Using non-Latin alphabets in Blaise

Rob Groeneveld, Statistics Netherlands

1. Basic techniques with fonts

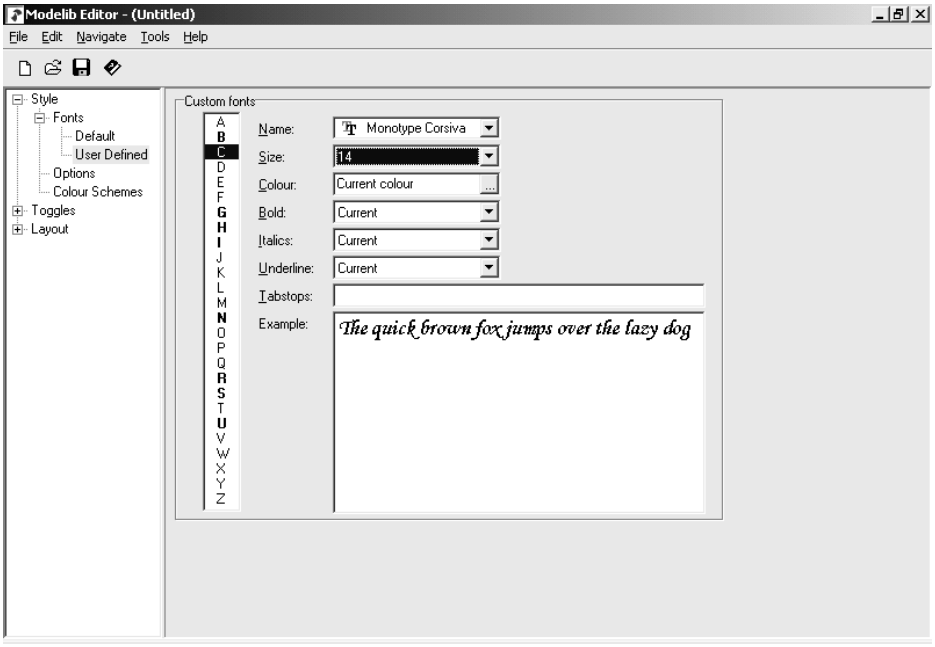
In the Data Entry Program in Blaise, it is possible to use different fonts. Here, we show an example with one font. From the Control Centre in Blaise, open the Modelib Editor:

Tools | Modelib Editor | File | New

Now choose:

Style | Fonts | User Defined

Select a free letter, for example 'C'. We can associate a font with this letter, e.g., Monotype Corsiva. We also change the size to let's say 14. The Mode Library Editor screen will look like this:



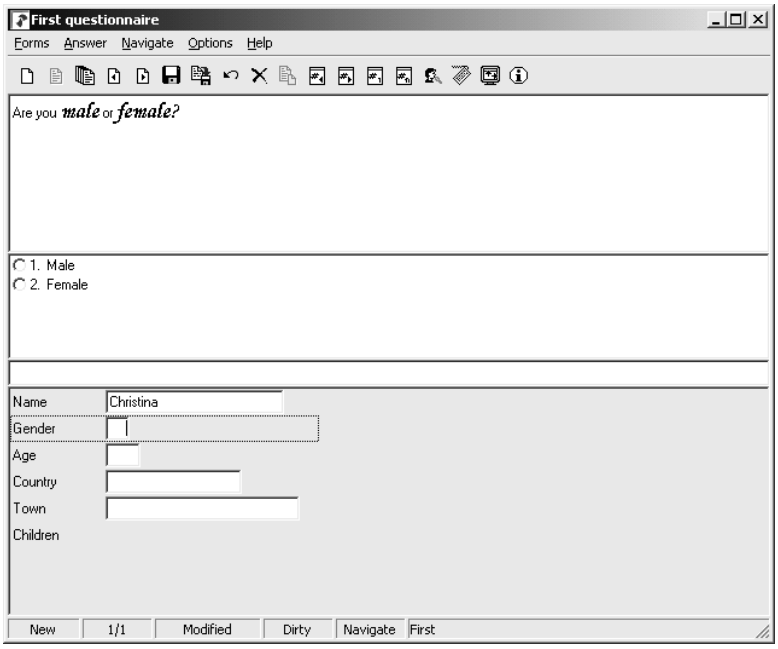
Now we can use the font character C in the text for the questions in a Blaise questionnaire. A question about the gender of the respondent could be, for example:

