Typesetting Catalan Texts with TeX

Gabriel Valiente Feruglio
Universitat de les Illes Balears
Departament de Ciències Matemàtiques i Informàtica
E-07071 Palma de Mallorca (Spain)
Internet: dmigva0@ps.uib.es

Robert Fuster
Universitat Politècnica de València
Departament de Matemàtica Aplicada
Camí de Vera, 14. E-46071 València (Spain)
Internet: mat5rfc@cci.upv.es

Abstract

As with other non-American English languages, typesetting Catalan texts imposes some special requirements on TeX. These include a particular set of hyphenation patterns and support for a special ligature: unlike other Romanic languages, Catalan incorporates the middle point in the l·l digraph. Hyphenation rules for Catalan are reviewed in this paper, after a short introduction to hyphenation by TeX. A minimal set of hyphenation patterns covering all Catalan accents and diacritics is also presented. A discussion about the l·l ligature concludes the paper. This work represents a first step towards the Catalan TLP (TeX Language Package), under development within the TWGMLC (Technical Working Group on Multiple Language Coordination), where the first author is chairing the Catalan linguistic subgroup.

Resum

Així com en altres llengües, la composició de textos escrits en català demana requeriments especials al TeX. Aquests inclouen un conjunt particular de patrons de guionat, així com suport per a un lligam especial ja que, a diferència d'altres llengües romàniques, el català incorpora el punt volat al dígraf l·l. En aquest paper es fa una introducció al guionat amb TeX i es revisen les regles de guionat per al català. Tanmateix, es presenta un conjunt mínim de patrons de guionat que cobreix tots els accents i marques diacrítiques del català. El paper acaba amb una discussió sobre el lligam l·l. Aquest treball representa un primer pas cap al TLP (Paquet de Llengua TeX) català que s'està desenvolupant dins el TWGMLC (Grup Tècnic de Treball sobre Coordinació de Múltiples Llengües), on el primer autor presideix el subgrup lingüístic català.

Hyphenation by TeX

Background on hyphenation by TeX is first presented, following the ninth edition of *The TeXbook* (Knuth, 1990) and the exposition in Haralambous (*TUGboat*, 1990). The actual hyphenation algorithm used by TeX is due to Liang (1983).

When TeX creates a format file like `plain.fmt`, `lplain.fmt` or `amsplain.fmt`, it reads information from a file called `hyphen.tex` (or `**hyphen.tex`, where `**` is a two-letter language code[1] (see Haralambous, *TeX and TUG NEWS*, 1992)) that contains the *hyphenation patterns* for a specific language. Using TeX3+, a format file can include more than one (up to 256) sets of patterns and, so, INITEX produces multilingual versions of TeX. In this case, language-switching mechanisms like those of the Babel system by Johannes Braams allow TeX to typeset every language according to its own rules. A syntax

[1] In the Catalan case, the name of this file will be `cahyphen.tex`.

for language-switching commands has not yet been standarized, but it is expected to be something like

```
\language{catalan}{...Catalan text...}
```

for short inserts and

```
\begin{language}{catalan}
...Catalan text...
\end{language}
```

for longer inserts.

Hyphenation patterns are clusters consisting of letters separated by digits, like x1y2z (more exactly, a pattern has the form

number/letter/number/letter/.../number

like 0x1y2z0, but the number 0 can be suppressed), meaning that:

- If the set of patterns is empty, no hyphenation takes place.
- If there is a pattern x1y, then hyphenation "x-y" will be possible in every occurrence of the cluster "xy". If the pattern is x1yzw, then the sequence of letters "xy" will be hyphenated only when followed by "zw".
- If there is a pattern x1y and a pattern x2yabc then the sequence "xy" will be hyphenated, as long as it is not followed by "abc". The digit 2 indicates therefore an exception to the rule "separate x and y" expressed by the digit 1.
- The same holds for greater numbers. Patterns with number 3 will be exceptions to patterns with number 2, and so on: odd numbers allow and even numbers disallow hyphenation, and the maximum decides.
- A dot in front of (or behind) a pattern, such as .x1y or xy2z. specifies that the pattern is valid only at the beginning (or at the end) of a word.

In this context, a *letter* is a character of category 11 or 12 whose \lccode is nonzero. Because, for almost all Latin-alphabet languages, some diacriticized characters are letters for which we need a mechanism, including these special characters as letters. Using TEX3+, which allows 8-bit input, this problem disappears.

