

The

TRANSLITERATOR

for ConT_EXt

MANUAL

Contents

1	Usage and Functionality	2
1.1	Overview	2
1.2	Module loading and configuration	2
2	Introduction	4
3	Examples	7
3.1	Cyrillic scripts	7
3.1.1	ISO 9 and derivatives	7
3.1.2	“Scientific” transliteration	9
3.1.3	Legacy national transcriptions	10
3.2	Glagolitic	12
3.3	Greek	12
4	References	13

1 Usage and Functionality

1.1 Overview

The Transliterator provides two commands: `\setupTranslit` 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.

```
\setupTranslit [...,*,...]
```

```
*   mode          = ru_old ru ru_transcript_de ru_transcript_en all iso9_ocs ocs  
      hyphenate    = cs sk  
                        ocs_gla ru_cz ocs_cz gr gr_n
```

```
\transliterate [...,*,...]
```

OPTIONAL

```
*   inherits from \setupTranslit
```

1.2 Module loading and configuration

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 `\setupTranslit[#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:

mode	description
ru	ISO 9 Russian
ru_old	ISO 9 Russian plus pre-1918 chars (the default)
all	ISO 9 complete
ru_transcript_de	German transcription for Russian
ru_transcript_en	English transcription for Russian
iso9_ocs	== all plus non-ISO additions for Old (Church) Slavonic
ocs	“scientific” transliteration for Old (Church) Slavonic
ocs_gla	“scientific” transliteration for Old (Church) Slavonic / Glagolitic alphabet
ru_cz	Czech transcription for Russian
ocs_cz	Czech transcription for Old (Church) Slavonic
gr	transliteration for Greek
gr_n	transliteration for Greek obeying nasalizations
mode	description

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

Through the **hyphenate** argument it is possible to adjust the language that is used for hyphenation. Specifying `\setupTranslit[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.

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.

2 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 T_EX 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 and 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_de]`).

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

¹ ISO (1995).

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.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 ѡ, ѕ, ꙗ, ѣ, ѿ, ѹ, Ѻ, ѻ, Ѽ, ѽ, Ѿ, ѿ, ѿ, ѿ, ѿ and ѿ 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.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 , κ , χ and ξ ; `[mode=gr_n]`).

Concerning the hyphenation within transliterated passages the default is set 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.⁴ However, as there is no hyphenation pattern I know of that closely resembles the transliteration of

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.

³ Birnbaum/Schaeken (1999) p. 77 [<http://www.schaeken.nl/lu/research/online/publications/akslstud/index.htm>].

⁴ You'll have to specify this through `\setupTranslit` or locally because the default hyphenation is not the same as your documents'.

Greek you might have to resort to putting `\discretionary` hyphens when line breaking does not satisfy.

To conclude this, let me have a word on the way the Transliterator works. Basically, it is a bunch of dictionaries containing substitution rules for elements that may occur in the text. These elements may be single characters or strings of more than one character. As 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.

	mode GNU ConT_EXt	
<code><none></code>	8.89	8.75
<code>ru</code>	9.66	9.45
<code>all</code>	12.18	12.04
<code>ru_cz</code>	10.17	10.02
<code>ru_transcript_en</code>	11.24	11.11
<code>ru_transcript_de</code>	56.58	56.42

Table 2.1 Processing speed according to GNU time(1) and the ConT_EXt stats.

The transliteration itself is, admittedly, extremely inefficient as it uses global substitution iteratively on the whole string for every rule in the dictionary. (Maybe this could be replaced by a faster implementation using look ahead that goes through the string only once, but for now it'll stay as it is until I find time to care for speed.) In ordinary use when transliterating single words or short phrases only the Transliterator should have little impact on document processing time at large, with

the exception of the German transcription mode, perhaps. For sake of completeness, here are some numbers: Transliterating (and typesetting in MKIV) Aleksandr Puškin's verse novel Evgenij Onegin, a corpus of about 27000 words, took only 9.7 seconds in `[mode=ru]`, compared to 8.9 seconds without transliteration.⁵

⁵ On an 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.

3 Examples

3.1 Cyrillic scripts

3.1.1 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`).⁶

В ворота гостиницы губернского города NN въехала довольно красивая рессорная небольшая бричка, в какой ездят холостяки: отставные подполковники, штабс-капитаны, помещики, имеющие около сотни душ крестьян, — словом, все те, которых называют господами средней руки. В бричке сидел господин, не красавец, но и не дурной наружности, ни слишком толст, ни слишком тонок; нельзя сказать, чтобы стар, однако ж и не так чтобы слишком молод.

