

Denis Moyogo Jacquerye <moyogo@gmail.com>  
2013-07-22

## Summary

The document provides comments on the recent discussion about cedilla and comma below and the proposal to add characters with invariant cedilla. With the evidence found, our conclusion is that the cedilla can have several forms, including that of the comma below or similar shapes. Adding characters with invariant cedilla would only increase the confusion and could lead to instable data as is the case with Romanian. The right solution to this problem is the same as for other preferred glyph variation: custom fonts or multilingual smart fonts, better application support for language tagging and font technologies.

## Background

In the Unicode Standard the cedilla and comma below are two different combining diacritical marks both inherited from previous character encodings. This distinction between those marks has not always been made, like in ISO 8859-2, or before computer encodings where they were used interchangeably based on the font style. The combining cedilla character is de facto a cedilla that can take several shapes or forms, and the combining comma below is a non contrastive character only representing a glyph variant of that cedilla as historically evidence shows.

Following [discussion](#) about the Marshallese forms used in LDS publications on the mpeg-OT-spec mailing list, Eric Muller's document (L2/13-037) raises the question of the differentiation between the cedilla and the comma below as characters in precomposed Latvian characters with the example of Marshallese that uses these but with a different preferred form.

In the Ad Hoc document (L2/13-128), it is concluded that non decomposable characters matching the description provided by Muller should be encoded. Everson's proposal extends the list of non decomposable characters based on issues raised on the Unicode mailing list.

This solution is inadequate, unnecessary and would create further confusion. It is inadequate because it only guarantees that the non decomposable characters will have a specific cedilla, any character canonically using the combining cedilla might have a different shape. So in Marshallese the O and M with cedilla might have a different cedilla than the proposed L and N with cedilla. In transliterations the C with cedilla might have a different one than the proposed non decomposable characters. It is unnecessary as there is evidence that Marshallese allows for variation, and that the comma-like cedilla is acceptable as it is used in various publications. Several solutions are already being used for Marshallese or transliterations (adequate fonts or addition of Marshallese language tag to the OpenType specification).

At this point, it does not make sense for the Latvian community to change the characters they use as Robert Rozis points (L2/13-127). In fact, historically the classic cedilla was the diacritic used in the early 20th century in books and laws that defined the Latvian spelling rules, and its comma-like shape became more popular only because of style preferences. Following today's Latvian preferred form the Unicode reference glyphs and the majority of system fonts use the comma-like cedilla for the Latvian letters but generally use a more classical form for other letters. Marshallese and other communities use fonts that have a cedilla they prefer. The lack of support for language tagging is not an excuse to encode unnecessary and confusing characters, and can have disastrous consequences as with Romanian. Unicode should instead encourage its members and implementors to provide better support for language tagging and smart fonts, already required by many language communities. There are several cases where more than one form is needed, like the Polish kreszka instead of the acute, the Navajo hook instead of the ogonek, or the Lazuri caron instead of the Czech and Slovak caron. All these can only be supported if the whole software stack and fonts let the user choose what variants they want to display.

## Encodings

A distinction between cedilla and comma below can be found in the ISO 5426:1980 where character 0x50 is cedilla, and 0x52 is hook to left (comma below); and in ANSEL (ANSI/NISO Z39.47-1985) where 0xF0 is cedilla, and 0xF7 is left hook (comma below). Unicode inherited these characters, which in practice are never used with distinctive meaning but only represent style or glyph variation.

## Marshallese

Following adoption by the MRI Ministry of education and the Marshallese Language Orthography (Standard Spelling) Act of 2010, Marshallese uses the spelling rules of the Marshallese-English Dictionary (1976). These rules follow the recommendations of a Language Committee on spelling Marshallese made in the 1970s, which did not specify what diacritic was to be used below. The MED uses cedillas attached to the rightmost stem of *m* and *n* and centered under *l* and *o* in its introduction (pages i-xxvii), but it uses centered cedillas under those letters in the dictionary parts (pages 1-589). The Marshallese-English Online Dictionary uses the dot below since the Language Committee did not specify what mark was to be used and because dot below character already exists as a precomposed character in Unicode.