Despite the existence of some fundamental rules, hyphenation of a particular language can be very complicated. There are two methods to handle this complexity: hidden mechanisms of hyphenation can be investigated and patterns made to correspond to the analytical steps of manual hyphenation, or patterns can be induced from a sufficiently representative set of already hyphenated words, using inductive inference tools tailored to this particular problem such as PATGEN.

The choice of method depends on the nature of the language and on the size of the available set of hyphenated words. Although in theory such a pattern generator would produce an *exhaustive* set of patterns from a file containing *all* words of a particular language in hyphenated form, it is more probable to have partial sets of hyphenated words, and the pattern generator will only produce more or less accurate approximations.

The authors have chosen the first method for Catalan. Besides hyphenation patterns, the effort resulted in more systematic and exhaustive rules for Catalan hyphenation than those found in grammar textbooks.

Catalan Hyphenation Rules and Patterns

Modern Catalan normative grammar was established by Pompeu Fabra and ratified by the *Institut d'Estudis Catalans* (Catalan Studies Institute) in 1917. Orthography (and in particular syllabification and hyphenation rules) can be found in many texts: Bruguera (1990), Fabra (1927), Mira (1974), Pitarch (1983), Salvador (1974), and many others. The official normative dictionary is *Diccionari general de la llengua catalana* (Fabra, 1974) and *Diccionari ortogràfic i de pronúncia* (Bruguera, 1990) is a hyphenation dictionary. A very interesting study of some difficulties in the Catalan orthography can be found in Solà (1990). Some of our observations on Spanish, Italian or French hyphenation were suggested by the preceding references, but also by Lázaro (1973) and Beccari (1992).

Catalan, like other Romanic languages, bases its hyphenation rules on the syllabic structure of words. This structure, as far as Catalan is concerned, is closely related to Spanish, Portuguese or Italian. But there exist a number of differences: for example, the Catalan word *València* has four syllables and the Spanish *Valencia* has only three.

Of course, the Catalan alphabet follows the standard Latin alphabet. The letters k and w never appear (except in foreign words), the letter y is only used to form the digraph ny and letter q only appears followed by letter u.

In general a Catalan word has as many syllables as it has vowels, either separated by consonants or contiguous but not forming diphthongs. In fact, a Catalan word has *exactly* as many syllables as it has vowels, but in some special cases, letters i, u are not vowels (Catalan vowels are a, e, i, o and u). Word stress, however, determines how a Catalan word breaks up into syllables and, in some polysyllabic words, is expressed by an accent on the vowel of the stressed syllable. In this way, accents in Catalan are used in nearly the same way as in Spanish. Also as in Spanish, the accents perform another diacritic function (to distinguish some homophones, as *dona* = woman and *dóna* = he/she gives). However, the kind of accent, grave (`) or acute (´), marks the difference between open and closed vowels, as in French or Italian: so, all accented vowels are à, è, é, í, ò, ó and ú. The diaeresis (¨), over i or u, splits a diphthong or causes the letter u to be pronounced when g or q precede it.

The cedilla under the letter c (ç) and the apostrophe (') are usual in Catalan, with the same use as in French. Virtually all European languages have their own particularities: Catalan has the special construction l·l.

Syllabification. Basic rules for word division into syllables include the following (v, $v_n, n \geq 1$ will be vowels and c, $c_n, n \geq 1$ consonants).

1. A single consonant between two vowels forms a syllable with the vowel that follows it: v_1-cv_2. Actually it suffices to consider patterns of the form -cv, because if another consonant (instead of the first vowel) precedes c the pattern would also be c_1-c_2v (see rules 2, 3 and 5 below). The necessary patterns will be:

   ```
   1ba 1be 1bi 1bo 1bu
   1ca 1ce 1ci 1co 1cu
   1ça         1ço 1çu
   1da 1de 1di 1do 1du
   1fa 1fe 1fi 1fo 1fu
   1ga 1ge 1gi 1go 1gu
   1ha 1he 1hi 1ho 1hu
   1ja 1je 1ji 1jo 1ju
   1la 1le 1li 1lo 1lu
   1ma 1me 1mi 1mo 1mu
   1na 1ne 1ni 1no 1nu
   1pa 1pe 1pi 1po 1pu
   1ra 1re 1ri 1ro 1ru
   1sa 1se 1si 1so 1su
   1ta 1te 1ti 1to 1tu
   1va 1ve 1vi 1vo 1vu
   1xa 1xe 1xi 1xo 1xu
   1za 1ze 1zi 1zo 1zu
   ```

