UTF-8 SAMPLER

¥ · £ · € · $ · ¢
· ₡ · ₲ · ₣ · ₤ · ₥ · ₦ · ₱ · ₨ · ₩ · ₪ · ₫ · ₭ · ₮ · ₯

Frank da Cruz
The Kermit Project - Columbia University
New York City
fdc@columbia.edu

Last update: Mon Feb 7 16:06:48 2005

[PEACE] [Poetry] [I Can Eat Glass] [The Quick Brown Fox] [HTML Features] [Credits, Tools, Commentary]

UTF-8 is an ASCII-preserving encoding method for Unicode (ISO 10646), the Universal Character Set (UCS). The UCS encodes most of the world's writing systems in a single character set, allowing you to mix languages and scripts within a document without needing any tricks for switching character sets. This web page is encoded directly in UTF-8.

As shown HERE, Columbia University's Kermit 95 terminal emulation software can display UTF-8 plain text in Windows 95, 98, ME, NT, XP, or 2000 when using a monospace Unicode font like Andale Mono WT J or Everson Mono Terminal, or the lesser populated Courier New, Lucida Console, or Andale Mono. C-Kermit can handle it too, if you have a Unicode display. As many languages as are representable in your font can be seen on the screen at the same time.

This, however, is a Web page. Some Web browsers can handle UTF-8, some can't. And those that can might not have a sufficiently populated font to work with (some browsers might pick glyphs dynamically from multiple fonts; Netscape 6 seems to do this). CLICK HERE for a survey of Unicode fonts for Windows.

The subtitle above shows currency symbols of many lands. If they don't appear as blobs, we're off to a good start!

Poetry

From the Anglo-Saxon Rune Poem (Rune version):

ᚠᛇᚻ᛫ᛒᛦᚦ᛫ᚠᚱᚩᚠᚢᚱ᛫ᚠᛁᚱᚪ᛫ᚷᛖᚻᚹᛦᚳᛚᚾᛞᛗ
ᛋᚳᛖᚪᛚ᛫ᚦᛖᚪᚻ᛫ᛗᚪᚠᛏᚾᛏᚠ᛫ᚷᛖᚻᚹᛦᚳᛚᚪ᛫ᛗᛁᚳᛚᚢᚾᛏ᛫ᚻᛦᛚᛏ᛫ᛞᚫᚱᚠᛏ
ᚷᛁᚠ᛫ᚻᛖᛗ᛫ᚹᛁᛚᛖᛗ᛫ᚠᚹᛖᚱ᛫ᛞᚹᛟᚻᛏᚾᛖᛗ᛫ᛞᚠᚹᛗᛁᛋ᛫ᚻᚩᛚᛏ᛫ᚹᛖᛟᛞ᛬

From La3amon's Brut (The Chronicles of England, Middle English, West Midlands):

An preost wes on leoden, La3amon was ihoten
He wes Leovenaðes sone -- liðe him be Drihten.
He wonede at Ernle3e at æðelen are chirechen,
Uppen Sevarne staþe, sel þar him þuhte,
Onfest Radestone, þer he bock radde.

(The third letter in the author's name is Yogh, missing from many fonts; CLICK HERE for another Middle English sample with some explanation of letters and encoding).

From the Tagelied of **Wolfram von Eschenbach** (Middle High German):

> Sîne klâwen durh die wolken sint geslagen,
> er stîget ûf mit grôzer kraft,
> ich sih in grâwen tägelîch als er wil tagen,
> den tac, der im geselleschaft
> erwenden wil, dem werden man,
> den ich mit sorgen în verliez.
> ich bringe in hinnen, ob ich kan.
> sîn vil manegiu tugent michz leisten hiez.

Some lines of **Odysseus Elytis** (Greek):

> Τη γλώσσα μου έδωσαν ελληνική
> το σπίτι φτωχικό στις αμμουδιές του Ομήρου.
> Μονάχη έγνοια η γλώσσα μου στις αμμουδιές του Ομήρου.
>
> από το Άξιον Εστί
> του Οδυσσέα Ελύτη

The first stanza of **Pushkin**'s Bronze Horseman (Russian):

> На берегу пустынных волн
> Стоял он, дум великих полн,
> И вдаль глядел. Пред ним широко
> Река неслася; бедный чёлн
> По ней стремился одиноко.
> По мшистым, топким берегам
> Чернели избы здесь и там,
> Приют убогого чухонца;
> И лес, неведомый лучам
> В тумане спрятанного солнца,
> Кругом шумел.

Šota Rustaveli's Vep̣xis Ṭq̇aosani, Th, The Knight in the Tiger's Skin (Georgian):

> ვეპხისტყაოსანი შოთა რუსთაველი
>
> ღმერთსი შემვედრე, ნუთუ კვლა დამხსნას სოფლისა შრომასა, ცეცხლს წყალსა და მიწასა, ჰაერთა თანა მრომასა; მომცნეს ფრთენი და აღვფრინდე, მივჰხვდემას ჩემსა ნდომასა, დღისით და ღამით ვჰხედვიდემზისა ელვათა კრთომაასა.

