```
UTF-8 SAMPLER ¥・£・€・$・¢・¢・G・F・£・メカ・トサ・トサ・トサ・トサ・トサ・ン P・トサ・ン Frank da Cruz The Kermit Project - Columbia University
New York City fdc@columbia.edu Last update: Wed Oct 22 12:36:23 2008 [ 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): سالم a good start! Poetry From the Anglo-Saxon Rune Poem (Rune version): שואלי און אינויים אוויים און אינויים אווייים אוויים אוויים אוויים אוויים אוויים אוויים אוויים אוויים אוויים א
West Midlands): An preost wes on leoden, Lagamon was ihoten He wes Leovenaões sone -- liõe him be Drihten. He wonede at Ernlege at æõelen are
chirechen, Uppen Sevarne stabe, sel bar him buhte, Onfest Radestone, ber 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): Monotonic: 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
Polytonic: Τὴ γλῶσσα μοῦ ἔδωσαν ἑλληνικὴ τὸ σπίτι φτωχικὸ στὶς ἀμμουδιὲς τοῦ Ὁμήρου. Μονάχη ἔγνοια ἡ γλῶσσα μου στὶς ἀμμουδιὲς τοῦ
Όμήρου. ἀπὸ τὸ Ἄξιον ἐστί τοῦ Ὀδυσσέα Ἐλύτη 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 Veṗxis Ţġaosani,
`Th, The Knight in the Tiger's Skin (Georgian): ვეპხის ტყაოხანი შოთა რუსთაველი ღმერთხი შემვედრე, ნუთუ კვლა დამხსნას სოფლისა შრომასა,
ცეცხლს, წყალსა და მიწასა, ჰაერთა თანა მრომასა: მომცნეს ფრთენი და აღვფრინდე, მიგჰხედე მას ჩემსა ნდომასა, დღისით და ღამით გჰხედვიდე
მზისა ელვათა კრთომაასა. Tamil poetry of Cupiramaniya Paarathiyar, சுப்ரமணிய பாரதியார் (1882–1921): யாமறிந்த மரெழிகளிலઉ
தமிழ்மரெழி பரேல் இனிதாவத எங்கும் காணரேம், பாமரராய் விலங்குகளாய், உலகனதைதம் இகழ்ச்சிசரெலப் பான்மகை கடெட்டு,
நாமமது தமிழரனெக் கரெண்ட இங்கு வாழ்ந்திடதல் நன்றரே? சரெல்லீர்! தமேதரத் தமிழரேசை உலகமலொம் பரவும்வகசையெதல்
வணேட்டம். I Can Eat Glass And from the sublime to the ridiculous, here is a certain phrase<sup>1</sup> in an assortment of languages: Sanskrit: कार्च
शक्नोम्पर्तुत्म् । नोपहनिस्ति माम् || Sanskrit (standard transcription): kācam śaknomyattum; nopahinasti mām. Classical Greek: ὕαλον φαγεῖν δύναμαι·
τοῦτο οὔ με βλάπτει. Greek (monotonic): Μπορώ να φάω σπασμένα γυαλιά χωρίς να πάθω τίποτα. Greek (polytonic): Μπορῶ νὰ φάω
σπασμένα γυαλιὰ χωρὶς νὰ πάθω τίποτα. Etruscan: (NEEDED) Latin: Vitrum edere possum; mihi non nocet. Old French: Je puis mangier del
voirre. Ne me nuit. French: Je peux manger du verre, ça ne me fait pas de mal. Provençal / Occitan: Pòdi manjar de veire, me nafrariá pas.
Québécois: J'peux manger d'la vitre, ça m'fa pas mal. Walloon: Dji pou magnî do vêre, çoula m' freut nén må. Champenois: (NEEDED) Lorrain:
(NEEDED) Picard: Ch'peux mingi du verre, cha m'foé mie n'ma. Corsican: (NEEDED) Jèrriais: (NEEDED) Kreyòl Ayisyen: Mwen kap manje vè, li pa
blese'm. Basque: Kristala jan dezaket, ez dit minik ematen. Catalan / Català: Puc menjar vidre, que no em fa mal. Spanish: Puedo comer vidrio, no
me hace daño. Aragones: Puedo minchar beire, no me'n fa mal . Galician: Eu podo xantar cristais e non cortarme. European Portuguese: Posso
comer vidro, não me faz mal. Brazilian Portuguese (8): Posso comer vidro, não me machuca. Caboverdiano/Kabuverdianu: M' podê cumê vidru, ca
ta maguâ-m'. Papiamentu: Ami por kome glas anto e no ta hasimi daño. Italian: Posso mangiare il vetro e non mi fa male. Milanese: Sôn bôn de
magnà el véder, el me fa minga mal. Roman: Me posso magna' er vetro, e nun me fa male. Napoletano: M' pozz magna' o'vetr, e nun m' fa mal.
