

Unicode for Slavic Medievalists

David J. Birnbaum (Pittsburgh, USA)

djbpitt+@pitt.edu

Ralph Cleminson (Portsmouth, UK)

ralph.cleminson@port.ac.uk

Pomorie, Bulgaria, September 2002

Part 2

Outline, Part 2

- Lumpers and Splitters
- Early Cyrillic Writing and Unicode
- The Text Encoding Initiative (TEI) Writing System Declaration (WSD)
- What's New: Variation Selectors and Early Cyrillic Writing

Lumpers and Splitters

- A few types of “e”: € € € € € € € €
- Handwritten letter forms *always* differ
- Approaches to classification
 - *Splitter*: If they’re at all different, they’re different, and should be encoded differently
 - Example: Retain difference in rendering
 - *Lumper*: If they’re at all similar, they’re the same, and should be encoded identically
 - Example: Conflate differences for querying, collation

Early Cyrillic and Unicode

- “The historic form of the Cyrillic alphabet is treated as a font style variation of modern Cyrillic because the historic forms are relatively close to the modern appearance and because some of them are still in modern use”
- Early Cyrillic “а” and modern Cyrillic “a” are both u+0430

Source Set Rule

Legacy Characters

| <i>Character</i> | <i>Image</i> | <i>Decomposition</i> |
|------------------|--------------|--|
| u+0477 | ṽ | u+0475 (v) u+030f (¨) |
| u+047f | Ṽ | u+0461 (w) u+0442 (̄) (superscript) |
| u+047d | Ṽ | u+0461(w) u+0483 (̄) |

Complementary Distribution (Non-Slavic)

| <i>Language</i> | <i>Phoneme</i> | <i>Spelling (lower case only)</i> |
|-----------------|---|-----------------------------------|
| Greek | /s/ | ς / ___# (u+03c2) |
| | | σ / elsewhere (u+03c3) |
| Hebrew | Final and nonfinal consonants | |
| Arabic | Initial, medial, final, and isolated consonants | |

Complementary Distribution (Slavic)

| <i>Language</i> | <i>Phoneme</i> | <i>Spelling (lower case only)</i> |
|---------------------|-------------------------|--|
| Rusian (some) | /ja/ ~ /҃a/ (~ /Cä/) | а / C __ ѡ / elsewhere |
| Rusian (some) | /o/ | о / C __ ѡ / elsewhere |
| Pre-1918 Russian | /i/ | і / __ V, ѣ (exc. мірѣ) и / elsewhere |

Visual Ambiguity

- Run-of-the-Mill Unicode Visual Ambiguity
 - u+002d HYPHEN-MINUS
 - u+2010 HYPHEN
 - u+2212 MINUS SIGN
- Super-Duper Early Cyrillic Visual Ambiguity
 - One character mapped to more than one glyph
 - One glyph associated with more than one character
 - u+0486 COMBINING CYRILLIC PSILI PNEUMATA (’)
 - u+0311 NON-SPACING INVERTED BREVE (ˆ)

Jotation

| <i>Character</i> | <i>Image</i> | <i>Notes</i> |
|------------------|--------------|---|
| u+0465 | ℥ | |
| u+044f | ƒ | |
| u+0469 | ƒ̣ | |
| u+046d | ƒ̥ | |
| u+044e | ƒ̇ | Jotation plus u+043e (◊) (!) |
| (none) | ƒ̈ | Jotation plus u+0463 (ƒ̣) |
| (none) | ƒ̉ | Δ not present in Unicode Variant of u+0469 (ƒ̣)? |

/i/

| <i>Character</i> | <i>Image</i> | <i>Numerical</i> | <i>Notes</i> |
|------------------|--------------|------------------|---|
| u+0438 | И | 8 | |
| u+0456 | і | 10 | |
| u+0457 | ї | 10 | |
| (none) | ı | 10 | Variant? Of what? |
| (none) | Ӏ | 10 (?) | Invented for transcriptions of Glagolitic |

