Albanian hyphenation

Claudio Beccari

Abstract After a short historical review of the Albanian language the procedure used to create the Albanian hyphenation pattern file is described.

Sommario Dopo una breve esposizione storica della lingua albanese viene descritta la procedura per crearne il file di pattern per la cesura.

1. Introduction

Sometimes in mid 2020 the TEX Hyphen group received a message asking instructions to create the Albanian hyphenation patterns, since this language, at that time, was the only one among those written by means of the extended Latin alphabet that was missing its hyphenation patterns and was typeset with the "nohyphen" metalanguage settings, therefore no hyphenation at all.

I offered the Albanian young man asking for instructions about patterns all the support I could; I sent him some example procedures and many papers on hyphenation, but after that I did not see any Albanian pattern file in the TEX Live distribution and its upgrade procedures. May be they are stil under processing by the TEX Hyphen group, but I tried to experiment myself with a language that I do not know at all.

Albanian has nothing to do with western European languages, and very few people know that language, unless they are Albanian themselves. There are a little more than three million people in Albania who use that language; several hundred thousands autochthonous Albanian in the neighbouring countries, even across the Adriatic see; there are about three million recent emigrants resident all over the world; approximately there are a total of about eight million people around the world who speak, read, and write Albanian.

The Albanian literature is quite recent, compared with the literature of the other mediterranean countries; according to the linguists the first recorded writings in Albanian, or in "proto Albanian" date back to the XIII century. Printed records are very few until the XIX century. Therefore typographical hyphenations is quite recent.

Modern grammars, as usual[1] are very soft on syllabification[2]. Some modern dictionaries often offer both the pronunciation and syllabification of each lemma. But hyphenation dictio-

1. I already complained several times about the fact that grammars generally skip hyphenation and syllabification; if they do, that address elementary school children, that are starting to work with written language; at maximum they repeat the same rules for junior high school students, but always on an elementary bases that are not sufficient for a correct hyphenation practice. Difficult languages such as English, with its numerous varieties, have available huge lists of hyphenated words; romance languages have phonetic rules that are more or less complicated and may refer to the word etymology, but this often requires higher linguistic competence. Unfortunately linguistic scholars consider syllabification a topic too simple for their studies and generally discard it. There are exceptions, though: a famous Italian linguist was the chairman of the Italian Standardisation Office that published the regulation UNI 6461 (1969) some 50 years ago. Such regulation is still valid to day, but it fails with the many new words that entered modern Italian: they have a foreign stem or are compound with foreign words. Any daily newspaper reader can spot hyphenation errors almost every day with such "anomalous" words.
2. According to a trend among the TEX Hyphen Team, "syllabification" is a grammatical notion, while "hyphenation" refers to word breaking at the end of printed lines in order to justify the lines in a printed text.

naries, containing only hyphenated words are very rare, and do exist when the rules are too many to be applied by average educated people; the motto "too many rules equals no rule" is valid with these languages. Of course the Albanian situation appears to be in line with the above description.

At the beginning of TEX a PhD student, Liang (1983), who was working with Knuth (1996), made his doctoral thesis on computer operated hyphenation; his original PhD thesis is not available any more, but it is possible to find on Internet a scanned copy, where the reader can find a lot of information on the data structure of the pattern hash, but very little on the actual process of building a pattern set suitable for a computer: a procedure to process a hyphenation dictionary in order to obtain a set of patterns.

The program Liang wrote was named `patgen` and apparently is available also with TEX Live and the other TEX system distributions, but even with TEX Live the documentation for its use is either absent or so specific that it can't be used with languages different than English; let me remind that English is the only Western Language for which the 26 + 26 lower and upper case ascii glyphs are sufficient; I would say that all other western language use diacritics; but even English writers have to use them when citing foreign names, but probably the standard TEX accent commands are sufficient for such rare instances of foreign names; unfortunately those accent macros highjack the hyphenation process.

Liang initially worked with a 25 000 entry hyphenation dictionary, just to prove the validity of his thesis approach; the actual pattern file contained in any TEX distribution was prepared with a larger hyphenation dictionary that contained about 100 000 entries; in spite of this, tug (the international TEX Users Group) maintains a list of English hyphenation exception that are not handled by the default pattern set.

Nevertheless nowadays the Unicode encoding, and its transcode UTF-8, is the standard to input source TEX files even when working with pdfLATEX, irrespective of the language being used. Such encodings are very useful with the totality of the other western languages that use the extended Latin alphabet. Albanian is one of such languages. The original `patgen` program cannot be used with non-ascii glyphs, but there exists a version, `patgen2`, that was used when the Omega typesetting program was still maintained by Haralambous (2009); the instructions are not so simple to apply to Unicode encoded fonts[3] glyphs.

