Entry #2525 [older|newer] / Entrée #2525 [précédente|suivante]:

2018-06-11 (lundi)

Club Contexte : Communications Papier (ou pensées en vrac sur les lettres et leur forme)

Les « communications papier » dont je veux parler c'est, bien sûr, l'écriture, mais le titre de cette entrée s'abrège en CCCP, ce qui n'a absolument rien à voir avec CCCP qui est l'abréviation de « Coopérative Communiste Consommateurs Perpétuelle » en russe, et de l'URSS. Selon votre navigateur et les polices installées sur votre ordinateur, le 'C' de « Club Contexte » (U+0043 LATIN CAPITAL LETTER C) et le 'C' de « Coopérative Cosoz » (U+0421 CYRILLIC CAPITAL LETTER ES) ne se ressemblent pas du tout. Ce sont absolument des caractères différents. Ce sont absolument des caractères différents pour Unicode, comme vous pouvez le vérifier en recherchant 'C' dans cette page, ou en copiant-collant le caractère et en le recherchant dans Google, ou quelque chose de ce genre. (Hum, à propos de Google, je ne sais pas si c'est un bug ou quoi, mais si je cherche « CCCP » avec le « C » cyrillique, ça ne trouve rien, mais si je cherche « CCCP » avec le « C » latin, ça trouve des pages avec des « CCCP » cyrilliques — pas « C » cyrillique, mais comme le 'C' cyrillique est plutôt une 'S', enfin, se prononce quelque chose ressemblant à *ès*, je devrais dire « l'C' cyrillique » : il va falloir que j'interroge le Club Contexte en Section Politique pour faire les détails.)

Bref, vous aurez compris l'idée : ce qu'est une lettre est quelque chose d'assez délicat. Un mathieu a envie de dire qu'une lettre est une classe d'équivalence pour la relation d'équivalence « être la même lettre », mais qu'est-ce que c'est que cette relation d'équivalence, justement.

Unicode doit régulièrement prendre des décisions *to conflate or to disunify?*, autrement dit, décider si on doit considérer que deux machins sont ou non la même lettre. Et ça crée des problèmes, mais ça en résout aussi. Et ça crée des problèmes parce que ça en résout aussi. Et ça en résout aussi parce que ça crée des problèmes. Et ainsi de suite. Notamment parce que la relation « être la même lettre » a furieusement tendance à ne pas être une relation d'équivalence : elle n'est pas systématiquement transitive. Je considère que Eos et Aar sont la même lettre, une autre que Aar et aun sont la même lettre, et une troisième que Eos et aun ne sont pas du tout la même lettre).

Voici quelques unes des choses, en vrac, que j'ai apprises en me documentant à gauche et à droite sur Unicode et sur l'histoire des formes d'écriture. On peut le faire chez Unicode, mais ça demande un sacré investissement, et ça ne rapporte pas grand-chose, et ça ne rapporte pas grand-chose parce qu'on ne peut être là indépendamment des autres : j'ai essayé de trouver un fil directeur, mais c'est trop difficile, il n'y a juste aucune logique dans cette collection de faits.)

Commençons par question basique : fait-il considérer que la première lettre des alphabets latin, grec et cyrillique, c'est-à-dire A, 'A' et 'А' sont la même lettre ? Dans Unicode, ce sont U+0041 LATIN CAPITAL LETTER A, U+0391 GREEK CAPITAL LETTER ALPHA et U+0410 CYRILLIC CAPITAL LETTER A, c'est-à-dire qu'Unicode a tranché pour dire non.

D'un côté, considérer que non (comme le fait Unicode), ie., que ce sont trois lettres différentes, est très confusant pour les gens qui ne voient aucune différence (notamment sur leur écran), cela peut être la cause de toutes sortes de problèmes informatiques, notamment d'attriages délibérés. Inversement, si on considère que oui, ie., qu'il s'agit de la même lettre, on a du coup un sacré problème de lisibilité : le lecteur moyen ne reconnaîtra pas les gens qui pensent aller à google.com — et de toute façon votre navigateur vous montrera probablement le Punycode xn--google-5ba.com si vous y allez — et, surtout, il ne reconnaîtra pas les gens qui essaient de le tromper en utilisant des homographes. Ce sont absolument des considérations différentes, ie., identificateurs en Unicode, et j'attends le moment où quelqu'un aura malicieusement introduit un trou de sécurité quelque part en nommant une variable 'А' (U+0410 CYRILLIC CAPITAL LETTER A) pour « cacher » une variable 'A' (U+0041 LATIN CAPITAL LETTER A) dans quelque logiciel open-source (ou bien, bien sûr, dans le logiciel de quelque politicien.) Ou à défaut de bugs, simplement de petites tracasseries : si je veux vérifier que je ne me suis pas trompé dans l'ordre de mes 'A', 'A' et 'А', c'est compliqué, ie., je dois taper 'A', 'A' et 'А' à la main.