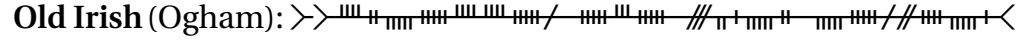
Tamil poetry of Cupiramaniya Paarathiyar சுப்ரமணிய பாரதியார் (1882-1921):

> யாமறிந்த மொழிகளிலே தமிழ்மொழி போல் இனிதாவது எங்கும்
> காணோம்,
> பாமரராய் விலங்குகளாய், உலகனைத்தும் இகழ்ச்சிசொலப்
> பான்மை கெட்டு,
> நாமமது தமிழரெனக் கொண்டு இங்கு வாழ்ந்திடுதல் நன்றோ?
> சொல்லீர்!

I Can Eat Glass

And from the sublime to the ridiculous, here is a certain phrase[1] in an assortment of languages:

1. **Sanskrit**: काचं शक्नोम्यत्तुम् । नोपहिनस्ति माम् ॥
2. **Sanskrit** (standard transcription): kācaṃ śaknomyattum; nopahinasti mām.
3. **Classical Greek**: ὕαλον φαγεῖν δύναμαι· τοῦτο οὔ με βλάπτει.
4. **Greek**: Μπορώ να φάω σπασμένα γυαλιά χωρίς να πάθω τίποτα.
   Etruscan: (NEEDED)
5. **Latin**: Vitrum edere possum; mihi non nocet.
6. **Old French**: Je puis mangier del voirre. Ne me nuit.
7. **French**: Je peux manger du verre, ça ne me fait pas de mal.
8. **Provençal / Occitan**: Pòdi manjar de veire, me nafrariá pas.
9. **Québécois**: J'peux manger d'la vitre, ça m'fa pas mal.
10. **Walloon**: Dji pou magnî do vêre, çoula m' freut nén må.
    Champenois: (NEEDED)
    Lorrain: (NEEDED)
11. **Picard**: Ch'peux mingi du verre, cha m'foé mie n'ma.
    Corsican: (NEEDED)
12. **Kreyòl Ayisyen**: Mwen kap manje vè, li pa blese'm.
13. **Basque**: Kristala jan dezaket, ez dit minik ematen.
14. **Catalan**: Puc menjar vidre que no em fa mal.
15. **Spanish**: Puedo comer vidrio, no me hace daño.
16. **Aragones**: Puedo minchar beire, no me'n fa mal .
17. **Galician**: Eu podo xantar cristais e non cortarme.
18. **Portuguese**: Posso comer vidro, não me faz mal.
19. **Brazilian Portuguese**: Posso comer vidro, não me machuca.
20. **Caboverdiano**: M' podê cumê vidru, ca ta maguâ-m'.
21. **Papiamentu**: Ami por kome glas anto e no ta hasimi daño.
22. **Italian**: Posso mangiare il vetro e non mi fa male.
23. **Milanese**: Sôn bôn de magnà el véder, el me fa minga mal.
24. **Roman**: Me posso magna' er vetro, e nun me fa male.
25. **Napoletano**: M' pozz magna' o'vetr, e nun m' fa mal.
26. **Sicilian**: Puotsu mangiari u vitru, nun mi fa mali.
27. **Venetian**: Mi posso magnare el vetro, no'l me fa mae.
28. **Zeneise** (Genovese): Pòsso mangiâ o veddro e o no me fà mâ.
    Rheto-Romance / Romansch: (NEEDED)
    Romany / Tsigane: (NEEDED)
29. **Romanian**: Pot să mănânc sticlă și ea nu mă rănește.
30. **Esperanto**: Mi povas manĝi vitron, ĝi ne damaĝas min.
    Pictish: (NEEDED)
    Breton: (NEEDED)
31. **Cornish**: Mý a yl dybry gwéder hag éf ny wra ow ankenya.
32. **Welsh**: Dw i'n gallu bwyta gwydr, 'dyw e ddim yn gwneud dolur i mi.
33. **Manx Gaelic**: Foddym gee glonney agh cha jean eh gortaghey mee.
34. **Old Irish** (Ogham):
35. **Old Irish** (Latin): Con·iccim ithi nglano. Ním·géna.
36. **Irish**: Is féidir liom gloinne a ithe. Ní dhéanann sí dochar ar bith dom.
37. **Scottish Gaelic**: S urrainn dhomh gloinne ithe; cha ghoirtich i mi.
38. **Anglo-Saxon** (Runes): ᛁᛕ·ᛗᚨᚷ·ᚷᛁᚩᛋ·ᛖᚨᛏᚨ·ᚨᛏᚪᚾ·ᚺᛁᛏ·ᛏᛖ·ᚺᛖᚨᚱᛗᛁᚹ·ᛗᛖ:
39. **Anglo-Saxon** (Latin): Ic mæg glæs eotan ond hit ne hearmiað me.