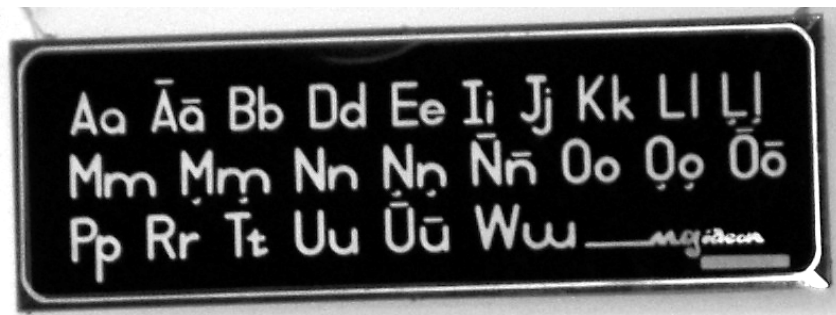


Figure 1: *Marshallese Alphabet* photo taken in a Public Library in Delap-Uliga-Darrit (Majuro, RMI) by Wikimedia Commons User:Enzino in 2002, showing a half-ring shaped cedilla.

# Marshallese Alphabet

a, ā, b, d, e, i, j, k, l, l̄, m, m̄, n, ñ, o,  
ō, ō, p, r, t, u, ū, w

Figure 2: Lori Philips, *Marshallese Alphabet* 2004, inside coverpage, showing a detached cedilla, half-ring shaped.

*abōl̄* (habel̄). From Engl. 3. Apple.  
*Abōl̄* (habel̄). 6(-i). A plant, banana variety.  
*abōn* (haben). 1(*abōna-*). Spiritual power. *Ekajoor wōt abōnān rūanijnij eo*. The sorcerer has such great powers.  
*abōne-* (habene-). 1, 6. Characteristics.  
*abōntoun* (habentawin). From Engl. *up and down*. 1(-i), 2(inf, tr -i), 3, 4. Teeter-totter; see-saw.  
*abwin* (habin). Dial. W, E: *abōb* (habēb). 2, 4. Refuse; dislike to do.  
*abwin bōk* (habin bek). Reject. *Ear abwin bōk mōñā*. He rejected food.

Figure 3: Takaji Abo et al., *Marshallese English Dictionary* 1976, page 3, showing a centered cedilla under Marshallese letters.

jeļā	know, know how to, find out Ex. ljeļā = I know Ex. ljab lukkuun jeļā = I don't really know/I'm not sure
jaje or ñak	not know, not know how to
kōñaan	want, like, do often
maroñ	can, may, might, possible
ban	cannot, will not, impossible
mejeje	understand, disentangled, meaning, information Ex. Mejeje in 'laddik', 'boy' = 'Laddik' means 'boy' Ex. Ta mejeje in 'laddik'? = What does 'laddik' mean?
dike	hate
ṁakoko (in)	unwilling (to), refuse (to), really not want (to)
meļokļok	forget

Figure 4: Peter Rudiak-Gould, *Practical Marshallese* 2004, page 16, showing an acute shaped cedilla.



Figure 5: Marshalles stamps, 1998, showing a centered cedilla.



Figure 6: Marshalles stamps, 2010, showing a centered cedilla.

Ł	Roman capital letter L with cedilla	A6 (4), A8 (10), CF (13*)	013B
ł	Roman small letter l with cedilla	B6 (4), B8 (10), EF (13*)	013C
Ł	(Roman capital letter M with cedilla)	-	004D+0327
ł	(Roman small letter m with cedilla)	-	006D+0327
Ñ	(Roman capital letter N with macron)	-	004E+0304
ñ	(Roman small letter n with macron)	-	006E+0304
Ŋ	Roman capital letter N with cedilla	D1 (4,10), D2 (13*)	0145
ņ	Roman small letter n with cedilla	F1 (4,10), F2 (13*)	0146
Ō	Roman capital letter O with macron	D2 (4,10), D4 (13*)	014C
ō	Roman small letter o with macron	F2 (4,10), F4 (13*)	014D
Œ	(Roman capital letter O with cedilla)	-	004F+0327
œ	(Roman small letter o with cedilla)	-	006F+0327
Ū	Roman capital letter U with macron	DE (4), AE (10), DB (13*)	016A
ū	Roman small letter u with macron	FE (4), BE (10), FB (13*)	016B

Figure 7: United Nations Group of Experts on Geographical Names (UNGEGN), *Technical reference manual for the standardization of geographical names*, 2007, showing a centered cedilla.

