The
TRANSLITERATOR
for ConTEXt
MANUAL
The Transliterator module and mini-manual,
by Philipp Gesang, Radebeul.
Mail any patches or suggestions to
philipp-dot-gesang-at-alumni-dot-uni-heidelberg-dot-de
https://phi-gamma.net
## CONTENTS

**USAGE AND FUNCTIONALITY** .............................................. 2

* Overview 2
* Loading and Configuring the Module 2

**INTRODUCTION** .......................................................... 5

**EXAMPLES** ............................................................... 8

* Cyrillic scripts 8
  * iso 9 and derivatives 8
  * “Scientific” transliteration 10
* Serbian 11
* Bulgarian 12
* Legacy national transcriptions 12

* Glagolitic 14
* Greek 14

**REFERENCES** ............................................................... 16
USAGE AND FUNCTIONALITY

OVERVIEW

The Transliterator provides two commands: \setuptransliterator preferably goes into the preamble and allows for global configuration. The Transliterator is invoked locally by \transliterate which does the actual transliteration of text passages.

\setuptransliterator [...,*.*...]
* mode = ru_old ru ru_transcript_de ru_transcript_de_exp
   ru_transcript_en all sr_tolt sr_tocy bg_de iso9_ocs ocs
   ocs_gla ru_cz ocs_cz gr gr_n
hyphenate = cs sk hr
deficient_font = yes no
hinting = yes no
sr_exceptions = yes no

\transliterate [...,*.*...]
* inherits from \setuptransliterator

LOADING AND CONFIGURING THE MODULE

In order to use the Transliterator in a document we put the following somewhere before \starttext.

\usemodule[transliterator]

Although it has some defaults already set at this point they will most likely not correspond to what is needed in the document. To override the presets we use the command \setuptransliterator[#1]. It takes a comma separated list of two key-value pairs: mode and hyphenate. Through mode we specify the transliteration method. By the time of this writing this can be one of the following set:

Nota bene: The description at this point only serves as a placeholder as the transliteration modes are discussed in detail later in this document.
### Table 1  Transliteration modes.

<table>
<thead>
<tr>
<th>mode</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>ISO 9 complete</td>
</tr>
<tr>
<td>bg_de</td>
<td>Bulgarian, German „scientific“ transliteration</td>
</tr>
<tr>
<td>gr</td>
<td>transliteration for Greek</td>
</tr>
<tr>
<td>gr_n</td>
<td>transliteration for Greek obeying nasalizations</td>
</tr>
<tr>
<td>iso9_ocs</td>
<td>== all plus non-ISO additions for Old (Church) Slavonic</td>
</tr>
<tr>
<td>ocs</td>
<td>“scientific” transliteration for Old (Church) Slavonic</td>
</tr>
<tr>
<td>ocs_cz</td>
<td>Czech transcription for Old (Church) Slavonic</td>
</tr>
<tr>
<td>ocs_gla</td>
<td>“scientific” transliteration for Old (Church) Slavonic / Glagolitic alphabet</td>
</tr>
<tr>
<td>ru</td>
<td>ISO 9 Russian</td>
</tr>
<tr>
<td>ru_cz</td>
<td>Czech transcription for Russian</td>
</tr>
<tr>
<td>ru_old</td>
<td>ISO 9 Russian plus pre-1918 chars (the default)</td>
</tr>
<tr>
<td>ru_transcript_de</td>
<td>German transcription for Russian</td>
</tr>
<tr>
<td>ru_transcript_en</td>
<td>English transcription for Russian</td>
</tr>
<tr>
<td>sr_tocy</td>
<td>Serbian, Latin to Cyrillic</td>
</tr>
<tr>
<td>sr_tolt</td>
<td>Serbian, Cyrillic to Latin</td>
</tr>
</tbody>
</table>

Through the **hyphenate** argument it is possible to adjust the language that is used for hyphenation. Specifying `\setuptransliterator[hyphenate=nl]` will let every transliterated part of the document be processed according to dutch rules, leaving the overall `\language[#1]` configuration unchanged for the rest of the content.