40. **Middle English**: Ich canne glas eten and hit hirtiþ me nowiht.
41. **English**: I can eat glass and it doesn't hurt me.
42. **English** (IPA): [aɪ kæn iːt glɑːs ænd ɪt dʌz nɒt hɜːt miː] (Received Pronunciation)
43. **English** (Braille): ⠊⠀⠉⠁⠝⠀⠑⠁⠞⠀⠛⠇⠁⠎⠎⠀⠁⠝⠙⠀⠊⠞⠀⠙⠕⠑⠎⠝⠄⠞⠀⠓⠥⠗⠞⠀⠍⠑⠲
44. **Lalland Scots / Doric**: Ah can eat gless, it disnae hurt us.
    Glaswegian: (NEEDED)
45. **Gothic** (4): 𐌲𐌿𐍄𐌰 𐌷𐌰𐍅𐌹 𐌾𐌿𐌺𐌹𐌸 𐌷𐍂𐌹𐌽𐌽𐌰𐌽 𐌰𐌽𐌰𐍄 𐌹𐍃𐍄, 𐌽𐌹 𐌼𐌹𐍃 𐌺𐍅𐌴𐌸𐌹𐌸
    𐌽𐌹 𐌼𐌹𐌺 𐌱𐌰𐌻𐌲𐌾𐌰𐌽𐌽𐌰𐌽 𐌱𐌰𐌽𐌰𐌽𐌰𐌽𐌰𐌽𐌰𐌽𐌰𐌽𐌰𐌽.
46. **Old Norse** (Runes): ᛖᚴ ᚷᛖᛏ ᛖᛏᛁ ᚦ ᚷᚱᛖᚱ ᛘᚾ ᚦᛖᚴᚴᚱ ᚠᚦ ᚹᛖᛁ ᚱᚦᚠ ᚴᚠᚱ
47. **Old Norse** (Latin): Ek get etið gler án þess að verða sár.
48. **Norsk / Norwegian (Nynorsk)**: Eg kan eta glas utan å skada meg.
49. **Norsk / Norwegian (Bokmål)**: Jeg kan spise glass uten å skade meg.
    Føroyskt / Faroese: (NEEDED)
50. **Íslenska / Icelandic**: Ég get etið gler án þess að meiða mig.
51. **Svenska / Swedish**: Jag kan äta glas utan att skada mig.
52. **Dansk / Danish**: Jeg kan spise glas, det gør ikke ondt på mig.
53. **Soenderjysk**: Æ ka æe glass uhen at det go mæ naue.
54. **Frysk / Frisian**: Ik kin glês ite, it docht me net sear.
55. **Nederlands / Dutch**: Ik kan glas eten, het doet mij geen kwaad.
56. **Kirchröadsj/Bôchesserplat**: Iech ken glaas èèse, mer 't deet miech jing pieng.
57. **Afrikaans**: Ek kan glas eet, maar dit doen my nie skade nie.
58. **Lëtzebuergesch / Luxemburgish**: Ech kan Glas iessen, daat deet mir nët wei.
59. **Deutsch / German**: Ich kann Glas essen, ohne mir weh zu tun.
60. **Ruhrdeutsch**: Ich kann Glas verkasematuckeln, ohne dattet mich wat jucken tut.
61. **Lausitzer Mundart** ("Lusatian"): Ich koann Gloos assn und doas dudd merr ni wii.
62. **Sächsisch / Saxon**: 'sch kann Glos essn, ohne dass'sch mer wehtue.
63. **Pfälzisch**: Isch konn Glass fresse ohne dasses mer ebbes ausmache dud.
64. **Schwäbisch / Swabian**: I kå Glas frässa, ond des macht mr nix!
65. **Bayrisch / Bavarian**: I koh Glos esa, und es duard ma ned wei.
66. **Allemannisch**: I kaun Gloos essen, es tuat ma ned weh.
67. **Schwyzerdütsch**: Ich chan Glaas ässe, das tuet mir nöd weeh.
68. **Hungarian**: Meg tudom enni az üveget, nem lesz őle bajom.
69. **Suomi / Finnish**: Voin syödä lasia, se ei vahingoita minua.
70. **Sami (Northern)**: Sáhtán borrat lása, dat ii leat bávčas.
71. **Erzian**: Мон ярсан суликадо, ды зыян эйстэнзэ а ули.
    Karelian: (NEEDED)
    Vepsian: (NEEDED)
    Votian: (NEEDED)
    Livonian (NEEDED)
72. **Estonian**: Ma võin klaasi süüa, see ei tee mulle midagi.
73. **Latvian**: Es varu ēst stiklu, tas man nekaitē.
74. **Lithuanian**: Aš galiu valgyti stiklą ir jis manęs nežeidžia
    Old Prussian: (NEEDED)
    Sorbian (Wendish): (NEEDED)
75. **Czech**: Mohu jíst sklo, neublíží mi.
76. **Slovak**: Môžem jesť sklo. Nezraní ma.
77. **Polska / Polish**: Mogę jeść szkło i mi nie szkodzi.
78. **Slovenian**: Lahko jem steklo, ne da bi mi škodovalo.
79. **Croatian**: Ja mogu jesti staklo i ne boli me.
80. **Serbian** (Latin): Mogu jesti staklo a da mi ne škodi.
81. **Serbian** (Cyrillic): Могу јести стакло а да ми не шкоди.