Gender "Are you male or female?": (Male, Female)

If we change this into

Gender "Are you @Cmale@C or @Cfemale?@C": (Male, Female)

the words 'male' and 'female' will be shown in the font represented by the letter C. Running this instrument shows the following data entry screen:



The words ‘male’ and ‘female’ now appear in Monotype Corsiva. Languages like English, French, Dutch, German, Swedish, Italian, Spanish, Indonesian, Turkish and a lot of other languages are written in an alphabet technically known as the Latin alphabet because it derives from the alphabet the Latin language was (and is) written in. Readers and writers of these languages will perhaps be surprised to hear that their alphabet is the “Latin” alphabet, because they never thought about it that way. For them, their alphabet is simply the “normal” (English etc.) alphabet.

When using computers, the Latin alphabet is nowadays coded in the so-called ANSI code in MS Windows and other operating systems. This is a coding system which uses a set of 8 bits to represent each character. Hence there is room for 2^8 = 256 different symbols. A number of codes are reserved for punctuation marks, the space, arithmetic symbols, mathematical symbols, typographic symbols, a few currency codes and other characters. Letters are represented in both uppercase and lowercase. In addition, there is room in the ANSI code for some variations of the Latin letters used in other Latin-alphabet languages.

These are the symbols represented in an often-used Windows font, Courier New:

```
-     ! " # $ % & ' ( ) * + , - . / 0 1
2 3 4 5 6 7 8 9 : ; < = > ? @ 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 [ \ ] ^ _ ` 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 { | } ~ □ □ □
, ƒ „ … † ‡ ˆ ‰ Š ‹ Œ □ □ □ □ ‘ ’ “ ” •
– — ˜ ™ š › œ □ □ Ÿ   ¡ ¢ £ ¤ ¥ ¦ § ¨ ©
ª « ¬ - ® ¯ ° ± ² ³ ´ µ ¶ · ¸ ¹ º » ¼ ½
¾ ¿ À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ð Ñ
Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã ä å
æ ç è é ê ë ì í î ï ð ñ ò ó ô õ ö ÷ ø ù
ú û ü ý þ ÿ
```

Characters that are not printable are represented by a □.
As an illustration of special “Latin” letters, see the eth ð (second line from below) and the thorn þ (last line) as used in Icelandic.

2. Problems when using non-Latin alphabets

2.1 Left to right writing

Writers of the Latin alphabet write from left to right. Some other-language writers write from right to left (historically also other ways were used, for example from top to bottom).

2.2 Ligatures

Other alphabets can contain combinations of letters called *ligatures*: two letters (rarely more) are combined into one character. Writing such a combination is easy: you just use your pen to write, for example, the second letter inside the first letter. In printing, you need a separate symbol. When printing books in the Latin alphabet, sometimes the combination "fi" is treated as a ligature (the 'i' losing its dot and appearing under the 'f').

2.3 Combinations of consonants and vowels

Some alphabets write vowels in combination with consonants. The vowels are usually symbols like strokes or dots. Because one consonant can have various vowels, in printing you need separate symbols for all possible consonant – vowel combinations. Basically this is still an alphabetic system.

2.4 Ideographies

Other 'alphabets' (more rightly called 'writing systems') use symbols for whole words or syllables. Examples of these are Chinese, Japanese and Korean. Because many words are in common use, each represented by a different symbol, the number of symbols becomes very large, in the order of thousands or tens of thousands. These writing systems are called 'ideographies'.