   ```
   1bà 1bè 1bé 1bí 1bò 1bó 1bú
   1cà 1cè 1cé 1cí 1cò 1có 1cú
   1çà             1çò 1çó 1çú
   1dà 1dè 1dé 1dí 1dò 1dó 1dú
   1fà 1fè 1fé 1fí 1fò 1fó 1fú
   1gà 1gè 1gé 1gí 1gò 1gó 1gú
   1hà 1hè 1hé 1hí 1hò 1hó 1hú
   1jà 1jè 1jé 1jí 1jò 1jó 1jú
   1là 1lè 1lé 1lí 1lò 1ló 1lú
   1mà 1mè 1mé 1mí 1mò 1mó 1mú
   1nà 1nè 1né 1ní 1nò 1nó 1nú
   1pà 1pè 1pé 1pí 1pò 1pó 1pú
   1rà 1rè 1ré 1rí 1rò 1ró 1rú
   1sà 1sè 1sé 1sí 1sò 1só 1sú
   1tà 1tè 1té 1tí 1tò 1tó 1tú
   1và 1vè 1vé 1ví 1vò 1vó 1vú
   1xà 1xè 1xé 1xí 1xò 1xó 1xú
   1zà 1zè 1zé 1zí 1zò 1zó 1zú
   ```

2. Of two consonants standing between two vowels, the first forms a syllable with the preceding vowel and the second forms a syllable with the vowel that follows it: v_1c_1-c_2v_2. Because the preceding patterns allow this break, we do not need special patterns for this rule. But one exception to this rule is that the liquid consonants, l and r, when preceded by certain consonants, form a syllable with this consonant and the vowel that follows. Another exception is that there are some special combinations, called *digraphs*, that represent only one phoneme or a geminated one. The complete list is: ig, ix, ll, l·l, ny, rr, ss, tg, tj, tl, tll, tx, tz. The digraph ig only occurs at the end of a word (and in plural form, igs).

   The two following rules exactly define these exceptions.

3. The combinations c-l and c-r that cannot be hyphenated are bl, cl, fl, gl, pl, br, cr, dr, fr, gr and pr. The necessary patterns will be:

   ```
   1b2l 1c2l      1f2l 1g2l 1p2l
   1b2r 1c2r 1d2r 1f2r 1g2r 1p2r 1t2r
   ```

   The combination vr is another one that cannot be hyphenated, but it appears only in a few toponymies.

4. The digraphs ll and ny are not split (following an etymological criterium). The pattern

   ```
   l2l
   ```

   voids the effect of the first rule. No analogous pattern is necessary for ny. In fact, ny and ll correspond to single consonant sounds and therefore rule 1 above applies to them as well. The necessary patterns will be:

   ```
   1lla 1lle 1lli 1llo 1llu
   1llà 1llè 1llé 1llí 1llò 1lló 1llú
   1nya 1nye 1nyi 1nyo 1nyu
   1nyà 1nyè 1nyé 1nyí 1nyò 1nyó 1nyú
   ```

   All other digraphs can be split. The l·l ligature is also a digraph and can be divided, replacing the middle dot with a hyphen. Hyphenation of the l·l ligature is discussed in the next section.

5. Of three or more consecutive consonants followed by a vowel, the last consonant forms a syllable with that vowel: c_1c_2-c_3v, et cetera, unless the last two consonants belong to those in the two rules above. No additional patterns are necessary for this rule.

6. Compound words with one of the following prefixes

```
an, con, des, en, ex, in, sub, trans
```

are divided according to components and therefore often constitute exceptions to the previous rules. These differ from prefix to prefix and present an evident problem: it is impossible, unless you make an exhaustive classification by scanning a dictionary, to determine if a certain combination is or is not a prefix.[2] For example, you must hyphenate *in-a-pe-tèn-ci-a* (inappetence) but *e-nò-leg* (an expert in wine) instead of *en-ò-leg*. For instance, using Bruguera (1990), we find the following patterns for `.ex`:

```
.e2x1a .e2x1à
.e3x2ag .e3x2am .e3x2àm
.e2x1on .e2x1or .e2x1osm
.e3x2orc .e3x2ord
.e2x1ulc
```

In all words starting with `trans` — except in `transit` and its derivatives — `trans` is a prefix. Then, the corresponding patterns will be

```
.tran2s1 .tran3s2i
```

Because these prefixes are very frequent in practice and — specially in technical languages — frequently used to create new words, this is a dangerous solution. Another possible solution — more conservative, but completely secure — consists of inhibiting the splitting of such a group whenever it is present at the beginning of a word (except in the case of `trans`, because it is a very long prefix):

```
.a2n .co2n .de2s .e2n .e2x
.i2n .su2b .tran2s .tran3s2i
```

To choose between these two options is still an open question.