V vorota gostinicy gubernskogo goroda NN v"ehala dovol'no krasivaâ ressonnaâ nebol'shaâ brička, v kakoj ezďat holostâki: otstavnye podpolkovniki, štabs-kapitany, po-mešiki, imeûšie okolo sotni duš krest'ân, — slovom, vse te, kotoryh nazyvaûť gospodami srednej ruki. V bričke sidel gospodin, ne krasavec, no i ne durnoj naružnosti, ni sliškom tolst, ni sliškom tonok; nel'zâ skazat', čtoby star, odnako Ź i ne tak čtoby sliškom molod.

Figure 3.1 [`mode=ru,hyphenate=cs`] *Transliteration rules for the contemporary russian alphabet.*

⁶ Unfortunately there are not yet any language files for some of them so please excuse the inadequate hyphenation in these cases.

А сведетса віра, убьють сотцкого в сель, ино тебѣ взяти полтина, а не сотцкого, ино четырь гривны, а намъ вірь не тaitи в Новѣгородѣ; а о убиствѣ вірь нѣтъ. А что волости, честны король, новгородцкіе, ино тебѣ не держати своими мужи, а держати мужми новгородцкими. А что пошлина в Торжку и на Волоцѣ, тівунъ свои держати на свои чѣсти, а Новугороду на свои чѣсти посадника держаті. А се волости новгородцкіе: Волокъ со всѣми волостми, Торжокъ, Бѣжіцѣ, Городецъ Палець, Шіпинъ, Мелеця, Егна, Заволочье, Тиръ, Пермь, Печера, Югра, Вологда с волостми.

A svedetsâ vira, ub'ût" sotckogo v selë, ino tebě vzâti poltina, a ne sotckogo, ino četyrë grivny, a nam" vîr" ne taiti v Novëgorodë; a o ubistvë vîr" nët". A čto volosti, čestny korol', novgorodckie, ino tebě ne deržati svoimi mužî, a deržati mužmi novgorodckimi. A čto pošlina v Toržku i na Volocë, tivun" svoi deržati na svoei časti, a Novugorodu na svoei časti posadnika deržati. A se volosti novgorodckie: Volok" so vsëmi volostmi, Toržok", Bëžici, Gorodec' Palec', Šipin", Melecâ, Egna, Zavoloč'e, Tir", Perm', Pečera, Ŭgra, Vologda s volostmi.

Figure 3.2 [mode=ru_old,hyphenate=cs] *With additional characters for pre-1981 Russian orthography (100 per cent ISO 9).*

Беларуская мова, мова беларусаў, уваходзіць у сям'ю індаеўрапейскіх моў, яе славянскай групы і ўсходнеславянскіх моваў падгрупы, на якой размаўляюць у Беларусі і на ўсім свеце, галоўным чынам у Расіі, Украіне, Польшчы. Б. м. падзяляе шмат граматычных і лексічных уласцівасцяў з іншымі ўсходнеславянскімі мовамі (гл. тaksama: Іншыя назвы беларускай мовы і Узаемныя ўплывы усходнеславянскіх моваў).

Belaruskaâ mova, mova belarusai, uvahodzic' u sâmû indaeûrapejskikh moû, âe slavânskaï grupy i ûshodneslavânskikh movai padgroupy, na âkoï razmaûlâiûc' u Belarusi i pa ûšim svece, galoûnym čynam u Rasiï, Ukraïne, Pol'sčy. B. m. padžâlâe šmat gramatyčnyh i leksičnyh ulascivascâû z inšymi ûshodneslavânskimi movami (gl. taksama: Înšyâ nazvy belaruskaj movy i Uzaemnyâ ûplyvy ushodneslavânskikh movai).

Figure 3.3 [mode=all,hyphenate=cs] *The complete cyrillic mapping from ISO 9; transliterating Belarusian.*

Українська мова (застарілі назви – руська мова, проста мова [...]) – слов'янська мова, державна в Україні та одна з трьох «офіційних мов на рівних засадах» у невизнаній Придністровській Молдавській Республіці. За різними оцінками загалом у світі українською мовою говорить від 41 млн. до 45 млн. осіб, вона входить до третього десятка найпоширеніших мов світу.