Another argument, **deficient_font** can be used in combination with the modes all, ru_old and iso9_ocs. It lets you circumvent the deficiency that some fonts show concerning the characters that ISO 9 assigns to cyrillic “ъ” and “ь”. Set it to true to enable it.

The actual transliteration is done using the macro `\transliterate[#1] {#2}`. The second argument takes the raw string in the original language that we want to process, while the first, optional argument accepts local adjustments for **mode** and **hyphenate**. Thus, we would typeset one of Epicuros’ sayings like this:

```
\transliterate[mode=gr]{κακὸν ἀνάγκη, ἀλλ’ οὐδεμία ἀνάγκη ζῆν μετὰ ἀνάγκης}
```

which yields “kakon anagkē, all’ udemia anagkē zēn meta anagkēs” in the PDF output. Alternatively there is an environment, `\starttransliterate[#1]`, as well, that takes the same arguments.

There are two special switches for the **Serbian** patterns, **hinting** and **sr_exceptions**, allowing for a little more fine-tuning. If activated, hinting
provides the special character “*” as a means to indicate positions, where the sequences “lj” and “nj” are to be treated as separate consonants. E. g. \transliterate[mode=sr_tocy]{in*jekcija} is correctly transliterated as инјекција, and not ињекција. Likewise, further exceptions that are internally represented as a lookup table can be toggled off or on by the sr_exceptions switch. This pertains to words like “nadživeti” (result: надживети) but may lead to accidental false positives in cases that the module author didn’t foresee. By default both hinting and lexical exceptions are set to yes.

For orientation purposes the Transliterator comes with two macros that allow for closer inspection of the internal tables. \showOneTranslitTab{#1} outputs, obviously, a single table; their identifiers can be found in the trans_tables_*.lua files in the transliterator directory. The lazy alternative is \showTranslitTabs which prints all registered tables in a row nicely formatted as indexable sections. (Be warned, this may take some time.)
INTRODUCTION

What’s all this, then?

GRAHAM CHAPMAN

At the first glance, transliteration – the accurate representation of letters from one alphabet in another – seems obsolete after the advent of Unicode which made its way even into TEX lately. Why not just go on and write down everything in the original script? But still there are lots of situations where transliteration is desirable, e.g. some scholarly habits might prescribe it in the main text with citations in footnotes left in the original alphabet; or transliteration might alleviate comparison within one language that happens to be written in different scripts; finally, including text in a foreign script might be impossible if there is no appropriate font which fits the main text. However, it is still most convenient for the writer to keep the untransliterated original in the document source as this allows for reusing it in another context where different transliterations rules might apply. The Transliterator module is meant to provide both: have the original in the source and a transliteration only in the final document.

Another way of handling foreign languages is transcription. It aims at producing some representation that does not rely on symbolisms alien to the language and thus to be at least “pronouncable” without further knowledge. As transcription methods are language specific and highly idiosyncratic they complicate the restoration of the original phrase because information may be lost. The Transliterator provides means of transcription as well but in most cases you should refrain from using them ([mode=ru_transcript_en], [mode=ru_transcript_del]).

For Cyrillic scripts the best quality is achieved using the standardized transliteration according to ISO 9.¹ This method not only covers all contemporary languages that are written in a variety of Cyrillic but provides a bijective mapping on latin characters as well. Consequently, you can unambiguously revert the transliteration into its original form which was impossible with previous versions of ISO 9 because they contained several exceptions depending on the original language. Although fifteen years old it has not yet made its way

into scholarly publications at large so it might not immediately look familiar.\textsuperscript{2} The diacritics are not identical to the “scientific” transliteration used in Slavic studies but as long as your editor does not enforce its traditional method you should always prefer ISO 9 ([mode=ru], [mode=ru_old], [mode=all]).