7. Personal pronouns *nosaltres* (we) and *vosaltres* (you) are etymologically composed words. They must, therefore, be hyphenated *nos-al-tres, vos-al-tres.*[3]. The necessary patterns are:

```
.no2s1al .vo2s1a
```

Exceptions to the syllabification rules above are certain groups of vowels where `i` or `u` are not really vowels. The next sections explain these exceptions.

Descending Diphthongs. When a vowel is followed by an unstressed `i` or *u*, this second letter is a semivowel and forms a syllable with the preceding vowel. These diphthongs are `ai`, `ei`, `oi`, `ui`, `au`, `eu`, `iu`, `ou` and `uu`.

All other combinations of two vowels are divided. The necessary patterns will be:

```
a1a a1à a1e a1è a1é     a1í a1o a1ò a1ó a1ú
e1a e1à e1e e1è e1é     e1í e1o e1ò e1ó e1ú
i1a i1à i1e i1è i1é i1i i1í i1o i1ò i1ó i1ú
o1a o1à o1e o1è o1é     o1í o1o o1ò o1ó o1ú
u1a u1à u1e u1è u1é     u1í u1o u1ò u1ó u1ú

à1a è1a é1a í1a ò1a ó1a ú1a
à1e è1e é1e í1e ò1e ó1e ú1e
            í1i
à1o è1o é1o í1o ò1o ó1o ú1o
```

Ascending Diphthongs and silent `u`. When the letters `g` or `q` come before the vowel `u` and another vowel, then either the `u` is not pronounced or the two vowels compose an ascending diphthong (and the `u` is a semiconsonant).[4] In both cases, the three letters belong to the same syllable and the combination cannot be hyphenated. Then, we need to make void some of the preceding patterns. For the `g` the patterns will be:

```
gu2a gu2e gu2o
gu2à gu2è gu2é gu2í gu2ò gu2ó
```

The letter `q` is used only in this context and always starting a syllable. Then the only necessary pattern is

```
1qu2
```

Triphthongs. Yet another exception to the syllabification rules above is a group of three vowels (actually, a semiconsonant/vowel/semivowel combination) that together constitute a single syllable. The only triphthong in Catalan is `uai` after `g` or

[2] PATGEN allows an *adjustment of an existing set of patterns*; it will read both a set of already existing patterns and a collection of hyphenated words, and will create a new set of patterns. This method can be used as a combination of the analytical and the raw PATGEN methods. For example, one could extract *all* words starting with one of the prefixes `an`, `con`, `des`, `en`, `ex`, `in`, `sub` and `trans`, from a dictionary, and run PATGEN on these and on the existing patterns. The test function of PATGEN will immediately evaluate if the new set of patterns is more powerful.

[3] In 1959 the *Academia Española de la Lengua* (Spanish Language Academy) revoked a similar prescriptive rule. So, in Spanish you can hyphenate *nos-otros* or *no-sotros* This applies also to the Spanish hyphenation of the prefix `des`.

[4] These are the only cases of ascending diphthongs in Catalan. It differs from Spanish and Italian: in these two languages all combinations of `i` or `u` with a vowel are diphthongs.

q, but no special patterns are necessary because the preceding patterns gu2a 1qu2 apply and the combination ai is not hyphenated.

Letter i or u as consonant. Unstressed i before a, e or o, however, becomes a consonant when situated at the beginning of a word (even when preceded by h), except in ió and its derivatives. The necessary patterns will be:

```
.i2a .i2à .i2e .i2è .i2é .i2o .i2ò
.hi2a .hi2e .hi2o
.hi2à .hi2è .hi2é .hi2ò .hi2ó
.i3on
```