Cyrillic Palatal Glides

- U+0458 CYRILLIC SMALL LETTER JE
(j)
 - Serbian, Macedonian
- U+0439 CYRILLIC SMALL LETTER
SHORT I (й)
 - Russian, Ukrainian, Belarusian, Bulgarian

Jers

| <i>Character</i> | <i>Image</i> | <i>Notes</i> |
|------------------|--------------|------------------------------------|
| u+044a | Რ | |
| u+044c | Ტ | |
| (none) | Უ | “Neutral” jer Variant? Of what? |

Jery

| <i>Part</i> | <i>Images</i> |
|-------------|----------------|
| First | Ƶ, Ъ |
| Second | н, і, і̇, л, ʟ |

- Non-ligated and ligated
- (Modern: u+044b Ы)

Other Special Problems

- Upper and Lower Case
 - Modern and early Cyrillic are the same script
 - Modern Cyrillic distinguishes case
 - Early Cyrillic typically does not distinguish case
- Ligation is productive
- Superscription may require different glyphs
 - “Recumbant r” (ѣ)

/u/

| <i>Character</i> | <i>Image</i> | <i>Note</i> |
|------------------|--------------|------------------------------------|
| u+0443 | ŷ | Modern image y |
| u+0475 | ʋ | Variously /ü/ or /u/ |
| u+0479 | oŷ | Horizontal digraph |
| u+043e | o | Sequence of two characters |
| u+0443 | ŷ | Second may alternatively be u+0475 |
| (none) | 8 | Vertical digraph (ligated) |

/o/

| <i>Character</i> | <i>Image</i> | <i>Notes</i> |
|------------------|--------------|----------------------------|
| u+043e | o | Narrow |
| u+047b | ○ | Broad |
| (none) | ⊙⊙∞ | Ocular (also “polyocular”) |

Cf. u+0461 ω (omega)

Cf. /e/

Greek

| <i>Greek</i> | | <i>Cyrillic</i> | |
|------------------|--------------|------------------|--------------|
| <i>Character</i> | <i>Image</i> | <i>Character</i> | <i>Image</i> |
| u+03c8 | ψ | u+0471 | Ψ |
| u+03be | ξ | u+046f | Ξ |
| u+03b8 | θ | u+0473 | Θ |
| u+03b1 | α | (none) | α |
| u+03b5 | ε | (none) | ε |

/e/

| <i>Character</i> | <i>Image</i> | <i>Notes</i> |
|------------------|--------------|---|
| u+0435 | e (modern) | CYRILLIC SMALL LETTER IE Most common modern <i>glyph</i> |
| u+0454 | є (modern) | CYRILLIC SMALL LETTER UKRAINIAN IE Most common early <i>glyph</i> |
| u+044d | э (modern) | |
| u+0465 | ѣ | |
| (none) | Є Ǝ Ɔ Ǝ ǎ ǎ | Variants? Of what? |

Palatal Consonants

| <i>Character</i> | <i>Image</i> | <i>Notes</i> |
|------------------|-----------------|--|
| u+0459 | љ (modern) | CYRILLIC SMALL LETTER LJE |
| u+045a | њ (modern) | CYRILLIC SMALL LETTER NJE |
| u+04a5 | ҥ | CYRILLIC SMALL LIGATURE EN GHE (Altay, Mari, Yakut) |
| (none) | Ӏ | Palatal /l/ |
| u+04a5 (?) | ӆ | Palatal /n/ |
| (none) | Ӈ | Palatal /d/ |
| (none) | (not available) | Palatal /m/ |

Old Church Slavonic and Russian

| <i>Character</i> | <i>Image</i> | <i>Sound</i> | |
|------------------|--------------|----------------------------|---------------|
| | | <i>Old Church Slavonic</i> | <i>Rusian</i> |
| u+044f | Ѧ | /e/ ~ /je/ | /ä/ |
| u+0469 | Ѧ | | |
| u+0467 | Ѧ | /ja/ | |