## Latvian

The current Latvian orthography has its roots in the work done at the end of the 19th century, the Latvian Language Commission started reforming the spelling rules in 1908. Several laws were adopted to redefine the spelling rules, in several publications these are clearly using the classic cedilla. The use of the classic cedilla is nowadays inadequate.

# LIKUMI UN RIKOJUMI.

## Rīkojums par ortogrāfijas reformu.

1. Ne vēlāk kā 1921. gada 1. jūlijā visās valdības un komunālās iestādēs un skolās gotu burtu vietā jāieved latīņu burti.
2. Gotu burtu **w**, **f**, **ſ**, **z**, **ch**, **ch**, **ſch**, **ſch**, **ſch**, **ſch**, vietā jālieto **v**, **s**, **z**, **c**, **š**, **ž**, **č**, **dz**, **dž**.
3. Mikstinātie **l**, **r**, **n**, **k**, **g** — jāraksta **l**, **r**, **n**, **k**, **g**.
4. Svešvārdos jāizšķir **h** (piem. Hamburga) no **ch** (piem. Čaņkova).
5. Saliktā patskaņa „**ee**“ vietā jāraksta „**ie**“.
6. Patskaņu gaŗumu apzīmēšanai vārdu saknēs, piedēkļos un galotnēs jālieto horizontāla svītriņa (—), pie kam jāapzīmē gaŗums visiem gaŗi izrunājamiem patskaņiem un ne tikai daŗiem.

Rīgā, 30. decembrī 1920. g.

Ministŗu prezidents **U l m a n i s**.  
Izglītības ministris **J. Plāķis**.

Figure 7: *Izglītības Ministrijas Mēnešraksts*, Nr.2 (01.02.1921), page 218, showing clear classic cedillas attached to the rightmost stem.

## Noteikumi par latviešu ortografiju

1. Pamatojoties uz 1922. g. 24. maija likuma par ortografijas jautājuma nokārtošanu, ortografijas komisija pie izglītības ministrijas noteikusi sekošu latviešu ortografiju, kuŗa lietojama visās skolās, valsts un pašvaldības iestādēs;

- a) gotu burtu vietā jāieved latīņu burti;
- b) gotu burtu **w**, **f**, **ſ**, **z**, **ch**, **ch**, **ſch**, **ſch**, **ſch**, **ſch**, vietā jālieto **v**, **s**, **z**, **c**, **š**, **ž**, **č**, **dz**, **dž**;
- c) mikstinātie **l**, **r**, **n**, **k**, **g** — jāraksta **l**, **r**, **n**, **k**, **g**;
- d) svešvārdos jāizšķir **h** (piem. Hamburga) no **ch** (piem. Čaņkava);
- e) saliktā patskaņa „**ee**“ vietā jāraksta „**ie**“;
- f) patskaņu gaŗumu apzīmēšanai vārdu saknēs, piedēkļos un galotnēs jālieto horizontāla svītriņa (—), pie kam jāapzīmē gaŗums visiem gaŗi izrunājamiem patskaņiem un ne tikai daŗiem.

2. Pārveidotā ortografija obligatoriski ievadama visās skolās sākot ar 1922. g. 1. augustu, bet valsts un pašvaldības iestādēs sākot ar 1923. g. 1. janvāri.

Rīgā, 1922. g. 18. jūlijā.  
(V. V. № 157.)

Ministŗu prezidents **Z. Meierovics**.  
Izglītības ministris **A. Dauge**.

Figure 8: *Izglītības Ministrijas Mēnešraksts*, Nr.2 (18.07.1922), page 672, showing clear classic cedillas attached to the rightmost stem.

(Pieņemti uz Izglītības un kultūras nolikuma 731.  
panta pamata.)

1. Latviešu alfabētā ir šādi burti: a, b, c, č, d, e, f, g, ģ, h, i, j, k, ķ, l, ļ, m, n, ņ, o, p, r, ŀ, s, š, t, u, v, z, ž.

