UNICODE

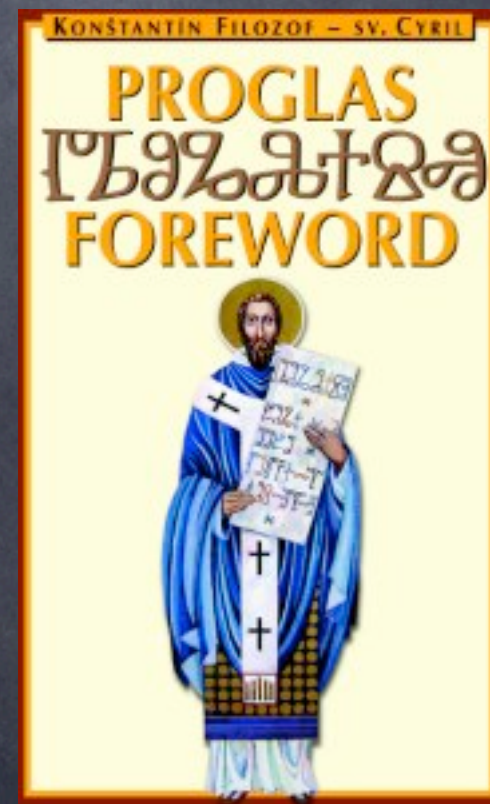## FONTS • PROBLEMS • SOLUTIONS

SEBASTIAN KEMPGEN
UNIVERSITY OF BAMBERG, GERMANY

Overview

- Introduction: Unicode Basics

- Do´s and Don´ts

- Problems and Solutions

1. Introduction

What is Unicode?

Unicode – a **system** to encode all the world's language (their characters)

Unicode – an **organisation** which takes care of the process

→

1. Introduction

Unicode - the encoding system

- ASCII: 2^7 characters = 128
- 'Code Pages': 2^8 characters = 256
- **Unicode**: 2^{16} characters = 65.536
- Versions: v. 1.0 1991, v. 4.1 2005
- Version 4.1: Glagolica ✓
- Version 5.1: Cyrillic Additions ✓

→

1. Introduction

Benefits of using Unicode

- Vendor independence
- Open standard
- Compatibility between platforms
- Compatibility between fonts
- Long-term readability of texts
- ...

✓

→

1. Introduction

Problems in using Unicode

- No solution yet for all problems
- A standard under development
- Not an ad-hoc solution
- ...

Unicode

Let's Have a Look...

→

UN **Unicode Home Page**

"Without Unicode, Java wouldn't be Java..." — James Gosling (more)

New to Unicode?
- What is Unicode?
- How to Use this Site
- FAQ
- Glossary of Unicode Terms

General Information
- Where is my Character?
- Display Problems?
- Useful Resources
- Unicode Enabled Products
- Mail Lists
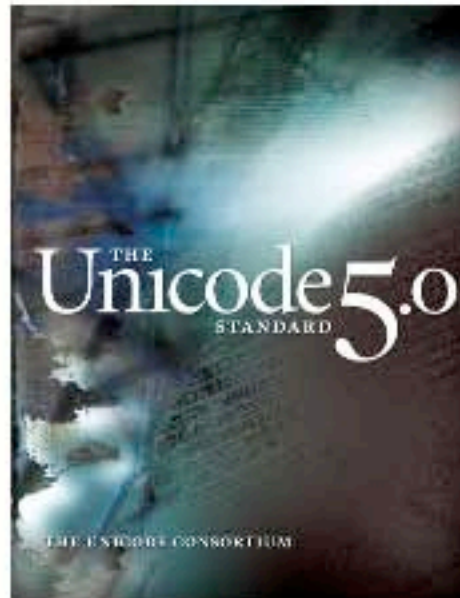- Subset of the Unicode website in French

The Consortium
- Who we are
- Our Members
- How to Join
- Conferences
- Job Postings (2007.10.29)
- Press Info
- Policies & Positions
- Contact Us

For Members Only
- Member Resources
- Working Documents

THE Unicode 5.0 STANDARD

THE UNICODE CONSORTIUM

Buy the Book

Some of Our Members

 digicert®

Announcements

Unicode CLDR Committee releases Unicode Transliteration Guidelines (2008.01.19)

Unibook character browser update for Version 5.1.0 Beta now available (2008.01.16)

Unicode announces new membership level for students (2007.12.18)

Public Review

The Unicode Standard
- Start Here
- Latest Version
- Code Charts
- Unicode Character Database
- Unihan Database

Key Specifications
- Unicode Collation (UCA)
- Bidirectional Algorithm (Bidi)
- Normalization (NFC, NFD, ...)
- Locale Data (CLDR)
- Script Codes (ISO 15924)
- Security Considerations

Technical Publications
- Technical Reports & Standards
- Technical Notes
- Online Data Tables
- The Unicode Guide
- Updates & Errata

Work in Progress
- Calendar of Meetings
- Proposals for Public Review
- Unicode Technical Committee
- UTC Meeting Minutes
- Proposed Characters
- Submitting Proposals
- CLDR Technical Committee
- UDHR in Unicode

Unicode – the organisation

UC 5.0 Cyrillic

	040	041	042	043	044	045	046	047	048	049	04A	04B	04C	04D	04E	04F
0	Ѐ	А	Р	а	р		Ѡ	Ѱ	Ҁ	Г	Қ	Ұ	І	Ӑ	Ӡ	Ӿ
1	Ё	Б	С	б	с		ѡ	ѱ	ҁ	г	қ	ұ	Ж	ӑ	ӡ	ӿ
2	Ђ	В	Т	в	т		Ѣ	Ѳ	҂	Ғ	Ң	Х	ж	Ӓ	Ӣ	
3	Ѓ	Г	У	г	у		ѣ	ѳ	҃	ғ	ң	х	Қ	ӓ	ӣ	
4	Є	Д	Ф	д	ф		Ѥ	Ѵ	҄	Ҕ	Ҥ	Ц	ҝ	Æ	Ӥ	Ҵ
5	Ѕ	Е	Х	е	х		ѥ	ѵ	҅	ҕ	ҥ	ц	Л	æ	ӥ	ҵ
6	І	Ж	Ц	ж	ц		Ѧ	Ѷ	҆	Ж	Ҧ	Ч	л	Ӗ	Ӧ	Г
7	Ї	З	Ч	з	ч		ѧ	ѷ		ж	ҧ	ч	Ӈ	ӗ	ӧ	г
8	Ј	И	Ш	и	ш		Ѩ	Оу		Ҙ	Ҩ	Ҷ	ӈ	Ә	Ө	Ӹ
9	Љ	Й	Щ	й	щ		ѩ	оу		ҙ	ҩ	ҷ	Ӊ	ә	ө	ӹ
A	Њ	К	Ъ	к	ъ		Ѫ	Ѻ	Й	Қ	Ҫ	Һ	ӊ	Ӛ	Ӫ	Ӻ
B	Ћ	Л	Ы	л	ы		ѫ	ѻ	й	қ	ҫ	һ	Ӌ	ӛ	ӫ	ӻ
C	Ќ	М	Ь	м	ь		Ѭ	Ѽ	Ҍ	Қ	Ҭ	Ҽ	ӌ	Ӝ	Ӭ	Х
D	Ѝ	Н	Э	н	э		ѭ	ѽ	ҍ	қ	ҭ	ҽ	Ӎ	ӝ	ӭ	х
E	Ў	О	Ю	о	ю		Ѯ	Ѿ	Ҏ	Ҟ	Ү	Ҿ	ӎ	Ӟ	Ӯ	Х
F	Џ	П	Я	п	я		ѯ	ѿ	ҏ	ҟ	ү	ҿ	І	ӟ	ӯ	ӿ

Cyrillic extensions

- 0400 Ѐ CYRILLIC CAPITAL LETTER IE WITH GRAVE
  - ≡ 0415 Е 0300 ◌̀
- 0401 Ё CYRILLIC CAPITAL LETTER IO
  - ≡ 0415 Е 0308 ◌̈
- 0402 Ђ CYRILLIC CAPITAL LETTER DJE
- 0403 Ѓ CYRILLIC CAPITAL LETTER GJE
  - ≡ 0413 Г 0301 ◌́
- 0404 Є CYRILLIC CAPITAL LETTER UKRAINIAN IE
- 0405 Ѕ CYRILLIC CAPITAL LETTER DZE
- 0406 І CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I
  - → 006I i latin small letter i
  - → 0456 і cyrillic small letter byelorussian-ukrainian i
  - → 04CF ӏ cyrillic small letter palochka
- 0407 Ї CYRILLIC CAPITAL LETTER YI
  - ≡ 0406 І 0308 ◌̈
- 0408 Ј CYRILLIC CAPITAL LETTER JE
- 0409 Љ CYRILLIC CAPITAL LETTER LJE
- 040A Њ CYRILLIC CAPITAL LETTER NJE
- 040B Ћ CYRILLIC CAPITAL LETTER TSHE
- 040C Ќ CYRILLIC CAPITAL LETTER KJE
  - ≡ 041A К 0301 ◌́
- 040D Ѝ CYRILLIC CAPITAL LETTER I WITH GRAVE
  - ≡ 0418 И 0300 ◌̀
- 040E Ў CYRILLIC CAPITAL LETTER SHORT U
  - ≡ 0423 У 0306 ◌̆
- 040F Џ CYRILLIC CAPITAL LETTER DZHE

Basic Russian alphabet

- 0410 А CYRILLIC CAPITAL LETTER A
- 0411 Б CYRILLIC CAPITAL LETTER BE
  - → 0183 ƃ latin small letter b with topbar
- 0412 В CYRILLIC CAPITAL LETTER VE
- 0413 Г CYRILLIC CAPITAL LETTER GHE
- 0414 Д CYRILLIC CAPITAL LETTER DE
- 0415 Е CYRILLIC CAPITAL LETTER IE
- 0416 Ж CYRILLIC CAPITAL LETTER ZHE
- 0417 З CYRILLIC CAPITAL LETTER ZE
- 0418 И CYRILLIC CAPITAL LETTER I
- 0419 Й CYRILLIC CAPITAL LETTER SHORT I
  - ≡ 0418 И 0306 ◌̆
- 041A К CYRILLIC CAPITAL LETTER KA
- 041B Л CYRILLIC CAPITAL LETTER EL
- 041C М CYRILLIC CAPITAL LETTER EM
- 041D Н CYRILLIC CAPITAL LETTER EN
- 041E О CYRILLIC CAPITAL LETTER O
- 041F П CYRILLIC CAPITAL LETTER PE
- 0420 Р CYRILLIC CAPITAL LETTER ER
- 0421 С CYRILLIC CAPITAL LETTER ES
- 0422 Т CYRILLIC CAPITAL LETTER TE
- 0423 У CYRILLIC CAPITAL LETTER U
  - → 0478 Ѹ cyrillic capital letter uk
  - → 04AF ү cyrillic small letter straight u
- 0424 Ф CYRILLIC CAPITAL LETTER EF
- 0425 Х CYRILLIC CAPITAL LETTER HA
- 0426 Ц CYRILLIC CAPITAL LETTER TSE
- 0427 Ч CYRILLIC CAPITAL LETTER CHE
- 0428 Ш CYRILLIC CAPITAL LETTER SHA
- 0429 Щ CYRILLIC CAPITAL LETTER SHCHA
- 042A Ъ CYRILLIC CAPITAL LETTER HARD SIGN
- 042B Ы CYRILLIC CAPITAL LETTER YERU
- 042C Ь CYRILLIC CAPITAL LETTER SOFT SIGN
- 042D Э CYRILLIC CAPITAL LETTER E
- 042E Ю CYRILLIC CAPITAL LETTER YU

- 042F Я CYRILLIC CAPITAL LETTER YA
- 0430 а CYRILLIC SMALL LETTER A
- 0431 б CYRILLIC SMALL LETTER BE
- 0432 в CYRILLIC SMALL LETTER VE
- 0433 г CYRILLIC SMALL LETTER GHE
- 0434 д CYRILLIC SMALL LETTER DE
- 0435 е CYRILLIC SMALL LETTER IE
- 0436 ж CYRILLIC SMALL LETTER ZHE
- 0437 з CYRILLIC SMALL LETTER ZE
- 0438 и CYRILLIC SMALL LETTER I
- 0439 й CYRILLIC SMALL LETTER SHORT I
- 043A к CYRILLIC SMALL LETTER KA
- 043B л CYRILLIC SMALL LETTER EL
- 043C м CYRILLIC SMALL LETTER EM
- 043D н CYRILLIC SMALL LETTER EN
- 043E о CYRILLIC SMALL LETTER O
- 043F п CYRILLIC SMALL LETTER PE
- 0440 р CYRILLIC SMALL LETTER ER
- 0441 с CYRILLIC SMALL LETTER ES
- 0442 т CYRILLIC SMALL LETTER TE
- 0443 у CYRILLIC SMALL LETTER U
- 0444 ф CYRILLIC SMALL LETTER EF
- 0445 х CYRILLIC SMALL LETTER HA
- 0446 ц CYRILLIC SMALL LETTER TSE
- 0447 ч CYRILLIC SMALL LETTER CHE
- 0448 ш CYRILLIC SMALL LETTER SHA
- 0449 щ CYRILLIC SMALL LETTER SHCHA
- 044A ъ CYRILLIC SMALL LETTER HARD SIGN
- 044B ы CYRILLIC SMALL LETTER YERU
- 044C ь CYRILLIC SMALL LETTER SOFT SIGN
  - → 0185 ƅ latin small letter tone six
- 044D э CYRILLIC SMALL LETTER E
- 044E ю CYRILLIC SMALL LETTER YU
- 044F я CYRILLIC SMALL LETTER YA

Cyrillic extensions

- 0450 ѐ CYRILLIC SMALL LETTER IE WITH GRAVE
  - • Macedonian
  - ≡ 0435 е 0300 ◌̀
- 0451 ё CYRILLIC SMALL LETTER IO
  - • Russian, ...
  - ≡ 0435 е 0308 ◌̈
- 0452 ђ CYRILLIC SMALL LETTER DJE
  - • Serbian
  - → 0111 đ latin small letter d with stroke
- 0453 ѓ CYRILLIC SMALL LETTER GJE
  - • Macedonian
  - ≡ 0433 г 0301 ◌́
- 0454 є CYRILLIC SMALL LETTER UKRAINIAN IE
  - = Old Cyrillic yest
- 0455 ѕ CYRILLIC SMALL LETTER DZE
  - = Old Cyrillic zelo
  - • Macedonian
- 0456 і CYRILLIC SMALL LETTER BYELORUSSIAN-UKRAINIAN I
  - = Old Cyrillic i
- 0457 ї CYRILLIC SMALL LETTER YI
  - • Ukrainian
  - ≡ 0456 і 0308 ◌̈
- 0458 ј CYRILLIC SMALL LETTER JE
  - • Serbian, Azerbaijani, Altay
- 0459 љ CYRILLIC SMALL LETTER LJE
  - • Serbian, Macedonian
  - → 01C9 lj latin small letter lj

Left page (045A–049C)

045A	њ	CYRILLIC SMALL LETTER NJE
		• Serbian, Macedonian
		→ 01CC nj latin small letter nj
045B	ћ	CYRILLIC SMALL LETTER TSHE
		= Old Cyrillic derv
		• Serbian
		→ 0107 ć latin small letter c with acute
		→ 0127 ħ latin small letter h with stroke
		→ 040B Ћ cyrillic capital letter tshe
		→ 210F ℏ planck constant over two pi
045C	ќ	CYRILLIC SMALL LETTER KJE
		• Macedonian
		≡ 043A к 0301 ́
045D	ѝ	CYRILLIC SMALL LETTER I WITH GRAVE
		• Macedonian, Bulgarian
		≡ 0438 и 0300 ̀
045E	ў	CYRILLIC SMALL LETTER SHORT U
		• Byelorussian, Uzbek
		≡ 0443 у 0306 ̆
045F	џ	CYRILLIC SMALL LETTER DZHE
		• Serbian, Macedonian, Abkhasian

Historic letters

0460	Ѡ	CYRILLIC CAPITAL LETTER OMEGA
0461	ѡ	CYRILLIC SMALL LETTER OMEGA
0462	Ѣ	CYRILLIC CAPITAL LETTER YAT
0463	ѣ	CYRILLIC SMALL LETTER YAT
0464	Ѥ	CYRILLIC CAPITAL LETTER IOTIFIED E
0465	ѥ	CYRILLIC SMALL LETTER IOTIFIED E
0466	Ѧ	CYRILLIC CAPITAL LETTER LITTLE YUS
0467	ѧ	CYRILLIC SMALL LETTER LITTLE YUS
0468	Ѩ	CYRILLIC CAPITAL LETTER IOTIFIED LITTLE YUS
0469	ѩ	CYRILLIC SMALL LETTER IOTIFIED LITTLE YUS
046A	Ѫ	CYRILLIC CAPITAL LETTER BIG YUS
046B	ѫ	CYRILLIC SMALL LETTER BIG YUS
046C	Ѭ	CYRILLIC CAPITAL LETTER IOTIFIED BIG YUS
046D	ѭ	CYRILLIC SMALL LETTER IOTIFIED BIG YUS
046E	Ѯ	CYRILLIC CAPITAL LETTER KSI
046F	ѯ	CYRILLIC SMALL LETTER KSI
0470	Ѱ	CYRILLIC CAPITAL LETTER PSI
0471	ѱ	CYRILLIC SMALL LETTER PSI
0472	Ѳ	CYRILLIC CAPITAL LETTER FITA
0473	ѳ	CYRILLIC SMALL LETTER FITA
		→ 0275 ɵ latin small letter barred o
		→ 03B8 θ greek small letter theta
0474	Ѵ	CYRILLIC CAPITAL LETTER IZHITSA
0475	ѵ	CYRILLIC SMALL LETTER IZHITSA
0476	Ѷ	CYRILLIC CAPITAL LETTER IZHITSA WITH DOUBLE GRAVE ACCENT
		≡ 0474 Ѵ 030F ̏
0477	ѷ	CYRILLIC SMALL LETTER IZHITSA WITH DOUBLE GRAVE ACCENT
		≡ 0475 ѵ 030F ̏
0478	Ѹ	CYRILLIC CAPITAL LETTER UK
		• basic Old Cyrillic uk is unified with cyrillic letter u
		→ 0423 У cyrillic capital letter u
0479	ѹ	CYRILLIC SMALL LETTER UK
047A	Ѻ	CYRILLIC CAPITAL LETTER ROUND OMEGA
047B	ѻ	CYRILLIC SMALL LETTER ROUND OMEGA
047C	Ѽ	CYRILLIC CAPITAL LETTER OMEGA WITH TITLO