Therefore it was necessary to find a different approach to create the Albanian hyphenation patterns. This paper describes such an approach. I anticipate that I followed the same approach I used to test Latin hyphenation while I was trying to create a specific hyphenation pattern set to be used in the composition of Gregorian chants.[4]

2. Short historical description of the Albanian language

The Albanian language has been classified as an Indo-European one by the German linguist Franz Bopp in mid XVIII century. Before, it was considered a language by itself.

3. Translation: I was not able to use it with Unicode encoded glyphs.
4. Eventually I gave up, because of the different points of view between German, French and Italian coworkers, that were unable to find a common base. I admit that I was unsuited for this work because I was working with hyphenation in mind, while the other team members had not only syllabification in mind, but also the chant music; the German-French team continued and is still continuing that work.

Table 1. The 36 letters of the Albanian alphabet

A a	B b	C c	Ç ç	D d	Dh dh	E e	Ë ë	F f	G g	Gj gj	H h
I i	J j	K k	Ll	LL ll	M m	N n	Nj nj	O o	P p	Q q	R r
RR rr	S s	Sh sh	T t	Th th	U u	V v	X x	Xh xh	Y y	Z z	Zh zh

Linguists found several varieties in the actual Albanian language territory that includes also Kosova and some neighbouring areas in Montenegro, Northern Makedonia, Greece, and Italy. At the same time the Slavic languages and Greek are the obvious adstrate ones from where many words entered into the common Albanian language.

Other adstrate languages derived from occupations by other people, particularly important Turkish and Italian. This influenced also the way of writing; it is reported that Albanian in the past was written with Latin, Greek, Cyrillic, and Turko-Arabic alphabets. On the modern Albanian territory two main regional languages are spoken and used to be written: the northern Gheg and the southern Tosk. Outside the country the Arbëresh variety survives in central southern Italy; it is not due to a recent Albanian immigration; it is used by the descendants of the followers of Gjergj Kastrioti, known as Skanderbeg, a national hero who fought against the Ottomans in the XV century; eventually he was awarded some counties in central southern Italian peninsula by Ferdinand II, king of Naples. This variety of Albanian is the one that was used in the last part of medieval times and is spoken still today in some areas of the Italian Adriatic coast.

In 1909 a new spelling of Albanian and in 1972 a new Albanian "koinè" language was agreed on for the whole nation; this koinè does not supersede the regional languages, but is generally understood by everybody while its orthography is the official one and is used all over the country.

3. The Albanian letters

The alphabet of modern Albanian contains 36 '"letters" shown in table 1.

The word "letter" is quoted because it conflicts with the meaning we usually associate to this name; we usually call the two character signs with the name of digraphs; such single or two character signs ("letters" according to the Albanian terminology) represent phonemes. In the following, I will try to avoid the use of the word letter, in order to avoid misunderstandings, and use the words character or digraph depending on how many characters form each sign.

Apparently there are no diphthongs as in other languages: the diphthongs generally are groups of two vowels of which either the first or the second is an unstressed closed or semi closed vowel, which in these cases plays the role of a semiconsonant. The character "j" plays the role of a semiconsonant when preceded or followed by a vowel, but it plays the role of a diacritical sign when it is part of a digraph; therefore sometimes the role of a "j" is ambiguous for what concerns a computer; this device does not know the pronunciation of each word but operates only on its spelling, therefore on the character string that forms a word; for creating pattern files this ambiguity is a problem. Furthermore the language uses the apocope or elision that is handled with the apostrophe without any spaces before or after it.

Anybody who would like to know which phonemes the characters and the digraphs represent, should consult a bilingual dictionary to and from Albanian; more simply the

blog at `http://www.albanianlanguage.net/` (in English) or the Wikipedia page at `https://it.wikipedia.org/wiki/Lingua_albanese` (in Italian, but pretty well done) offer lots of information.

4. The approach to make Albanian patterns

What is explained above partially explain the premises of my approach.