D'un autre côté, considérer que 'A', 'A' et 'А' sont la même lettre serait très gênant quand il s'agit de passer en minuscules, par exemple (a 'et' n'a pas d'différent certainement, et 'a' différent peut-être aussi), ou si on veut développer des polices spécifiques à l'un ou l'autre des alphabets. (L'affichage du grec sur mon ordinateur est quasiment illisible, par exemple, parce que mes polices qui ont quelque chose de correct pour le latin ne connaissent pas le grec, et inversement, je ne sais pas pourquoi mon navigateur tend à préférer cette police, je vois souvent cette seule lettre dans une police visiblement différente. Imagine à quel point ça va être compliqué si on commence à utiliser des polices différentes pour le latin étaient prises dans une police prévue pour l'alphabet latin et le grec dans une autre.) Et je vais revenir plus bas sur la question de l'écriture cursive.

Et encore, au moins 'A', 'A' et 'А' sont historiquement la même origine : ce sont des lettres cognates ou dérivées les unes des autres. (Disons que le alpha majuscule grec est un A majuscule qui a pris un accent circonflexe — modeste mais extrêmement visible — et que le А cyrillique est un A majuscule qui a pris un LETTER AU phénicien et a donné naissance au 'А', enfin, au 'A' (U+0100 DJI ITALIC LETTER a italique puis au 'A' latin et bien plus tard au 'A' cyrillique.) Bref, c'est quand même la même lettre. Mais ça ne suffit pas à en faire la même lettre, sinon on aurait U+0100 LATIN CAPITAL LETTER A WITH CIRCUMFLEX, U+0391 U+0302 GREEK CAPITAL LETTER ALPHA WITH CIRCUMFLEX et U+0410 U+0302 CYRILLIC CAPITAL LETTER A WITH CIRCUMFLEX, U+0041 GREEK CAPITAL LETTER RHO et U+0410 CYRILLIC CAPITAL LETTER A), il y a clairement un intrus, et c'est le premier, parce que l'accent circonflexe latin a été historiquement ajouté au latin bien plus tôt que le grec ou le cyrillique. (Et, bien sûr, Unicode a tranché pour dire que U+0391 U+0302 GREEK CAPITAL LETTER ALPHA WITH CIRCUMFLEX est un grec, U+0410 U+0302 CYRILLIC CAPITAL LETTER A WITH CIRCUMFLEX est un cyrillique et U+0100 LATIN CAPITAL LETTER A WITH CIRCUMFLEX est un latin.)

Et vous voulez des exemples plus épineux de choix à faire entre analogueur et désunifier, il y en a au sein de l'alphabet latin : 'I' et 'i' sont-ils la majuscule et la minuscule de la même lettre ? Pour la plupart des langues du monde, oui, mais pas pour le turc, qui considère fort logiquement que 'I' est la majuscule de 'ı' (U+0131 LATIN SMALL LETTER DOTLESS I) et que 'i' est la majuscule de 'İ' (U+0130 LATIN CAPITAL LETTER I WITH DOT ABOVE). Bon, Unicode a tranché pour Unicode, « être la majuscule et la minuscule de la même lettre » est plus formellement défini (pas de façon normative, en tout cas), mais ça a quand même fait beaucoup de remous, et ça a quand même nécessité un sacré investissement de la part du Consortium. Pour ma part, j'aurais quand même préféré d'avoir trois lettres différentes (six caractères en tout) : un 'I' dont la minuscule est 'i', qui sert dans beaucoup de langues, un 'I' dont la minuscule est 'ı', qui sert dans quelques langues, et un 'I' dont la minuscule est 'İ', qui sert dans quelques autres langues. Et, bien sûr, Unicode a opportunément proposé U+0049 LATIN SMALL LETTER DOTLESS I, U+0130 LATIN CAPITAL LETTER I WITH DOT ABOVE, U+0131 LATIN SMALL LETTER I WITH DOT BELOW et U+0132 LATIN CAPITAL LETTER I WITH TOP BORDER qui manquait de place pour mettre trop de fantaisies et de variantes, mais qui permettrait par compatibilité avec un standard antérieur (typiquement, ISO-8859-9) qui manquait de place pour mettre trop de fantaisies et de variantes, de ne pas mettre trop de fantaisies et de variantes. Qui turcs qui turcs, ils auraient pas compris pourquoi il y avait deux 'I' différents et deux 'i' différents et la logique foin un pour le turc et un pour toutes les autres langues). Bref, ce n'est pas facile !

