

Werkgroep 13: Nederlandstalige T_EX The Right of ij to be a Ligature

David van Leeuwen

david@rulkol.leidenuniv.nl

Ever since calligraphy (beautiful writing) is used as a way to express ones thoughts writers (and later printers) were concerned about how certain letters clashed. In fact, people thought of ways to camouflage these collisions, and they had at least two methods for doing so: One is known as *kerning*, the slightly shifting the letters with respect to one another, the other is the use of *ligatures*, the replacement of the characters by one single but bigger symbol. What combinations of letters were ugly enough to qualify for a ligature has probably always been influenced by fashion. Everyone who has an eye for typesetting must have noticed how in the earlier centuries manuscript had this weird ‘ct’ ligature. Although our current Roman style of printing looks in some ways much like the older handwritings, a ct-ligature seems completely superfluous. On the other hand with computer typesetting systems like T_EX now readily available, people seem to have a tendency to go overboard in ‘inventing’ new ligatures.

The computer roman family of fonts, designed by Donald E. Knuth after the fonts used in the first editions of his books on Computer Algorithms, is strongly provided with f-ligatures. The fonts appear a little ‘old’, because most books and newspapers are printed nowadays in the popular Times-Roman set of characters. These seem to need the f-ligatures much less, although there might be a correspondence between the decline of the number of skilled lead-typesetters and the use of ligatures. Now the fonts of characters of the computer modern family are based on English/American use, where these f-combinations appear strikingly often. But in other languages different letter combinations will occur frequently enough to apply for a ligature.

My knowledge of languages is very poor, the only one I can say meaningful things about is Dutch. Dutch is language with lots of ‘open’ vowels. So many, that we come short of double combinations *ee*, *aa*, *uu*, and *oo* and need to use *ei*, *ij*, *eu*, *ui* for typically Dutch sounds. (Of course, we do have the ‘normal’ *a*, *e*, *i*, *o* and *u*, and we use *ie* for a long *i*, but the vowel *y* is not used in normal Dutch words.) An odd thing to a foreigner might be that we use the consonant *j* in a vowel combination. To make things even stranger, the pronunciation of the *ij* is identical to that of *ei*, and the letter *y* in the

alphabet is pronounced the same way! Sometimes, the *ij* is treated as a single letter. When one capitalizes a word starting with an *ij*, *both* will be written in upper case, as in ‘IJmuiden’ (impossible to pronounce by non Dutchmen, by the way). In some cases the *ij* leads to inconsistencies: some alphabetization programs use the . . . xyz-alphabet, others the . . . xyijz-alphabet.

One of the worst things of running an English T_EX version with Dutch text, is that words with an *ij* are very likely to break *between* the two characters, which is the worst choice. So having a pattern `i4j` in the `\pattern` sequence is the change that has the highest yield in improving the English version for Dutch. Well, nowadays there are commercial and non-commercial hyphenation patters for Dutch available, so that problem has been solved. Somewhat more delicate is the fact that we would prefer to position the *i* and the *j* just somewhat closer together than cm-fonts would naturally do. So a kerning of one `k#` between the *i* and *j* would considerably improve the readability of Dutch texts.

On this subject, whether the *ij* deserves a kerning or even a ligature and more importantly how this should be implemented on T_EX systems, there has been quite some discussion at the Dutch T_EX bulletin board. Generally people agree that the *ij* has a special status, but how we should deal with it is not a question that is solved. Roughly there are two camps. The first thinks that everything should be done using the standard set of cm-fonts and a huge amount of fairly difficult T_EX-source spaghetti, so that everybody can use its standard T_EX-setup. The second camp has the opinion that special fonts should be used that respect the *ij* as a ligature, so that typists need no special training for identifying *ij*’s. The difference may be clear: the first group makes it easier for system operators, the second group for the users. Moreover, the first solution is much more ‘international’ than the second one, while that will run faster and have better readable source texts.