047D	ѽ	CYRILLIC SMALL LETTER OMEGA WITH TITLO
		• the exact identity of these broad omegas is unclear and may require revision of glyphs
047E	Ѿ	CYRILLIC CAPITAL LETTER OT
047F	ѿ	CYRILLIC SMALL LETTER OT
0480	Ҁ	CYRILLIC CAPITAL LETTER KOPPA
0481	ҁ	CYRILLIC SMALL LETTER KOPPA

Historic miscellaneous

0482	҂	CYRILLIC THOUSANDS SIGN
0483	҃	COMBINING CYRILLIC TITLO
0484	҄	COMBINING CYRILLIC PALATALIZATION
		→ 033E ̾ combining vertical tilde
0485	҅	COMBINING CYRILLIC DASIA PNEUMATA
		→ 0314 ̔ combining reversed comma above
0486	҆	COMBINING CYRILLIC PSILI PNEUMATA
		→ 0313 ̓ combining comma above
0487	◌	<reserved>
0488	҈	COMBINING CYRILLIC HUNDRED THOUSANDS SIGN
		• use 20DD ◌ for ten thousands sign
0489	҉	COMBINING CYRILLIC MILLIONS SIGN

Extended Cyrillic

048A	Ҋ	CYRILLIC CAPITAL LETTER SHORT I WITH TAIL
048B	ҋ	CYRILLIC SMALL LETTER SHORT I WITH TAIL
		• Kildin Sami
048C	Ҍ	CYRILLIC CAPITAL LETTER SEMISOFT SIGN
048D	ҍ	CYRILLIC SMALL LETTER SEMISOFT SIGN
		• Kildin Sami
048E	Ҏ	CYRILLIC CAPITAL LETTER ER WITH TICK
048F	ҏ	CYRILLIC SMALL LETTER ER WITH TICK
		• Kildin Sami
0490	Ґ	CYRILLIC CAPITAL LETTER GHE WITH UPTURN
0491	ґ	CYRILLIC SMALL LETTER GHE WITH UPTURN
		• Ukrainian
0492	Ғ	CYRILLIC CAPITAL LETTER GHE WITH STROKE
0493	ғ	CYRILLIC SMALL LETTER GHE WITH STROKE
		• Azerbaijani, Bashkir, ...
		• full bar form preferred over half-barred "F"-type
0494	Ҕ	CYRILLIC CAPITAL LETTER GHE WITH MIDDLE HOOK
0495	ҕ	CYRILLIC SMALL LETTER GHE WITH MIDDLE HOOK
		• Abkhasian, Yakut
0496	Җ	CYRILLIC CAPITAL LETTER ZHE WITH DESCENDER
0497	җ	CYRILLIC SMALL LETTER ZHE WITH DESCENDER
		• Tatar, ...
0498	Ҙ	CYRILLIC CAPITAL LETTER ZE WITH DESCENDER
0499	ҙ	CYRILLIC SMALL LETTER ZE WITH DESCENDER
		• Bashkir
		• letterforms with right hooks are preferred, although occasional variants with left hooks occur
049A	Қ	CYRILLIC CAPITAL LETTER KA WITH DESCENDER
049B	қ	CYRILLIC SMALL LETTER KA WITH DESCENDER
		• Abkhasian, Tajik, ...
049C	Ҝ	CYRILLIC CAPITAL LETTER KA WITH VERTICAL STROKE

Right page (049D–04DA)

049D	ҝ	CYRILLIC SMALL LETTER KA WITH VERTICAL STROKE
		• Azerbaijani
049E	Ҟ	CYRILLIC CAPITAL LETTER KA WITH STROKE
049F	ҟ	CYRILLIC SMALL LETTER KA WITH STROKE
		• Abkhasian
04A0	Ҡ	CYRILLIC CAPITAL ... ABKHASIAN CHE
04A1	ҡ	CYRILLIC S... ...BKHASIAN CHE
04A2	Ң	CYRILLIC CAPITAL ... ...ABKHASIAN CHE DESCENDE...
04A3	ң	CYRILLIC S... ...BKHASIAN CHE DESCENDE...
		• Bashkir, ...
04A4	Ҥ	CYRILLIC CAPITAL ... ...A
04A5	ҥ	CYRILLIC S... ...casian languages
		• Altay, Ma... ...formal lowercase
		• this is not ...
04A6	Ҧ	CYRILLIC C... HOOK
04A7	ҧ	CYRILLIC S... ...r byelorussian-HOOK
		• Abkhasia...
04A8	Ҩ	CYRILLIC S... ...ZHE WITH BREVE
04A9	ҩ	CYRILLIC S...
04AA	Ҫ	CYRILLIC C... ...E WITH BREVE DESCENDE...
04AB	ҫ	CYRILLIC SMALL LETTER ES WITH DESCENDER
		• Bashkir, Chuvash
		• letterforms with right hooks are preferred, although occasional variants with left hooks occur
		• in Chuvashia, letterforms identical to or similar in form to 00E7 ç regularly occur
04AC	Ҭ	CYRILLIC CAPITAL LETTER TE WITH DESCENDER
04AD	ҭ	CYRILLIC SMALL LETTER TE WITH DESCENDER
		• Abkhasian
04AE	Ү	CYRILLIC CAPITAL LETTER STRAIGHT U
04AF	ү	CYRILLIC SMALL LETTER STRAIGHT U
		• stem is straight, unlike LETTER U
		• Azerbaijani, Bashkir, ...
		→ 0423 У cyrillic capital letter u
04B0	Ұ	CYRILLIC CAPITAL LETTER STRAIGHT U WITH STROKE
04B1	ұ	CYRILLIC SMALL LETTER STRAIGHT U WITH STROKE
		• Kazakh
04B2	Х	CYRILLIC CAPITAL LETTER HA WITH DESCENDER
04B3	х	CYRILLIC SMALL LETTER HA WITH DESCENDER
		• Abkhasian, Tajik, Uzbek
04B4	Ҵ	CYRILLIC CAPITAL LIGATURE TE TSE
04B5	ҵ	CYRILLIC SMALL LIGATURE TE TSE
		• Abkhasian
		• this is not a decomposable ligature
04B6	Ҷ	CYRILLIC CAPITAL LETTER CHE WITH DESCENDER
04B7	ҷ	CYRILLIC SMALL LETTER CHE WITH DESCENDER
		• Abkhasian, Tajik
04B8	Ҹ	CYRILLIC CAPITAL LETTER CHE WITH VERTICAL STROKE
04B9	ҹ	CYRILLIC SMALL LETTER CHE WITH VERTICAL STROKE
		• Azerbaijani
04BA	һ	CYRILLIC CAPITAL LETTER SHHA

04BB	һ	CYRILLIC SMALL LETTER SHHA
		• originally derived from Latin "h", but uppercase form 04BA һ is closer to an inverted che (0427 Ч)
		• Azerbaijani, Bashkir, ...
		→ 0068 h latin small letter h
04C3	҃	CYRILLIC CAPITAL LETTER KA WITH HOOK
04C4	҄	CYRILLIC SMALL LETTER KA WITH HOOK
		• Khanty, Chukchi
04C5	Ӆ	CYRILLIC CAPITAL LETTER EL WITH TAIL
04C6	ӆ	CYRILLIC SMALL LETTER EL WITH TAIL
		• Kildin Sami
04C7	Ӈ	CYRILLIC CAPITAL LETTER EN WITH HOOK
04C8	ӈ	CYRILLIC SMALL LETTER EN WITH HOOK
		• Khanty, Chukchi, Nenets
04C9	Ӊ	CYRILLIC CAPITAL LETTER EN WITH TAIL
04CA	ӊ	CYRILLIC SMALL LETTER EN WITH TAIL
		• Kildin Sami
04CB	Ҷ	CYRILLIC CAPITAL LETTER KHAKASSIAN CHE
04CC	ҷ	CYRILLIC SMALL LETTER KHAKASSIAN CHE
04CD	Ӎ	CYRILLIC CAPITAL LETTER EM WITH TAIL
04CE	ӎ	CYRILLIC SMALL LETTER EM WITH TAIL
		• Kildin Sami
04CF	ӏ	CYRILLIC SMALL LETTER PALOCHKA
04D0	Ӑ	CYRILLIC CAPITAL LETTER A WITH BREVE
		≡ 0410 А 0306 ̆
04D1	ӑ	CYRILLIC SMALL LETTER A WITH BREVE
		≡ 0430 а 0306 ̆
04D2	Ӓ	CYRILLIC CAPITAL LETTER A WITH DIAERESIS
		≡ 0410 А 0308 ̈
04D3	ӓ	CYRILLIC SMALL LETTER A WITH DIAERESIS
		≡ 0430 а 0308 ̈
04D4	Ӕ	CYRILLIC CAPITAL LIGATURE A IE
04D5	ӕ	CYRILLIC SMALL LIGATURE A IE
		• this is not a decomposable ligature
		→ 00E6 æ latin small letter ae
04D6	Ӗ	CYRILLIC CAPITAL LETTER IE WITH BREVE
		≡ 0415 Е 0306 ̆
04D7	ӗ	CYRILLIC SMALL LETTER IE WITH BREVE
		≡ 0435 е 0306 ̆
04D8	Ә	CYRILLIC CAPITAL LETTER SCHWA
04D9	ә	CYRILLIC SMALL LETTER SCHWA
		→ 0259 ə latin small letter schwa
04DA	Ӛ	CYRILLIC CAPITAL LETTER SCHWA WITH DIAERESIS
		≡ 04D8 Ә 0308 ̈

UC 5.0 Cyrillic

The Unicode Standard 5.0, Copyright © 1991-2006 Unicode, Inc. All rights reserved.

Unicode Docs

names and numbers

Historic letters

0460	Ѡ	CYRILLIC CAPITAL LETTER OMEGA
0461	ѡ	CYRILLIC SMALL LETTER OMEGA
0462	Ѣ	CYRILLIC CAPITAL LETTER YAT
0463	ѣ	CYRILLIC SMALL LETTER YAT
0464	Ѥ	CYRILLIC CAPITAL LETTER IOTIFIED E
0465	ѥ	CYRILLIC SMALL LETTER IOTIFIED E
0466	Ѧ	CYRILLIC CAPITAL LETTER LITTLE YUS
0467	ѧ	CYRILLIC SMALL LETTER LITTLE YUS
0468	Ѩ	CYRILLIC CAPITAL LETTER IOTIFIED LITTLE YUS
0469	ѩ	CYRILLIC SMALL LETTER IOTIFIED LITTLE YUS
046A	Ѫ	CYRILLIC CAPITAL LETTER BIG YUS
046B	ѫ	CYRILLIC SMALL LETTER BIG YUS
046C	Ѭ	CYRILLIC CAPITAL LETTER IOTIFIED BIG YUS
046D	ѭ	CYRILLIC SMALL LETTER IOTIFIED BIG YUS
046E	Ѯ	CYRILLIC CAPITAL LETTER KSI
046F	ѯ	CYRILLIC SMALL LETTER KSI
0470	Ѱ	CYRILLIC CAPITAL LETTER PSI
0471	ѱ	CYRILLIC SMALL LETTER PSI
0472	Ѳ	CYRILLIC CAPITAL LETTER FITA
0473	ѳ	CYRILLIC SMALL LETTER FITA

→ 0275 ɵ latin small letter barred o

→ 03B8 θ greek small letter theta

Unicode

- **Unicode: defines numbers**
- **PostScript & pdf: likes names**
- **Names are not defined in UC**
- **Names are recommended**
- **Fonts should have both correct numbers and names**

Unicode Block

uni040	Iocyril	Djecyr	Gjecyr	Ecyrill	Dzecyr	Icyrilli	Yicyril	Jecyri	Ljecyr	Njecyr	Tshecy	Kjecyr	uni040	Ushort	Dzhecy
È	Ë	Ђ	Ѓ	Є	Ѕ	І	Ї	Ј	Љ	Њ	Ћ	Ќ	Й	Ў	Џ

Acyrill	Becyri	Vecyri	Gecyri	Decyri	Iecyril	Zhecyr	Zecyri	Iicyril	Iishort	Kacyri	Elcyri	Emcyri	Encyri	Ocyrill	Pecyri
А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П

Ercyri	Escyri	Tecyri	Ucyrill	Efcyri	Khacyr	Tsecyr	Checyr	Shacyr	Shchac	Hardsig	Yericy	Softsig	Erever	Iucyri	Iacyri
Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я

acyrill	becyri	vecyri	gecyri	decyri	iecyril	zhecyr	zecyri	iicyril	iishort	kacyri	elcyri	emcyri	encyri	ocyrill	pecyri
а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п

ercyri	escyri	tecyri	ucyrill	efcyri	khacyr	tsecyr	checyr	shacyr	shchac	hardsig	yericy	softsig	erever	iucyri	iacyri
р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я

uni045	iocyril	djecyr	gjecyr	ecyrill	dzecyr	icyrilli	yicyril	jecyri	ljecyr	njecyr	tshecy	kjecyr	uni045	ushort	dzhecy
è	ё	ђ	ѓ	є	ѕ	і	ї	ј	љ	њ	ћ	ќ	й	ў	џ

Omegac	omegac	Yatcyr	yatcyr	Eiotifie	eiotifie	Yuslitt	yuslitt	Yuslitt	yuslitt	Yusbig	yusbig	Yusbigi	yusbigi	Ksicyr	ksicyr
Ѡ	ѡ	Ѣ	ѣ	Ѥ	ѥ	Ѧ	ѧ	Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ	Ѯ	ѯ

Psicyr	psicyr	Fitacyr	fitacyr	Izhitsa	izhitsa	Izhitsa	izhitsa	Ukcyri	ukcyri	Omega	omega	Omega	omega	Otcyri	otcyri
Ѱ	ѱ	Ѳ	ѳ	Ѵ	ѵ	Ѷ	ѷ	Ѹ	ѹ	Ѻ	ѻ	Ѽ	ѽ	Ѿ	ѿ

Koppac	koppac	thousar	titlocy	palatal	dasiapr	psilipne	uni048	uni048	uni048	lishort	iishort	Semiso	semiso	Ercyri	ercyri
Ҁ	ҁ	҂	҃	҄	҅	҆				Ҋ	ҋ	Ҍ	ҍ	Ҏ	ҏ

Unicode Blocks

- Basic Latin, Latin-1, Latin Extended-A, Latin Extended-B, Latin Extended Additional; Latin Extd C (5.0)

- Cyrillic, Cyrillic Extended (= Non-Sl.)