Unstressed i or u standing between vowels are consonants and form a syllable with the vowel that follows it. The necessary patterns will be:

```
a1i2a a1i2e a1i2i a1i2o a1iu
e1i2a e1i2e e1i2i e1i2o e1iu
ii2a  ii2e        ii2o
o1i2a o1i2e o1i2i o1i2o o1iu
u1i2a u1i2e u1i2i u1i2o u1iu

a1u2a a1u2e a1ui a1u2o a1uu
e1u2a e1u2e e1ui e1u2o e1uu
i1u2a i1u2e i1ui i1u2o i1uu
o1u2a o1u2e o1ui o1u2o o1uu
u1u2a u1u2e u1ui u1u2o

à1i2a à1i2e à1i2i à1i2o à1iu
à1u2a à1u2e à1ui  à1u2o à1uu
è1i2a è1i2e è1i2i è1i2o è1iu
è1u2a è1u2e è1ui  è1u2o è1uu
ò1i2a ò1i2e ò1i2i ò1i2o ò1iu
ò1u2a ò1u2e ò1ui  ò1u2o ò1uu
é1i2a é1i2e é1i2i é1i2o é1iu
é1u2a é1u2e é1ui  é1u2o é1uu
í1i2a í1i2e       í1i2o
í1u2a í1u2e í1ui  í1u2o í1uu
ó1i2a ó1i2e ó1i2i ó1i2o ó1iu
ó1u2a ó1u2e ó1ui  ó1u2o ó1uu
ú1i2a ú1i2e ú1i2i ú1i2o ú1iu
ú1u2a ú1u2e ú1ui  ú1u2o

a1i2à a1i2è a1i2ò a1u2à a1u2è a1u2ò
e1i2à e1i2è e1i2ò e1u2à e1u2è e1u2ò
ii2à  ii2è  ii2ò  i1u2à i1u2è i1u2ò
o1i2à o1i2è o1i2ò o1u2à o1u2è o1u2ò
u1i2à u1i2è u1i2ò u1u2à u1u2è u1u2ò
a1i2é a1i2í a1i2ó a1i2ú
a1u2é a1u2í a1u2ó a1u2ú
e1i2é e1i2í e1i2ó e1i2ú
e1u2é e1u2í e1u2ó e1u2ú
o1i2é o1i2í o1i2ó o1i2ú
o1u2é o1u2í o1u2ó o1u2ú
```

Diaereses. In Catalan the diaeresis is used in two different contexts: first, if an i or u — following a vowel or between two vowels — is a real vowel (and in consequence does not belong to the same syllable). But, second, in the combinations qüe, güe, qüi, güi it indicates that the u is pronounced (forming a diphthong with the following vowel).

The corresponding patterns are:

```
a1ï e1ï i1ï o1ï u1ï a1ü e1ü i1ü o1ü u1ü
ï1a ï1e ï1i ï1o ï1u ü1a ü1e ü1i ü1o ü1u
1gü2 1qü2
ü3ï
```

The last pattern applies to a very special case: in *argüïen* and other related words appear two consecutive diaereses (Valor (1983), p. 20).

Breaks. Catalan words may be broken into syllables containing just one letter. Actually, only vowels can form a syllable on their own, but some learned words or words of foreign origin, like *psicòleg* or *show* start with a pair of consonants: the possible combinations are gn, mn, pn, ps, sc, sh, sl, sm, sn, sp, st, ts; the only occurrence of a digraph beginning a word is in the word *txec* and its derivatives (as *txecoslovac*). Then, the following patterns are necessary in order to make void the effect of the first rule and to prevent separating single consonants at the beginning of words:

```
.g2 .m2 .p2 .s2 .t2
```

Also combinations like cl and br can start a word, but then rule 3 applies and no special patterns are required.

Finally, to prevent hyphenation of an apostrophe, we only need the pattern

```
'2h
```

Now we have a complete set of hyphenation patterns, even if the parameters \lefthyphenmin and \righthyphenmin are set to 1. Regarding this question, we suggest the values

```
\lefthyphenmin=1 \righthyphenmin=3
```

because long ending syllables are frequent in Catalan words and then, with the default values, very frequent words like *aquests* (the plural masculine demonstrative) must not be hyphenated. So, the macros involved in the Catalan language LaTeX environment should include:

```
\language=2 % or the appropriate value
\lccode`\'=`\'
\nonfrenchspacing
\lefthyphenmin=1
\righthyphenmin=3
```

The l·l Ligature

All Catalan characters belong to the ISO 8859-1 coding scheme, known as ISO Latin-1, with only one exception. Double *ll* also exhibits a geminated form, *l·l*. Let us take a look at its etymology.

While some Romanic languages preserve the phonetic distinction between |λ| and |*ll*|, in particular French, Italian and Catalan, it is only in

Catalan where this phonetic distinction finds a corresponding orthographic distinction. For instance, Latin INTELLIGENTIA derives into French *intelligence* and Italian *intelligenza*, while Latin SELLA derives into French *selle* and Italian *sella*. Then these languages use the same orthography for two different phonemes.