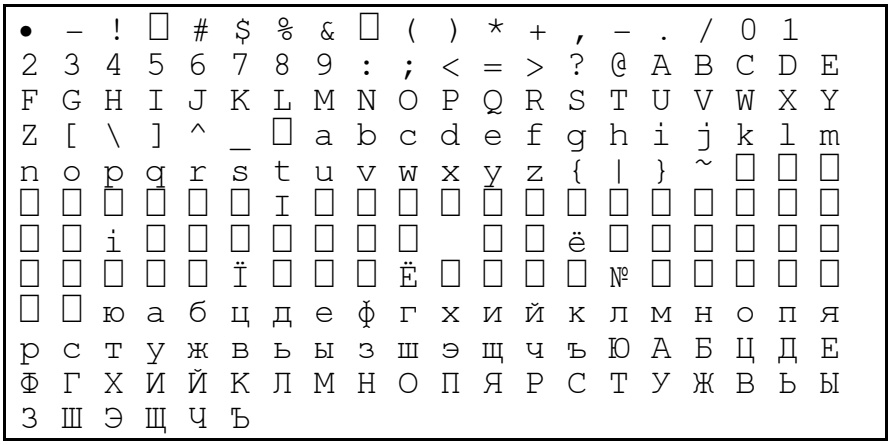
3. The Arabic alphabet

The Arabic alphabet has 28 consonants and is written from right to left. The vowels are written as little symbols (strokes and curls) above or below the consonants. There are also symbols for doubling consonants and for denoting the absence of a vowel, there is a consonant written only in combination with vowels (the 'hamza') and a few other special symbols. Moreover, the consonants can have up to four different forms: when written initially, in the middle, at the end or isolated (not all consonants have all variations). So here we see a problem not mentioned before: different forms of the letters depending on their position relative to other letters, in addition to problems Nos. 2.1 and 2.3.

All this is really not very complicated and a foreigner can learn to read and write with a pen with not too much trouble, but in printing or typing on a computer one needs many symbols (for the different forms of the consonants, each of them in combination with all possible vowels, etc.). A very elegant system of writing thus leads to a complicated printing and typing task. There is one simplification, however: usually the vowels are not printed or written, only the consonants. The reader knows from the context how to read the consonant-only words. Notable exceptions are the Koran, in which vowels are always written in order to make sure that the pronunciation is always identical, and dictionaries, in which words with the same consonants but different vowels must be distinct. Some other languages, not cognate with Arabic, are also written in the Arabic alphabet with a few additions, for example Farsi, Pashto and Urdu.

4. A solution for Cyrillic alphabets

We can use fonts in Blaise to represent non-Latin alphabets that are sufficiently close to the Latin writing system. This applies mainly to Greek and Cyrillic alphabets. The Greek alphabet stands at the origin of the Latin alphabet and is hence similar (more rightly we should say “the Latin alphabet is similar to the Greek”), but uses different letter forms (of course, it would not be a different alphabet if it didn’t). The Cyrillic alphabet has elements from the Greek, Latin and Hebrew alphabets. It has some variations for different Slavonic and other languages. A freeware font called K8 Kurier Fixed implements the Cyrillic alphabet for Russian. Its symbols are:



Non-printable characters are represented by □. The first 127 characters (up to and including the ~) are the same as in Courier New and many other fonts. From the character ю (4th line from below) Cyrillic letters are represented, both lowercase and uppercase. So we must use these characters when we want to write something in Cyrillic letters in Blaise. However, these characters do not appear on the keyboard. When we want to type them in an MS Word document such as this one, we can insert them via **Insert | Symbol**, or type some key combination like ALT + 0192 for ю or ALT + CTRL + SHIFT + Z for ф.

This is inconvenient, although it can be used for small amounts of text. To simplify typing I designed a Manipula program which takes ordinary typed Latin alphabet letters and transforms them into Cyrillic characters. The basic idea is to have a conversion table between Latin letters (single letters or up to four-letter combinations) and Cyrillic characters. A short idea of the conversion table:

A - А
B - Б
V - В
G – Г
D – Д
E – Е
EX - Ё

And so on. The complete table is available upon request. This is not an official transliteration system, by the way. I used characters not appearing in the Cyrillic alphabet like X and H to distinguish letters from one another if necessary. One can then type one’s Russian text on the normal keyboard in the Blaise editor, for instance:

Kak Vashe imax? (‘What is your name?’)

The principle behind the Manipula program is to recode the Latin letters to the corresponding Cyrillic characters in the Kurier Fixed font, like this:

```
CASE TransChar OF
‘A’: RusChar := CHAR(223)
‘B’: RusChar := CHAR(193)
‘D’: RusChar := CHAR(196)
 .
 .
 .
ENDCASE
```

Here, TransChar is the character to be recoded and RusChar is the transcribed Cyrillic character.