- Greek, Greek Extended (= Polytonic)

- IPA, IPA Extensions + Supplement

- Spacing & Combining Diacritics

→

Latin-1

nbspace	exclam	cent	sterling	currenc	yen	brokenl	section	dieresis	copyrig	ordfem	guillem	logicaln	sfthyph	registe	oversc
	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	–	®	¯
degree	plusmin	twosup	threesu	acute	micro	paragr	middled	cedilla	onesup	ordmas	guillem	onequar	onehalf	threequ	questio
°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
Agrave	Aacute	Acircum	Atilde	Adieres	Aring	AE	Ccedilla	Egrave	Eacute	Ecircum	Edieres	Igrave	Iacute	Icircum	Idieres
À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Eth	Ntilde	Ograve	Oacute	Ocircum	Otilde	Odieres	multipl	Oslash	Ugrave	Uacute	Ucircum	Udieres	Yacute	Thorn	german
Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
agrave	aacute	acircum	atilde	adieres	aring	ae	ccedilla	egrave	eacute	ecircum	edieres	igrave	iacute	icircum	idieres
à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
eth	ntilde	ograve	oacute	ocircum	otilde	odieres	divide	oslash	ugrave	uacute	ucircum	udieres	yacute	thorn	ydieres
ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Latin-Extended A

Amacron	amacron	Abreve	abreve	Aogonek	aogonek	Cacute	cacute	Ccircum	ccircum	Cdotacc	cdotacc	Ccaron	ccaron	Dcaron	dapostr
Ā	ā	Ă	ă	Ą	ą	Ć	ć	Ĉ	ĉ	Ċ	ċ	Č	č	Ď	ď
Dmacron	dslash	Emacron	emacron	Ebreve	ebreve	Edotacc	edotacc	Eogonek	eogonek	Ecaron	ecaron	Gcircum	gcircum	Gbreve	gbreve
Đ	đ	Ē	ē	Ĕ	ĕ	Ė	ė	Ę	ę	Ě	ě	Ĝ	ĝ	Ğ	ğ
Gdotacc	gdotacc	Gcomma	gcomma	Hcircum	hcircum	Hbar	hbar	Itilde	itilde	Imacron	imacron	Ibreve	ibreve	Iogonek	iogonek
Ġ	ġ	Ģ	ģ	Ĥ	ĥ	Ħ	ħ	Ĩ	ĩ	Ī	ī	Ĭ	ĭ	Į	į
Idotacc	dotless	IJ	ij	Jcircum	jcircum	Kcomma	kcomma	kra	Lacute	lacute	Lcomma	lcomma	Lapostr	lapostr	Ldotmid
İ	ı	Ĳ	ĳ	Ĵ	ĵ	Ķ	ķ	ĸ	Ĺ	ĺ	Ļ	ļ	Ľ	ľ	Ŀ
Idotmid	Lslash	lslash	Nacute	nacute	Ncomma	ncomma	Ncaron	ncaron	napostr	Eng	eng	Omacron	omacron	Obreve	obreve
Ŀ	Ł	ł	Ń	ń	Ņ	ņ	Ň	ň	ŉ	Ŋ	ŋ	Ō	ō	Ŏ	ŏ
Odblacu	odblacu	OE	oe	Racute	racute	Rcomma	rcomma	Rcaron	rcaron	Sacute	sacute	Scircum	scircum	Scedilla	scedilla
Ő	ő	Œ	œ	Ŕ	ŕ	Ŗ	ŗ	Ř	ř	Ś	ś	Ŝ	ŝ	Ş	ş
Scaron	scaron	Tcedilla	tcedilla	Tcaron	tapostr	Tbar	tbar	Utilde	utilde	Umacron	umacron	Ubreve	ubreve	Uring	uring
Š	š	Ţ	ţ	Ť	ť	Ŧ	ŧ	Ũ	ũ	Ū	ū	Ŭ	ŭ	Ů	ů
Udblacu	udblacu	Uogonek	uogonek	Wcircum	wcircum	Ycircum	ycircum	Ydieres	Zacute	zacute	Zdotacc	zdotacc	Zcaron	zcaron	longs
Ű	ű	Ų	ų	Ŵ	ŵ	Ŷ	ŷ	Ÿ	Ź	ź	Ż	ż	Ž	ž	ſ

Latin-Extended B

Cyrillic

Cyrillic Extended

IPA Extensions

aturned	ascript	ascript	bhook	oopen	ccurl	dtail	dhook	erevers	schwa	schwah	eopen	eopenr	eopenr	eopenr	jdotless
ɐ	ɑ	ɒ	ɓ	ɔ	ɕ	ɖ	ɗ	ɘ	ə	ɚ	ɛ	ɜ	ɝ	ɞ	ɟ

ghook	gscript	Gsmall	gamma	ramsho	hturned	hhook	hengho	istroke	iotalat	Ismalli	lmiddle	lbelt	lhookre	lezh	mturne
ɠ	ɡ	ɢ	ɣ	ɤ	ɥ	ɦ	ɧ	ɨ	ɩ	ɪ	ɫ	ɬ	ɭ	ɮ	ɯ

mlongle	mhook	nhookle	nhookr	Nsmall	obarred	OEsmal	omegal	philatin	rturned	rlongle	rhooktu	rlongle	rhook	rfishho	rfishho
ɰ	ɱ	ɲ	ɳ	ɴ	ɵ	ɶ	ɷ	ɸ	ɹ	ɺ	ɻ	ɼ	ɽ	ɾ	ɿ

Rsmall	Rsmall	shook	esh	dotless	eshsqu	eshcurl	tturned	tretrof	ubar	upsilon	vhook	vturned	wturney	yturned	Ysmall
ʀ	ʁ	ʂ	ʃ	ʄ	ʅ	ʆ	ʇ	ʈ	ʉ	ʊ	ʋ	ʌ	ʍ	ʎ	ʏ

zretrof	zcurl	ezh	ezhcurl	glottals	glottals	glottals	cstretc	bilabial	Bsmall	eopenc	Gsmall	Hsmall	jcrosse	kturned	Lsmalli
ʐ	ʑ	ʒ	ʓ	ʔ	ʕ	ʖ	ʗ	ʘ	ʙ	ʚ	ʛ	ʜ	ʝ	ʞ	ʟ

qhook	glottals	glottals	dzalton	dezh	dzcurl	ts	tesh	tccurl	feng	ls	lz	bilabial	bidenta	hturned	hturned
ʠ	ʡ	ʢ	dz	dʒ	dʑ	ts	tʃ	tɕ	fŋ	ls	lz	ʬ	ʭ	ʮ	ʯ

Combining Diacritics

Spacing Modifiers

02B0	02B1	02B2	02B3	02B4	02B5	02B6	02B7	02B8	02B9	02BA	02BB	02BC	02BD	02BE	02BF
ʰ	ʱ	ʲ	ʳ	ʴ	ʵ	ʶ	ʷ	ʸ	ʹ	ʺ	ʻ	ʼ	ʽ	ʾ	ʿ

02C0	02C1	02C2	02C3	02C4	02C5	02C6	02C7	02C8	02C9	02CA	02CB	02CC	02CD	02CE	02CF
ˀ	ˁ	˂	˃	˄	˅	ˆ	ˇ	ˈ	ˉ	ˊ	ˋ	ˌ	ˍ	ˎ	ˏ

02D0	02D1	02D2	02D3	02D4	02D5	02D6	02D7	02D8	02D9	02DA	02DB	02DC	02DD	02DE	02DF
ː	ˑ	˒	˓	˔	˕	˖	˗	˘	˙	˚	˛	˜	˝	˞	˟

02E0	02E1	02E2	02E3	02E4	02E5	02E6	02E7	02E8	02E9	02EA	02EB	02EC	02ED	02EE	02EF
ˠ	ˡ	ˢ	ˣ	ˤ	˥	˦	˧	˨	˩	˪	˫	ˬ	˭	ˮ	˯

02F0	02F1	02F2	02F3	02F4	02F5	02F6	02F7	02F8	02F9	02FA	02FB	02FC	02FD	02FE	02FF
˰	˱	˲	˳	˴	˵	˶	˷	˸	˹	˺	˻	˼	˽	˾	˿

That was…

1. Unicode Basics

2. Do's: Watch your Language!

Do's: Watch your Language!

Don't mix languages (and scripts) ✓

→

Example 1:
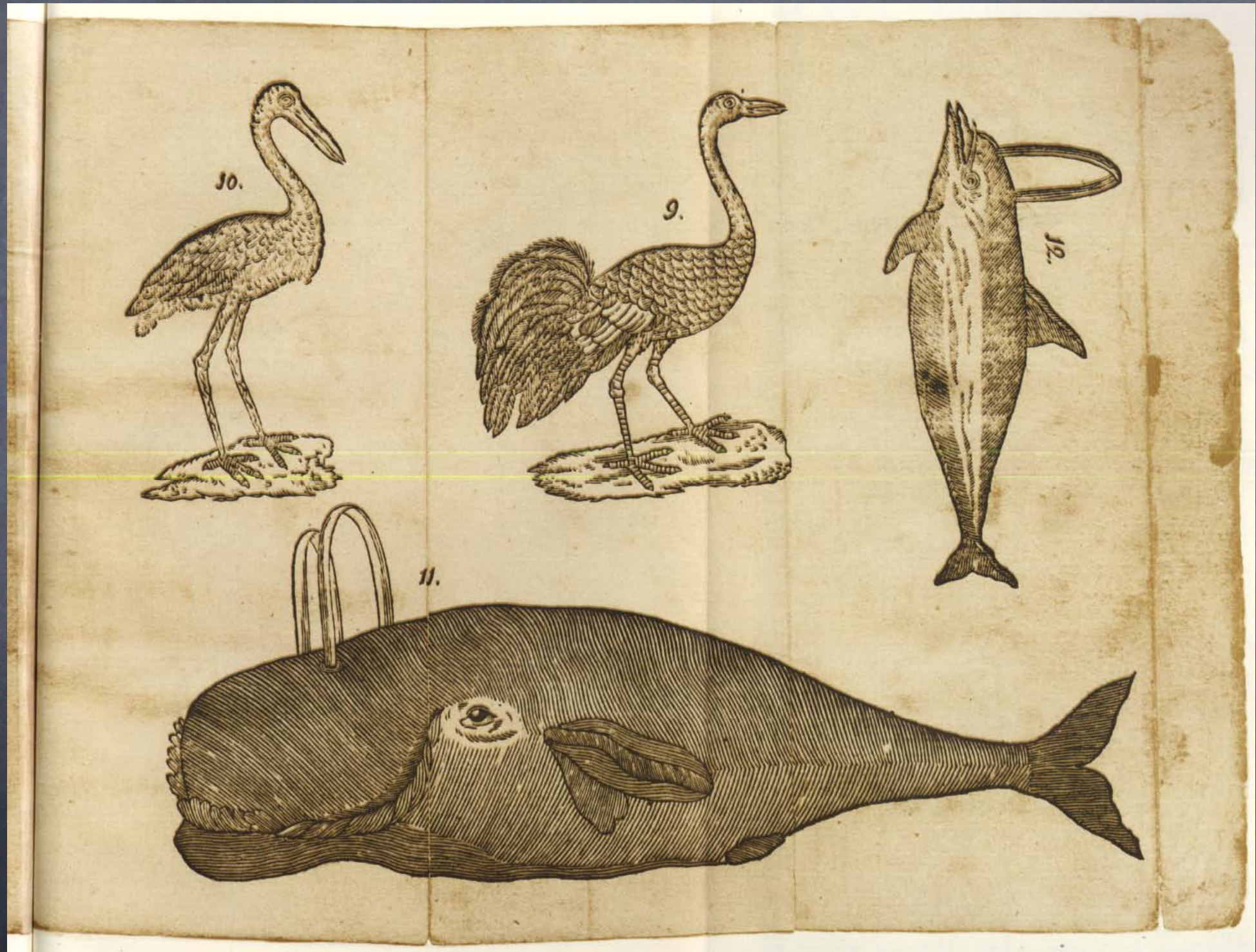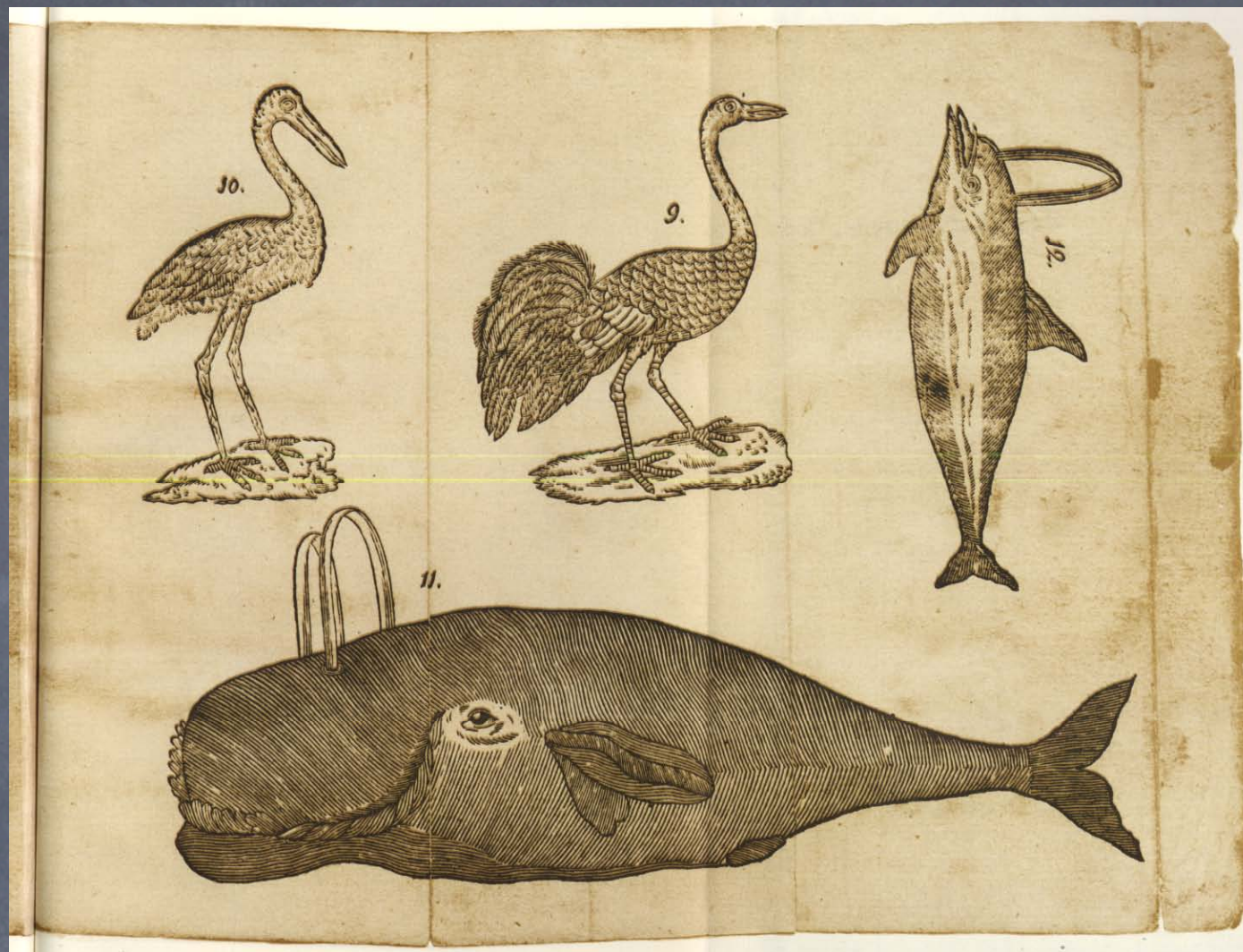
Petăr Beron,
Riben Bukvar
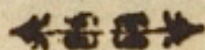
Моли́тва предъ да наче́натъ да чета́тъ дѣца́-та (а).

Пребла́гій Го́споди пода́й намъ благода́ть сватагѡ Д҃ха твоегѡ, да пода́ри намъ смы́слъ и да оутверди́ дꙋше́вны тѣ наши си́лы, за да внима́имъ на оуче́нїе-то кое́то ни прида́ватъ: и та́и да порасте́мъ та да просла́вимъ и́ма-то твое́ Созда́телю нашъ, и да по́льзовамы наше-то оте́чество.

Моли́тва като совершатъ.

Благодари́мъ тебе́ Созда́телю нашъ, чи ны̀ споди́би съ твоя-та благода́ть да внима́имъ на оуче́нїе-то. Благослови́ наши-тѣ нача́лницы, роди́тели и оучи́тели, кои́то пока́зватъ намъ пать-атъ на и́стина-та, и пода́й намъ си́ла и я́кость за да продолжи́мъ оуче́нїе-то наше.

До-

(а) Поди́ръ като и́зпеятъ досто́йно е́стъ: тре́ба да чите́ тѣзи моли́твы ка́та де́нь е́дно момче́, а дрꙋги тѣ да стоя́тъ гологлави и да слы́шатъ съ внима́нїе.

III.
До́бри сові́ти.

Што̀ не ще́шъ да ти стрꙋ́ватъ дрꙋги-тѣ ни́то ты̀ да го стрꙋ́вашъ на дрꙋги-тѣ.

Я́кѡ мно́гѡ хортꙋ́вашъ, мно́го погрѣ́шки стрꙋ́вашъ.
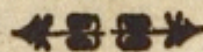
Я́кѡ ит-а́жишъ е́динъ па́ть, не та вѣ́роватъ вто́рый па́ть.

По до́брѣ прїа́тель вѣ́ренъ, а не ка́мень безъ цѣ́ненъ.

На голя́мцы да не са проти́влшъ, и съ боля́рцы да не са ска́рвашъ.

Дрꙋ́гомꙋ не са присми́вай, а́ми те́бе си ꙋбгле́двай.

Стрꙋ́вай добро́, да наме́ришъ добро́.

Да не са присми́вашъ нико́мꙋ кога́ испа́ни, защо̀ нико́й не зна́й ка́къ ще ꙋса́мни.

Кой́то дне́сь е боля́рецъ, ꙋтрѣ мо́жи да е слѣпе́цъ.

И кой́то дне́сь е голя́мецъ, ꙋтрѣ мо́жи да е мъртве́цъ.

Почита́й ста́ри-тѣ доде́ младе́йшъ, да почита́тъ и те́бе кога́ ꙋстарѣ́йшъ.

Ща

Молитва предъ да начёнатъ да четаатъ дѣца-та (а).

Преблагїй Господи подай намъ благодать сватагѡ Духа твоегѡ, да подари намъ смыслъ и да оутверди душёвны тѣ наши силы, за да внимаймъ на оучёнїе-то коёто ни придаватъ: и тѫи да порастёмъ та да прославимъ има-то твоё Создателю нашъ, и да пользовамы наше-то отечество.

(а) Подиръ като изпѣатъ достойно ёсть: трѣба да читѐ тѣзи молитвы ката дёнь ёдно момчё, а други тѣ да стоятъ гологлави и да слышатъ съ вниманїе.

III.
Добри совѣти.

Щото не щешъ да ти струватъ други-тѣ нито ты да го струвашъ на други-тѣ.

Якѡ многѡ хортувашъ, много погрѣшки струвашъ.

Якѡ итѧжишъ единъ пѫть, не та вѣроватъ вторый пѫть.

По добрѣ прїятель вѣренъ, а нё камень безъ цѣненъ.

... да ё мёртвецъ.

Почитай стари-тѣ доде младёйшъ, да почитатъ и тебе кога оустарѣйшъ.

Ща

Watch your Language!