Ukraïns'ka mova (zastarilî nazvi – rus'ka mova, prosta mova [...]) – slov'âns'ka mova, deržavna v Ukraïni ta odna z tr'oh «oficiïjnih mov na rîvnih zasadah» u neviznaniï Pridnistrov's'kij Moldav's'kij Respublîci. Za rîznimi ocînkami zagalom u svîti ukraïns'koïu movoû govorit' vid 41 mln. do 45 mln. osîb, вона vhodit' do tret'ogo desâtka najpoširenîših mov svîtu.

Figure 3.4 [mode=all,hyphenate=cs] *The complete cyrillic mapping from ISO 9; transliterating Ukrainian.*

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

Srpski jezik je jedan od slovenskih jezika iz porodice indoevropskih jezika. Prvi pisani 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 gdje žive Srbi, među ostalima i u Hrvatskoj.

Figure 3.5 [mode=all,hyphenate=cs] *The complete cyrillic mapping from ISO 9; transliterating Serbian.*

Что сѧ дѣѣтъ по вѣремънемъ : то ѡидѣто по вѣр-
мънемъ : приказано будѣте добрымъ людѣмъ :
а любо грамотою оутвѣрдатъ : како то будѣте
всемъ вѣдомъ : или кто послѧ живыи ѡстанѣтъ
сѧ : того лѣтъ коли алѣбрахътъ : влѣка ризкии
оумѣраъ : оуздоумалъ князѣ смольнескыи : мѣсти-
салавъ : двѣдѣ снѣ : прислалъ въ ригу сѡего лѡчѣ-
шего попа : ѣрмеѧ : и сѧ нимъ оумѣна мужа пан-
тельѧ : искоеного горда смольнеска : та два была по-
салавъ оу ризѣ : из ригы ѣхали на гоцкыи верѣго :
тѧмо твѣрдити миръ :

Čto se dějetě po věrem'nem' : to otideto po
věr'm'nem' : prikazano budete dobrým' lū-
dēm'' : a lūbo gramotoū utvĕrdet' : kako to
budete všem' vědom'' : ili kto posl' živyi ōs-
tanĕt' se : togo lĕť koli al'braht'' : vldka rizkii
um'rl'' : uzdumal'' knĕzĕ smol'neskyi : m'sti-
slav'' : dvĕdv'' sn'' : prislal'' v'' rigu svojego
luč'sĕgo popa : ĵer'meā : i s'' nim' um'na muža
pantel'ā : isvojego gorda smol'neska : ta dva
byla posl'm' u rizĕ : iz rigy jĕhali na gočkyi
ber'go : tamo tverditĭ mir'' :

Figure 3.6 [mode=iso9_ocs,hyphenate=cs] *Transliteration rules according to ISO 9 with additions for Old (Church) Slavonic.*

3.1.2 “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.

Се научемъ повѣсть сѧю. По потопѣ . первие снѣ-
гѡви . раздѣлиша землю . Сѧмъ . Хамъ . Афетъ .
и ѧса вѣстокъ . Сѧмови Персида . Батръ . тоже и до
Индѧкиѧ в долготу и в ширину [и до Нѧрокуриѧ]
ѧкоже реши ѡ вѣстока и до полудѧныѧ . и Сурѧна
и Индиѧ по Ефратъ рѣку . Вавилонъ . Кордунъ
Асурѧне . Мисопотѧмира . Арѧвѧ . старѣишѧ .
Елѧмѧиъ . Инди . Рѧвѧна . на всѧ Д.

Se načnem' povĕst' siju. Po potopĕ . pervie
snve Noevi . razdĕliša zemlju . Sim' . Ham' .
Afet' . i jasĕ v'stok' . Simovi Persida .
Vatr' . toĵe i do Indikiĵa v dolgotu i v širinu
[i do Nirokuria] jakoĵe reši ot v'stoka i do
poludĕn'ĵa . i Suriĵa . i India po Efrat' rĕku
Vavilon' . Korduna . Asurene . Misopot-
amira . Aravija . starĕišaja . Elmais' . Indi .
Ravija . na vsĕ D.

Figure 3.7 [mode=ocs,hyphenate=cs] *Transliteration for Old Slavonic used in Slavic studies, taken from the excellent book of Birnbaum/Schaeken (1999).*

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; подготавливались причины образования и разложения государств, перемещений народов.