Modern Catalan uses **ll** for phoneme |λ| and **l·l** for phoneme |*ll*|. Then Latin INTELLIGENTIA derives into Catalan *intel·ligència* and Latin SELLA derives into Catalan *sella*.

This correspondence between phonetics and orthography is a debt to the normalization process to which Catalan has been subject to, where Pompeu Fabra (1984) has played a fundamental role. Early grammar texts, however, use **ł** for **ll** and **ll** for **l·l** (Fabra, 1912). See Fabra (1983, 1984) and Segarra (*Història de l'ortografia catalana*, 1985) for more details on these orthographic distinctions[6].

The l·l ligature and DC fonts. This section is a revised excerpt from discussions held between Gabriel Valiente Feruglio and Yannis Haralambous during 1992, while contributing to Haralambous' efforts in incorporating national requirements from different countries into the design of DC fonts.

Q. Is it necessary or facultative (like the fi ligature)?

A. It is mandatory.

Q. Is there also an "ll" without dot?

A. Yes, there is. The "ll" without dot corresponds to a palatal sound, while the "l·l" with middle dot corresponds to the "gemination" or duplication of the "l" sound.

Q. What is its uppercase counterpart?

A. The l·l ligature cannot appear at the beginning of a word, only joining two syllables. Therefore, the only way in which the l·l must be shown in uppercase is when the whole word is in uppercase, and in such a case both L's are capitalized, as the word INTEL·LIGENCIA shows.

Q. How do you create it using TEX and/or other word processors?

A. Detailed definitions for TEX are given and discussed in the next section. Many WYSIWYG word processors actually support the l·l ligature, that is obtained by joining two characters: an *l* with middle dot (l·) and another *l*. When hyphenation takes place, the *l·* gets replaced by a normal *l*.

Q. Can it be hyphenated?

A. The function of l·l can be seen as that of joining two syllables, one ending in "l" and the other beginning with "l". Therefore, it can be hyphenated, and the right hyphenation is "l-" and "l". For instance, the word *intel·ligència* would be hyphenated as: `in-tel-li-gèn-ci-a`. It is therefore a ligature instead of a single character. This justifies the lack of an l·l character in DC fonts, although a middle dot other than TEX's centered dot `$\cdot$` could also be useful, besides Catalan, for other languages as well.

Q. What is its alphabetical order?

A. It does not appear in the alphabetical order, because it has no extra sound, just the mere duplication of the "l" sound. [Comment of R. Fuster: Colomer (1989), a Catalan-English/English-Catalan dictionary, and Bruguera (1990) arrange cella before cel·la. But Fabra (1974), Ferrer (1973) and Romeu et al. (1985) give this order: cel·la, cella.]

Q. What are the local encoding schemes used? Are there Catalan keyboards with ·, l· or l·l support?

A. A centered dot appears in ISO 8859-3 as character 0xB7, and the character combinations LATIN CAPITAL LETTER L WITH MIDDLE DOT and LATIN SMALL LETTER L WITH MIDDLE DOT appear in positions 0x3F and 0x40 of row 01 (EXTENDED LATIN A) of ISO IEC DIS 10646-1.2. Besides these ISO codes for middle dot, character sets for Personal Computers happen to include a special "l·" character, often in the Danish or Norwegian code pages.

Q. Can it appear in ligatures, like fl·l or ffl·l ?

A. No, it cannot. For morphological reasons l·l has to be preceded and followed by vowel sounds.

Q. Are there special spacing rules? Is the dot special?

A. Yes, the l's are closer to the dot than other letters, and the dot is a normal dot but raised approximately half the height of a vowel from the baseline for lowercase and three times that height for uppercase[7].

Q. When did this letter appear in Catalan printing?

[6] The two last paragraphs demonstrate the hyphenation of the l·l ligature, which is discussed in detail in the next section.

[7] More reasonable spacing can be achieved by raising the dot exactly the height of a lowercase vowel, and this is precisely what has been coded in the macro for the l·l ligature presented below. Thanks to Marek Ryćko and Bogusław Jackowski for their comments on that particular spacing convention during the 1993 TEX Users Group Annual Meeting.

A. Although it was Pompeu Fabra who always supported the idea of an orthographic distinction in correspondence with the phonetic distinction between $|\lambda|$ and $|ll|$, his approach consisted of leaving *ll* for $|ll|$ and looking for a new symbol for $|\lambda|$. The actual ligature *l·l* is due to Mossèn Alcover in his amendment to the fourth writing of the *Normes Ortogràfiques* (Orthographic Norms) by the *Institut d'Estudis Catalans* (Catalan Studies Institute) (Segarra, 1985). The *l·l* ligature appeared therefore in Catalan printing for the first time in 1913 in *Normes Ortogràfiques*.