Sicilian: Puotsu mangiari u vitru, nun mi fa mali. Venetian: Mi posso magnare el vetro, no'l me fa mae. Zeneise (Genovese): Pòsso mangiâ o
veddro e o no me fà mâ. Romansch (Grischun): Jau sai mangiar vaider, senza che quai fa donn a mai. Romany / Tsigane: (NEEDED) Romanian: Pot
să mănânc sticlă și ea nu mă rănește. Esperanto: Mi povas manĝi vitron, ĝi ne damaĝas min. Pictish: (NEEDED) Breton: (NEEDED) Cornish: Mý a yl
dybry gwéder hag éf ny wra ow ankenya. Welsh: Dw i'n gallu bwyta gwydr, 'dyw e ddim yn gwneud dolur i mi. Manx Gaelic: Foddym gee
Ním: géna. Irish: Is féidir liom gloinne a ithe. Ní dhéanann sí dochar ar bith dom. Scottish Gaelic: S urrainn dhomh gloinne ithe; cha ghoirtich i
mi. Anglo-Saxon (Runes): Ik-MFX-XIF4-MF1F1-F+W-M1-+M-MMFRMIFP-MM: Anglo-Saxon (Latin): Ic mæg glæs eotan ond hit ne hearmiað me. Middle English:
Ich canne glas eten and hit hirtip me nou3t. English: I can eat glass and it doesn't hurt me. English (IPA): [aɪ kæn iɪt glaːs ænd ɪt dɐz nɒt hɜːt
(Latin): Ek get etið gler án þess að verða sár. Norsk / Norwegian (Nynorsk): Eg kan eta glas utan å skada meg. Norsk / Norwegian (Bokmål): Jeg
kan spise glass uten å skade meg. Føroyskt / Faroese: Eg kann eta glas, skaðaleysur. Íslenska / Icelandic: Ég get etið gler án þess að meiða mig.
Svenska / Swedish: Jag kan äta glas utan att skada mig. Dansk / Danish: Jeg kan spise glas, det gør ikke ondt på mig. Sønderjysk: Æ ka æe glass
uhen at det go mæ naue. Frysk / Frisian: Ik kin glês ite, it docht me net sear. Nederlands / Dutch: Ik kan glas eten, het doet mij geen kwaad.
Kirchröadsj/Bôchesserplat: Iech ken glaas èèse, mer 't deet miech jing pieng. Afrikaans: Ek kan glas eet, maar dit doen my nie skade nie.
Lëtzebuergescht / Luxemburgish: Ech kan Glas iessen, daat deet mir nët wei. Deutsch / German: Ich kann Glas essen, ohne mir zu schaden.