P i e z ī m e. Latgaliešu uzvārdu rakstīšanai izloksnes formā var lietot arī burtu y.

2. Patskaņu a, e, i un u gaŗuma apzīmēšanai lietojami šie paši burti ar gaŗuma zīmi — ā, ē, ī un ū.

Figure 9: *Ceļš un Satiksme. Oficālā daļa*, Nr.15 1939-08-01, page 1, showing an inprecise cedilla but a clear upturned cedilla above g.

**Aģitatori sāks darbu**

Gatavošanās vietējo Darb- laužu deputatu padomju vē- lēšanām jaunu rosmi izraisī- jusi Rumbas ciema kolchoz- zos. Kolchozā „Kopdarbība” šajās dienās notika aģitatoru sanāksme. Tajā viņiem ieda- līja apkalpojamās mājas, un aģitkolektīva vadītāja b. Fi- šere viņus iepazīstināja ar tuvākajā laikā veicamajiem uzdevumiem.	Rumbas ciema bibliotēkā izraudzītas telpas aģitpunk- tam. Tā vadītāja pienākumi uzticēti komunistam b. Jan- sonam. Uzsākta aģitpunkta iekārtošana. Bibliotekare b. Krūmiņa jau savākusi vien- kopus literatūru, kas nepie- ciešama vēlēšanām veltītā stenda iekārtošanai.
---	--

**A. Llepa**

Figure 10: *Padomju Kuldīga*, Nr.7 (1896), 1957-01-16, page 2, showing an upturned cedilla on g instead of an upturned comma.

## Cedilla's various forms

In languages where the classic cedilla is the preferred form, variation is usually well acceptable. Several text body typefaces have comma shaped cedilla, or other variations. In display typefaces the variation can be even greater. Even in Latvian some variation is acceptable, and more used to be acceptable. So a Latvian font with classic cedilla is correct but only for documents of another era.



I. Moore  
1768  
Stephenson Blake & Company Ltd.  
H. Berthold AG

---

ABCDEFGHIJKLMNO  
PQRSTUVWXYZ  
abcdefghijklmnopqrstuvwxyz  
1/234567890 %  
(.,-;:;!i?c-).['‘,,”“»«]  
+-=/\$£†\*&§  
ÄÅÆÖØŒÛäåæıöøœßü  
ÁÀÂÃÇĆÉÈÊËÏÎĬŁŃŇÓÒÔÕ  
ŘŔŠŤÚÛÛŴŶŸŹ  
áâãçćéèêëïîĭłńňóòôõ  
řŕššťúúûŵŵŷŷžž

Figure 12: Götz Gorissen, *Berthold Fototypes*, volume 1, 1982, s.v. Baskervill Old Face, page 173

## ÉDITORIAL

---

François Hollande menaçait de bloquer le lancement des négociations. Soucieux du sort des dispositifs de soutien à l’audiovisuel, comme les

Figure 13: *Le Monde*, 18 June 2013, page 1, showing a curved-tick shaped cedilla different from the comma, with TheAntique typeface used in French newspaper.

■ Yüksek yargı üyeliğine 20 yıl kıdem şartı getiren yasayı onaylayan Cumhurbaşkanı Gül’den bir de tekzip geldi. Köşk “Sözcü gazetesinin ‘Gül’ün partisi hazır’ başlıklı haberi tamamıyla asılsızdır” açıklaması yaptı. ● 12

Figure 14: *Bügun*, 7 July 2013, page 1, showing a tick shaped cedilla, used in Turkish newspaper.

## na para oito anos e inos na rede pública

hoje pelo governo federal lança o  
o pacto da saúde a fim de melhorar a  
o ações polêmicas recebidas com duras  
lo setor, cautela no meio acadêmico e  
mpliação do curso de medicina para  
os alunos a trabalharem os dois últimos  
Nesse período, eles terão registro  
olsa com valor ainda indefinido.  
le médicos estrangeiros para atuar em  
rasileiros que vivem no país ou no  
que estão abertas a partir de hoje.