Choosing a macro for the l·l ligature. When it comes to choosing the best character sequence for the TeX macro producing the l·l ligature we realize that perhaps we Catalan TeX users have arrived too late, because most short combinations already have a definition in plain TeX. Among the interesting ones are `\l` and `\L`, assigned to Polish letters ł and Ł and `\ll`, assigned to the "much less than" relation $\ll$, whereas `\LL` is undefined in plain TeX.

It must be noted, however, that $\ll$ only occurs in math mode, while the l·l ligature is not supposed to be typed in math mode. We have therefore chosen `\ll` and `\LL` as character sequences for the macro definition producing the l·l ligature, and have included a test for math mode in the definition in order to restore the original $\ll$ relation when in math mode for lowercase `\ll`, as explained in the next section.

The macro name `\ll` is submitted to the TWGMLC for standarization.

Typesetting the l·l ligature. No normative exists for typesetting the *l·l* ligature and therefore quite different kernings between the middle dot and the two consonants can be found in modern Catalan writings. The definitions

```
\newskip\zzz
\def\allowhyphens{\nobreak\hskip\zzz}
\def\ll{\allowhyphens%
  \discretionary{l-}{l}{l\hbox{$\cdot$}l}%
  \allowhyphens}
\def\LL{\allowhyphens%
  \discretionary{L-}{L}{L\hbox{$\cdot$}L}%
  \allowhyphens}
```

constitute a good starting point because, besides achieving an easy-to-read spacing, such as in il·lusió and IL·LUSIÓ, they produce the right hyphenation. Middle dot is lost and l·l is hyphenated `l-l`.

Explicit kerning can be added between middle dot and the two consonants. Because kern is font-dependent, some character height, width, and depth values for the actual font in use are taken into account in the following definitions in order to set appropriate kerning.

```
\newskip\zzz
\def\allowhyphens{\nobreak\hskip\zzz}
\newdimen\leftkern
\newdimen\rightkern
\newdimen\raisedim

\def\ll{\relax\ifmmode \mathchar"321C
  \else
    \leftkern=0pt\rightkern=0pt%
    \raisedim=0pt%
    \setbox0\hbox{l}%
    \setbox1\hbox{l\/}%
    \setbox2\hbox{x}%
    \setbox3\hbox{.}%
    \advance\raisedim by -\ht3%
    \divide\raisedim by 2%
    \advance\raisedim by \ht2%
    \ifnum\fam=7 \else
      \leftkern=-\wd0
      \divide\leftkern by 4%
      \advance\leftkern by \wd1
      \advance\leftkern by -\wd0
      \rightkern=-\wd0
      \divide\rightkern by 4%
      \advance\rightkern by -\wd1
      \advance\rightkern by \wd0
    \fi
    \allowhyphens\discretionary{l-}{l}%
    {\hbox{l}\kern\leftkern%
      \raise\raisedim\hbox{.}%
    \kern\rightkern\hbox{l}}\allowhyphens
  \fi}

\def\LL{\setbox0\hbox{L}%
    \leftkern=0pt\rightkern=0pt%
    \raisedim=0pt%
    \setbox1\hbox{L\/}%
    \setbox2\hbox{x}%
    \setbox3\hbox{.}%
    \advance\raisedim by -\ht3%
    \divide\raisedim by 2%
    \advance\raisedim by \ht2%
    \ifnum\fam=7 \else
      \leftkern=-\wd0
      \divide\leftkern by 8%
      \advance\leftkern by \wd1
      \advance\leftkern by -\wd0
      \rightkern=-\wd0
      \divide\rightkern by 6%
      \advance\rightkern by -\wd1
      \advance\rightkern by \wd0
    \fi
    \allowhyphens\discretionary{L-}{L}%
    {\hbox{L}\kern\leftkern%
      \raise\raisedim\hbox{.}%
    \kern\rightkern\hbox{L}}\allowhyphens
  }
\endinput
```

The definitions produce the following result:

`\rm` l·l	intel·ligència	L·L	COL·LECCIÓ
`\it` *l·l*	*intel·ligència*	*L·L*	*COL·LECCIÓ*
`\sl` *l·l*	*intel·ligència*	*L·L*	*COL·LECCIÓ*
`\bf` **l·l**	**intel·ligència**	**L·L**	**COL·LECCIÓ**
`\tt` `l·l`	`intel·ligència`	`L·L`	`COL·LECCIÓ`

Availability

Besides the `\patterns` described in this paper, two other sets of hyphenation patterns exist for Catalan. They have been developed by Gonçal Badenes and Francina Turon (`badenes@imec.be`), and by Francesc Carmona (`franc@porthos.bio.ub.es`).