Ruhrdeutsch: Ich kann Glas verkasematuckeln, ohne dattet mich wat jucken tut. Langenfelder Platt: Isch kann Jlaas kimmeln, uuhne datt mich
datt weh dääd. Lausitzer Mundart ("Lusatian"): Ich koann Gloos assn und doas dudd merr ni wii. Odenwälderisch: Iech konn glaasch voschbachteln
ohne dass es mir ebbs daun doun dud. Sächsisch / Saxon: 'sch kann Glos essn, ohne dass'sch mer wehtue. Pfälzisch: Isch konn Glass fresse ohne
dasses mer ebbes ausmache dud. Schwäbisch / Swabian: I kå Glas frässa, ond des macht mr nix! 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. Hungarian: Meg
tudom enni az üveget, nem lesz tőle bajom. Suomi / Finnish: Voin syödä lasia, se ei vahingoita minua. Sami (Northern): Sáhtán borrat lása, dat ii
leat bávččas. Erzian: Мон ярсан суликадо, ды зыян эйстэнзэ а ули. Northern Karelian: Mie voin syvvä lasie ta minla ei ole kipie. Southern
Karelian: Minä voin syvvä st'oklua dai minule ei ole kibie. Vepsian: (NEEDED) Votian: (NEEDED) Livonian: (NEEDED) Estonian: Ma võin klaasi süüa,
see ei tee mulle midagi. Latvian: Es varu ēst stiklu, tas man nekaitē. Lithuanian: Aš galiu valgyti stiklą ir jis manęs nežeidžia Old Prussian:
```

```
(NEEDED) Sorbian (Wendish): (NEEDED) Czech: Mohu jíst sklo, neublíží mi. Slovak: Môžem jesť sklo. Nezraní ma. Polska / Polish: Mogę jeść szkło i
mi nie szkodzi. Slovenian: Lahko jem steklo, ne da bi mi škodovalo. Croatian: Ja mogu jesti staklo i ne boli me. Serbian (Latin): Ja mogu da jedem
staklo. Serbian (Cyrillic): Ја могу да једем стакло. Macedonian: Можам да јадам стакло, а не ме штета. Russian: Я могу есть стекло, оно мне не
вредит. Belarusian (Cyrillic): Я магу есці шкло, яно мне не шкодзіць. Belarusian (Lacinka): Ja mahu jeści škło, jano mne ne škodzić. Ukrainian: Я
можу їсти шкло, й воно мені не пошкодить. Bulgarian: Мога да ям стъкло, то не ми вреди. Georgian: მინას ეჭამ და არა მტკივა. Armenian: Կրնամ
ապակի ուտել և ինծի անհանգիստ չըներ։ Albanian։ Unë mund të ha qelq dhe nuk më gjen gjë. Turkish։ Cam yiyebilirim, bana zararı
dokunmaz. Turkish (Ottoman): بهنور الله عوي باع Bangla / Bengali: আমি কাঁচ খতেে পারন্ তাততে আমার কণেনণে ক্ষত হিয় না। Marathi: मी काच
खाऊ शकतो, मला ते दुखत नाही. Hindi: मैं काँच खा सकता हूँ और मुझे उससे कोई चोट नहीं पहुंचती. Tamil: நான் கண்ணாடி சாப்பிடிவனே், அதனால் எனக்கு
Aramaic: (NEEDED) Maltese: Nista' niekol il-ħģieġ u ma jagħmilli .ينم ل ؤي ال اذه و ج إجزل الكأ يلع رداق انأ . Aramaic (neede il-ħģieġ u ma jagħmilli
xejn. Hebrew(2): ייוו טשינ רימ טוט סע זוא זָאלג זסע זעק דיא Yiddish(2): אל קיזמ אל הזו תיכוכז לוכאל לוכי ינא. Judeo-Arabic: (NEEDED) Ladino:
(NEEDED) Gə'əz: (NEEDED) Amharic: (NEEDED) Twi: Metumi awe tumpan, snys me hwee. Hausa (Latin): Inā iya taunar gilāshi kuma in gamā lāfiyā.
Hausa (Ajami) (2): كوت كوا إنوا (2) Yoruba(3): Mo lè ję dígí, kò ní pa mí lára. Lingala: Nakokî kolîya biténi bya milungi, ekosála ngáî
mabé té. (Ki)Swahili: Naweza kula bilauri na sikunyui. Malay: Saya boleh makan kaca dan ia tidak mencederakan saya. Tagalog: Kaya kong kumain
nang bubog at hindi ako masaktan. Chamorro: Siña yo' chumocho krestat, ti ha na'lalamen yo'. Javanese: Aku isa mangan beling tanpa lara.
Burmese: ကိယ်ဝန်တင်္ခေါက်ယ်ဝန်မ မိယက်စားနိုင်သည်။ ၎င်းက်ရငာင့် ထိခိုက်မ်ဟု မရ်ဟိပာ။ (9) Vietnamese (quốc ngữ): Tôi có thể ăn thủy tinh mà không hại gì.