1. Without any available document written in a language that I could decipher, this first obstacle is over if I could find somebody who not only knows the Albanian language, but also its grammar. Luckily enough, I have been knowing Sabina Kolici for many years; she was born and raised in Albania, where she got a university degree[5] in Albanian Literature; She has been living in Italy for more than 20 years. I proposed her to help me in this task and she accepted with enthusiasm.
2. We needed a text to work on; Ms Kolici chose a couple of chapters from an Albanian novel that she could get as a computer file, so as to use it as a test file; it is not important the name of the novel, because any book would have been usable for our purpose; Ms Kolici selected it because that novel did not contain special names connected to this or that professional discipline. Ms Kolici processed for me an initial file where the words were rearranged one per line.
3. I further processed the initial file in order to get rid of any hyphens that could have remained in such a file; I got rid of quotation marks, parentheses, punctuation signs, and the like; I lowercased all words and entered this long list of about 6000 words, into a spreadsheet column; with the facilities of the spreadsheet processing program, I sorted all words and eliminated all duplicates; the list was reduced to about 2600 words.
4. I created an initial `hyph-sq.tex`[6] where I introduced an initial set of pattern concerned with each and all single or double consonant signs at the beginning or at the end of words: i.e. a list of `.b2 2b. .c2 2c. .ç2 2ç.)....y2 2y. .z2 2z. .zh2 2zh. -- a very elementary set.`
5. I created a LuaLaTeX source file that would load the above pattern file and process the word list, in order to print out all input words in the hyphenated form. More on this file in the following section.
6. I would pass the hyphenated word list to Ms Kolici who would correct the wrong hyphen points and would send me the corrected list; I would modify the `hyph-sq.tex` file and repeat the process from the preceding step.

Of course this process would terminate when the corrected hyphenated word list would not contain any errors.

Too much work done by hand? Well, yes and no; missing an intelligent program that could compare the hyphenated word list with an initial hyphenated word list (as `patgen` or `patgen2` could have done) it is difficult to do it in a different way; moreover with this

5. Her degree is equivalent to a master degree of the European 3+2+3 university levels; this agreement was undersigned by the European member countries about 20 years ago; other European countries, although not members of the European Union, adopted the same system.
6. The string "sq" is the ISO code for Albanian: *Shqip* in Albanian.

do-rësh-kri-min	du-ro-i	ë-mës	fat-ke-që-si-a
do-re-zës	dy	em-rin	fat-ke-qe-ve
do-run-ti-na	dy-fish-të	en-de	fe-je-së
do-run-ti-në	dy-ja	ën-dë-rron-te	fe-je-sën
do-run-ti-nën	dy-ja-ve	ëndrre	fe-mër
do-run-ti-nës	dy-ja-vor	e-nësh	fë-mi-jët

Figure 1. A specimen of the output hyphenated word list

procedure the corrected word list and the corrected pattern list are processed by humans, who may also produce errors, but certainly are more intelligent than a program. Moreover `patgen` and `patgen2` require repetitions that do not reach the zero error solution, but a solution with the least number of errors. Our procedure stops when errors are absent.

I would not claim that our procedure is infallible; certainly it is not applicable to languages that do not mark the pronunciation; for example it is not applicable to English that contains homographs that are pronounced differently according to the role the word plays in a sentence; noun vs. verb; noun vs. adjective, and so on. A typical example is "to record" and "the record", but I could list many more. If accents were used, the words would not be homographs: "to recòrd" vs. "the rècord". Of course I am not invoking the use of stress accents in English (although this procedure was taken with ecclesiastic Latin); I am just remarking that a program that hyphenates words basing its rules on the spelling, cannot correctly hyphenate homographs that are pronounced and hyphenated in a different way depending on the role they play in a sentence.