Unicode traite parfois des pieds pour accepter de désunifier des lettres (dans l'autre sens, une fois qu'elles sont désunifiées, c'est impossible de les réunifier). Par exemple, le grec U+03BC GREEK SMALL LETTER MU a été créé pour pouvoir supprimer le U+00B5 MICRO SIGN parce que « µ » et « μ », ce n'est pas pareil (LETTRE MU) pour des lettres qui servent dans l'écriture cyrillique du turkmène : ils datent d'Unicode 5.1 c'est-à-dire d'avril 2008. Avant ça, je me souviens d'avoir vu passer des requêtes pour U+0131 LATIN SMALL LETTER DOTLESS I pour pouvoir utiliser le 'ı' turc dans des logiciels informatiques, mais je ne me souviens plus comment savoir ? Est-ce que la question a même un sens ?

La source du plus grand nombre de maux de tête vient certainement de l'Alphabet Phonétique International (API). Celui-ci a été créé par des linguistes qui voulaient noter toutes les langues du monde de façon systématique et rationnelle, et ils ont rencontré un problème : il leur a fallu inventer plein de nouveaux caractères (c'était plus facile que de faire créer des symboles entièrement nouveaux). Il y a un nombre incroyable de bizarreries. Le caractère 'ɡ' (U+0261 LATIN SMALL LETTER SCRIPT G) est un caractère spécifiquement conçu pour l'API, mais il ne ressemble pas du tout au 'g' qui est en gros le 'g' (ou le hongrois) était évidemment à l'origine un 'T' retourné, et quelqu'un a trouvé malin de le réinterpréter, au moins pour ce qui est de l'Unicode. Mais « ɡ » ressemble quand même plus au « g » qu'un « g » ne ressemble à un « T » retourné, et quelqu'un a trouvé que ça valait le coup qu'il ait été représenté ; par contre, de façon décevante, le caractère 'ɡ' (U+0245 LATIN SMALL LETTER TURNED V), qui représente la semi-voyelle notée par la lettre 'w' dans le mot français « fouet », ne ressemble pas du tout à un 'v' retourné. Il ressemble à un 'y' inversé, qu'à peu près n'importe quoi d'autre, et avec une lettre allongée à droite. Bon, mais l'alphabet phonétique distingue, par exemple, 'e' et 'ɛ' (pour noter des voyelles plus ou moins antérieures, par exemple à différence de « pour » et « pas »), et, bien sûr, Unicode a tranché pour dire que ce sont deux lettres différentes, et, bien sûr, Unicode a opportunément trouvé à inventer un U+025A LATIN SMALL LETTER ALPHA, un « alpha latin », ce qui est quand même un concept original. De même, le symbole 'ɣ' qui peut la prononcer « gamma » en grec, mais qui représente le phonème qui ressemble à un « r » grasseyé en U+0263 LATIN SMALL LETTER GAMMA, ce « gamma latin », que le Club Contexte vous rappelle qu'il ne faut d'ailleurs pas confondre avec la voyelle 'ʏ' (U+0254 LATIN SMALL LETTER RAMS HORN), les « cornes de bélier » Alphabetique : « bélier » en latin, « ram » en anglais, « mouton » en français, « ovin » en espagnol, « ovino » en italien, « ovino » en portugais, « ovin » en roumain, « ovin » en polonais ; mais bizarrement, le symbole 'ʏ' qui note le son français de « bête » ou de « fait » par opposition à « fête » quand ce n'est pas un minéral qui parle est, dans Unicode, une « corne de bélier » ; je ne sais pas pourquoi ce n'est pas un « epsilon latin », ce qu'il semble bien être, et qui a pour une unité avec le 'ɛ' de l'alphabet grec (U+0395 GREEK SMALL LETTER EPSILON). Il y avait aussi une lata (U+1D6A LATIN SMALL LETTER GAMMA) mais il ne fait plus partie de l'API (en gros, il a été remplacé par 'ɣ', U+0263 LATIN SMALL LETTER GAMMA), sans doute pour des raisons de lisibilité ; en qui le 'ɣ' a pas ressemblé à un 'γ' (U+03B3 GREEK SMALL LETTER GAMMA) parce que, visiblement, la minuscule n'est pas le seul caractère qui ait pu rester à l'esprit du dessinateur de caractères qui a pris le mandat de créer le U+0263 LATIN SMALL LETTER GAMMA. Et, bien sûr, le U+1D6A LATIN SMALL LETTER GAMMA, qui a été décidé de ne pas amalgamer avec le 'ɣ' (pour U+03A3 GREEK CAPITAL LETTER SIGMA) parce que, visiblement, la minuscule n'est pas le seul caractère qui ait pu rester à l'esprit du dessinateur de caractères qui a pris le mandat de créer le U+03A3 GREEK CAPITAL LETTER SIGMA (ou le version MAJ de l'alphabet africain de référence publié par l'UNESCO, et le seb lui-même a servi en bambara mais a été remplacé par l'orthographe 'sh'. Cependant, il semble qu'au moins certaines orthographes du mbembe utilisent le alpha latin minuscule et majuscule, d'où le 'Ɑ' (U+0186 LATIN CAPITAL LETTER OPEN O) Quel chaos !