Vietnamese (nôm) (4): 些 👨 世 咹 水 晶 👊 空 🙎 🗈 害 咦 Khmer: ខ្សាំអាចព្យុំកញ្ញចក់ឋាន ដលោយគុមានបញ្ជហារ Lao: 2อ้อภิ่มแท้อได้โดยที่มัมปีได้เอัดใช้เออัยเจ้บ. Thai:
ฉันกินกระจกได้ แต่มันไม่ทำให้ฉันเจ็บ Mongolian (Cyrillic): Би шил идэй чадна, надад хортой биш Mongolian (Classic) (5): 🕫 キペラペ ๙- รุง๙- ร
我能吞下玻璃而不伤身体。 Chinese (Traditional): 我能吞下玻璃而不傷身體。 Taiwanese(6): Góa ē-tàng chiảh po-lê, mā bē tiỏh-siong. Japanese:
私はガラスを食べられます。それは私を傷つけません。 Korean: 나는 유리를 먹을 수 있어요. 그래도 아프지 않아요 Bislama: Mi save kakae glas, hemi no save
katem mi. Hawaiian: Hiki ia'u ke 'ai i ke aniani; 'a'ole nō lā au e 'eha. Marquesan: E ko'ana e kai i te karahi, mea 'ā, 'a'e hauhau. Chinook Jargon:
Naika məkmək kakshət labutay, pi weyk ukuk munk-sik nay. Navajo: Tsésǫ' yisháago bííníshghah dóó doo shił neezgai da. Cherokee (and Cree,
Ojibwa, Inuktitut, Náhuatl, Quechua, and other American languages): (NEEDED) Garifuna: (NEEDED) Gullah: (NEEDED) Lojban: mi kakne le nu citka
le blaci .iku'i le se go'i na xrani mi Nórdicg: Ljœr ye caudran créneþ ý jor cẃran. (Additions, corrections, completions, gratefully accepted.) For
testing purposes, some of these are repeated in a monospace font . . . Euro Symbol: €. Greek: Μπορώ να φάω σπασμένα γυαλιά χωρίς να πάθω
τίποτα. Íslenska / Icelandic: Ég get etið gler án þess að meiða mig. Polish: Mogę jeść szkło, i mi nie szkodzi. Romanian: Pot să mănânc sticlă și ea
nu mă rănește. Ukrainian: Я можу їсти шкло, й воно мені не пошкодить. Armenian: Կրնամ ապակի ուտել և ինծի անհանգիստ չրներ: Georgian:
მანას გჭამ და არა მტკივა. Hindi: मैं काँच खा सकता हूँ, मुझे उस से कोई पीडा नहीं होती. Hebrew(2): דיא יל קיזמ אל הזו תיכוכז לוכאל לוכי ינא
ער ן ועק (ידימ טוט סע דוא אַ אַרלג ן סע דעק Arabic(2): יידור טשינ רימ טוט סע דוא אַ אַרלג ן סע דעק. Tapanese: 私はガラスを食べられます。それは私を傷つけません。 Thai:
ฉันกินกระจกได้ แต่มันไม่ทำให้ฉันเจ็บ Notes: The "I can eat glass" phrase and initial translations (about 30 of them) were borrowed from Ethan
Mollick's I Can Eat Glass page (which disappeared on or about June 2004) and converted to UTF-8. Since Ethan's original page is gone, I should
mention that his purpose was to offer travelers a phrase they could use in any country that would command a certain kind of respect, or at
least get attention. See Credits for the many additional contributions since then. When submitting new entries, the word "hurt" (if you have a
choice) is used in the sense of "cause harm", "do damage", or "bother", rather than "inflict pain" or "make sad". In this vein Otto Stolz
comments (as do others further down; personally I think it's better for the purpose of this page to have extra entries and/or to show a
greater repertoire of characters than it is to enforce a strict interpretation of the word "hurt"!): 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 read: "Ich kann Glas essen ohne mir zu schaden." rather than: "Ich kann Glas essen, ohne mir weh zu tun." (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. That was from July 2004. In December 2007, Otto writes again:
Hello Frank, in days of yore, I had written: > "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. The latest revision (2006) of the official German orthography has revived
the comma around infinitive clauses commencing with ohne, or 5 other conjunctions, or depending from a noun or from an announcing
demonstrative (http://www.ids-mannheim.de/reform/regeln2006.pdf, §75). So, it's again: Ich kann Glas essen, ohne mir zu schaden. Best
wishes, Otto Stolz 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
wanted, not just the ones that say (NEEDED). Correct right-to-left display of these languages depends on the capabilities of your browser. The
period should appear on the left. In the monospace Yiddish example, the Yiddish digraphs should occupy one character cell. Yoruba: The third
word is Latin letter small 'j' followed by small 'e' with U+0329, Combining Vertical Line Below. This displays correctly only if your Unicode
font includes the U+0329 glyph and your browser supports combining diacritical marks. The Lingala and Indic examples also include combining
sequences. Includes Unicode 3.1 (or later) characters beyond Plane 0. The Classic Mongolian example should be vertical, top-to-bottom and
left-to-right. But such display is almost impossible. Also no font yet exists which provides the proper ligatures and positional variants for the
characters of this script, which works somewhat like Arabic. 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 well enough to read. More to the point, though, a
```

written standard using Chinese characters has never developed, so a significant minority of words are represented with different candidate

```
characters, depending on one's personal preference or etymological theory. In this sentence, for example, "-tang", "chiah", "ma" and "be" are
problematic using Chinese characters. "Góa" (I/me) and "po-lê" (glass) are as written in other Sinitic languages (e.g. Mandarin, Hakka)." Wagner
Amaral of Pinese & Amaral Associados notes that the Brazilian Portuguese sentence for "I can eat glass" should be identical to the Portuguese
one, as the word "machuca" means "inflict pain", or rather "injuries". The words "faz mal" would more correctly translate as "cause harm".