5. The hyphenation testing source file

Here is the code for the LuaLaTeX testing file.

```
% !TEX TS-program = LuaLaTeX
% !TEX encoding = UTF-8 Unicode
\documentclass[12pt]{article}
\usepackage{fontspec}
 \defaultfontfeatures{Ligatures={NoCommon,
 NoDiscretionary, NoHistoric, NoRequired,
 NoContextual}}
\setmainfont{CMU serif}

\usepackage{luacode}
\usepackage{testhyphens}
\usepackage{multicol}

\begin{luacode}
local patfile = io.open('./hyph-sq.tex')
langobject = lang.new()
lang.patterns(langobject, patfile:read('*all'))
patfile:close()
\end{luacode}
```

```
\patterns{
2’2
.a1jo. a1a
1b .b2 2b. b2l 2bsh
1c .c2 2c. 2cj 2cn 2ct
1ç .ç2 2ç. 2çs ç2k
1d d2h .d2 2d. d2j 2dn d2r 2drr 2dt d2shmd2h 2dh. 2dhj2dht 2dhsh 2dhj 2dht
   dh2r dh2j
e1a e3ll e1u
ë1a
1f .f2 2f. f2l f2r 2fs 2ft 3f2sh 2f2t.
1g .g2 2g. g2j 2gj. 2gjv 2gl 2gm 2gr 2gt
1h .h2 2h. 2hd 2hj 2hm 2hn 2ht 2hrr
i1a i1e i1u .i2k3i .i2k3j
1j2 .j2 2j. 2j3c2 2j3d 2j3m 2j3p 2j3r 2j3t 2j3v 2j3s 2jf. j4tp 2jt. j3sh2m
1k .k2 2k. k2j 2kl 2km 2kth. k2r 2kt 2ks 2ksh
1l .l2 2l. 2lb 2lç 2lf 2lj 2lm 2ln l3n2g 2ls 2lt
l2l2 4ll. 2ll3s 4ll3z 2ll3k 4ll3gj 2ll3n 2ll3t 4ll3z
1m .m2 2m. m2b mb2j mb2l mb2r m2j 2m3n2d 2mt 2mr 2m3sh2 2m4sh. 2m1v
1n .n2 2n. .ng2r 2nc 2nd n2dm n2dv n2d3sh 2ng 2nk 2nsp 2nsh n3sh2m 2nt 2nv
   2nx 2nz n2j 2njt 2nj. 2njv
o1i
1p .p2 2p. p2j 2pn 2pt p2je. 2ps p2r pa2s3her .pe2r3af .që2r3af
1q .q2 2q. 2qj 2qk 2qm 2qn 2qt q2v
1r .r2 2r. 2rt 2rb2 2r2b3r 2rc 2rç 2rd 2rc2rd 2rf 2rg  2rh .ri3n2d 2rk 2rl
   2rm 2rn r2n3d2 2rp 2rq 2rs 2r3sh2m 2rdh r2dht 2r3dr 2rj 2rv 2rz
r2r .rr2 2rr. 2rrj 2rrk 2rrm 2rrn 2rrt 2rrs
1s .s2 2s. 2sb 2sc 2sd 2sf 2sg 2sj 2sk 2sm 2sn sn2k 2sp 2ssh 2st st2r
   3s2je2ll sk2ll s2ve.
s2h 2sh. .sh4 2shm 3sh2mj 2shj .sh2j sh2k 2shk. sh2n shn4d sh2p .sh2q
    4sh3k2r 2shq 2sh3nj 2shpr 3sh4pj 3sh4pr 2shr 2shs 2sht .sh2t 3sh2te.
1t .t2 4t. 2tk t2j 2tm 2tn 2tp 2t3sh2m t2r 2tv
t2h .th2 2th. 2thç 2ths 2thç 2thf 2thm 2tht
u1ar. u1a u1e
1v .v2 2v. 2vr v2j
1x .x2 2x.
x2h
y1
1z .zb2r .z2 2z. 2zm 2zn 2zj 2zs 2zt 2zv z3sh2m
z2h .zh2 2zhd
}
```

The obtained Albanian pattern file

```
20
21 \language=%
22 \directlua{tex.sprint(lang.id(langobject))}
23
24 \advance\textwidth 60mm
25 \advance\oddsidemargin -30mm
26 \advance\textheight 50mm
27 \advance\topmargin -20mm
28
29 \begin{document}
30
31 Language: \the\language
32
33 \bigskip
34
35 \begin{multicols}{4}
36 \lefthyphenmin=1
37 \righthyphenmin=1
38 \begin{checkhyphens}
39 acarime
40 acarimi
41 afërta
42 ...
43 zyrë
44 zyrtarë
45 \end{checkhyphens}
46 \end{multicols}
47 \end{document}
```

Some comments are in order.

1. Line 1 and 2 are the so called "magic lines" that inform the editor that the file has to be saved with the UTF-8 transcode encoding, and the the file has to be processed with LuaLaTeX. Those shell editors that understand such "magic lines" act as specified according to such lines, and use the LuaLaTeX typesetting engine even if their default engine is a different one. If the editor does not understand these lines, the user can understand them and s/he can act accordingly.
2. Line 3 is obvious.
3. Lines from 4 to 8 use fontspec to specify which fonts have to be used for the output; in this case we select the normal Computer Modern OpenType fonts for the main (serifed) font.
4. Lines from 10 to 12 specify the packages to load: luacode is to allow to use Luacode in this input file; testhyphens is a package to print hyphenated words according to the patterns of the current language; multicol prints the text in a specified number of columns.
5. Lines from 14 to 23 are the specific lines that involve the `hyph-sq.tex` Albanian language pattern file and lines 22 and 23 set such pattern file as the default. This is

the special feature of LuaLaTeX we are exploiting; pdfLaTeX and XeLaTeX require that all hyphenation pattern file are preloaded while constructing their specific kernel `.fmt` format file; LuaLaTeX, on the opposite, can load hyphen pattern files even at run time.