82. **Macedonian:** Можам да јадам стакло, а не ме штета.
83. **Russian**: Я могу есть стекло, оно мне не вредит.
84. **Belarusian** (Cyrillic): Я магу есці шкло, яно мне не шкодзіць.
85. **Belarusian** (Lacinka): Ja mahu jeści škło, jano mne neškodzić.
86. **Ukrainian**: Я можу їсти шкло, й воно мені не пошкодить.
87. **Bulgarian**: Мога да ям стъкло, то не ми вреди.
88. **Georgian**: მინას ვჭამ და არა მტკივა.
89. **Armenian**: Կրնամապակիուտելև ինծի անհանգիստչըներ:
90. **Albanian**: Unë mund të ha qelq dhe nuk më gjen gjë.
91. **Turkish**: Cam yiyebilirim, bana zararı dokunmaz.
92. **Turkish** (Ottoman): جام ییه بلورم بڭا ضرری طوقونمز
93. **Bangla / Bengali**: আমি কাঁচ খতেে পারি, তাতে আমার কোনো ক্ষতি হয় না।
94. **Marathi**: मी काच खाऊ शकतो, मला ते दुखत नाही.
95. **Hindi** मैं काँच खा सकता हूँ, मुझे उस से कोई पीडा नहीं होती.
96. **Tamil**: நான் கண்ணாடி சாப்பிடுவனே , அதனால் எனக்கு ஒரு கடேும் வராது.
97. **Urdu**(2): - میں کانچ کھا سکتا ہوں اور مجھے تکلیف نہیں ہوتی
98. **Pashto**(2): زه شیشه خوړلۍ شم، هغه ما نه خوږوي
99. **Farsi / Persian**: من می توانم بدونِ احساس درد شیشه بخورم.
100. **Arabic**(2): أنا قادر على أكل الزجاج و هذا لا يؤلمني.
Aramaic: (NEEDED)
101. **Hebrew**(2): אני יכול לאכול זכוכית וזה לא מזיק לי.
102. **Yiddish**(2): איך קען עסן גלאָז און עס טוט מיר נישט וויי.
Judeo-Arabic: (NEEDED)
Ladino: (NEEDED)
Gəʾəz: (NEEDED)
Amharic: (NEEDED)
103. **Twi**: Metumi awe tumpanɜ̃nyɜ me hwee.
104. **Hausa** (Latin): Inā iya taunar gilāshi kuma in gamālā fiyā.
105. **Hausa** (Ajami) (2): إِنا إِىَ تَونَر غِلَاشِ كُمَ إِن غَمَا لَافِيَا
106. **Yoruba**(3): Mo lè jẹ dígí, kò ní pa mí lára.
107. **(Ki)Swahili**: Naweza kula bilauri na sikunyui.
108. **Malay**: Saya boleh makan kaca dan ia tidak mencederakan saya.
109. **Tagalog**: Kaya kong kumain nang bubog at hindi ako masaktan.
110. **Chamorro**: Siña yo' chumocho krestat, ti ha na'lalamen yo'.
111. **Javanese**: Aku isa mangan beling tanpa lara.
Burmese: (NEEDED)
112. **Vietnamese (quốc ngữ)**: Tôi có thể ăn thủy tinh mà không hại gì.
113. **Vietnamese (nôm)** (4): 些 □□ 世 咹 水 晶 □□ 空 □□ 害 咦
Khmer: (NEEDED)
Lao: (NEEDED)
114. **Thai**: ฉันกินกระจกได้แต่มันไม่ทำให้ฉันเจ็บ
115. **Mongolian**(Cyrillic): Би шил идэй чадна, надад хортой биш
116. **Mongolian**(Classic) (5): [illegible]
Dzongkha: (NEEDED)
Nepali: (NEEDED)
117. **Tibetan**: ཤེལ་སྒོ་ཟ་ནས་ང་ན་གི་མ་རེད།
118. **Chinese**: 我能吞下玻璃而不伤身体。
119. **Chinese** (Traditional): 我能吞下玻璃而不傷身體。
120. **Taiwanese**(6): Góa ē-tàng chiah po-lê, mā bē tioh-siong.
121. **Japanese**: 私はガラスを食べられます。それは私を傷つけません。
122. **Korean**: 나는 유리를 먹을 수 있어요. 그래도 아프지 않아요

123. **Bislama**: Mi save kakae glas, hemi no save katem mi.
124. **Hawaiian**: Hiki iaʻu ke ʻai i ke aniani;ʻaʻole nō lā au e ʻeha.
125. **Marquesan**: E koʻana e kai i te karahi, meaʻā, ʻaʻe hauhau.
126. **Chinook Jargon**:Naika məkməkkakshət labutay, pi weykukukmunk-sik nay.
127. **Navajo**: Tsésǫʼ yishą́ągo bíníshghah dóó doo shiłneezgai da.
Cherokee (and Cree, Ojibwa, Inuktitut, and other Native American languages): (NEEDED)
Garifuna: (NEEDED)
Gullah: (NEEDED)
128. **Lojban**: mi kakne le nu citka le blaci .iku'i le se go'i na xrani mi
129. **Nórdicg** Ljœr ye caudran créneþ ý jor cwran.

(Additions, corrections, completionsgratefully accepted)

For testing purposes, some of these are repeated in**monospace font**. . .