Pour le grec, il y a une autre chose intéressante à mentionner : certains caractères grecs ont plusieurs variantes assez différentes. Par exemple, le 'β' (phi) peut ressembler à un 'β' (beta) si on le tourne un peu et si on le penche un peu, sauf que beaucoup de polices grecques ne le représentent pas comme ça, mais verticalement ; du point de vue de la langue grecque, ces deux glyphes sont interchangeables, et il faudrait les amalgamer, mais les mathématiques, et plus précisément les gens qui font des équations en LaTeX, en font un usage différent. Unicode a donné une définition formelle à la notion d'équivalence pour la variante d'écriture du 'ϕ' qui ressemble à la majuscule, 'Φ' (U+03D5 GREEK PHI SYMBOL) ; en plus, comme le Club Contexte a encore frappé, sous TeX, la forme minuscule ressemble à un « φ » et la forme majuscule ressemble à un « Φ » (mais pas systématiquement, ça dépend des polices) et, bien sûr, Unicode a tranché pour d'ailleurs Unicode a commencé par prendre la convention de TeX avant de changer d'avis. Il y a de même deux formes d'écriture du thêta, la forme « θ » (U+03B8 GREEK SMALL LETTER THETA) et la forme « ϑ » (U+03D1 GREEK THETA SYMBOL), qui ressemble un peu à un 'ν'. Il y a aussi deux formes d'écriture du kappa, la forme « κ » (U+03BA GREEK SMALL LETTER KAPPA) et la forme « ϰ » (U+03F0 GREEK KAPPA SYMBOL), qui ressemble un peu à un 'p'. Il y a même deux formes d'écriture du rho, la forme « ρ » (U+03C1 GREEK SMALL LETTER RHO) et la forme « ϱ » (U+03F1 GREEK RHO SYMBOL), qui ressemble un peu à un 'o'. Il y a aussi deux formes d'écriture du sigma, la forme « σ » (U+03C3 GREEK SMALL LETTER SIGMA) et la forme « ς » (U+03C2 GREEK SMALL LETTER FINAL SIGMA) et la forme finale, qui ressemble un peu à un 'ν' surmonté d'un tilde. Parfois on distingue aussi deux formes d'écriture du kappa, la forme normale 'κ' (kappa en TeX) et la variante 'ϰ' (varkappa en TeX) et la variante 'ϰ' ressemble à un « κ » qui a pris un accent circonflexe, mais pas toujours, ça dépend des polices. Il y a aussi deux formes d'écriture du thêta, la forme normale 'θ' (vtheta en TeX) et la variante 'ϑ' (U+03D1 GREEK THETA SYMBOL, VARTHETA en TeX) qui ressemble un peu à un 'ν'. Il y a aussi deux formes d'écriture du rho, la forme normale 'ρ' (U+03C1 GREEK SMALL LETTER RHO SYMBOL, VARTHETA en TeX) qui ressemble un peu à un 'o'. Il y a aussi deux formes d'écriture du sigma, la forme normale 'σ' (U+03C3 GREEK SMALL LETTER SIGMA) qui ressemble un peu à un 'o'. Il y a aussi deux formes d'écriture du phi, la forme « φ » (U+03C6 GREEK SMALL LETTER PHI) et la forme « ϕ » (U+03D5 GREEK PHI SYMBOL) qui ressemble un peu à un 'o'. Il y a aussi deux formes