But ISO 9, too, has its shortcomings. It has no definitions for historical forms of the cyrillic script like pre-XVIII-century Russian and Old (Church) Slavonic while those are covered by the scholarly transliterations. To amend the situation the Transliterator provides an extension to ISO 9 for Old Slavonic containing the glyphs я, ў, ј, є, ћ, ќ, і, ў, ї and љ taken from the scientific transliteration ([mode=iso9_ocs]). If you prefer more coherency you might want to use pure “scientific” transliteration ([mode=ocs]). This method is complemented by [mode=ocs_gla], the only option the Transliterator offers for the Glagolitic alphabet; they can be used consistently along each other as they were taken from the same book.\textsuperscript{3}

As far as I know there is no standardized transliteration for Greek so I had to resort to the one that is used in scholarly literature. Its main drawback is that it has no representation for diacritics apart from (rough) breathing, but it respects specific rules for diphthongs and vowels in initial positions ([mode=gr]). There is one alternative mode for those who prefer their γ phonetically resolved to /n/ before velars (γ, ξ, χ and ξ; [mode=gr_n]).

Concerning the hyphenation within transliterated passages the default is set to to [hyphenate=cs] (Czech) which produces reasonable results when using all, iso9_ocs or ru_cz. For stuff like the English and German transcription use their respective native hyphenation.\textsuperscript{4} However, as there is no hyphenation pattern I know of that closely resembles the transliteration of Greek you might have to resort to putting \discretionary hyphens when line breaking does not satisfy.

The Transliterator as a whole is nothing more than a bunch of dictionaries containing substitution rules for tokens that may occur in the text. These tokens may be single characters or strings of more than one character. As

\textsuperscript{2} A hasty glance at the latest issues of around 20 journals in a local library revealed that 2 of them actually are using ISO 9, these are Przegląd wschodni as of Nr. X, 3 (2008) and Kwartalnik historyczny as of CXVI, 3 (2009); the latter even contains a table on p. 218 showing a subset of the ISO 9 transliteration rules.


\textsuperscript{4} You’ll have to specify this through \setuptransliterator or locally because the default hyphenation is not the same as your documents'.
there is no simple way to impose order onto those dictionaries the rules for one transliteration method are, if needed, distributed over more than one table which will be applied successively to ensure that multi-character rules are processed first.

<table>
<thead>
<tr>
<th>mode</th>
<th>time(1) in s</th>
<th>ConTeXt</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>8.98</td>
<td>8.82</td>
</tr>
<tr>
<td>all</td>
<td>8.37</td>
<td>8.25</td>
</tr>
<tr>
<td>ru_cz</td>
<td>8.61</td>
<td>8.48</td>
</tr>
<tr>
<td>ru_transcript_en</td>
<td>9.26</td>
<td>9.10</td>
</tr>
<tr>
<td>ru_transcript_de</td>
<td>14.83</td>
<td>14.71</td>
</tr>
</tbody>
</table>

Table 1 Processing time for corpus Evgenij Onegin according to GNU time(1) and the ConTeXt stats.

Following suggestions from the mailing list, the Transliterator uses \texttt{L Peg} when substituting. This means a huge speed improvement for most substitution modes when compared to the older mechanism that used \texttt{string.gsub} iteratively. In ordinary use when transliterating single words or short phrases the Transliterator should have little impact on document processing time at large, with the exception of the German transcription mode, perhaps.\footnote{The problem lies within the rule set for the German transcription which dictates different instructions depending on the environment of a character; these may conflict, i.e. it is impossible to substitute a character stream in a single run as some rules may apply only to the result of previous rule. Let me know if there’s a way to tell \texttt{L Peg} to backtrack to the last character of a match and not to continue on the next.}

Transliterating (and typesetting in MKIV) Aleksandr Puškin’s verse novel Evgenij Onegin, a corpus of about 27000 words, in \texttt{[mode=all]} shows little to no delay at all. In fact, typesetting cyrillic letters with russian hyphenation seems slow things down so much that transliteration may be faster and uses slightly less memory.\footnote{On an \texttt{IBM T43}: 2.6.32-ARCH #1 SMP PREEMPT Tue Feb 9 14:46:08 UTC 2010 i686 Intel(R) Pentium(R) M processor 1.60GHz GenuineIntel GNU/Linux.}
EXAMPLES