The best thing to do would probably be using virtual fonts. My own knowledge about it is almost nothing (I don’t have any written paper on virtual fonts nor do I have any access to one), but I understand that they come in two parts: a .`tfm` file that instructs T_EX, and an other file that tells a dvi driver how to compose charac-

ters from which .pk files. In any case, this means that every font that is specific to Dutch needs font space in T_EX. For plain-T_EX users this is not a large problem, nor will singly Dutch LaT_EX documents be hard to process. But multilingual LaT_EX users should have to enlarge T_EX's font space, since LaT_EX uses almost all of the standard capacity of T_EX's font memory, and if we were to double the amount of fonts this would become a serious thing. But suppose the system wizards that find room to store two (or more) sets of hyphenation patters will also know do deal with the font storage. Then a language-change macro would have to do two things: change the value of `\language` and change the current font. The latter can actually be done if the font naming conventions of the foreign country has the same structure as the cm-family.

Jörg Knappen noticed that a family of fonts exists that has this property. These are called the dc-family, presumably meaning 'Deutsche Computer modern'. The fonts were actually accepted by the international T_EX community at the conference in Cork. In order to see if these would be suitable for the ij (and correlated) problem(s) I ftp-ed one of them, dcr10 (cf. cmr10), in .tfm and .pk form. It consists of 255 characters, among them ones that I have never seen before, but are claimed to be used somewhere in Europe (which I have no doubt about). Most normal Latin letters appear on the same place as in cmr10, but ligatures, accents en special symbols are located elsewhere. There are no Greek capitals in it. A fairly large amount of the upper 128 characters are letters that could be formed by plain T_EX with normal accents, in this sense the font can be used with 8 bit input where characters are mapped on identical places, so that the French could use this font directly. The ij and IJ appear as one-symbol letters in positions 188 and 156, but these are *not* declared in the .tfm file as being formed by a ligature ij. So in this sense the dcr10 font merely acts as a bitmap source for the drivers, unless Dutch keyboards are supplied with an ij-key. In fact, a dcr10 user could never do without the cmr10 font, because some characters are missing. Also the ligatures !` and ?` are not recognized, although the characters appear in the font.

It is not clear to me how the dcr10 font should be used. It does have the f-ligatures and ligatures for the french double quotes that can be entered as << and >>, but when one types a single < one doesn't get the single french quote (which the font contains) but a smaller-than sign <, which makes no sense outside mathmode. Also it supports the double quote ligatures “ and ”, but

does not recognize the double-comma opening quote „ (which the font does contain), which the German and Dutch would like to use. What is the use of an ij-symbol if it is not recognized as an ij-ligature?

Re-reading the discussion on the ij-ligature I see that Gerben Wierda and Johannes Braams have reached a similar conclusion much earlier than me, and Johannes (and also Yannis) noticed that non-dutch people would probably not like an ij-ligature. I am not so sure about that. In Dutch, we do not need the ffl-ligature as much as the English do, but do we bother to have one? I don't think the ij combination occurs much in foreign languages, or that it would matter much that they are kerned a little closer. On Johannes' remark that virtual fonts would be a solution Nico Poppelier argues that not all dvi-drivers supports virtual fonts. In some aspect this is just the same as with 256-character fonts. Some older dvi-drivers didn't support 8 bit fonts although even T_EX 2, *x* did, but since the cm-fonts had no more than 128 characters dvi-driver-makers did not bother to support 8 bit fonts. If dvi-drivers do not yet support virtual fonts we must keep asking the programmers to update their drivers.

I think that we must seek for a solution where T_EX itself should take care of finding ligatures like ij, double-comma, french quote etcetera. This can be done by either using virtual fonts and the standard cm family, or by a complete new set of fonts like dcr but with language specific ligatures. (One could think of making the 'i' an active character, checking for the next to be a 'j', but let us call this an academic solution since it will not make life very much easier.) Typing "y for a ij ligature could be acceptable in English texts where the ligature hardly ever will be needed, but is completely unacceptable for Dutch texts. It would be comparable to asking the English to write "ff whenever the ff ligature should be used. Of the two alternatives (virtual fonts and a different family) the former seems more elegant, using less .pk fonts. On the other hand the latter alternative can be more widely be implemented since not all dvi-drivers accept virtual fonts yet.

For either solution people should get together and decide which languages deserve what ligatures. The dc-family may be a good scheme to start from. For a typical national problem like this international portability is less important than ease to read and write the source texts.

I hope that this can be a positive contribution to the question of language dependent ligatures.