Front Nasal (Unicode)

| <i>Character</i> | <i>Image</i> | <i>Notes</i> |
|------------------|--------------|--|
| u+0467 | Ѧ | CYRILLIC SMALL LETTER LITTLE YUS |
| u+0469 | ѧ | CYRILLIC SMALL LETTER IOTIFIED LITTLE YUS |

Front Nasal (Manuscripts) 1

| <i>Manuscript</i> | <i>Nonjotated</i> | <i>Jotated</i> |
|--------------------------------|-------------------|----------------|
| Savvina Kniga | Ɱ, Ɱ (rarely) | Ɱ̣ |
| Zograph Folia | Ɱ | Ɱ̣ |
| Suprasliensis Šuck Psalter | Ɱ | Ɱ̣ |
| Hilandar Folia | Ɱ̣ | Ɱ |
| Ostromir Gospel | Ɱ̣ | Ɱ̣̣ |
| Preslav ceramic inscription | Ɱ | Ɱ̣̣ |

Front Nasal (Manuscripts) 2

Manuscripts with a Single Front Nasal Letter

| <i>Manuscript</i> | <i>Image</i> | <i>Notes</i> |
|----------------------------|--------------|----------------|
| Undol'skij Folia | ʌ | |
| Cyrillic Macedonian Folium | Δ | ʌ twice (/C__) |

ʌ may represent etymological front or back nasal

Mixed Corpora: Geographic

Old Church Slavonic u+0467 (Ɑ) may correspond to:

| <i>Character</i> | <i>Image</i> | <i>Recension</i> |
|------------------|--------------|------------------|
| u+044f | Ɑ | Rusian |
| u+0454 | € | Serbian |
| u+046b | Ɱ | Middle Bulgarian |

But: Not all Rusian Ɑ, Serbian €, and Middle Bulgarian Ɱ correspond to one another.

Diachronic Paleography: Ѧ and ꙗ

| <i>Image</i> | <i>Character</i> | <i>Period</i> |
|--------------|------------------|---------------|
| Ѧ | u+0467 | Early |
| я | u+044f | Modern |
| ꙗ | (none) | Early |

| | <i>East Slavic</i> | <i>South Slavic</i> |
|-------------|--------------------|---------------------|
| Sound | я = ꙗ = Ѧ | я = ꙗ (\neq Ѧ) |
| Paleography | я < Ѧ | |

The TEI WSD

- Text Encoding Initiative (TEI)
- Writing System Declaration (WSD)
 - Encoded as an “auxiliary SGML Document” (subdoc)
 - Used for two different purposes
 - Documentation
 - Processing
- Each text element is ...
 - Encoded as an *entity* in a document (e.g., &aos;)
 - Described in a `<form>` element in a WSD ...