36 ⟪❦⟫

Моли́тва пред̌ да наче́нат̌ да

чета̌т̌ д҆ьца́-та (а).

ЧЕТА̌Т̌ — Bulgarian, Cyrillic

→

Watch your Language!

36 Молитва предъ да начéнатъ да четăтъ дѣцá-та (а).

ЧЕТĂТЪ – Bulgarian, Cyrillic

Ă – not in Slavic Unicode

→

Watch your Language!

36 Молѝтва пред да наченат да чета́тъ д҃бца-та (а).

ЧЕТӐТЪ – Bulgarian, Cyrillic

ă – not in Slavic Unicode
   – but in Non-Slavic Cyr.
   – and also in Latin (RO)

→

Amacr Ā	amacr ā	Abreve Ă	abreve ă	Aogone Ą	aogonek ą	Cacute Ć	cacute ć	Ccircum Ĉ	ccircum ĉ	Cdotab Ċ	cdotab ċ	Ccaron Č	ccaron č	Dcaron Ď	dapostr ď
Dmacr Đ	dslash đ	Emacr Ē	emacr ē	Ebreve Ĕ	ebreve ĕ	Edotab Ė	edotab ė	Eogone Ę	eogonek ę	Ecaron Ě	ecaron ě	Gcircum Ĝ	gcircum ĝ	Gbreve Ğ	gbreve ğ
Gdotab Ġ	gdotab ġ	Gcomm Ģ	gcomm ģ	Hcircum Ĥ	hcircum ĥ	Hbar Ħ	hbar ħ	Itilde Ĩ	itilde ĩ	Imacr Ī	imacr ī	Ibreve Ĭ	ibreve ĭ	Iogonek Į	iogonek į
Idotab İ	dotless ı	IJ Ĳ	ij ĳ	Jcircum Ĵ	jcircum ĵ	Kcomm Ķ	kcomm ķ	kra ĸ	Lacute Ĺ	lacute ĺ	Lcomm Ļ	lcomm ļ	Lapostr Ľ	lapostr ľ	Ldotrig
Idotrig Ŀ	Lslash Ł	lslash ł	Nacute Ń	nacute ń	Ncomm Ņ	ncomm ņ	Ncaron Ň	ncaron ň	napostr ŉ	Eng Ŋ	eng ŋ	Omacr Ō	omacr ō	Obreve Ŏ	obreve ŏ
Odblacu Ő	odblacu ő	OE Œ	oe œ	Racute Ŕ	racute ŕ	Rcomm Ŗ	rcomm ŗ	Rcaron Ř	rcaron ř	Sacute Ś	sacute ś	Scircum Ŝ	scircum ŝ	Scedilla Ş	scedilla ş
Scaron Š	scaron š	Tcedill Ţ	tcedilla ţ	Tcaron Ť	tapostr ť	Tbar Ŧ	tbar ŧ	Utilde Ũ	utilde ũ	Umacr Ū	umacr ū	Ubreve Ŭ	ubreve ŭ	Uring Ů	uring ů
Udblacu Ű	udblacu ű	Uogone Ų	uogonek ų	Wcircum Ŵ	wcircum ŵ	Ycircum Ŷ	ycircum ŷ	Ydieres Ÿ	Zacute Ź	zacute ź	Zdotab Ż	zdotab ż	Zcaron Ž	zcaron ž	longs ſ

Cyrillic Extended

Watch your Language!

ă - Latin + Cyrillic!

Presentation: четăтъ   ☺

Encoding: CY-CY-CY-CY-CY-CY   ☺

Watch your Language!

ă – Latin + Cyrillic!

Presentation: четăтъ 🙂

Encoding: CY-CY-CY-CY-CY-CY 🙂

Presentation: четăтъ 🙂

Encoding: CY-CY-CY-LT-CY-CY ☹

Encode Educated!

Fact: Presence of right Unicode characters does not force users to make the right choices.

→

Encode Educated!

Fact: Presence of correct Unicode characters does not force users to make the right choices.

Fact: Using a Unicode font does not automatically mean that the encoding of all characters in a text follows Unicode principles or is correct.

→

Encode Educated!

Ergo: We need educated users.

©Agencja Gazeta

Encode Educated!

Example 2: Transliteration of Soft Sign/Hard Sign

Encode Educated!

Example 2: Transliteration of Soft Sign/Hard Sign

Рыльский Монастырь

Ryl'skij Monastyr'

Encode Educated!

Example 2: Transliteration of Soft Sign/Hard Sign

Рыльский Монастырь
Ryl'skij Monastyr'

What's wrong here?

Encode Educated!

Рыльский Монастырь
Ryl'skij Monastyr'

single curly quotes,
not Unicode translit.
characters

Copyright © 2003 r.a. schwarz

Encode Educated!

Unicode:
- Punctuation [(...;.-+)]
- Characters ABCdef
- Dingbats, Symbols ... ∀≢■◎✔✂❦

Copyright © 2003 r.a. schwarz

Encode Educated!

Рыльский Монастырь

R y l ' s k i j

char-char-char-punct-char-char-char

Encode Educated!

Рыльский Монастырь
Ryl'skij Monastyr'

Presentation: fine 😊
Encoding: problematic ☹

Encode Educated!

Unicode has separate transliteration characters for these Cyrillic chars.

′

″

(Spacing Modifiers)

Miscellaneous phonetic modifiers

02B9 ′ MODIFIER LETTER PRIME
- primary stress, emphasis
- transliteration of mjagkij znak (Cyrillic soft sign: palatalization)
→ 0027 ' apostrophe
→ 00B4 ´ acute accent
→ 02CA ´ modifier letter acute accent
→ 0301 ◌́ combining acute accent
→ 0374 ʹ greek numeral sign
→ 2032 ′ prime

02BA ″ MODIFIER LETTER DOUBLE PRIME
- exaggerated stress, contrastive stress
- transliteration of tverdyj znak (Cyrillic hard sign: no palatalization)
→ 0022 " quotation mark
→ 030B ◌̋ combining double acute accent
→ 2033 ″ double prime

Encode Educated!

Рыльский Монастырь

Ryl′skij Monastyr′

Encoding: Correct ☺

Presentation: well... ☹ ☺

Choose Wisely!

Example 3: Confusing choices between similar characters

Choose Wisely!

Slavic Jat' vs. Semisoft sign (Sami)

That was...

2. Do's and Don'ts

3. Problems & Solutions
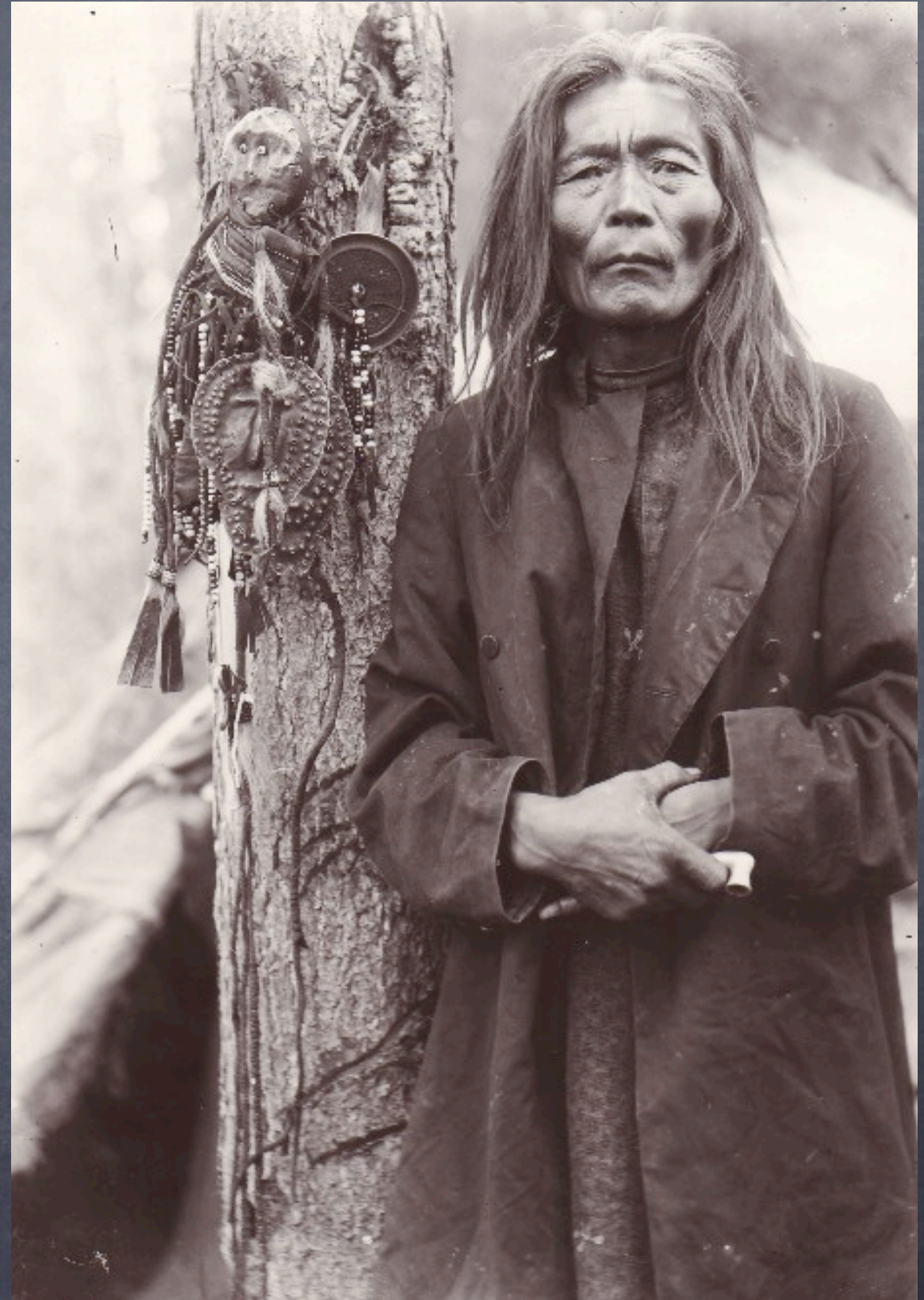
Missing Characters

Problem: Characters not part of Unicode (yet)
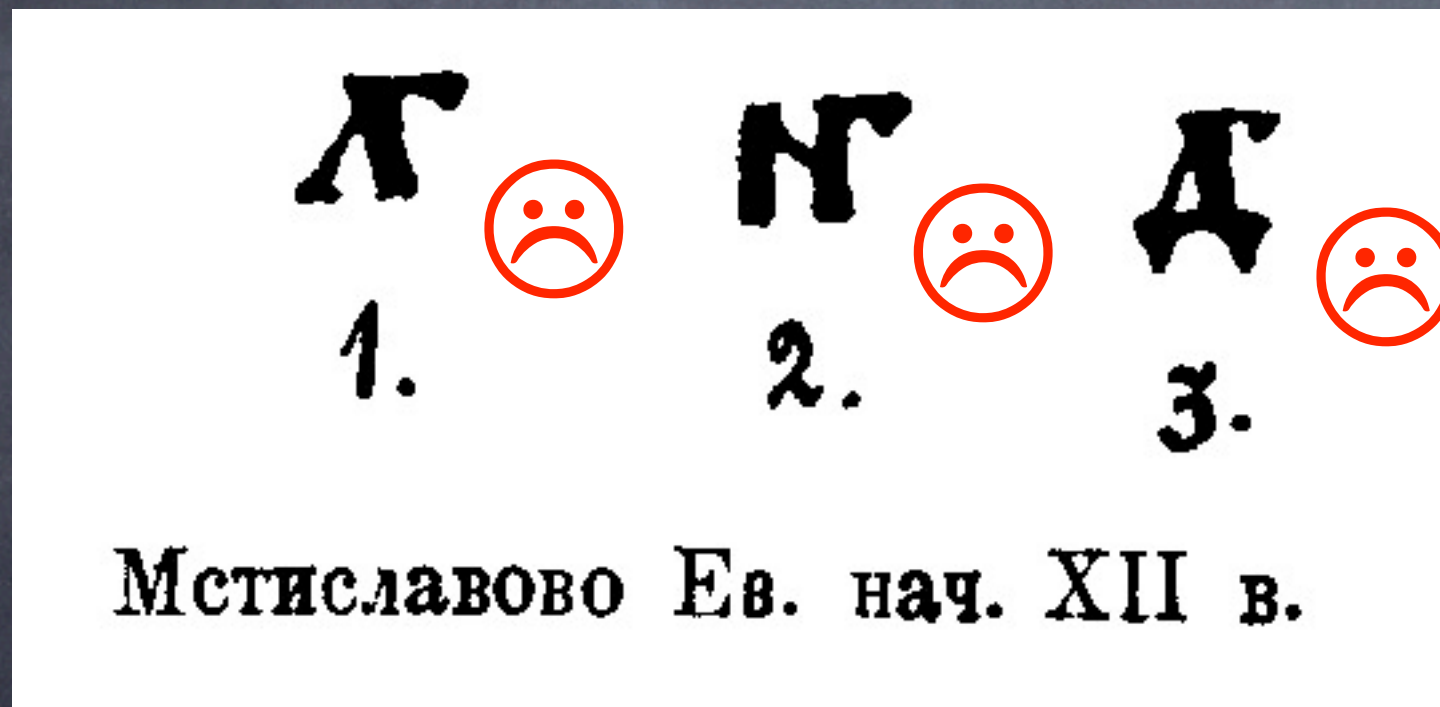
А.И. Соболевский, Славяно-русская палеография. Изд. 2-е, СПб. 1908, 52.
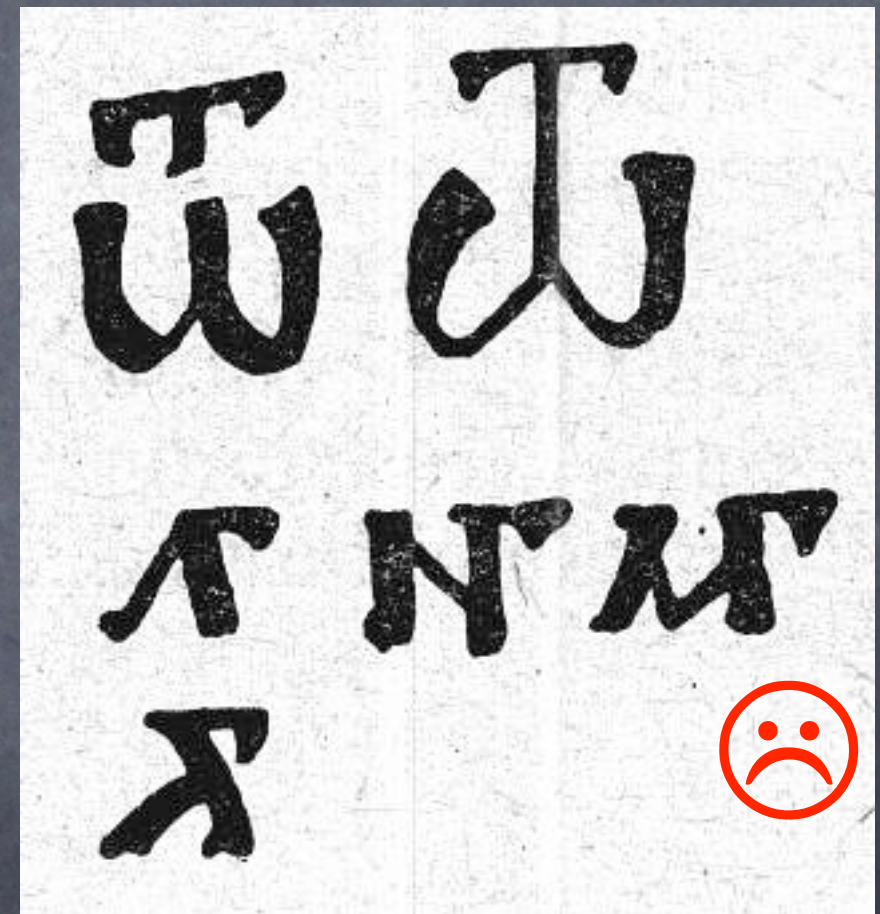
В.Н. Щепкин, Русская палеография, Москва 1967, 111.

Expanding Unicode

Solution 1:
- build consensus in scientific community
- write proposal to Unicode Inc.
- help making Unicode better, but...
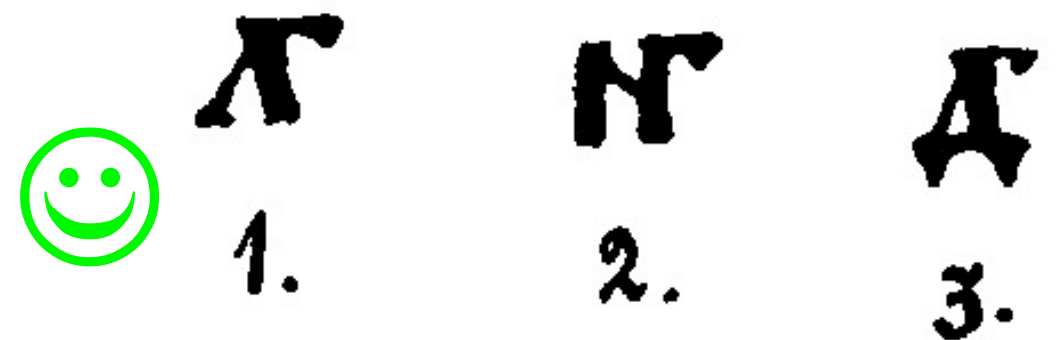
Мстиславово Ев. нач. XII в.