## Mobilização pelo fim do voto secreto

Deputados estaduais mineiros  
querem decidir ainda hoje o fim  
das votações sigilosas na  
Assembleia Legislativa, depois de  
duas tentativas frustradas por  
falta de quórum. Projeto de lei  
teve uma primeira versão  
aprovada em julho de 2012, mas  
foi ampliado para garantir voto  
aberto em todas as situações.

PÁGINA 7 E EDITORIAL, NA 8

Figure 15: *Estado de Minas*, 9 July 2013, page 1, showing a tick shaped cedilla and a curved-tick shaped cedilla, used in Brazilian Portuguese newspaper.

## Romanian

Romanian has always used the classic cedilla and the comma below glyphs to represent the same diacritic. At this moment Romanian data is still largely using the cedilla characters even if the Romanian Standardization body (ASRO) opted for the comma below characters in 1998, the Romanian Academy adopted the comma as the correct form in 2003 and law forces public institutions to use the comma below since 2006. By analyzing the [Romanian corpus](#) of Hans Christensen collected between 2011 and 2013, we found that on news websites, the common word *mulți* is encoded with no marks *multi* 3.28%, with the comma below *mulți* 5.52% and with the cedilla *mulți* 91.19% of the time, the common word *și* is encoded with no marks *si* 3.58%, with the comma below *și* 4.84% and with the cedilla *și* 91.56% of the time. Similar ratios occur in data collected from blogs and Twitter. Several of the most visited Romanian websites (Google.com, Yahoo.com, etc.) mix both the cedilla and comma below characters. Even if the decision to add and use comma below characters makes sense in 1998, when smart font technologies weren't widely available and only custom fonts were the solution, it has had disastrous consequences on stability of Romania data. It is completely confusing to common users, information specialist and font developers, as both older print data and electronic data uses the cedilla and the comma below interchangeably.

## Solutions to locale preferred glyphs

There are several situations where the same character has different preferred glyphs depending on language, context or era.

All these problems can be solved with custom fonts, as the Marshallese community is currently doing, or with smart multilingual fonts relying on applications to let users access preferred glyphs, which is already possible through advanced OpenType features and will be in the near future with the Marshallese locale settings with default OpenType features.

These problems are shared with Serbian preferred Cyrillic be, Polish preferred steep kreska over acute, various capital eng in different languages or countries, simplified or traditional Chinese, Latin letters with stroke (centered or high), or Latin letters with caron (Czech and Slovak changed form or transliteration and phonetic unchanged form)

See for example: [Cyrillic be and others](#), [Kreska](#), or [Eng](#).

Information for multilingual font development best practices can be found on sites such as [ScriptSource.org](#), Wikipedia, [Diacritics Projects](#) and Type forums.



## Bibliography

- Takaji Abo et al., Marshallese-English Dictionary (MED), Honolulu: University Press of Hawaii, 1976
- Götz Gorissen, Berthold Fototypes: Body Types, 2<sup>nd</sup> edition, volume 1, Berlin: Berthold, München: Callwey, 1982
- Dzintra Paegle, Pareizrakstības jautājumu kārtošana latvijas brīvvalsts pirmajos gados (1918–1922) (=The Standardization of the Orthography during the First Years of the Republic of Latvia (1918–1922)), Baltu filoloģija, XVII (1/2), 2008, [PDF](#), pp. 73–88
- Dzintra Paegle, Pareizrakstības jautājumu kārtošana Latvijā no 1937. līdz 1940. gadam (=The Standardization of the Orthography in Latvia from 1937 to 1940), Baltu filoloģija, XVII (1/2), 2008, [PDF](#), pp. 89–102
- Lori Philips, >Marshallese Alphabet, Honolulu: Pacific Resources for Education and Learning (PREL), 2004
- Peter Rudiak-Gould, Practical Marshallese, [www.peterrg.com/Practical Marshallese.pdf](http://www.peterrg.com/Practical_Marshallese.pdf)/span>, 2004
- Steve Trussel et al., Marshallese-English Online Dictionary (MOD), 2009, [www.trussel2.com/mod/index.htm](http://www.trussel2.com/mod/index.htm)
- United Nations Group of Experts on Geographical Names (UNGEGN), Technical reference manual for the standardization of geographical names, New York: United Nations, 2007, [PDF](#)
-