1. Euro Symbol: €.
2. Greek: Μπορώ να φάω σπασμένα γυαλιά χωρίς να πάθω τίποτα.
3. Íslenska / Icelandic: Ég get etið gler án þess að meiða mig.
4. Polish: Mogę jeść szkło, i mi nie szkodzi.
5. Romanian: Pot să mănânc sticlă și ea nu mă rănește.
6. Ukrainian: Я можу їсти шкло, й воно мені не пошкодить.
7. Armenian: Կրնամ ապակի ուտել և ինծի անհանգիստ չըներ:
8. Georgian: მინას ვჭამ და არა მტკივა.
9. Hindi: मैं काँच खा सकता हूँ, मुझे उस से कोई पीडा नहीं होती.
10. Hebrew(2): .אני יכול לאכול זכוכית וזה לא מזיק לי
11. Yiddish(2): .איך קען עסן גלאָז און עס טוט מיר נישט ווי
12. Arabic(2): .أنا قادر على أكل الزجاج و هذا لا يؤلمني
13. Japanese: 私はガラスを食べられます。それは私を傷つけません。
14. Thai: ฉันกินกระจกได้ แต่มันไม่ทำให้ฉันเจ็บ

Notes:

1. The "I can eat glass" phrase and initial translations were borrowed from Ethan Mollick's I Can Eat Glasspage (which disappeared on or about June 2004) and converted to UTF-8. See Creditsfor the many additional contributions since then. The word "hurt" is used in the sen of "cause harm", "do damage", or "bother", rather than "inflict pain" or "make sad". Otto Stolz comments:

> This is the meaning I have translated to the Swabian dialect. However, I just have noticed that most of the German variants translate the "inflict pain" meaning. The German example should rather read:
>
> > "Ich kann Glas essen ohne mir zu schaden."
>
> (The comma fell victim to the 1996 orthographic reform, cf. http://www.ids-mannheim.de/reform/e3-1.html#P76.
>
> You may wish to contact the contributors of the following translations to correct them:
>
> - Lëtzebuergescht / Luxemburgish: Ech kan Glas iessen, daat deet mir nët wei.
> - Lausitzer Mundart ("Lusatian"): Ich koann Gloos assn und doas dudd merr ni wii.
> - Sächsisch / Saxon: 'sch kann Glos essn, ohne dass'sch mer wehtue.
> - Bayrisch / Bavarian: I koh Glos esa, und es duard ma ned wei.
> - Allemannisch: I kaun Gloos essen, es tuat ma ned weh.
> - Schwyzerdütsch: Ich chan Glaas ässe, das tuet mir nöd weeh.

In contrast, I deem the following translations *alright*:

- Ruhrdeutsch: Ich kann Glas verkasematuckeln, ohne dattet mich wat jucken tut.
- Pfälzisch: Isch konn Glass fresse ohne dasses mer ebbes ausmache dud.
- Schwäbisch / Swabian: I kå Glas frässa, ond des macht mr nix!

(However, you could remove the commas, on account of http://www.ids-mannheim.de/reform/e3-1.html#P76 and http://www.ids-mannheim.de/reform/e3-1.html#P72, respectively.)

I guess, also these examples translate the wrong sense of "hurt", though I do not know these languages well enough to assert them definitely:

- Nederlands / Dutch: Ik kan glas eten; het doet mij geen pijn. (This one has been changed)
- Kirchröadsj/Bôchesserplat: Iech ken glaas èèse, mer 't deet miech jing pieng.

In the Romanic languages, the variations on "fa male" (it) are probably wrong, whilst the variations on "hace daño" (es) and "damaĝas" (Esperanto) are probably correct; "nocet" (la) is definitely right.

The northern Germanic variants of "skada" are probably right, as are the Slavic variants of "škodi/шкоди" (se); however the Slavic variants of " boli" (hv) are probably wrong, as "bolena" means "pain/ache", IIRC.

The numbering of the samples is arbitrary, done only to keep track of how many there are, and can change any time a new entry is added. The arrangement is also arbitrary but with some attempt to group related examples together. Note: All languages not listed are wante not just the ones that say (NEEDED).

2. Correct right-to-left display of these languages depends on the capabilities of your browse The period should appear on the left. In the monospace Yiddish example, the Yiddish digraphs should occupy one character cell.
3. Yoruba: The third word is Latin letter small 'j' followed by small 'e' with U+0329, Combinin Vertical Line Below. This displays correctly only if your Unicode font includes the U+0329 glyph and your browser supports combining diacritical marks. The Indic examples also include combining sequences.
4. Includes Unicode 3.1 (or later) characters beyond Plane 0.
5. The Classic Mongolian example should be vertical, top-to-bottom and left-to-right. But su display is almost impossible. Also no font yet exists which provides the proper ligatures an positional variants for the characters of this script, which works somewhat like Arabic.
6. Taiwanese is also known as Holo or Hoklo, and is related to Southern Min dialects such as Amoy. Contributed by Henry H. Tan-Tenn, who comments, "The above is the romanized version, in a script current among Taiwanese Christians since the mid-19th century. It was invented by British missionaries and saw use in hundreds of published works, mostly of a religious nature. Most Taiwanese did not know Chinese characters then, or at least not wel enough to read. More to the point, though, a written standard using Chinese characters ha never developed, so a significant minority of words are represented with different candida characters, depending on one's personal preference or etymological theory. In this senten for example, "-tàng", "chi̍a", "mā" and "bē" are problematic using Chinese characters. "Góa" (I/me) and "po-lê" (glass) are as written in other Sinitic languages (e.g. Mandarin, Hakka)."