Expanding Unicode

- ... educate yourself
- ... be realistic
- ... avoid extremes
- ... do not provoke rejection

ПРОСЬБА МУСОР ЗДЕСЬ НЕ БРОСАТЬ
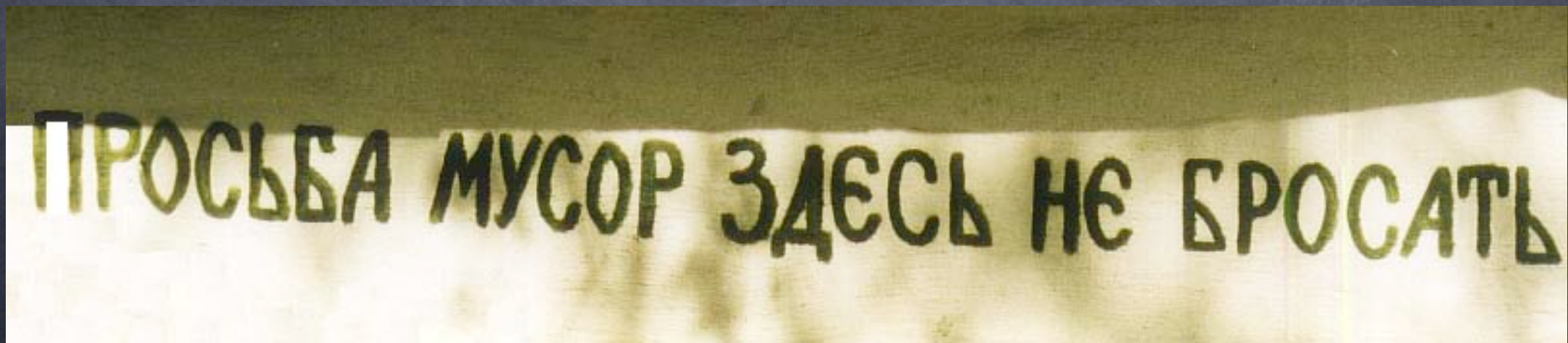
Unicode.org

- **Submissions, Proposals: Technical and Philological Aspects!!**

- **Pipeline of characters and scripts**

- **Rejected scripts and characters**

- **Published Errata**
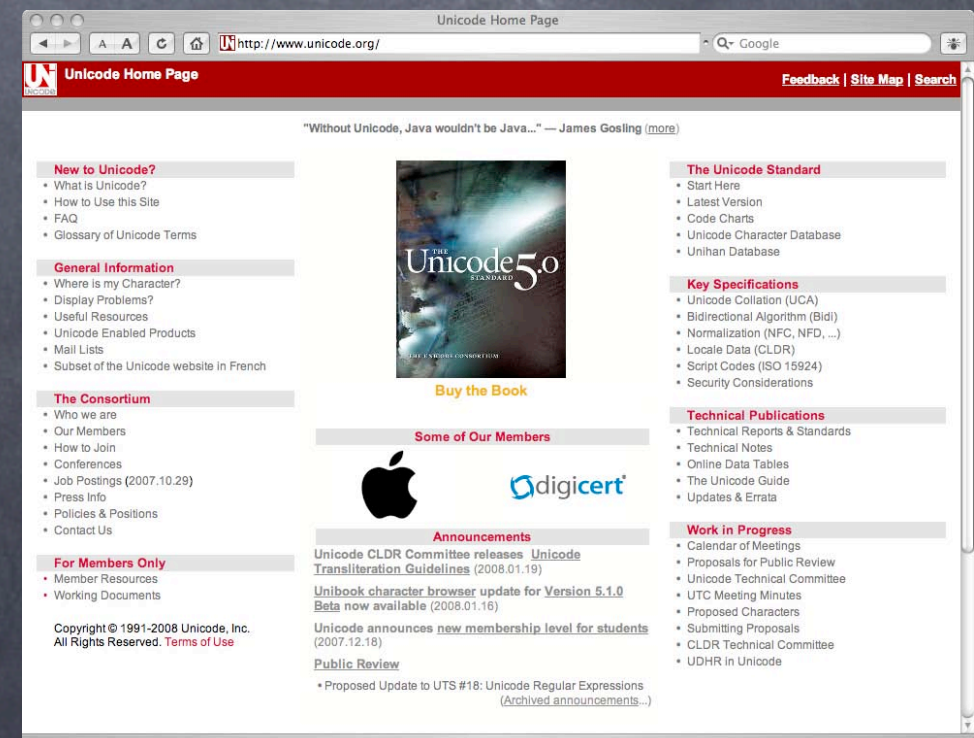
Unicode Policy

OCS Unicode

"The historical form of the Cyrillic alphabet is treated as a font style variation of modern Cyrillic because the historical forms are relatively close to the modern appearance" (UC book, ch. 7)

→

Unicode Policy

Extending Unicode OCS:

"If, at some future date, the old letterforms are adequately documented and the need for them demonstrated, then they can be added to this [the Cyrillic-Extended] block" (Unicode book v. 1, vol. 1, 45)

Proposal to Unicode

How to...

... write a proposal
(demo: Hist. additions Unicode 5.1)

Proposal to Unicode

That was... Solution 1: Proposals

pro:
- ★ everybody benefits
- ★ official
- ★ best solution for problems that fit

con:
- – patience needed
- – not an immediate solution

Missing Characters

Solution 2:
- use "private area" of Unicode

'Private Area'

- 6.400 Slots
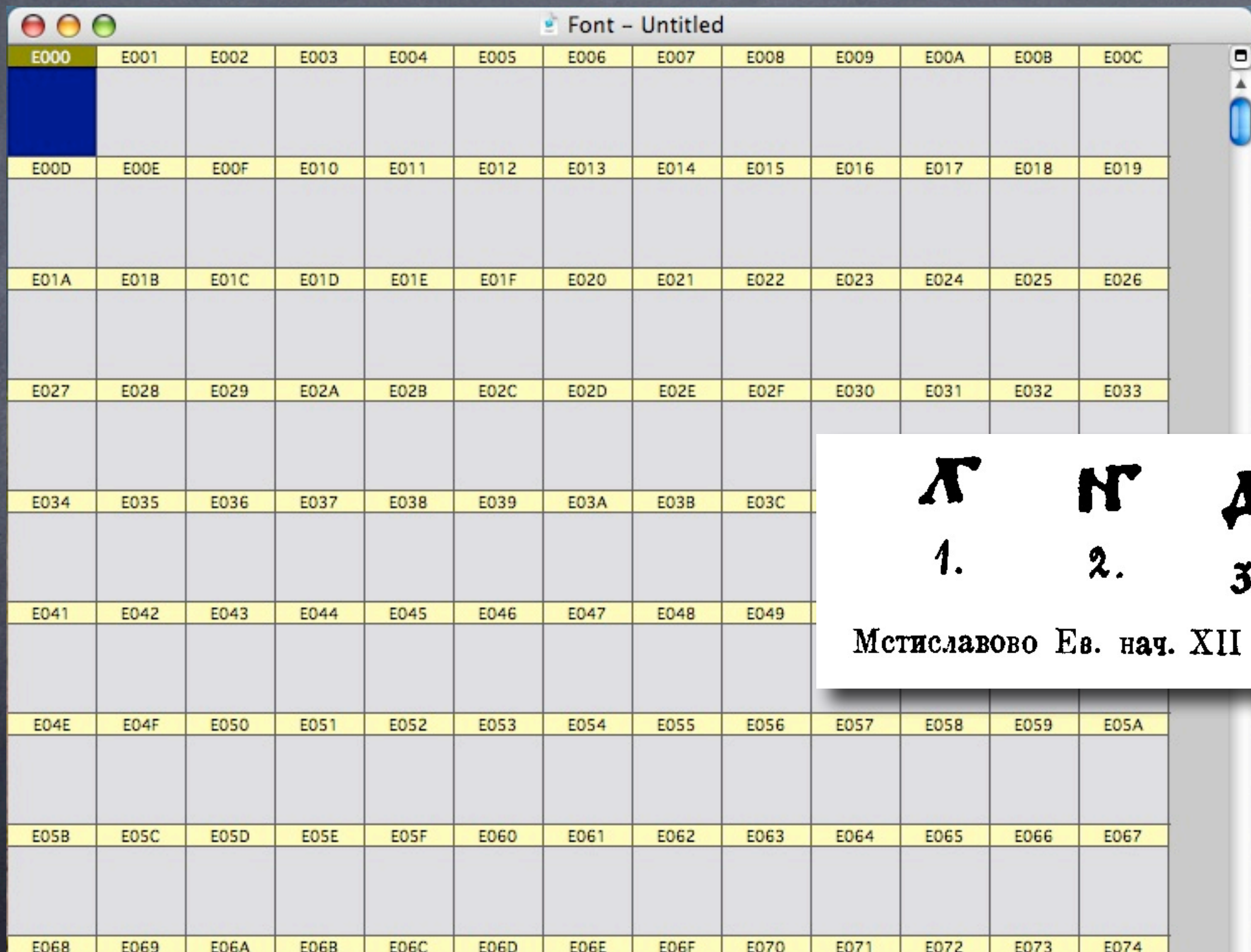
'Private Area'

- 6.400 Slots - waiting to be filled

Мстиславово Ев. нач. XII в.
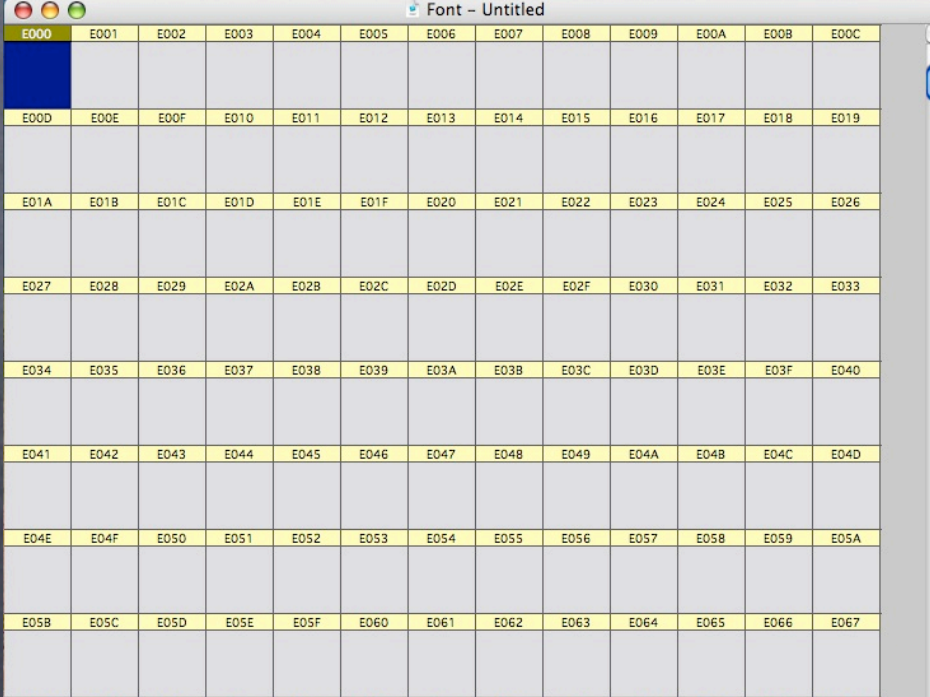
'Private Use Area'

Pro:
- Everything in one font
- Immediate solution
- ...
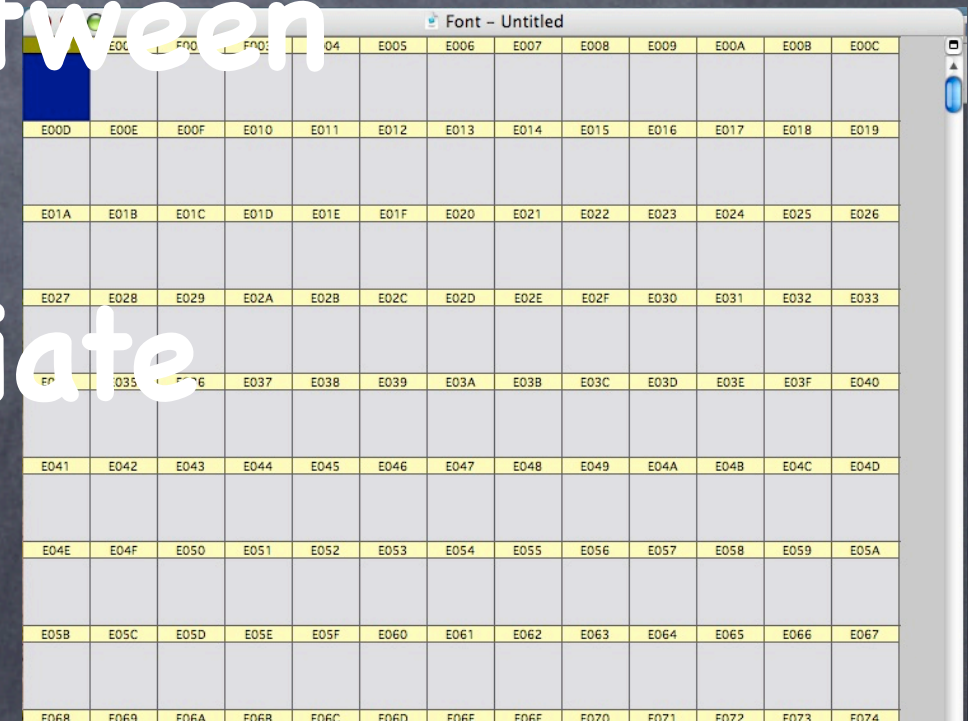
'Private Use Area'

Con:

- Compatibility between fonts and documents not guaranteed
- Slots may be in use by others
- Use not coordinated between philologies
- Sometimes an intermediate solution only

Private Use Area

Normal Example...

Private Use Area

Demo: Normal Example

Gandhari.ttf

Private Use Area

Bad Example...

http://gramoty.ru/

ДРЕВНЕРУССКИЕ БЕРЕСТЯНЫЕ ГРАМОТЫ

РУКОПИСНЫЕ ПАМЯТНИКИ ДРЕВНЕЙ РУСИ

Сайт разработан в рамках проекта "Birchbark Literacy from Medieval Rus: Contents and Contexts", финансируемого INTAS (№03513867)

О сайте

Об источнике

База данных

Библиотека

Карты

Библиография

Показать грамоту №... (1–956) [ ] (ОК)   [Список всех грамот]

Грамота 9
2-я половина XII в., Новгород

Жалоба прогнанной мужем жены
[Открыть]

Грамота 109
Рубеж XI–XII вв., Новгород

Дело о покупке краденой рабыни
[Открыть]

Грамота 906
2-я половина XI в., Новгород
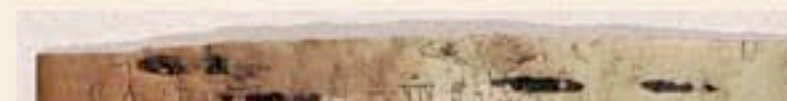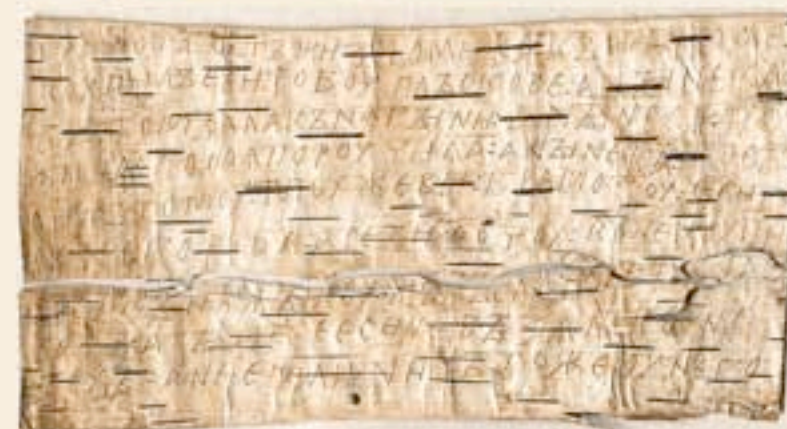
Private Use Area