6. Lines from 25 to 28 are low level commands to enlarge the typesetting area; by so doing and selecting to typeset the results in four columns we are pretty sure that no hyphenated word is going to be split on two or more lines.
7. Line 32 prints out the number of the current language; missing this line, one cannot be sure that the new pattern files have been correctly loaded and are the current ones in force.
8. On line 36 the molticols environment is opened with a specification of four columns; it shall be closed by the end of the file, in this numbered listing at line 47.
9. In lines 37 and 38, the minimum lengths of the first and last syllables of each word are both set to one character; actually these numbers shall not be the default values for actual typesetting; such actual values are specified by `albanian.ldf` or `gloss-albanian.ldf`. Here we set these values to 1 in order to verify that digraphs don't get split by the hyphenation process.
10. On line 39 the checkhyphens environment is opened and shall be closed at the end of the word list. This is the specific environment that hyphenates each word at *all its hyphen points*, and outputs the corresponding string so as to have the full word hyphenation to check with the "hand made" controlling process; human eyes are very good in noticing wrong break points.

A specimen of the output file is shown in figure 1 on page 5; the actual complete file is 15 pages long.

6. The final pattern file

The described procedure worked fine; the final hyphenation pattern file for Albanian turned out as shown on page 6.[7] The total number of patterns resulted to be 310, not so different from the corresponding number for some other romance languages, although Albanian does not belong to that language class.

It must be noticed that the pattern set shown on page 6 does not produce any errors with the 2600 words used to make them up. If tested with other texts it might still produce some hyphenation errors. However this is normal; at the beginning all the other pattern files produced errors and the corresponding pattern files had to be updated; nevertheless we are confident that this pattern set may produce correct results with the majority of texts that use common words. We hope that the community of Albanian LaTeX users continues to upgrade the Albanian pattern sets while they use the hyphenation utility with their actual documents. If they do not, each author has to create his/her hyphenation exception lists; of course this is not terrible, but it is what happens when a LaTeX user community is too small and/or does not work as a community. We hope that the Albanian community will be active in caring and maintaining their language facilities provided by the TeX system.

7. The pattern file is built by hand; the program that has to prove its correctness does not care if each line contains many patterns, but the pattern files distributed with any TeX installation have one pattern per line. Here, in order to save space, we selected to show several patterns per line.

7. Example

Here we show a short list of fully hyphenated Albanian words:

be-së-shke-lë-sin bu-da-lla-llëk di-nji-te-tin dëm-tu-a-ra do-rësh-kri-min fi-llu-an
go-di-tje-je ha-men-djesh he-rë-pas-her-shme kë-mbën-gul-ja ko-rri-do-rit mby-llu-ra
mi-rë-sje-lljes ngul-çi-mës një-ko-hë-sisht pshe-rë-ti-u s'u-lë-ri-u shp je-go-hej
u-dhë-hiq-nin zma-dhu-ar

Such words are a small sample of the results obtained with the used procedure where hyphenation was performed with both end syllable lengths set to unity. In the real usage of the Albanian language such parameters would be set to 2.

Conclusion

The `hyph-sq.tex` file has been submitted to the TEX Hyphen group; they are going to extract from this single file the necessary variants to be used with the other typesetting systems, as they usually do for all languages.

Acknowledgments

Creating Albanian pattern files requires a good knowledge of the language; I know how to make pattern files, but I do not know Albanian. My deep thanks are therefore due to Ms Sabina Kolici who contributed to this work with her constant support.

References

Beccari, Claudio (2014). «Greek and Latin hyphenation — Recent advances». *ArsTEXnica*, (18), pp. 87–96.

Carette, François e Arthur Reutenauer (2015). «Polyglossia: an alternative to Babel for XeLaTeX and LuaLaTeX». `http://mirrors.ctan.org/macros/latex/contrib/polyglossia/polyglossia.pdf`.

Haralambous, Yannis (2009). «A small tutorial on the multilingual features of PatGen2». PDF document. Readable with line command `texdoc patgen2`.

Knuth, Donald E. (1996). *The TEX book*. Addison Wesley, Reading, Mass., 16ª edizione.

Liang, Frank (1983). *Word Hy-phen-a-tion by Com-put-er*. Tesi di Laurea, Stanford University.

UNI 6461 (1969). *Divisione delle parole in fin di linea*. Ente Italiano di Unificazione, Milano.

Claudio Beccari
`claudio.beccari@gmail.com`