The Quick Brown Fox

The "I can eat glass" sentences do not necessarily show off the orthography of each language to

best advantage. In many alphabetic written languages it is possible to include all (or most) letter (or "special" characters) in a single (often nonsense) pangram These were traditionally used in typewriter instruction; now they are useful for stress-testing computer fonts and keyboard input methods. Here are a few examples (SEND MORE):

1. **English:** The quick brown fox jumps over the lazy dog.
2. **Irish:** "An bfuil do chroí ag bualadh ó fhaitíos an ghrá a mheall lena phóg éada ó shlí do leasa thú?" "D'fhuascail Íosa Úrmhac na hÓighe Beannaithe pór Éava agus Ádhaimh."
3. **Dutch:** Pa's wijze lynx bezag vroom het fikse aquaduct.
4. **German:** Falsches Üben von Xylophonmusik quält jeden größeren Zwerg. (1)
5. **German:** Im finſteren Jagdſchloß am offenen Felsquellwaſſer patzte der affig-flatterhafte kauzig-höfliche Bäcker über ſeinem verſifften kniffligen C-Xylophon. (2)
6. **Swedish:** Flygande bäckasiner söka strax hwila på mjuka tuvor.
7. **Czech:** Příliš žluťoučký kůň úpěl ďábelské kódy.
8. **Slovak:** Starý kôň na hŕbe kníh žuje tíško povädnuté ruže, na strope sa ďateľ učí kvákať novú ódu o živote.
9. **Russian:** В чащах юга жил-был цитрус? Да, но фальшивый экземпляр! ёъ.
10. **Bulgarian:** Жълтата дюля беше щастлива, че пухът, който цъфна, замръзна като гьон.
11. **Sami (Northern):** Vuol Ruoŧa geđggiid leat máŋga luosa jačuovžža.
12. **Hungarian:** Árvíztűrő tükörfúrógép.
13. **Spanish:** El pingüino Wenceslao hizo kilómetros bajo exhaustiva lluvia y frío, añoraba a su querido cachorro.
14. **Portuguese:** O próximo vôo à noite sobre o Atlântico, põe freqüentemente o único médico. (3)
15. **French:** Les naïfs ægithales hâtifs pondant à Noël où il gèle sont sûrs d'être déçus et de voir leurs drôles d'œufs abîmés.
16. **Esperanto:** Eĥoŝanĝo ĉiuĵaŭde.
17. **Hebrew:** זה כיף סתם לשמוע איך תנצח קרפד עץ טוב בגן.
18. **Japanese** (Hiragana):

    いろはにほへど□ちりぬるを
    わがよたれぞ□つねならむ
    うゐのおくやま□けふこえて
    あさきゆめみじ□ゑひもせず (4)

Notes:

1. Other phrases commonly used in Germany include: "Ein wackerer Bayer vertilgt ja bequem zwo Pfund Kalbshaxe" and, more recently, "Franz jagt im komplett verwahrlosten Taxi quer durch Bayern", but both lack umlauts and esszet. Previously, going for the shortest sentence that has all the umlauts and special characters, I had "Grüße aus Bärenhöfe (und Óechtringen)!" Acute accents are not used in native German words, so I was surprised to discover "Óechtringen" in the Deutsche Bundespost Postleitzahlenbuch (Vorsicht! 2.8MB JPG image). It's a small village in eastern Lower Saxony. The "oe" in this case turns out to be the Lower Saxon "lengthening e" (Dehnungs-e), which makes the previous vowel long (used in a number of Lower Saxon place names such as Soest and Itzehoe), not the "e" that indicates umlaut of the preceding vowel. Many thanks to the Óechtringen-Namenschreibungsuntersuchungskomitee (Alex Bochannek, Manfred Erren, Asmus Freytag, Christoph Päper, plus Werner Lemberg who serves as the Óechtringen-Namenschreibungsuntersuchungskomiteerechtschreibungsprüfer) for their relentless pursuit of the facts in this case. Conclusion: the accent almost certainly does not

belong on this (or any other native German) word, but neither can it be dismissed as dirt o the page. To add to the mystery, it has been reported that other copies of the same edition the PLZB do not show the accent!

2. From Karl Pentzlin (Kochel am See, Bavaria, Germany): "This German phrase is suited for display by a Fraktur (broken letter) font. It contains: all common three-letter ligatures: ffi ff and all two-letter ligatures required by the Duden for Fraktur typesetting: ch ck ff fi ſſch ſſi ſſt ſſt tz (all in a manner such they are not part of a three-letter ligature), one example of f-l where German typesetting rules prohibit ligating (marked by a ZWNJ), and all German lett a...z, ä,ö,ü,ß, ſ [long s] (all in a manner such that they are not part of a two-letter Fraktur ligature)." Otto Stolz notes that "'Schloß' is now spelled 'Schloss', in contrast to 'größer' (example 4) which has kept its 'ß'. Fraktur has been banned from general use, in 1942, and long-s (ſ) has ceased to be used with Antiqua (Roman) even earlier (the latest Antiqua-ſ have seen is from 1913, but then I am no expert, so there may well be a later instance." Late Otto confirms the latter theory, "Now I've run across a book “Deutsche Rechtschreibung” (edited by Lutz Mackensen) from 1954 (my reprint is from 1956) that has kept the Antiqua-ſ in its dictionary part (but neither in the preface nor in the appendix)."