d'écriture du pi, la forme « π » (U+03C0 GREEK SMALL LETTER PI) et la forme « ϖ » (U+03D6 GREEK PI SYMBOL) qui ressemble un peu à un 'o'. Il y a aussi deux formes d'écriture du upsilon, la forme « υ » (U+03C5 GREEK SMALL LETTER UPSILON) et la forme « ϒ » (U+03D2 GREEK UPSILON WITH HOOK SYMBOL) qui ressemble un peu à un 'o'. Il y a encore un sacré problème de lisibilité, comme on le voit, sauf que beaucoup de polices ne le représentent pas comme ça, mais verticalement ; du point de vue de la langue grecque, ces deux glyphes sont interchangeables, et il faudrait les amalgamer, mais les mathématiques, et plus précisément les gens qui font des équations en LaTeX, en font un usage différent. Unicode a donné une définition formelle à la notion d'équivalence pour la variante d'écriture du 'ϕ' qui ressemble à la majuscule, 'φ' (U+03D5 GREEK PHI SYMBOL) ; en plus, comme le Club Contexte a encore frappé, sous TeX, la forme minuscule ressemble à un « φ » et la forme majuscule ressemble à un « Φ » (mais pas systématiquement, ça dépend des polices) et, bien sûr, Unicode a tranché pour d'ailleurs Unicode a commencé par prendre la convention de TeX avant de changer d'avis. Il y a de même deux formes d'écriture du thêta, la forme « θ » (U+03B8 GREEK SMALL LETTER THETA) et la forme « ϑ » (U+03D1 GREEK THETA SYMBOL), qui ressemble un peu à un 'ν'. Il y a aussi deux formes d'écriture du kappa, la forme normale 'κ' (kappa en TeX) et la variante 'ϰ' (varkappa en TeX) et la variante 'ϰ' ressemble à un « κ » qui a pris un accent circonflexe, mais pas toujours, ça dépend des polices. Il y a aussi deux formes d'écriture du rho, la forme normale 'ρ' (vrho en TeX) et la variante 'ϱ' (U+03F1 GREEK RHO SYMBOL, VARTHETA en TeX) qui ressemble un peu à un 'o'. Il y a aussi deux formes d'écriture du sigma, la forme « σ » (U+03C3 GREEK SMALL LETTER SIGMA) et la forme « ς » (U+03C2 GREEK SMALL LETTER FINAL SIGMA) et la forme finale, qui ressemble un peu à un 'ν' surmonté d'un tilde. Parfois on distingue aussi deux formes d'écriture du kappa, la forme normale 'κ' (kappa en TeX) et la variante 'ϰ' (varkappa en TeX) et la variante 'ϰ' ressemble à un « κ » qui a pris un accent circonflexe, mais pas toujours, ça dépend des polices. Il y a aussi deux formes d'écriture du thêta, la forme normale 'θ' (vtheta en TeX) et la variante 'ϑ' (U+03D1 GREEK THETA SYMBOL, VARTHETA en TeX) qui ressemble un peu à un 'ν'. Il y a aussi deux formes d'écriture du rho, la forme normale 'ρ' (U+03C1 GREEK SMALL LETTER RHO SYMBOL, VARTHETA en TeX) qui ressemble un peu à un 'o'. Il y a aussi deux formes d'écriture du sigma, la forme « σ » (U+03C3 GREEK SMALL LETTER SIGMA) et la forme normale