Burmese: In English the first person pronoun "I" stands for both genders, male and female. In Burmese (except in the central part of Burma)
kyundaw (က်ယ်ဝန်တင်္ဘေ) for male and kyanma (က်ယ်ဝန်မ) for female. Using here a fully-compliant Unicode Burmese font -- sadly one and only
Padauk Graphite font exists -- rendering using graphite engine. CLICK HERE to test Burmese characters. 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) letters (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): English: The quick brown fox jumps over the lazy dog. Jamaican: Chruu, a kwik di kwik brong fox a jomp huova di liezi daag de, yu no
siit? Irish: "An bfuil do croí ag bualad ó faitíos an grá a meall lena póg éada ó slí do leasa tú?" "D'fuascail Íosa Úrmac na hÓige Beannaite pór Éava
agus Ádaim." Dutch: Pa's wijze lynx bezag vroom het fikse aquaduct. German: Falsches Üben von Xylophonmusik quält jeden größeren Zwerg. (1)
German: Im finsteren Jagdschloß am offenen Felsquellwasser patzte der affig-flatterhafte kauzig-höfliche Bäcker über seinem versifften
kniffligen C-Xylophon. (2) Swedish: Flygande bäckasiner söka strax hwila på mjuka tuvor. Icelandic: Sævör grét áðan því úlpan var ónýt. Polish:
Pchnąć w tę łódź jeża lub ośm skrzyń fig. Czech: Příliš žluťoučký kůň úpěl ďábelské kódy. Slovak: Starý kôň na hŕbe kníh žuje tíško povädnuté
ruže, na stĺpe sa ďateľ učí kvákať novú ódu o živote. Greek (monotonic): ξεσκεπάζω την ψυχοφθόρα βδελυγμία Greek (polytonic): ξεσκεπάζω τὴν
ψυχοφθόρα βδελυγμία Russian: Съешь же ещё этих мягких французских булок да выпей чаю. Russian: В чащах юга жил-был цитрус? Да, но
фальшивый экземпляр! ёъ. Bulgarian: Жълтата дюля беше щастлива, че пухът, който цъфна, замръзна като гьон. Sami (Northern): Vuol Ruoŧa
geðggiid leat mánga luosa ja čuovžža. Hungarian: Árvíztűrő tükörfúrógép. Spanish: El pingüino Wenceslao hizo kilómetros bajo exhaustiva lluvia y
frío, añoraba a su querido cachorro. Portuguese: O próximo vôo à noite sobre o Atlântico, põe freqüentemente o único médico. (3) French: Les
naïfs ægithales hâtifs pondant à Noël où il gèle sont sûrs d'être déçus et de voir leurs drôles d'œufs abîmés. Esperanto: Eĥoŝanĝo ĉiuĵaŭde.