The result of converting the string

Kak Vashe imax?

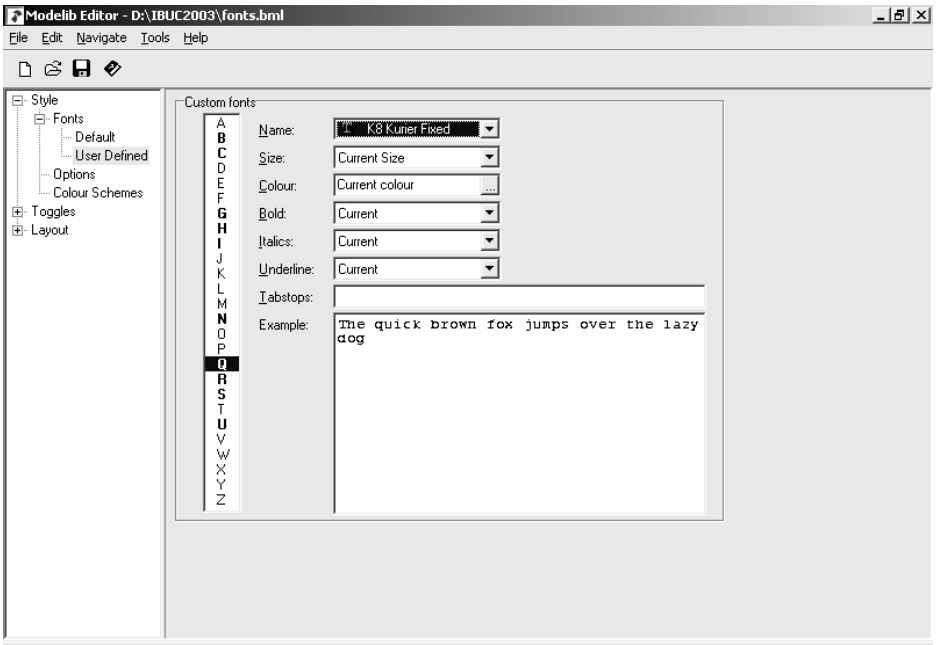
Is:

ëÁË ÷ÁÛÅ ÉÍÑ?

This converted text can be copied and pasted into a Blaise question text, for instance:

Name “ëÁË ÷ÁÛÅ ÉÍÑ?”: STRING[20]

Now the font Kurier Fixed must be associated with a font letter in the Mode Library. Let’s take Q:

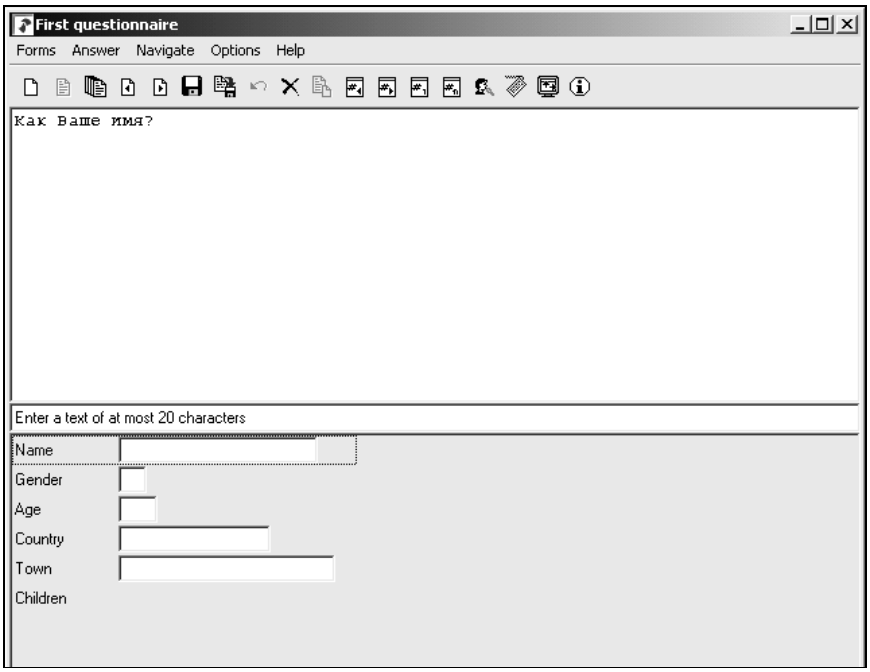


We add the @Q before and after the question text:

Name “@QëÁË ÷ÁÛÅ ÉÍÑ?@Q”: STRING[20]

When we run this questionnaire, the question text appears in Russian:

Как Ваше имя?