All three `cahyphen.tex` files are under beta test, and can be obtained from the respective authors. The authors have tested the version described here on a PC, using multilingual PCTeX 3.1, PCTeX 3.14 and emTeX 3.141 — using the primitive `\charsubdef` of Ferguson (1990) — and also on a Macintosh, using Euro-OzTeX and the Cork font scheme. Hopefully, a unified set of Catalan hyphenation patterns will soon be available by anonymous `ftp` from major TeX servers.

Acknowledgements

The authors wrote this joint work after two independent works submitted to *TUGboat* (Fuster) and to this conference (Valiente Feruglio). Barbara Beeton suggested this collaboration to us and the result is certainly better than the originals. Gonçal Badenes (IMEC, Leuven) and Francesc Carmona (Universitat de Barcelona) provided beta-test sets of hyphenation patterns. Joan Moratinos (Ajuntament de Palma de Mallorca) and Magdalena Ramon (Universitat de les Illes Balears) provided assistance in grammatical, linguistic and terminological issues. Joan Alegret (Universitat de les Illes Balears) provided guidance on etymological issues. Luz Gil and Eddy Turney (Universitat Politècnica de València) revised our English. Special thanks to all of them.

Bibliography

Beccari, C. "Computer Aided Hyphenation for Italian and Modern Latin". *TUGboat* **13**(1), 1992, pp. 23-33.

Bruguera i Talleda, J. *Diccionari ortogràfic i de pronúncia.* Enciclopèdia Catalana. Barcelona, 1990.

Colomer, J. *Nou diccionari anglès-català català-anglès.* Pòrtic, Barcelona, 1989.

Fabra, P. *Gramática de la lengua catalana.* Tipografia de "L'Avenç", Barcelona, 1912.

Fabra, P. *Ortografia catalana.* Barcino, Barcelona, 1927.

Fabra, P. *Diccionari general de la llengua catalana.* EDHASA, Barcelona, 1974.

Fabra, P. *La llengua catalana i la seva normalització.* Edicions 62 i "La Caixa", Barcelona, 1980.

Fabra, P. *Converses filològiques I.* EDHASA, Barcelona, 1983.

Fabra, P. *Converses filològiques II.* EDHASA, Barcelona, 1984.

Ferrer Pastor, F. *Vocabulari castellà-valencià valencià-castellà.* L'Estel, València, 1973.

Ferguson, M. J. *Documentation file for the use of charsublist.* June, 1990.

Haralambous, Y. 'TeX Conventions Concerning Languages". *TeX and TUG NEWS*, **1**(4), 1992, pp. 3-10.

Haralambous, Y. "Hyphenation patterns for ancient Greek and Latin". *TUGboat*, **13**(4), 1992, pp. 457-469.

Institut d'Estudis Catalans. *Normes Ortogràfiques.* Barcelona, 1913.

Knuth, D. E. *The TeXbook.* Addison-Wesley, Reading, Massachusetts, 9th printing, 1990.

Lázaro Carreter, F. *Lengua española: historia, teoría y práctica.* Anaya, Madrid, 1973.

Liang, F. M. "Word hy-phen-a-tion by com-pu-ter". Ph.D. thesis. Department of Computer Science. Stanford University, 1983.

Mira, J. F. *Som (llengua i literatura).* Edicions 3i4, València, 1974.

Pitarch, V. *Curs de llengua catalana.* Edicions 3i4, València, 1983.

Romeu, X. et al. *Diccionari Barcanova de la llengua.* Barcanova, Barcelona, 1985.

Salvador, C. *Gramàtica Valenciana (amb exercicis pràctics).* Lo Rat Penat, València, 1974.

Segarra, M. *Història de l'ortografia catalana.* Empúries, Barcelona, 1985.

Segarra, M. *Les set redaccions de les "Normes Ortogràfiques" de l'Institut d'Estudis Catalans.* In A. M. Badia i Margarit, Estudis de Llengua i Literatura Catalanes. X: Miscel·lània. Publicacions de l'Abadia de Montserrat, Barcelona, 1985.

Solà, J. *Lingüística i normativa.* Empúries, Barcelona, 1990.

Valor, E. *La flexió verbal.* Edicions 3i4, València, 1983.