3. Diaeresis is not used in Iberian Portuguese.

4. From Yurio Miyazawa: "This poetry contains all the sounds in the Japanese language and used to be the first thing for children to learn in their Japanese class. The Hiragana version particularly neat because it covers every character in the phonetic Hiragana character set." Yurio also sent the Kanji version:

色は匂へど 散りぬるを
我が世誰ぞ 常ならむ
有爲の奥山 今日越えて
浅き夢見じ 酔ひもせず

Accented Cyrillic:

(This section contributed by Vladimir Marinov.)

In Bulgarian it is desirable, customary, or in some cases required to write accents over vowels. Unfortunately, no computer character sets contain the full repertoire of accented Cyrillic letters. With Unicode, however, it is possible to combine any Cyrillic letter with any combining accent. T appearance of the result depends on the font and the rendering engine. Here are two examples.

1. Той видя бялата коса́по главата и́ и ко́са на рамото и́, и ре́че да и́ рече́: "Пара́та по́ па́ри от па́рата, не ща пари́и", но си поми́сли: "Хей, помисли́! А́ и́ река, а́ е скочила в тази река, която щеше да тече́, а не те́че."

2. По пъ́тя пъту́ват кю́рди и югослав́яни.

HTML Features

Here is the Russian alphabet (uppercase only) coded in three different ways, which should look identical:

1. АБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ(Literal UTF-8)
2. АБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ(Decimal numeric character reference)

3. АБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ (Hexadecimal numeric character reference)

In another test, we use HTML language tags to distinguish Bulgarian, Russian, and Serbian, which have different italic forms for lowercase б, г, д, п, and/or т:

Bulgarian:	[бгдпт]	[*бгдпт*]	*Мога да ям стъкло и не ме боли.*
Russian:	[бгдпт]	[*бгдпт*]	*Я могу есть стекло, это мне не вредит.*
Serbian:	[бгдпт]	[*бгдпт*]	*Могу јести стакло а да ми не шкоди.*

Credits, Tools, and Commentary

Credits:

The "I can eat glass" phrase and the initial collection of translations: Ethan Mollick. Transcription / conversion to UTF-8: Frank da Cruz. **Albanian:** Sindi Keesan. **Afrikaans:** Johan Fourie, Kevin Poalses. **Anglo Saxon:** Frank da Cruz. **Arabic:** Najib Tounsi. **Armenian:** Vaçe Kundakçı. **Belarusian:** Alexey Chernyak. **Bengali:** Somnath Purkayastha, Deepayan Sarkar. **Bislama:** Dan McGarry. **Braille:** Frank da Cruz. **Bulgarian:** Sindi Keesan, Guentcho Skordev, Vladimir Marinov. **Cabo Verde Creole:** Cláudio Alexandre Duarte. **Chinese:** Jack Soo, Wong Pui Lam. **Chinook Jargon:** David Robertson. **Cornish:** Chris Stephens. **Croatian:** Marjan Bać e. **Czech:** Stanislav Pecha, Radovan Garabík. **Dutch:** Peter Gotink. Pim Blokland, Rob Daniel, Rob de Wit. **Erzian:** Jack Rueter. **Esperanto:** Franko Luin, Radovan Garabík. **Estonian:** Meelis Roos. **Farsi/Persian:** Payam Elahi. **Finnish:** Sampsa Toivanen. **French:** Luc Carissimo, Anne Colin du Terrail, Sean M. Burke. **Galician:** Laura Probaos. **Georgian:** Giorgi Lebanidze. **German:** Christoph Päper, Otto Stolz, Karl Pentzlin, Frank da Cruz. **Gothic:** Aurélien Coudurier. **Greek:** Ariel Glenn, Constantine Stathopoulos, Siva Nataraja. **Hebrew:** Jonathan Rosenne, Tal Barnea. **Hausa:** Malami Buba, Tom Gewecke. **Hawaiian:** na Hauʻoli Motta, Anela de Rego, Kaliko Trapp. **Hindi:** Shirish Kalele. **Hungarian:** András Rácz, Mark Holczhammer. **Icelandic:** Andrés Magnússon. **International Phonetic Alphabet (IPA):** Siva Nataraja / Vincent Ramos. **Irish:** Michael Everson, Marion Gunn, James Kass, Curtis Clark. **Italian:** Thomas De Bellis. **Japanese:** Makoto Takahashi, Yurio Miyazawa. **Kirchröadsj:** Roger Stoffers. **Kreyòl:** Sean M. Burke. **Korean:** Jungshik Shin. **Lëtzebuergescht:** Stefaan Eeckels. **Lithuanian:** Gediminas Grigas. **Lojban:** Edward Cherlin. **Lusatian:** Ronald Schaffhirt. **Macedonian:** Sindi Keesan. **Malay:** Zarina Mustapha. **Manx:** Éanna Ó Brádaigh. **Marathi:** Shirish Kalele. **Marquesan:** Kaliko Trapp. **Middle English:** Frank da Cruz. **Milanese:** Marco Cimarosti. **Mongolian:** Tom Gewecke. **Napoletano:** Diego Quintano. **Navajo:** Tom Gewecke. **Nórdicg:** Yẃlyan Rott. **Norwegian:** Herman Ranes. **Old Irish:** Michael Everson. **Old Norse:** Andrés Magnússon. **Papiamentu:** Bianca and Denise Zanardi. **Pashto:** N.R. Liwal. **Pfälzisch:** Dr. Johannes Sander. **Picard:** Philippe Mennecier. **Polish:** Juliusz Chroboczek. **Portuguese:** "Cláudio" Alexandre Duarte, Bianca and Denise Zanardi, Pedro Palhoto Mat. **Québécois:** Laurent Detillieux. **Roman:** Pierpaolo Bernardi. **Romanian:** Juliusz Chroboczek, Ionel Mugurel. **Ruhrdeutsch:** "Timwi". **Russian:** Alexey Chernyak, Serge Nesterovitch. **Sami:** Anne Colin du Terrail, Luc Carissimo. **Sanskrit:** Siva Nataraja / Vincent Ramos. **Sächsisch:** André Müller. **Schwäbisch:** Otto Stolz. **Scots:** Jonathan Riddell. **Serbian:** Sindi Keesan, Ranko Narancic, Boris Daljevic, Szilvia Csorba. **Slovak:** G. Adam Stanislav, Radovan Garabík. **Slovenian:** Albert Kolar. **Spanish:** Aleida Muñoz, Laura Probaos. **Swahili:** Ronald Schaffhirt. **Swedish:** Christian Rose, Bengt Larsson. **Taiwanese:** Henry H. Tan-Tenn. **Tagalog:** Jim Soliven. **Tamil:** Vasee Vaseeharan. **Tibetan:** D. Germano, Tom Gewecke. **Thai:** Alan Wood's wife. **Turkish:** Vaçe Kundakçı, Tom Gewecke, Merlign Olnon. **Ukrainian:** Michael Zajac. **Urdu:** Mustafa Ali. **Vietnamese:** Dixon Au, [James] Đỗ Bá Phước 杜 伯 福. **Walloon:** Pablo