All other question texts can be transcribed in the same manner. So the steps to be taken are:

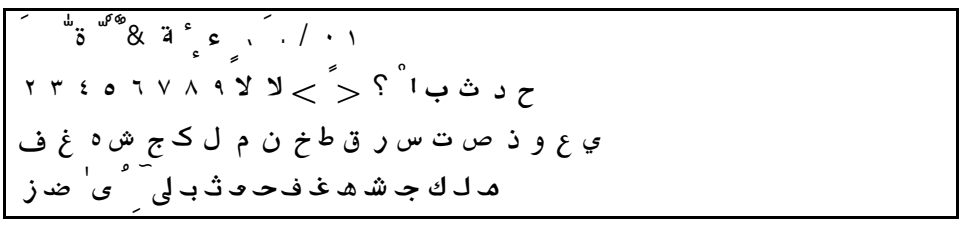
1. Associate the font letter Q in the Modelib Editor with the Kurier Fixed font
2. Type the question texts in normal keyboard characters using the conversion table
3. Transform the text using the Manipula program
4. Copy and paste the question text into the Blaise editor
5. Put the font letter to the left and right of the question text
6. Prepare the Blaise survey using the mode library with the Kurier Fixed font.
7. Run the survey. The Russian text appears in the Data Entry Program.

All this can be done from the Blaise Control Centre.

A similar conversion could be done for the Greek alphabet.

5. The Arabic alphabet

When we try to do the same for the Arabic alphabet, we first select a font. There is a freeware Arabic font called Amien 01. This is the character set:



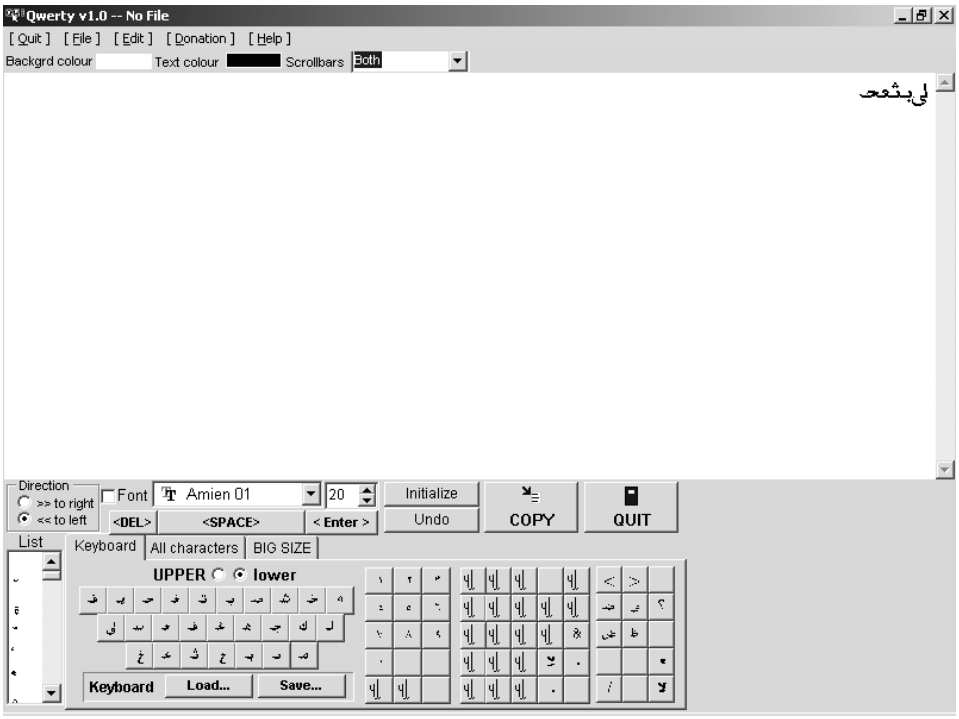
ن خ ه ق ف س ت ص ع ح ي ع ض غ ظ ً

¡ ¢ £ ¤ ¥ ¦ § ¨ ©
ª « ¬ - ® ¯ ° ± ² ³ ´ µ ¶ · ¸ ¹ º » ¼ ½
¾ ¿ À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ð Ñ
Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã ä å
æ ç è é ê ë ì í î ï ð ñ ò ó ô õ ö ÷ ø ù
ú û ü ý þ ÿ

The character seems to represent an unknown character (similar to the □ of other character sets). Also, some (variant of) Latin characters are present in this set. There seems to be an insufficient number of characters to represent all the variations of Arabic writing.