Prošlo sedm let posle 12-go goda. Vzvolnovanное istoričeskoe more Evropy ulegloсь v svoi berega. Ono kazaloсь zatičšim; no tainstvennye sily, dvigajuščie čelovečstvo (tainstvennye potomu, čto zakony, opredeljajuščie ich dviženie, neizvestny nam), prodolžali svoe dejstvie. Nesmotřja na to, čto poverchnostь istoričeskogo morja kazalaсь nepodvižnoju, tak že nepreryvno, kak dviženie vremeni, dvigaloсь čelovečstvo. Slagaliсь, razlagaliсь različnye grupy ljudskich sceplenij; podgotovljaliсь pričiny obrazovanija i razloženija gosudarstv, peremeščeniј narodov.

Figure 3.8 [mode=ru_cz,hyphenate=cs] *Czech phonetic transcription for contemporary Russian.*

Убѣтъ мужъ мужа, то мѣстити брату брата, или сыновѣ отца, лѣбо отцю сына, или братѹчаду, лѣбо сестринѹ сыновѣ; аще не бѹдетъ кто мѣстити, то 40 гривенѹ за голову; аще бѹдетъ русинѹ, лѣбо gridинѹ, лѣбо купчина, лѣбо ѡвѣтник, лѣбо мечникѹ, аще изъгон бѹдетъ, лѣбо словенинѹ, то 40 гривенѹ положити за нь.

Uběť muž muža, to městiti bratu brata, ili synovi otca, ljuo otcju syna, ili bratučadu, ljuo sestřinu synovi; ašče ne budeť kto městie, to 40 grivenų za głovu; ašče budeť rusinų, ljuo gridinų, ljuo kupčina, ljuo iěbetnik, ljuo mečnik, ašče izъgoi budeť, ljuo sloveninų, to 40 grivenų položiti za нь.

Figure 3.9 [mode=ocs_cz,hyphenate=cs] *Czech phonetic transcription for Old Slavonic (superset of the corresponding Russian transcription).*

3.1.3 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 unless when physically threatened by barbarians.

Прошло семь лет после 12-го года. Взволнованное историческое море Европы улеглось в свои берега. Оно казалось затихим; 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.

Proshlo sem let posle 12-go goda. Vzvolnovannoye istoricheskoye more Evropy uleglos v svoi berega. Ono kazalos zatikhshim; no tainstvennyye sily, dvigayushshiye chelovechestvo (tainstvennyye potomu, chto zakony, opredelyayushshiye ikh dvizheniye, neizvestny nam), prodolzhal svoye deystviye. Nesmotrya na to, chto poverkhnost istoricheskogo morya kazalas nepodvizhnoyu, tak zhe nepreryvno, kak dvizheniye vremeni, dvigalos chelovechestvo. Slagalis, razlagalis razlichnyye gruppy lyudskikh stsepleni; podgotovlyalis prichiny obrazovaniya i razlozheniya gosudarstv, peremeshheniy narodov.

Figure 3.10 [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 затихим; 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.

Proschlo sem let posle 12-go goda. Wswolnowannoje istoritscheskoje more Jewropy uleglos w swoi berega. Ono kasalos satichschim; no tainstwennyje sily, dwigajuschtschije tschelowetschestwo (tainstwennyje potomu, tschto sakony, opredelajajuschtschije ich dwischenije, neiswestny nam), prodolschali swoje deistwiye. Nesmotrja na to, tschto powerchnost istoritscheskogo morja kasalas nepodwischnoju, tak sche neprerywno, kak dwischenije vremeni, dwigalos tschelowetschestwo. Slagalis, raslagalis raslitschnyje gruppy ljudskikh szepteni; podgotowljalis pritschiny obrasowanija i rasloschenija gossudarstw, peremeschtscheni narodow.

Figure 3.11 [mode=ru_transcript_de,hyphenate=deo] *German transcription for contemporary Russian.*⁷

4 References

- 1 Drosdowski, Müller, Schulze-Stubenrecht and Wermke, editors (1991). *DUDEN Rechtschreibung der deutschen Sprache*. 20. edition Mannheim et al.
- 2 Bornemann and Risch (1978). *Griechische Grammatik*. 2. edition Frankfurt am Main.
- 3 Kirschbaum (2001). *Grammatik der russischen Sprache*. Berlin.
- 4 Birnbaum, H. and Schaeken, J. (1999). *Altkirchenslavische Studien*, volume 2. München.
- 5 DIN Deutsches Institut für Normung e. V., editor (2001). *Bibliotheks und Dokumentationswesen*. Berlin/Wien/Zürich.
- 6 ISO International Organization for Standardization, editor (1995). *Information and documentation – Transliteration of Cyrillic characters into Latin characters – Slavic and non-Slavic languages* 2. edition..