Cyrillic Scripts

ISO 9 and Derivatives

Several transliteration rules are either strictly ISO 9 compliant (ru, ru_old, all) or contain ISO 9 as a subset (iso9_ocs).7

Figure 1 [mode=ru,hyphenate=cs] Transliteration rules for the contemporary Russian alphabet.

Figure 2 [mode=ru_old,hyphenate=cs] With additional characters for pre-1981 Russian orthography (100 per cent ISO 9).

Unfortunately ConTeXt still lacks language files for some of them so please excuse the inadequate hyphenation in these cases.
Беларуская мова, мова беларусаў, уваходзіць у сям'ю індаеўрапейскіх моў, яе славянскай групы і ўсходнеславянскіх моў падгрупы, на якой размаўляюць у Беларусі і па ўсім свеце, галоўным чынам у Расіі, Украіне, Польшчы. Б. м. падзяляе шмат граматычных і лексічных уласціваццяў з іншымі ўсходнеславянскімі моўамі (гл. таксама: Іншыя назвы беларускай моўы і Узаемныя ўплывы ўсходнеславянскіх моў).
Что съ дѣятѣ по вѣремьнымъ : то отидето по вѣрьмымъ : приказано будѣте добрымъ людемъ :
а любо грамотоу утвердѣть : kako to budete konem vnedalka : nan kto poslo zhivymen metalen:
са тою лѣтъ como dileraucta : kladka rizsionky
съ: удадутыхъ knyte сильнымемъ : myti-
самъ : дадѣть си: приказалъ быв ризу своего аскети-
шаго попа : веролом : и съ нимъ сѣля вожжа пан-
телья : исконаго города сильнага : та два была по-
сладь вѣрѣ : изъ ризы нѣхали на гончей берцѣ :
такъ твердити мира :

Се начнемъ повѣсть сию. По потопѣ . первиє снве
Ноєви . раздѣлиша землю . Симъ . Хамъ . Афетъ .
и юса вѣстокъ . Симови Персида . Ватръ . тоже и до
Индикѣ в долоту и в ширину [и до Нирокуріа]
jakоже рещи о вѣстока и до полуведенія . и Сурия
и Индия по Ефратъ рѣкъ . Бавилонъ . Кардунъ .
Леванъ . Месопотамиа . Армія . старыиша :
Сионъ . Индъ . Равива . на вѣл Д.

Se načnem pověst’ siju. Po potopě . pervie sне
i ясю въстокъ . Simovi Persida . Vatrъ . tožе i do
Indikija в dolgotu i в ширину [и до Nirokuria]
jakožе rešti o вѣstoka и do poluvedenjа . i Surija
Mesopotamija . Armiя . старыишиа :
Siyonъ . Indъ . Ravniva . на всє D.

Figure 6 [mode=iso9_ocs,hyphenate=cs] Transliteration rules according to iso 9 with additions for Old (Church) Slavonic.

“SCIENTIFIC” TRANSLITERATION

These transliterations are widely used among scholars, mainly linguists and, to a lesser extent, historians. They comprise large character sets in order to represent the original text adequately and facilitate comparison of texts of the same language written in different scripts; they are not, however, as easily reversible as ISO 9.