If one wants to try this font in Blaise, one can assign the font letter A (for Arabic) to this font. There is a freeware utility called QWERTY for typing from right to left and using a font installed on the computer. When this is started, the font Amien can be selected and the virtual keyboard in QWERTY adapts itself. One can then type characters on the virtual keyboard and use the COPY key to copy the sentence to the clipboard. Then the clipboard contents can be pasted into the Blaise setup. Surround the question text with @A. When running the Data Entry Program, the Arabic letters will appear.

The following figure shows the QWERTY program with the Amien font and some letters typed from right to left.



Both the Amien font and the QWERTY utility can be found at
http://user.dtcc.edu/~berlin/font/arabic
The Internet address for the QWERTY utility is
http://logics.ghanima.org/en/qwerty.htm

6. Chinese, Japanese and Korean

For Chinese, Japanese and Korean (the so-called CJK languages) a custom solution is available as a commercial product called NJStar Communicator (http://www.njstar.com)The use of this product in combination with Blaise is discussed in the paper by Gina-Qian Y. Cheung and Youhong Liu, University of Michigan, entitled 'Displaying Chinese Characters in Blaise' (see the Proceedings of this IBUC). The solution is to use two bytes to represent one symbol. To show the CJK character on the computer screen, the product must be activated. When it is turned off, one sees the two bytes as separate (ANSI) characters.

7. Images in Blaise

Yet another solution is to make use of the Blaise capability of showing images in the DEP. One writes or prints the texts in the foreign writing system on paper and scans the images. The texts can then be shown in the DEP as images. This type of solution (along with a discussion of other solutions) is described by Leif Bochis in a paper for the previous Blaise User Conference.
(http://www.blaiseusers.org/ibucpdfs/2001/Bochis--IBUC_Paper.pdf).

8. Multi-byte character sets

The problem of representing other writing systems in computer programs was recognized many years ago. Various attempts at solving this problem have resulted in the 16-bit Unicode which is able to represent most writing systems, notably Chinese and related systems, the Arabic and Cyrillic alphabets along with a great number of variations of the Latin alphabet. The Internet, modern operating systems and software use the Unicode system. Hence one can type in MS Word a document in one of these writing systems. Unfortunately, what has been typed in an MS Word document cannot be copied and pasted into the Blaise editor. This has been the source of some confusion. Maybe future versions of Blaise will support Unicode, in which case the problems treated in this paper will be over.

9. Fonts in the Blaise editor

The Blaise editor supports only monospaced fonts. These can be True Type fonts or other fonts, but nowadays few fonts are `monospaced` (Courier is an example). *Proportional* fonts, for example the Times New Roman font this paper is written in, are more pleasing to the eye. In order for a font to even appear as a possible choice in the Blaise editor, it has to be monospaced. The Kurier Fixed font is monospaced, so it can be used in the Blaise editor. Once a font is available in the editor, the field names can be typed in the font. So we are able to make an all-Russian questionnaire.
The author of this paper was unable to find a monospaced Arabic font among available fonts on the Internet.

10. Other alphabets

Some alphabets not mentioned so far but which have turned up in discussions with the Blaise Support Group are the Hebrew, Thai and Vietnamese alphabets. It would be interesting to see how these alphabets can be represented in Blaise.

11. Summary

The problems of using non-Latin alphabets in Blaise were lined out. A solution for a few alphabets close to the Latin alphabet was presented, but many problems still exist for other alphabets, mainly because the number of letters and letter combinations is much greater than the number of characters in the single-byte ANSI character set. For Chinese-like systems, a custom solution is available. Such a solution is not readily available for other writing systems.