Saratxaga. **Welsh:** Geiriadur Prifysgol Cymru (Andrew). **Yiddish:** Mark David, **Zeneise:** Angelo Pavese.

Tools Used to Create This Web Page:

The UTF8-aware Kermit 95 terminal emulator on Windows, to a Unix host with the EMACS text editor. Kermit 95 displays UTF-8 and also allows keyboard entry of arbitrary Unicode BMP characters as 4 hex digits, as shown HERE. Hex codes for Unicode values can be found in The Unicode Standard (recommended) and the online code charts. When submissions arrive by email encoded in some other character set (Latin-1, Latin-2, KOI, various PC code pages, JEUC, etc), I use the TRANSLATE command of C-Kermit on the Unix host (where I read my mail) to convert the character set to UTF-8 (I could also use Kermit 95 for this; it ha the same TRANSLATE command). That's it -- no "Web authoring" tools, no locales, no "smart" anything. It's just plain text, nothing more. By the way, there's nothing special abo EMACS -- any text editor will do, providing it allows entry of arbitrary 8-bit bytes as text, including the 0x80-0x9F "C1" range. EMACS 21.1 actually supports UTF-8; earlier versions don't know about it and display the octal codes; either way is OK for this purpose.

Commentary:

Date: Wed, 27 Feb 2002 13:21:59 +0100
From: "Bruno DEDOMINICIS" <b.dedominicis@cite-sciences.fr>
Subject: Je peux manger du verre, cela ne me fait pas mal.

I just found out your website and it makes me feel like proposing an interpretation of the choice of this peculiar phrase.

Glass is transparent and can hurt as everyone knows. The relation between people and civilisations is sometimes effusional and more often rude. The concept of breaking frontier through globalization, in a way, is also an attempt to deny any difference. Isn't "transparen the flag of modernity? Nothing should be hidden any more, authority is obsolete, and the new powers are supposed to reign through loving and smiling and no more through coercion...

Eating glass without pain sounds like a very nice metaphor of this attempt. That is, frontier should become glass transparent first, and be denied by incorporating them. On the revers it shows that through globalization, frontiers undergo a process of displacement, that is, w they are not any more speakable, they become repressed from the speech and are therefor incorporated and might become painful symptoms, as for example what happens when or tries to eat glass.

The frontiers that used to separate bodies one from another tend to divide bodies from wit and make them suffer.... The chosen phrase then appears as a denial of the symptom that might result from the destitution of traditional frontiers.

Best,
Bruno De Dominicis, Paris, France

Other Unicode pages onsite:

- Peace in All Languages
- Frank's Compulsive Guide to Postal Addresses (especially the Index)
- Representing Middle English on the Web with UTF-8
- The Kermit Bibliography (in UTF-8)
- Interchange of Non-English Computer Text (UTF-8 math and box-drawing)

- Unicode Table (in UTF-8)

Unicode samplers offsite:

- Michael Everson's Bibliography of Typography and Scripts
- Sample Unicode Test Pages and Script Links
- I don't know, I only work here
- Anyone can be provincial!
- Transcriptions of "Unicode"
- Example Unicode Usage for Business Applications
- UTF-8 and Unicode FAQ for Unix/Linux

Unicode fonts:

- Unicode Fonts for Windows Computers (Alan Wood)
- Unicode Fonts and Tools for X11 (Markus Kuhn)
- Everson Mono (Michael Everson)
- Agfa Monotype

[Kermit 95] [K95 Screen Shots] [C-Kermit] [Kermit Home] [Display Problems?] [The Unicode Consortium]

UTF-8 Sampler / The Kermit Project / Columbia University / kermit@columbia.edu