Site uses a special font

Demo
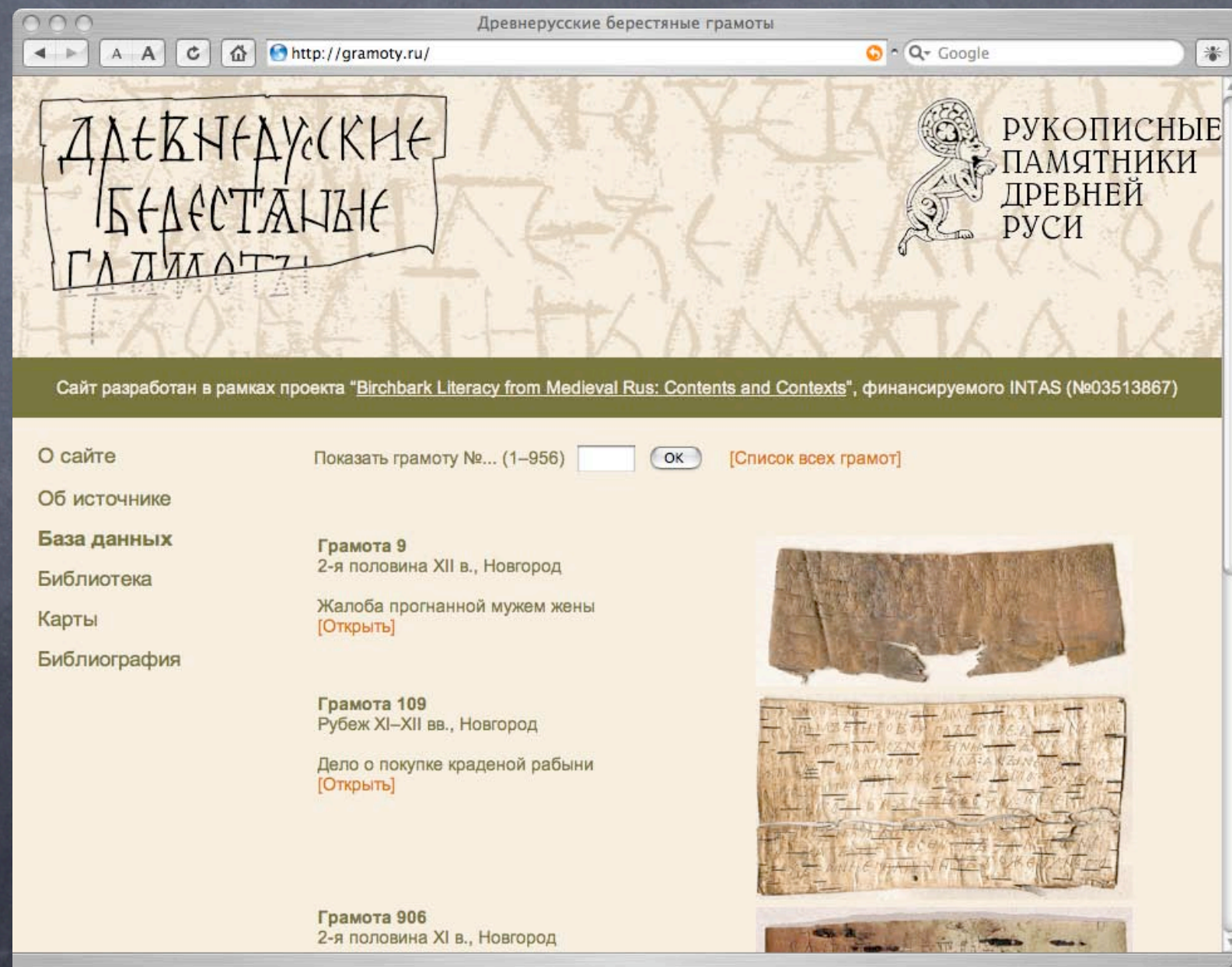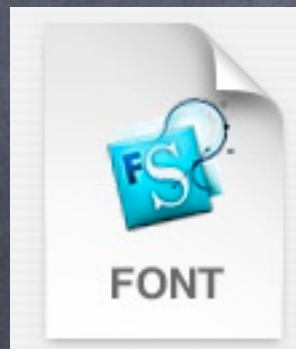
Problems & Solutions

Solutions so far:
- expanding Unicode itself
- using "private area" of Unicode

→

Problems & Solutions

Solutions so far:

– expanding Unicode itself

– using "private area" of Unicode

Next solution

– use alternate fonts

Character Variation

old, new, connected, Latin i

Solution: Alternate Fonts

	Font 1: standard	Font 2: alternate
character	4	Y
character	T	m

→

Solution: Alternate Fonts

Pro:
- encoding ☺
- presentation ☺

	Font 1: standard	Font 2: alternate
character	Ч	Ƴ
character	Т	ɯ

Con:
- some inconvenience, ☹
- some redundancy
- solves certain problems only

Problems & Solutions

Solutions so far:
– expanding Unicode itself
– using "private area" of Unicode
– using alternate fonts

Next solution
– using OpenType

What is OpenType?
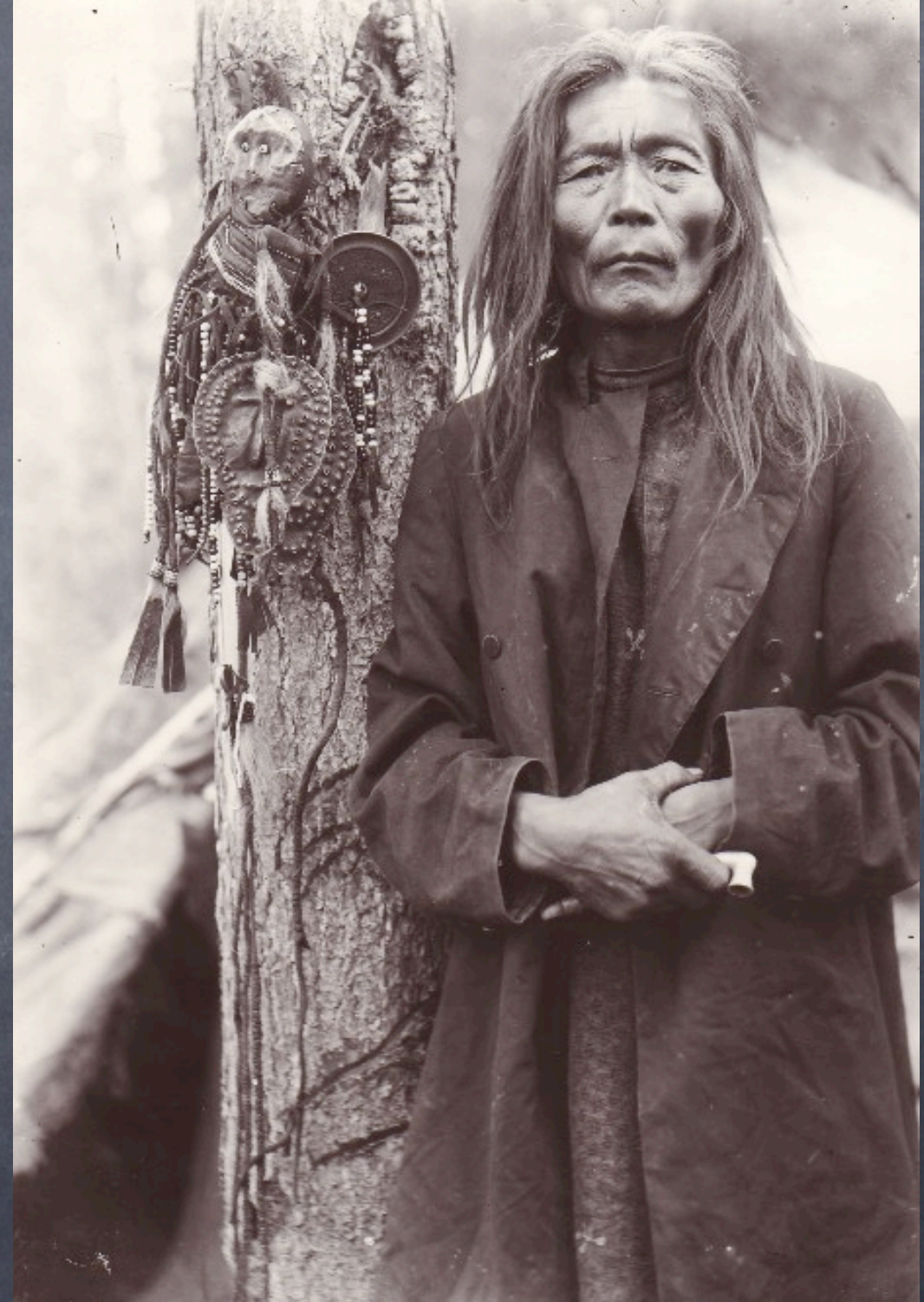
What is OpenType?
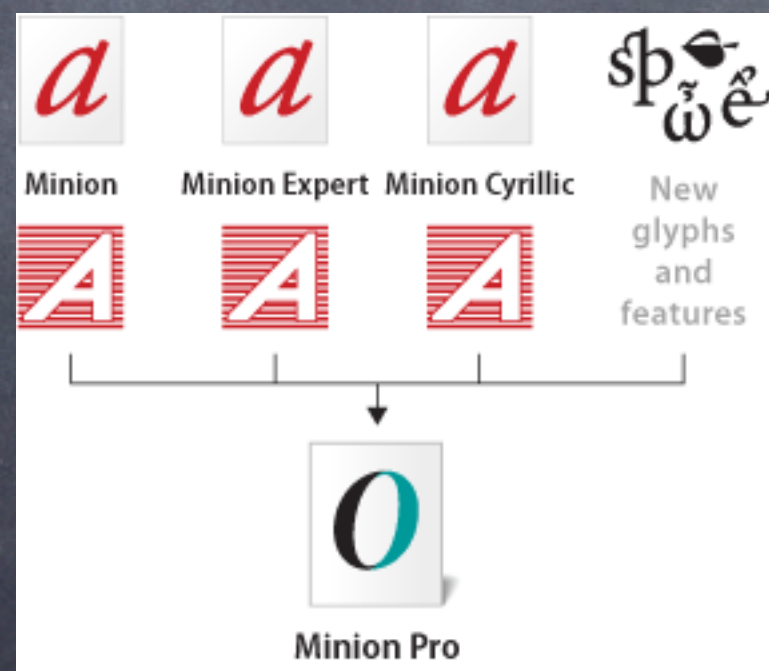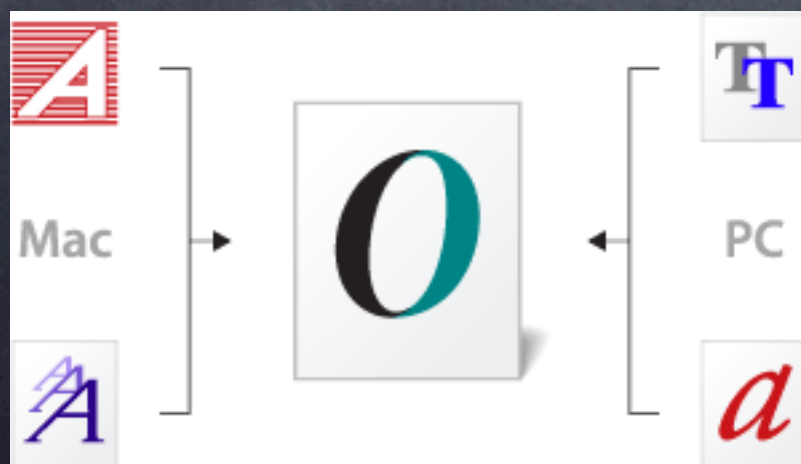
- **Font Technology**
- Cross-Platform
- Advanced Typography
- Multiple Languages
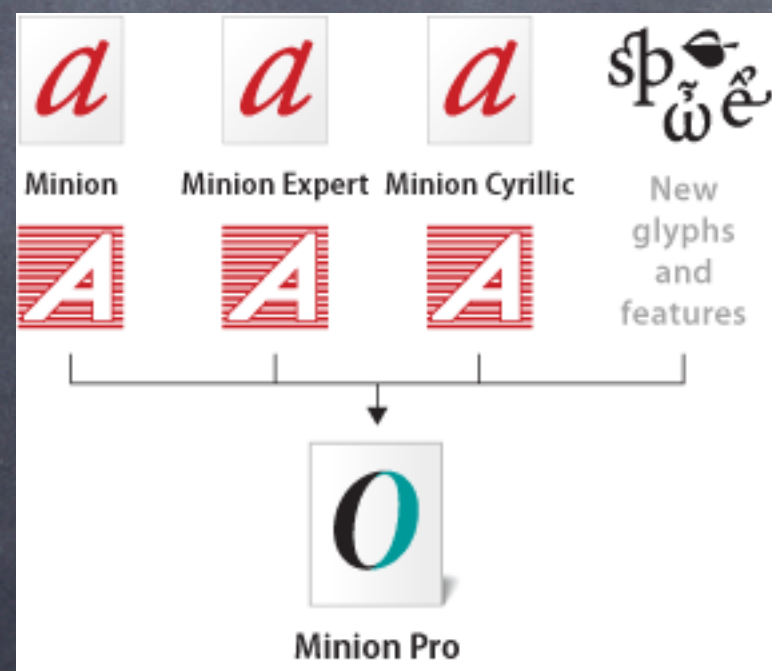
What is OpenType?

- **Font Technology**
- **Cross-Platform**
- Advanced Typography
- Multiple Languages

What is OpenType?

- Font Technology
- Cross-Platform
- Advanced Typography
- Multiple Languages
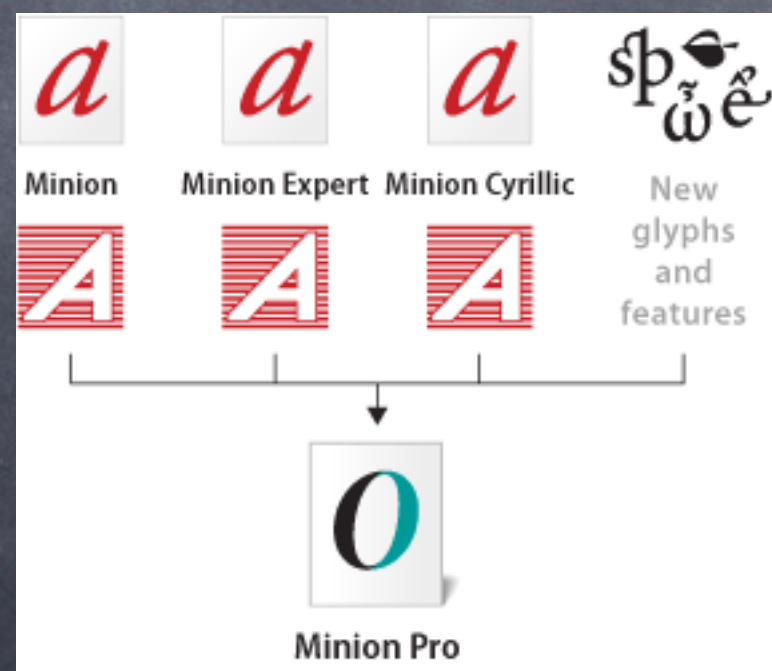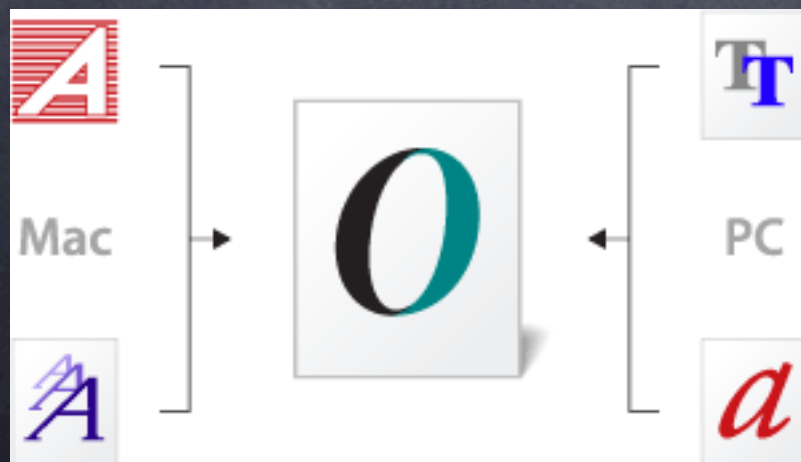
What is OpenType?

- **Font Technology**
- **Cross-Platform**
- **Advanced Typography**
- **Multiple Languages**

What OpenType isn't

- the same as "free fonts"
- or "public domain fonts"

→

How does OT work?

- scripting language inside fonts
- built-in programming
- developed by Adobe
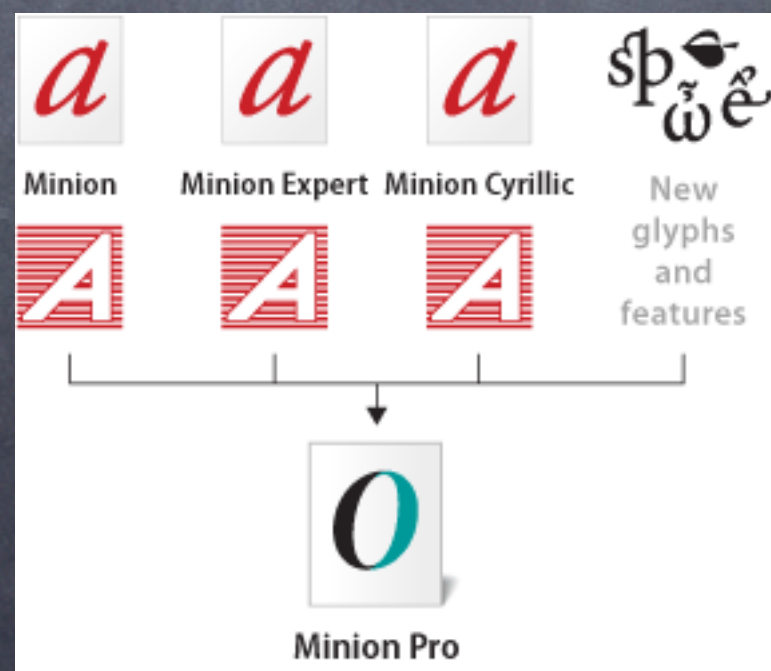
OpenType Fonts

Sample Solutions...

Unicode Ligatures

"Presentation forms"
- some for Latin
- many for Arabic

Unicode Ligatures

"Presentation forms"
- some for Latin
- many for Arabic
- **none for Cyrillic**
- **none for Glagolitic**

OpenType Features

Standard Ligatures
Live Demo ...