Figure 7 [mode=ocs,hyphenate=cs] Transliteration for Old Slavonic used in Slavic studies, taken from the excellent book of Birnbaum/Schaeken (1999).8

8 This one and both of the following Czech transliterations, although elegantly dealing with hard and weak signs by taking characters from the Cyrillic alphabet, are not unquestioned from a typographical point of view: “If contrasting faces are used for phonetic transcriptions and main text, each entire phonetic word or passage, not just the individual phonetic characters, should be set in the chosen phonetic face. Patchwork typography, in which the letters of a single word come from different faces and fonts, is a sign of typographic failure. [...] Such mixtures are almost sure to fail unless all the fonts involved have been designed as a single family.” (Bringhurst (2008)) From this follows that it is advisable to reconsider your font whether it indeed provides the needed glyphs from Russian as well.
Прошло семь лет после 12-го года. Возвол-нованное историческое море Европы улег-лось в свои берега. Оно казалось за- тишшим; но таинственные силы, двига-ющие человечество (таинственные пото-му, что законы, определяющие их движе-ние, неизвестны нам), продолжали свое действие. Несмотря на то, что поверх-ность исторического моря казалась непо- движною, так же непрерывно, как дви- жение времени, двигалось человечество. Слагались, разлагались различные групп-пы людских сцеплений; подготавливались причины образования и разложения го- сударств, перемещений народов.

**Figure 8** [mode=ru_cz,hyphenate=cs] *Czech phonetic transcription for contemporary Russian.*

Убьетъ мужа, то мстить брата брата, или сынови отца, любо отцю сына, или братчаду, любо сестрину сынови; аще не будеть кто мститъ, то 40 гривенъ за голову; аще будеть русинъ, любо гривинъ, любо купчина, любо и тебник, любо мечникъ, любо игои будеть, любо словенинъ, то 40 гривенъ положити за нь.

**Figure 9** [mode=ocs_cz,hyphenate=cs] *Czech phonetic transcription for Old Slavonic (superset of the corresponding Russian transcription).*

**SERBIAN**

The tables for converting Serbian text between Cyrillic and Latin alphabets are sr_tolt and sr_tocy.

Српски језик је један од словенских језика из породице индоевропских језика. Први писани споменици у српској редакцији старословенског језика потичу из XI и XII века.

Српски језик је стандардни језик у службеној употреби у Србији, Босни и Херцеговини и Црној Гори, а у употреби је и у другим земљама где живе Срби, међу осталима и у Хрватској.

**Figure 10** [mode=sr_tolt,hyphenate=hr] *Transliteration ћирилица → латиница.*
Srpski jezik je jedan od slovenskih jezika iz porodice indoevropskih jezika. Prvi pisan spomenici u srpskoj redakciji staroslovenskog jezika potiču iz XI i XII veka. Srpski jezik je standardni jezik u službenoj upotrebi u Srbiji, Bosni i Hercegovini i Crnoj Gori, a u upotrebi je i u drugim zemljama gde žive Srbi, među ostalima i u Hrvatskoj.

Srpski jezik je jedan od slovenskih jezika iz porodice indoevropskih jezika. Prvi pisan spomenici u srpskoj redakciji staroslovenskog jezika potiču iz XI i XII veka. Srpski jezik je standardni jezik u službenoj upotrebi u Srbiji, Bosni i Hercegovini i Crnoj Gori, a u upotrebi je i u drugim zemljama gde žive Srbi, među ostalima i u Hrvatskoj.

**Figure 11** [mode=sr_tocy, hyphenate=sr] *Transliteration latinica → cirilica.*

**BULGARIAN**

Българският език е индоевропейски език от групата на южнославянските езици. Той е официалният език на Република България и един от 23-те официални езици на Европейския съюз.

Българският език е индоевропейски език от групата на южнославянските езици. Той е официалният език на Република България и един от 23-те официални езици на Европейския съюз.

**Figure 12** [mode=bg_de, hyphenate=cs] *German scientific transliteration for Bulgarian (based on old iso 9 standard).*

**LEGACY NATIONAL TRANSCRIPTIONS**

At the moment there are tables for “old school” transcription into three languages: English (via ru_transcript_en), German (ru_transcript_de) and Czech (ocs_cz). At least the German one is almost unreadable if used with strings longer than two words. As we have the bijective ISO 9 mapping at hand there should be no reason at all to use any of them.
Прошло семь лет после 12-го года. Взволнованное историческое море Европы улеглось в свои берега. Оно казалось затихшим; но таинственные силы, двигающие человечество (таинственные потому, что законы, определяющие их движение, неизвестны нам), продолжали свое действие. Несмотря на то, что поверхность исторического моря казалась неподвижною, так же непрерывно, как движение времени, двигалось человечество. Слагались, разлагались различные группы людских сцеплений; готовились причины образования и разложения государств, перемещений народов.