The `<form>` Element

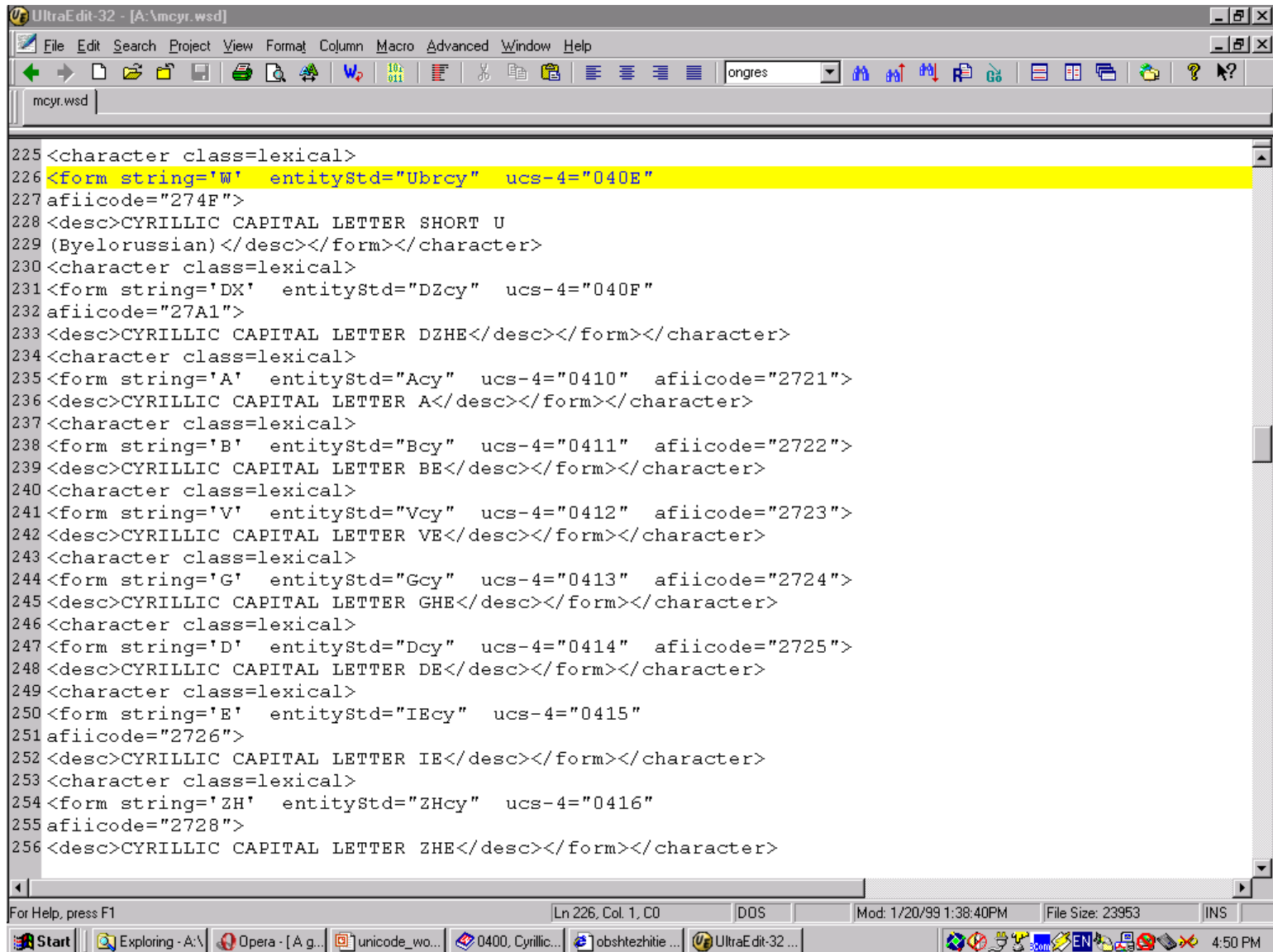
- Attributes of the `<form>` element
 - `string`: byte string
 - `codedCharSet`: base character set for string
 - `entityStd`: entity name for character
 - **`entityLoc`**: entity name for character
 - **`ucs-4`**: Unicode-related 32-bit identifier
- No glyph identifier
 - **`afiicode`** attribute removed from P4
 - RIP Association for Font Information Interchange (AFII)

TEI WSD (P3 Example)

```
<character class="lexical">  
<form    entityStd="aos"  
        ucs-4="0430"  
        afiicode="10993"  
</character>
```

(See *Computer Standards and Interfaces* 18 [1996]: 201–252)