OpenType Features

Presentation

Encoding

OpenType Features

"Unconditional Ligatures"

substitute f l **by** fl;
substitute f i **by** fi;

Some fonts add longer ligatures:

substitute f f i **by** f_f_i;

ffi ffi

OpenType Ligatures

OpenType Features

"Unconditional Ligatures"

OpenType Features

"Unconditional Ligatures"

i.e. always being used
(but may be script-specific)

OpenType Features

"Unconditional Ligatures"

#ᴀ# → ·Ꙇ·

OpenType Features

not "unconditional"

OpenType Features

"Unconditional Ligatures"

"Discretionary Ligatures"

i.e. if user selects them

OpenType Features

dlig	Discretionary Ligatures	Replaces a sequence of glyphs with a single glyph, which is preferred for typographic purposes. This feature covers those ligatures that may be used for special effect, at the user's preference

```
feature dlig {
  sub c t by c_t;
  sub s t by s_t;
  sub longs h by longs_h;
  sub longs i by longs_i;
  sub longs l by longs_l;
  sub longs t by logs_t;
  sub longs longs by longs_longs;
} dlig;
```

ct st → ĉt ŝt

Discretionary Ligature
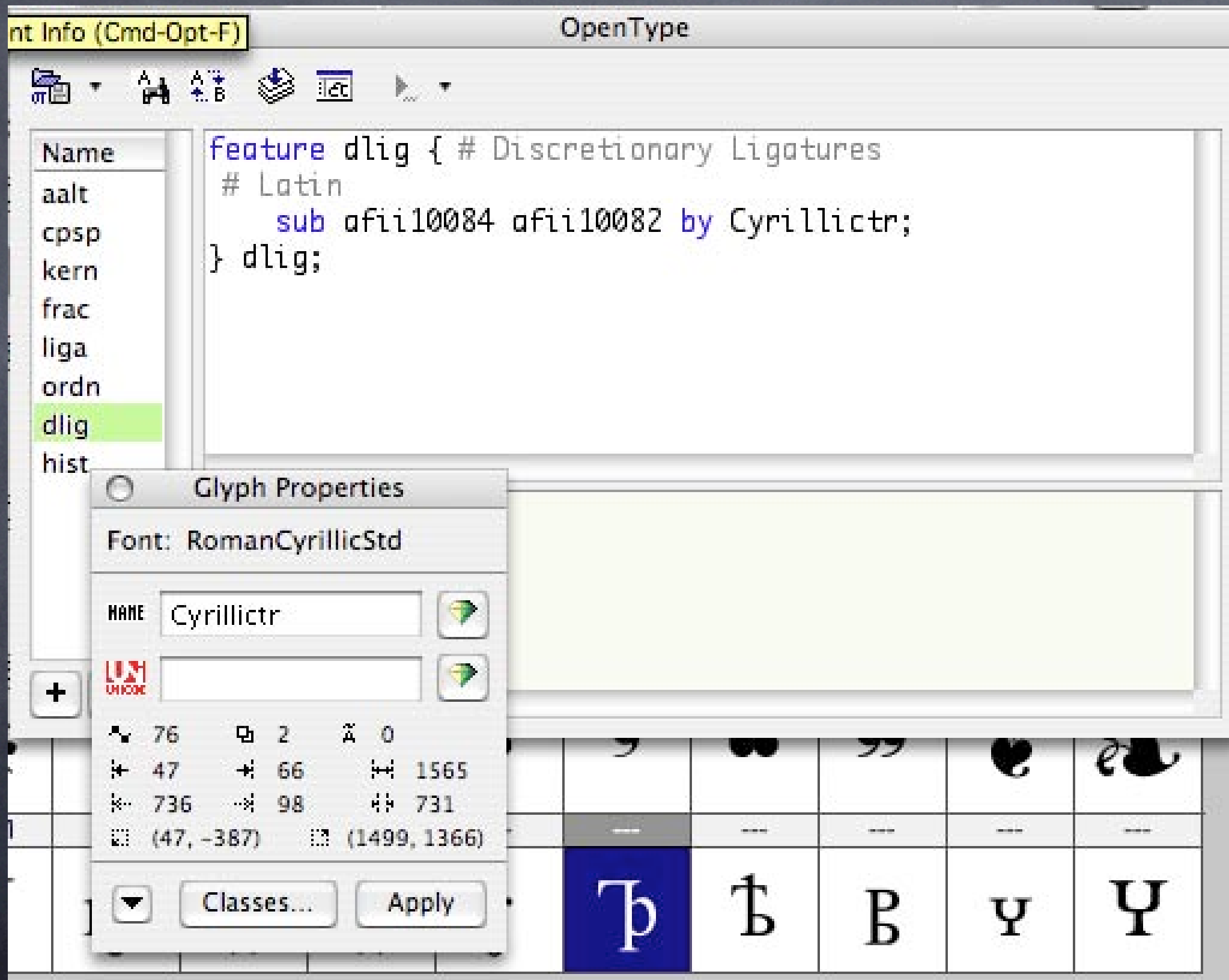
Discretionary Ligature
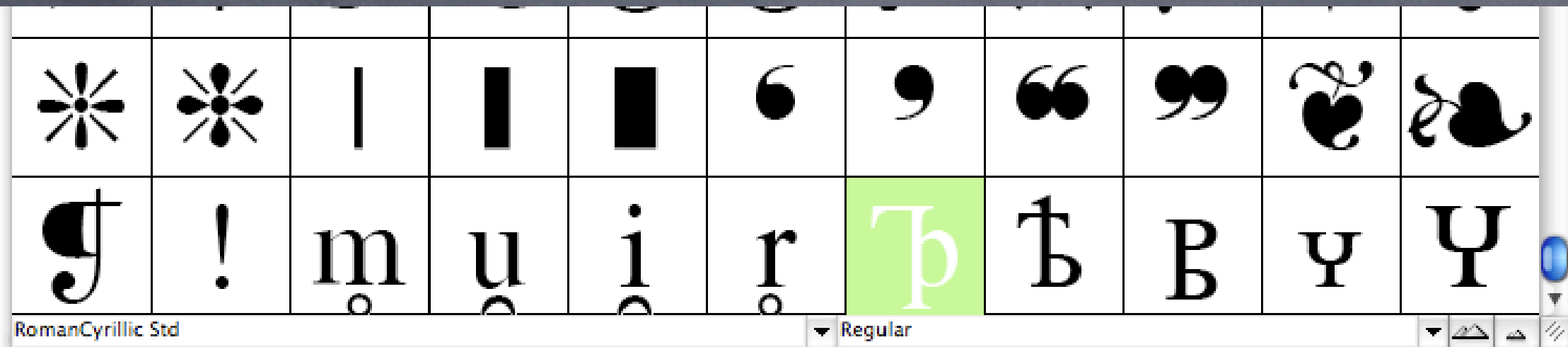
Presentation ☺

Encoding ☺

Discretionary Ligatures

LITERAE LIGATAE.

Historical char. shapes

hist	Historical Forms	Replaces glyphs with their historical forms, like long form of s or Fraktur form of k.

```
feature hist {
    sub s by longs;
} hist;
```

s → ſ

Historical char. shapes
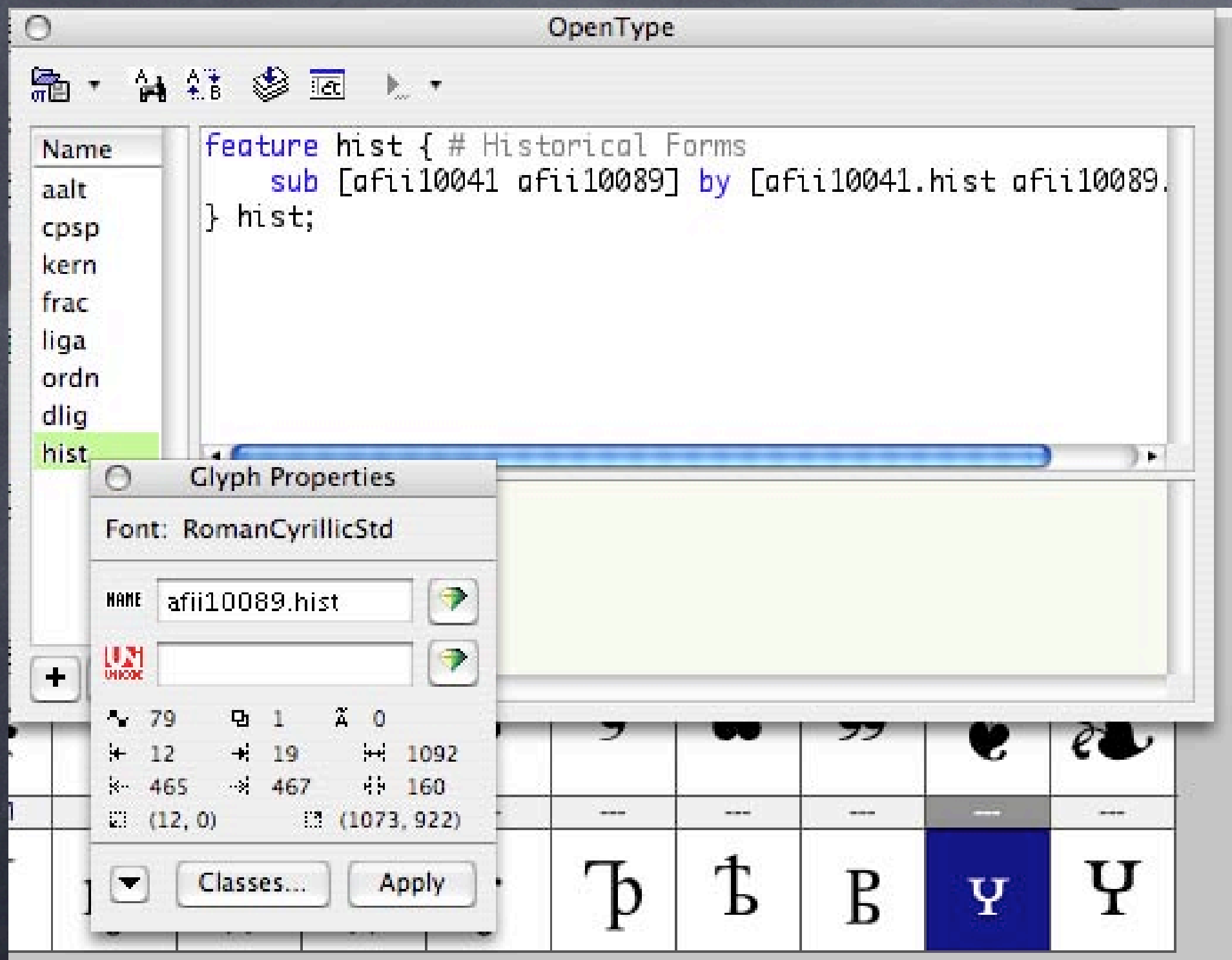
Presentation ☺

Encoding
Ч
☺

Alternate char. shapes
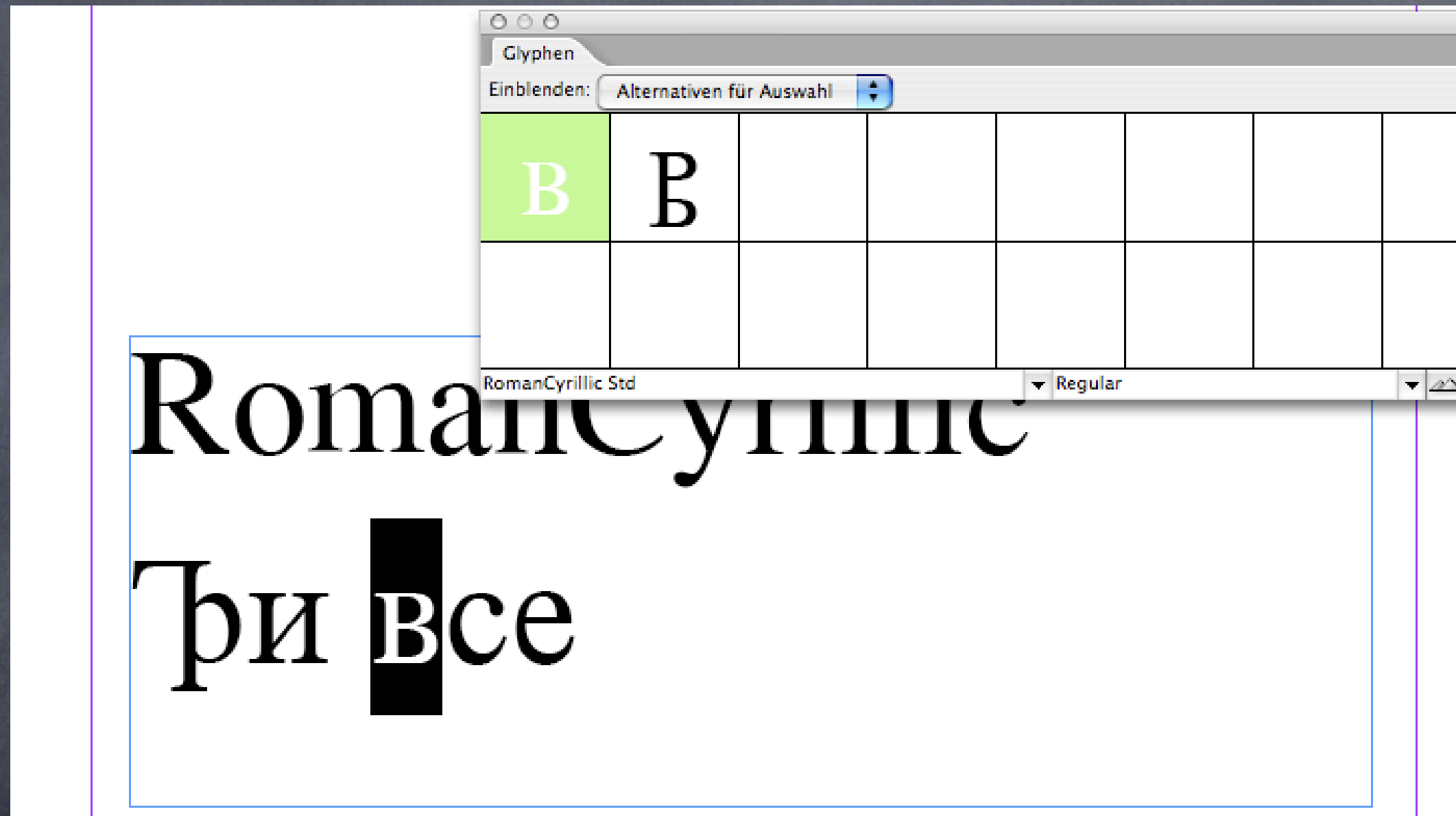
Presentation ☺   **Encoding** ☺

Alternate char. shapes

б = δ
б = δ

Serbian & Macedonian *cursive* variant forms.

Г = Г

г = ī

Russian norms on the left.
Serbian and Macedonian
norms on the right.

д = д

∂ = g

П = П

Note that only one variant,
the lowercase /be/ occurs
in both roman and italic styles.
The other variants occur
only in the italic style.

n = ū

Т = Т

m = ūū

Standard solution needed for cursive variants

OpenType Features

Pro...

- a lot

OpenType Features

Pro...

- a lot

Con...

