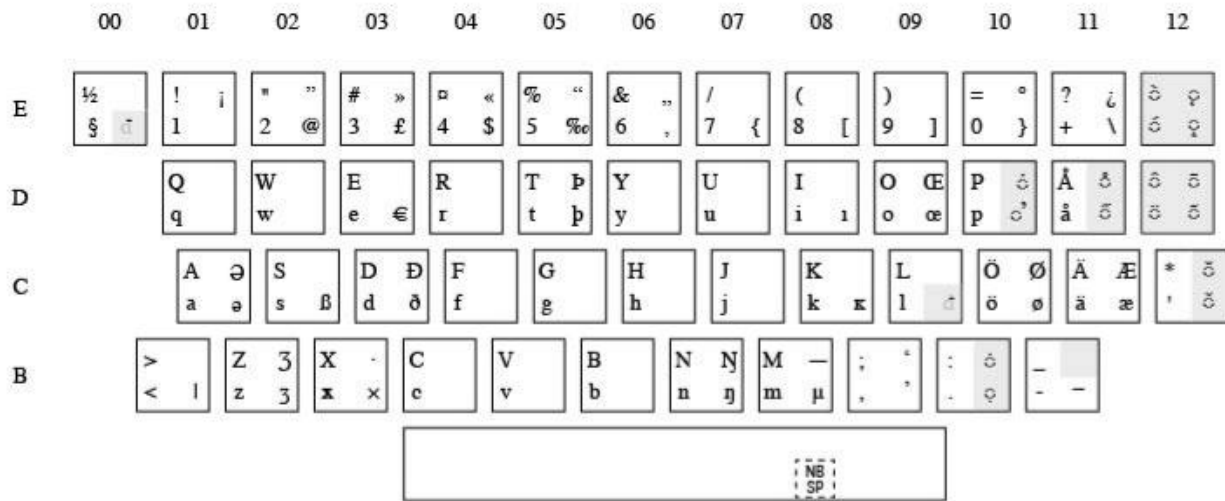
since pre-WWII time. Furthermore, extensive analysis would be required to come up with a specific layout for any other language.

Another sometimes stated requirement is the design for a Pan-European keyboard layout. Although this might make sense as a possibly more user friendly alternative to the US International layout, it is clearly outside the scope of this Workshop.

Other methods include special mnemonics for use in sequences to create characters with diacritics. These have been used widely to represent non-ASCII characters in system tables that had to be defined in ASCII, particularly in the UNIX environment, but they cannot be characterized as being easily comprehensible to non-technical users.

Appendix 1

A quick guide to the basic setting of the keyboard layout standard SFS 5966



Character layout (combining characters in grey)

The Finnish-Swedish general purpose keyboard layout has been enhanced with new capabilities. Its prior functionality, however, is not affected except in rare special situations. The new functionality is typically effected by pressing the AltGr key, situated to the right of the space bar.

Additions to the layout include several letters, punctuation marks and special characters used in many European languages. The goal has been to place them on intuitively recognizable key positions.

The diacritical marks are added to the base letters utilizing the so called dead letter key method. In this method, the diacritical mark is being entered ahead of the letter to which it will be combined. While all diacritical marks used to be positioned on the two specific diacritical keys, new diacritical marks have been added to other keys as well.

Also the letters with a stroke are being produced utilizing the dead letter key method as if the stroke were a diacritical mark.

Additional letters

The new letters are being produced by pressing the AltGr key together with the appropriate letter key. The capital letters are being produced by pressing also the shift key.

Key	Resulting letter	Glyph	Exemplar language
letter A	small and capital letter schwa	[ə Ə]	(Azerbaijani)
letter D	small and capital letter eth	[ð Ð]	Icelandic

letter I	small letter dotless i	[ı]	Turkish
letter K	small letter kra	[κ]	Kalaallisut
letter N	small and capital letter eng	[ŋ Ɲ]	Northern Sámi
letter O	small and capital ligature oe	[œ œ]	French
letter S	small letter sharp s	[ß]	German
letter T	small and capital letter thorn	[þ Þ]	Icelandic
letter Z	small and capital letter ezh	[ƶ Ʒ]	Skolt Sámi
letter Ä	small and capital letter ae	[æ Æ]	Danish
letter Ö	small and capital letter o with stroke	[ø Ø]	Danish

Letters with stroke

A stroke will be produced by pressing simultaneously the AltGr and § keys. Instead of the § key, also the L key may be used, since many laptops don't have the § key at all. The stroke will be followed by the letter to which it will be attached. For capital letters, the shift key will be pressed together with the letter key.

The stroke can be added to the following letters (with the exemplar language in parenthesis):

letter D [ď Ď] (Northern Sámi), letter G [ǥ Ğ] (Skolt Sámi), letter H [ħ Ħ] (Maltese), letter L [ł Ł] (Polish), letter O [ø Ø] (Danish) and letter T [ṭ Ṭ] (Northern Sámi).

Additional special characters

Key	Position	Resulting character	Glyph
# 5	AltGr	per mille sign	[‰]
# 0	AltGr+shift	degree sign	[°]
letter E	AltGr	euro sign	[€]
letter M	AltGr	micro sign	[μ]
letter X	AltGr	multiplication sign	[×
letter X	AltGr+shift	middle dot	[·]

Additional punctuation marks

Key	Position	Resulting character	Glyph	Exemplar language
# 1	AltGr+shift	inverted exclamation mark	[¡]	Spanish
# 2	AltGr+shift	quotation mark	[”]	Finnish
# 3	AltGr+shift	double angle quotation mark	[»]	Finnish
# 4	AltGr+shift	double angle quotation mark	[«]	Danish
# 5	AltGr+shift	quotation mark	[“]	Englishi
# 6	AltGr	single low quotation mark	[,]	German

In addition, the following position is reserved for use to support Romanian:

Comma below Hyphen-m [-] AltGr+shift ș Ș and ț Ț

END