Figure 13 [mode=ru_transcript_en,hyphenate=en] English transcription for contemporary Russian.

Прошло семь лет после 12-го года. Взволнованное историческое море Европы улеглось в свои берега. Оно казалось затихшим; но таинственные силы, двигающие человечество (таинственные потому, что законы, определяющие их движение, неизвестны нам), продолжали свое действие. Несмотря на то, что поверхность исторического моря казалась неподвижной, так же непрерывно, как движение времени, двигалось человечество. Слагались, разлагались различные группы людских сцеплений; готовились причины образования и разложения государств, перемещений народов.

Figure 14 [mode=ru_transcript_de,hyphenate=deo] German transcription for contemporary Russian.
GLAGOLITIC

[vl] advěko čk’ si xoštet’ raz[orit] i za-
kom manastirski: [iże] ti ustaviRecě že
igum[̑x] [m] kako xoštet’ razoriti zak[omu]
[.] [a] estu- črii i [m] [............] [.] m[ţ] ne
możem’ sego tr pä[v] [da] lubo sego im[ė]
sudė a ma o [ť]demč: lubo sego pasti: da
ot [id]et’ iże est’ prištiv: s[e]

Figure 15 [mode=ocs_gla,hyphenate=cs] “Scientific” transliteration for Old Slavonic written in the Glagolitic alphabet as used in Birn-

GREEK

The Transliterator offers two modes for handling Greek: gr and gr_n. They
differ only on one aspect. gr transliterates the canonical Greek alphabet as
well as the special glyphs Digamma, Quoppa and Sampi. gr_n behaves exactly
the same way except that nasalization is observed such that γ[γ̃κ] yields
n+[g̃k].

οἰνῷ δὲ κάρτα προσκέαται, καὶ σφι οὐκ ἐμέ-
σαι ἔξεστι, οὐκ οὔρησαι ἀντίον ἀλλο. ταύτα
μὲν νυν οὖσα φιλάσσεται, μεθυσκόμενοι δὲ
ἐκδίδουσι ποιμανθά τά σπουδαίστατά τών
γρηγόταν: τό δ’ ἂν καὶ οὔρησαι, χρέονται
τά σπουδαίστατά, καὶ ἂν μὴν καὶ νήφος
χρέονται: τά δ’ ἂν νήφος προβολεύονται,
μεθυσκόμενοι ἐπιδιαγινώσκουσι.

Figure 16 [mode=gr,hyphenate=de] Transliteration for Greek – stan-
dard.

9 Following Drosdowski/Müller/Scholze-Stubenrecht/Wermke (1952) p. 82; all the canonical rules
are implemented save one: -eso and -oso should resolve to -owo respectively if genitive
endings. As this is a grammatical rather than graphetical criterion writing a substitution algorithm
would amount to do natural language parsing. To make things worse this rule is phonetically
confused as it would not take care of other contexts where  ě in those patterns is articulated as
/v/ like for instance in cesodha (which is a historical genitive, though ...). So even if this could be
implemented it would not be advisable to use such a rule.
ταῦτα καὶ νεωτέρῳ καὶ πρεσβυτέρῳ ὅτῳ ἂν ἐν ἐντυγχάνω ποιήσω, καὶ ξένῳ καὶ ἀστῷ, μᾶλλον δὲ τοῖς ἀστοῖς, ὅσῳ μου ἐγγυτέρῳ ἐστέ γένει.

Figure 17  [mode=gr_n,hyphenate=de] Transliteration for Greek – alternative respecting nasalization.
REFERENCES