- advanced features need application support
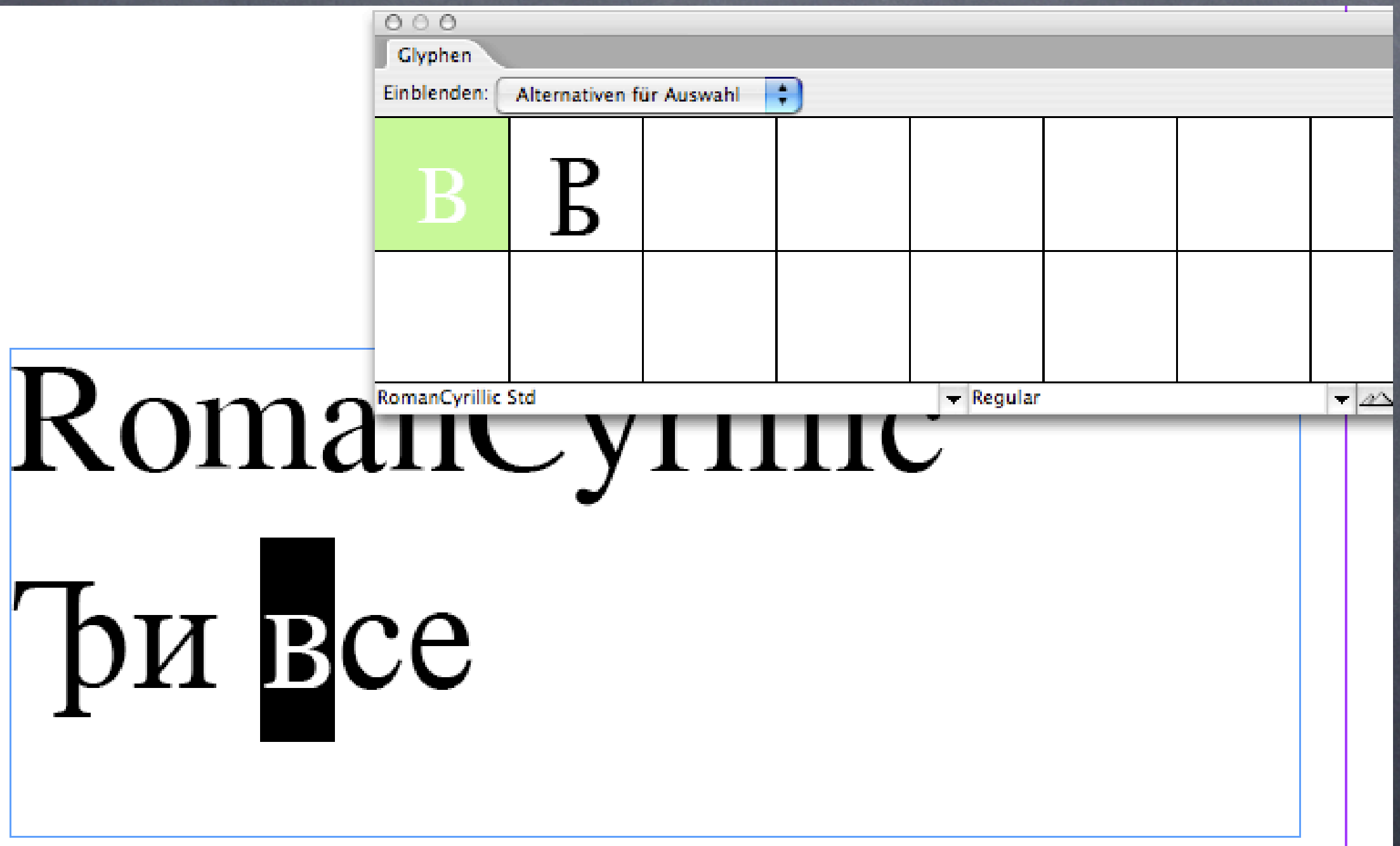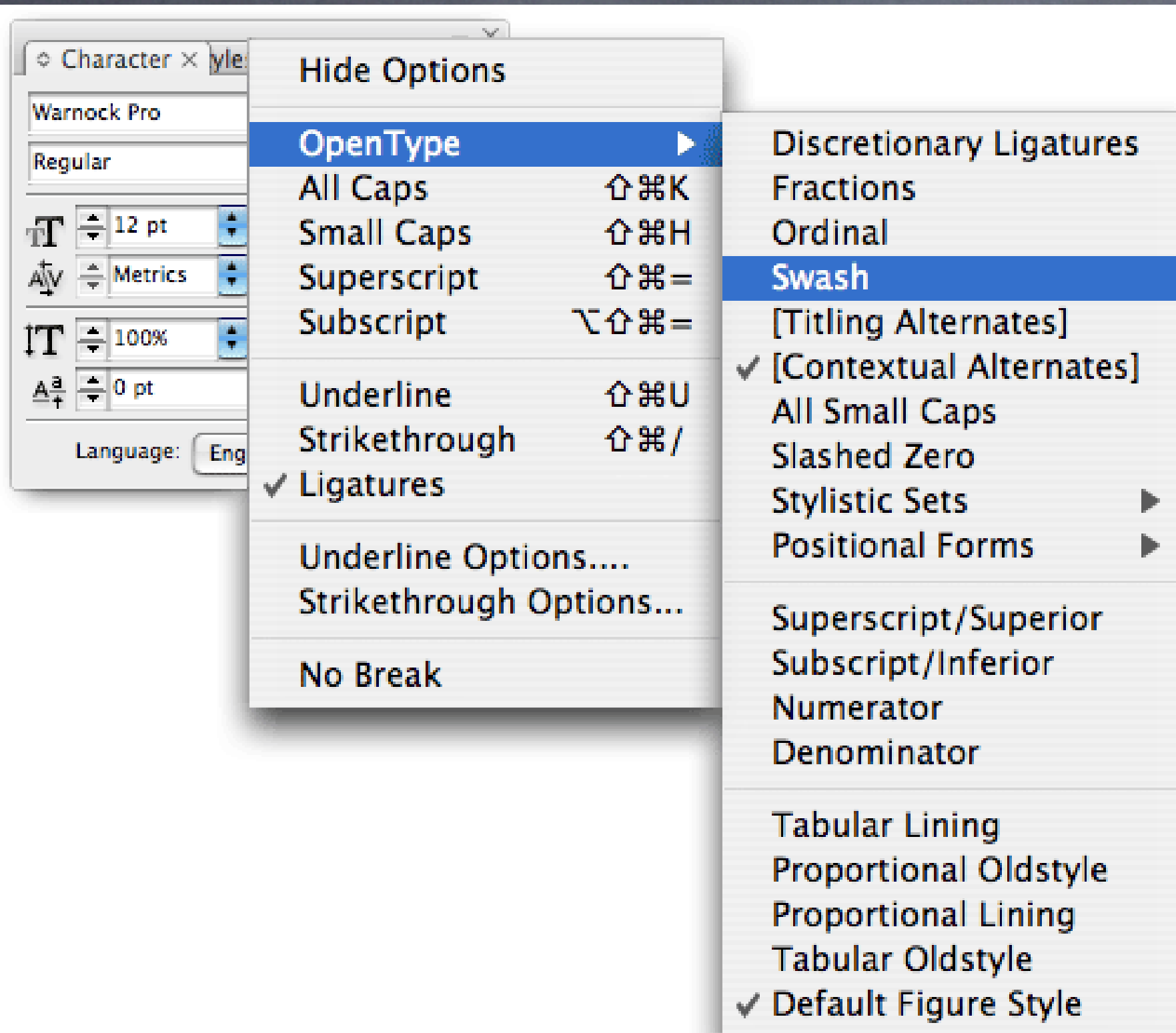
OT feature access
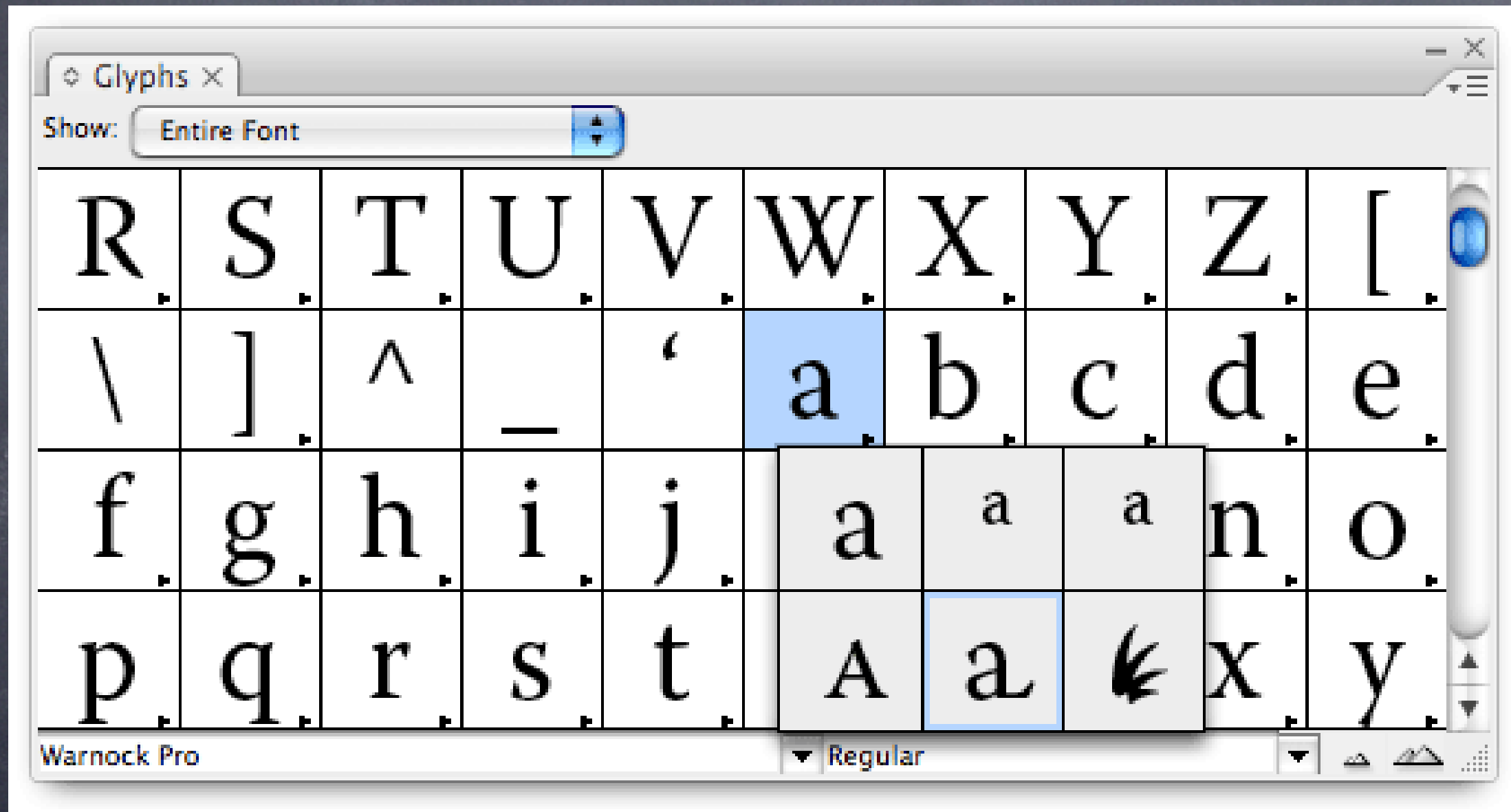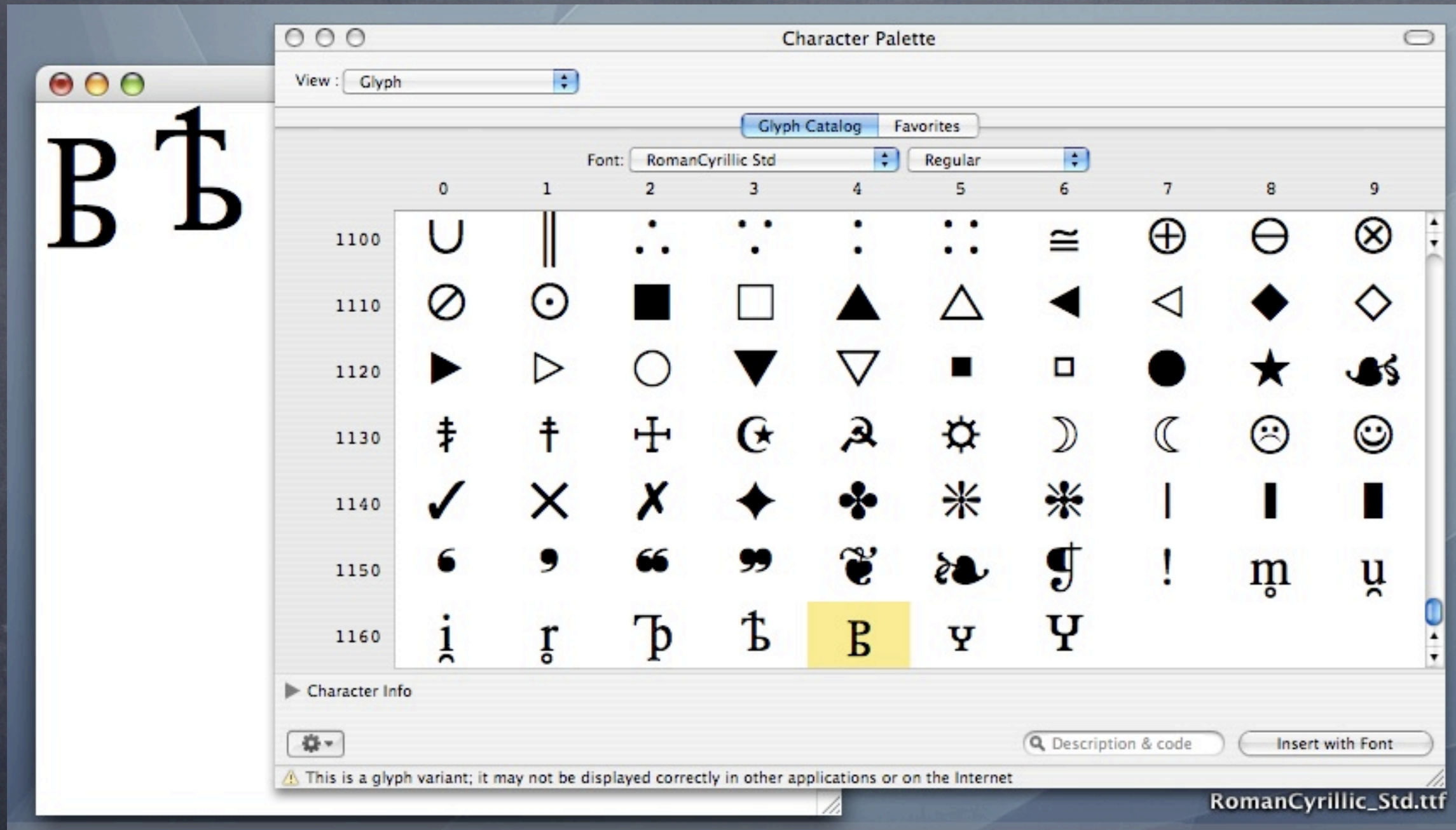

Adobe apps

OT feature access

OT feature access

Adobe apps

OT feature access

Character palette

OpenType Features

Good for...

★ **pure character variation**

★ presentation purposes

★ glyphs which have similar standard Unicode counterpart

★ **precise placement of diacritics**

OpenType Features

Good for...

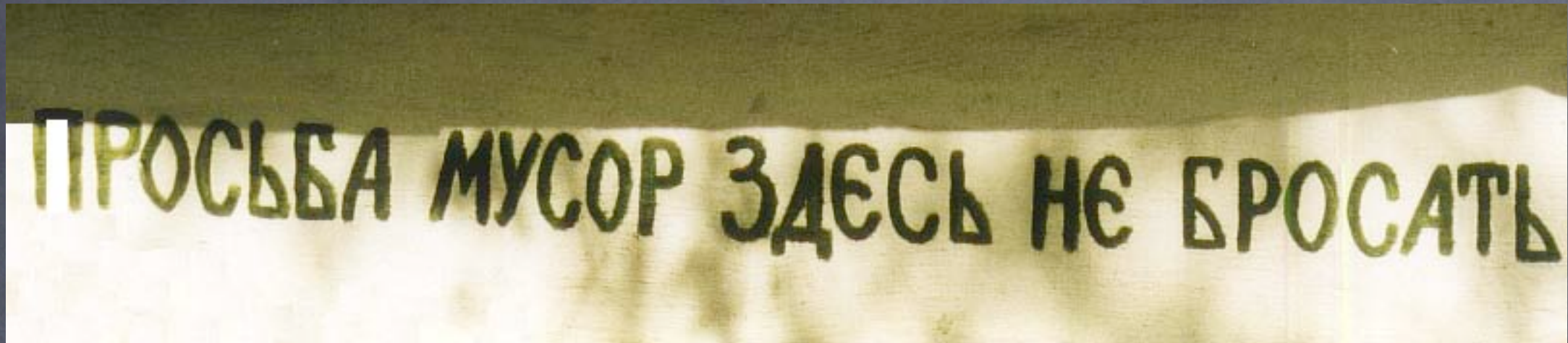**ONE Unicode diacritic – several shapes or positions realized**

OpenType Features

Not good for characters...

- which need their own Unicode identity
- which should have their own encoding

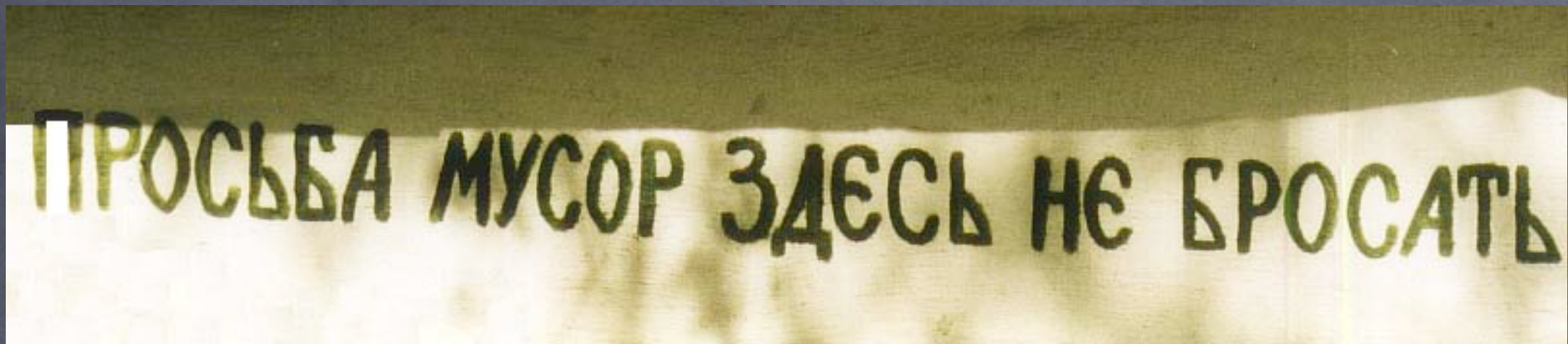# OpenType Technology

ПРОСЬБА МУСОР ЗДЕСЬ НЕ БРОСАТЬ

Solves some problems,
but not all!

OpenType Technology

Solves some problems,
but not all!

Solves some problems today,
but not all!

That was...

3. OpenType Features

Resume

★ **Basics**

★ **Do's and Don'ts**

★ **Expanding Unicode standard**

★ **Using alternate fonts**

★ **Use of 'Private Area'**

★ **Using OpenType Features (Ligatures, Diacritic Placement, Substitution, Character Variants etc.)**

Thankъ You!

Discussion

This presentation was given at a workshop on Unicode and Slavic Philology during the "Slovo" conference in Sofia in 2008.
The following paper was the printed outcome of this presentation:

Sebastian Kempgen: Unicode 5.1, Old Church Slavonic, Remaining Problems - and Solutions, including OpenType Features. In: Slovo. Towards a Digital Library of South Slavic Manuscripts. Proceedings of the International Conference, 21-26 February 2008, Sofia, Bulgaria. Sofia 2008, 200–219.

The paper is available online.