Hebrew: אַב רוט אַע דפרק חצנת דיא עומשל סתס דיכ הו , וגב בוט אַע דפרק הצנת דיא עומשל סתס דיכ הו , Hebrew: אונד בוט אָע דפרק הצנת דיא עומשל
あさきゆめみじ ゑひもせず (4) Notes: 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: 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 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 on the page. To add to the mystery, it has been reported that other copies of the same edition of the PLZB do not show
the accent! UPDATE (March 2006): David Krings was intrigued enough by this report to contact the mayor of Ebstorf, of which Oechtringen is
a borough, who responded: Sehr geehrter Mr. Krings, wenn Oechtringen irgendwo mit einem Akzent auf dem O geschrieben wurde, dann kann
das nur ein Fehldruck sein. Die offizielle Schreibweise lautet jedenfalls "Oechtringen". Mit freundlichen Grüssen Der
Samtgemeindebürgermeister i.A. Lothar Jessel 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 ffl fft and all two-letter ligatures required by the Duden
for Fraktur typesetting: ch ck ff fi fl ft ll ſch ſi ſſ ſt 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 letters 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-f I have seen is from 1913, but then I am no expert, so there may well be a later instance." Later 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-f in its dictionary part (but neither in the preface nor in the appendix)." Diaeresis is not used in Iberian Portuguese. 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 is 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. The appearance of the result depends on the font and the rendering engine. Here are two examples. Той видя
бялата коса́ по главата и́ и ко́са на рамото и́, и ре́че да и́ рече́: "Пара́та по́ па́ри от па́рата, не ща пари́!", но си поми́сли: "Хей, помисли́ си! А и́ река,
á е скочила в тази река, която щеше да течé, а не тéче." По пъ́тя пъту́ват кю́рди и югославя́ни. HTML Features Here is the Russian alphabet
(uppercase only) coded in three different ways, which should look identical: АБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫБЭЮЯ (Literal UTF-8)
АБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ (Decimal numeric character reference) 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. Burmese: "cetanapa". Cabo Verde Creole: Cláudio Alexandre Duarte. Catalán: Jordi
Bancells. Chinese: Jack Soo, Wong Pui Lam. Chinook Jargon: David Robertson. Cornish: Chris Stephens. Croatian: Marjan Baće. 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. Faroese: Jón Gaasedal. 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, David Krings,
Frank da Cruz. Gothic: Aurélien Coudurier. Greek: Ariel Glenn, Constantine Stathopoulos, Siva Nataraja, Christos Georgiou. Hebrew: Jonathan
Rosenne, Tal Barnea. Hausa: Malami Buba, Tom Gewecke. Hawaiian: na Hauʻoli Motta, Anela de Rego, Kaliko Trapp. Hindi: Shirish Kalele, Nitin
```

Dahra. Hungarian: András Rácz, Mark Holczhammer. Icelandic: Andrés Magnússon, Sveinn Baldursson. International Phonetic Alphabet (IPA): Siva

Nataraja / Vincent Ramos. Irish: Michael Everson, Marion Gunn, James Kass, Curtis Clark. Italian: Thomas De Bellis. Jamaican: Stephen J. Cherin. Japanese: Makoto Takahashi, Yurio Miyazawa. Karelian: Aleksandr Semakov. Khmer: Tola Sann. Kirchröadsj: Roger Stoffers. Kreyòl: Sean M. Burke. Korean: Jungshik Shin. Langenfelder Platt: David Krings. Lao: Tola Sann. Lëtzebuergescht: Stefaan Eeckels. Lingala: Denis Moyogo Jacquerye (Nkóta ya Kóngo míbalé). (Nkóta ya Kóngo míbal Lithuanian: Gediminas Grigas. Lojban: Edward Cherlin. Lusatian: Ronald Schaffhirt. Macedonian: Sindi Keesan. Malay: Zarina Mustapha. Maltese: Kenneth Joseph Vella. 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. Odenwälderisch: Alexander Heß. 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, Paweł Przeradowski. Portuguese: "Cláudio" Alexandre Duarte, Bianca and Denise Zanardi, Pedro Palhoto Matos, Wagner Amaral. Québécois: Laurent Detillieux. Roman: Pierpaolo Bernardi. Romanian: Juliusz Chroboczek, Ionel Mugurel. Romansch: Alexandre Suter. 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, O. Dag. 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 has 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 about 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 frontiers through globalization, in a way, is also an attempt to deny any difference. Isn't "transparency" 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, frontiers should become glass transparent first, and be denied by incorporating them. On the reverse, it shows that through globalization, frontiers undergo a process of displacement, that is, when they are not any more speakable, they become repressed from the speech and are therefore incorporated and might become painful symptoms, as for example what happens when one tries to eat glass. The frontiers that used to separate bodies one from another tend to divide bodies from within 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 and resources offsite: Unicode Code ConverterUnicode Code Conversion (converts among different Unicode encoding forms and notations). Michael Everson's Bibliography of Typography and Scripts Does your browser support Unicode English? (James Kass) 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: Code 2000 (James Kass) 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