A Real WSD



```
UltraEdit-32 - [A:\mcyr.wsd]
File Edit Search Project View Format Column Macro Advanced Window Help
mcyr.wsd
225 <character class=lexical>
226 <form string='W' entityStd="Ubrcy" ucs-4="040E">
227 afiicode="274F">
228 <desc>CYRILLIC CAPITAL LETTER SHORT U
229 (Byelorussian)</desc></form></character>
230 <character class=lexical>
231 <form string='DX' entityStd="DZcy" ucs-4="040F">
232 afiicode="27A1">
233 <desc>CYRILLIC CAPITAL LETTER DZHE</desc></form></character>
234 <character class=lexical>
235 <form string='A' entityStd="Acy" ucs-4="0410" afiicode="2721">
236 <desc>CYRILLIC CAPITAL LETTER A</desc></form></character>
237 <character class=lexical>
238 <form string='B' entityStd="Bcy" ucs-4="0411" afiicode="2722">
239 <desc>CYRILLIC CAPITAL LETTER BE</desc></form></character>
240 <character class=lexical>
241 <form string='V' entityStd="Vcy" ucs-4="0412" afiicode="2723">
242 <desc>CYRILLIC CAPITAL LETTER VE</desc></form></character>
243 <character class=lexical>
244 <form string='G' entityStd="Gcy" ucs-4="0413" afiicode="2724">
245 <desc>CYRILLIC CAPITAL LETTER GHE</desc></form></character>
246 <character class=lexical>
247 <form string='D' entityStd="Dcy" ucs-4="0414" afiicode="2725">
248 <desc>CYRILLIC CAPITAL LETTER DE</desc></form></character>
249 <character class=lexical>
250 <form string='E' entityStd="IEcy" ucs-4="0415">
251 afiicode="2726">
252 <desc>CYRILLIC CAPITAL LETTER IE</desc></form></character>
253 <character class=lexical>
254 <form string='ZH' entityStd="ZHcy" ucs-4="0416">
255 afiicode="2728">
256 <desc>CYRILLIC CAPITAL LETTER ZHE</desc></form></character>
For Help, press F1
Ln 226, Col. 1, CO DOS Mod: 1/20/99 1:38:40PM File Size: 23953 INS
Start Exploring - A:\ Opera - [A g... unicode_wo... 0400, Cyrillic... obshtezhitie... UltraEdit-32... 4:50 PM
```

How sdata Entities Work

1. Begin parsing document
2. Upon reaching WSD declaration, parse WSD and build three-column look-up table (entity, character, glyph) in memory or on disk
3. As each sdata entity is encountered during parsing:
 - a. Throw away regular entity replacement string
 - b. Use entity name associated with output sdata node as pointer into the table that you built at step #2
 - c. Retrieve and insert value from appropriate column

The WSD in Action

- Simple example
 - Encoding: `<p>Cyrillic &aos;</p>`
 - Rendering: Cyrillic а
- Complex example:
 - Encoding: `<p>Cyrillic &juos; &juros;</p>`
 - Rendering: Cyrillic 1 ъ
 - Same `ucs-4` value
(`u+044e` CYRILLIC SMALL LETTER YU)
 - Different `glyph (afii code)` values
 - Conflate or distinguish (lump or split), as needed

The WSD Solution

- `sdata` entities (e.g., `&aos;`) encode pointers into WSD, which records Unicode values and glyphs
- `sdata` encodings may represent many-to-many relationships among Unicode values and glyphs
 - Example: 1 ǃ are represented by different `sdata` entities, where the WSD `<form>` elements have the same `ucs-4` attribute but different `aficode` attributes
 - Similarity is encoded through shared character pointers
 - Difference is encoded through different glyph pointers

TEI and XML

- What happened to the WSD?
 - XML excludes subdoc
 - Alternatives are available: ndata entities, TEI header, in-line encoding
 - XML excludes sdata
 - Most types of entities have replacement text
 - SGML sdata entities represented by replacement text *in a special node type* in document object model (DOM) after parsing
 - XML resolves all entities to replacement text during parsing (no such special types)

The Problem

- The WSD encodes the fact that two items are simultaneously the same and different
- If the ability to encode that fact is to be retained under XML, it needs to be provided without subdoc and without sdata

Solutions

- Sdata
 - SGML only
- XML Character-Level Markup
 - Cumbersome
 - Doesn't work in attribute values
- Regular Characters (transliteration)
 - Ambiguous
- Unicode Variation Selector
- Unicode Private Use

Transliteration

- Published Early Cyrillic transliteration systems
 - *Polata* 1981
 - *Polata* 1987
 - Grünberg 1995
 - Cleminson 1997
 - Lazov 2000
- Requirements
 - Unambiguous
 - Reversible
 - Separation of content and markup

Unicode Variation Selector 1

- Introduced in Unicode Version 3.2
- Postposed combining character
- Specify glyphic variants of a common character (common semantics) in plain text
- Affects only appearance
- Uses must be codified, similarly to new characters
- Ignorable
- Commercial support?

Unicode Variation Selector 2

- Currently encoded for Mongolian and mathematics
 - Mongolian variation selectors
 - Sixteen generic variation selectors:
 - from: u+fe00 VARIATION SELECTOR-1
 - to: u+fe0f VARIATION SELECTOR-16
- Cyrillic proposal under preparation by Everson, Birnbaum, and Cleminson (based on Birnbaum 1996)

Unicode Private Use

- 6,400 16-bit private use characters
 - u+e000 to u+f8ff
- 131,068 surrogate pairs for private use
- “for defining user- or vendor-specific characters”
- Guaranteed never to be assigned by the Unicode consortium
- Commercial support?

Prior Agreement Required

- “Successful interchange requires agreement between sender and receiver regarding interpretation of private-use codes.”
- “These codes can be freely used for characters of any purpose, but successful interchange requires agreement between sender and receiver on their interpretation.”

Two Approaches to Private Use

1. Assert that text elements not (yet) present in Unicode are characters
 - PUA characters are intended to be rendered directly
 - Splitter Heaven
2. Replace old `sdata` entities as pointers into WSD
 - PUA characters are not intended to be rendered directly
 - PUA characters survive parsing, and are processed afterwards (e.g., by XSLT stylesheet) by mapping to some value in `<form>` element

Glagolitic and Unicode

“The Unicode Standard regards Glagolitic as a *separate* script from Cyrillic, not as a font change from Cyrillic. This position is taken primarily because Glagolitic appears unrecognizably different from Cyrillic, and secondarily because Glagolitic has not grown to match the expansion of Cyrillic. The Glagolitic script is not currently supported by the Unicode Standard.”

Notes on Glagolitic 1

- Round and square glagolitic are different “typographic variants” (fonts)
- **ꙗꙗ** is two characters, not one
- Stapić
 - Historically a variant of ꙗ
 - Graphically distinct
 - Treated as distinct by those who work with square Glagolitic

Notes on Glagolitic 2

- Round and square đerv
 - Historically the same letter
 - Different functions (đ and j)
 - Separate characters?
- Additional diacritics or punctuation?
- Upper and lower case?

Commission

- Special Commission to the Executive Council of the International Committee of Slavists for the Computer-Supported Processing of Slavic Manuscripts and Early Printed Books
- Birnbaum, Bojadžiev, Bojaniv'ska, Camuglia, Cleminson, Miltenova, et al.

Commission Projects

- Unicode Early Cyrillic Proposal
- Unicode Glagolitic Proposal
- Agreement (private standardization) of part of Unicode PUA for Slavic medievalist community

How to Participate

- **ОБЩЕЖИТІЄ Obštežitie Portal**
 - <http://www.ceu.hu/medstud/ralph/obsht.htm>
- **Mailing list for Early Slavic Written Sources**
 - slav-mss-list@port.ac.uk
 - Subscription information available at Obštežitie
- **Commission Web Site**
 - <http://clover.slavic.pitt.edu/~repertorium/commission/>
(after 15 October 2002)
- **Webmasters:**
 - David J. Birnbaum: djbpitt+@pitt.edu
 - Ralph Cleminson: ralph.cleminson@port.ac.uk