Changes to the IMPACT project Polish Ground-Truth texts

Janusz S. Bień

31 March 2012

1 File manipulation

1.1 Redundant files

The documents `00487409` and `00487410` (scans and XML files) has been removed because they represent the pages already encoded (in an equivalent or better way) as respectively `00487411` and `00487408` (pages 64 and 65 of *Zbior rytmow* …).

The documents `00436273.xml` and `00436274.xml` has been also removed for the same reason (duplicated pages 138 and 139 of *O cieplicach* …).

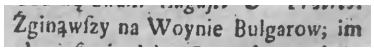
2 Characters

2.1 Single occurences

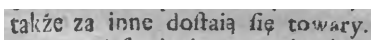
1. `00433129.xml`
   Encoding «*Bożym*» changed to «*Bożym*», i.e. the superflous character `COMBINING MACRON` (`0x000304`) removed, cf.

2. `00433408.xml`
   Encoding «*Z@ydom*» changed to «*Żydom*», cf.

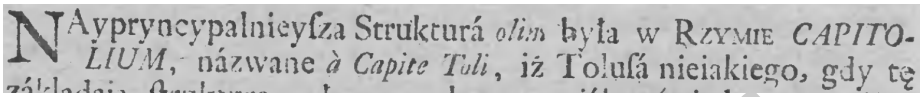
3. `00433461.xml`
   Encoding «*Z$ginąwſzy*» changed to «*Źginąwſzy*», cf.



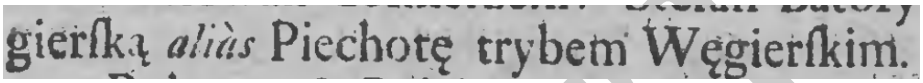
4. `00436088.xml`
   Encoding «*doſtai+*» changed to «*doſtaią*», cf.



5. 00434014.xml
   Removed OBJECT REPLACEMENT CHARACTER (0x00FFFC) occuring before «*LIUM*», cf.

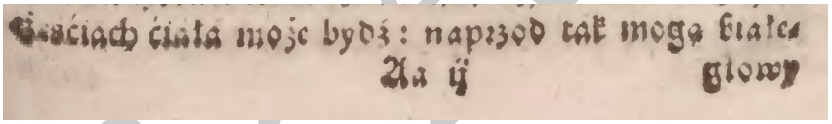


6. 00434343.xml
   Removed RIGHT-TO-LEFT MARK (0x00200F) occuring in «*aliàs*», cf.

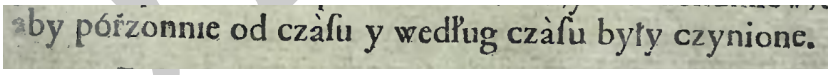


7. 00435950.xml
   The character LATIN SMALL LETTER D WITH CARON (0x00010F) replaced by the sequence of the characters LATIN SMALL LETTER D (0x000064) and APOSTROPHE (U+0027), cf.

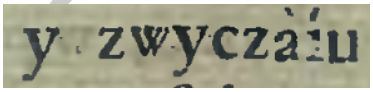
8. 00436324.xml
   Encoding «*mogą»*» changed to «*mogą*», i.e. the character RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK (0x0000BB) removed, cf.



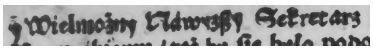
9. 00436368.xml
   The Private Use Area character 0x00F510 (LATIN SMALL LETTER R WITH MACRON ABOVE) replaced by the sequence of the characters LATIN SMALL LETTER R (0x000072) and COMBINING MACRON (0x000304), cf.



10. 00436453.xml
    Encoding «*zwyczàíu*» changed to «*zwyczàíu*», i.e. LATIN SMALL LETTER I (0x000069) replaced by LATIN SMALL LETTER DOTLESS I (U+0131), cf.



11. 00436772.xml
    The character CYRILLIC SMALL LETTER U WITH MACRON (0x0004EF) replaced by the sequence of the characters LATIN SMALL LETTER Y (U+0079) and COMBINING MACRON (0x000304), cf.

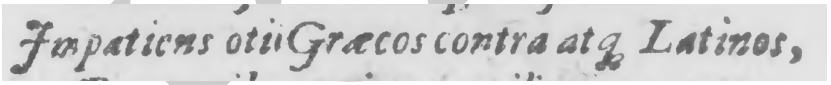


12. 00436776.xml
The Private Use Area character 0x00F509 (LATIN SMALL LETTER Q LIGATED WITH FINAL ET WITH OVERLINE) replaced by the sequence of the characters LATIN SMALL LETTER Q (0x000071), COMBINING MACRON (0x000304) and the MUFI character LATIN SMALL LETTER ET (A76B), cf.

The interpretation of the character has been discussed on the mailing list of Medieval Unicode Font Initiative (http://www.mufi.info/) on March 30, 2012. In particular Sebastian Kempgen explained that the proper reading of whole word is *quoque* and Susana Tavares Pedro confirmed that the proposed encoding is correct and noted:

> In fact, the abbreviation marks have both the same meaning: q+macron reads "que"; q+latin small letter et also reads "que". There is no such word as "quoqueue" so I would interpret this as a unnecessary duplication.

13. 00436848.xml
The character LATIN SMALL LETTER Q WITH DIAGONAL STROKE (0x00A759) replaced by the MUFI character LATIN SMALL LETTER Q LIGATED WITH FINAL ET (E8BF), cf.



The interpretation of the character has been discussed on the mailing list of Medieval Unicode Font Initiative (http://www.mufi.info/) on March 30, 2012. At first both Sebastian Kempgen and Susana Tavares Pedro suggested that this is a letter *q* with a tiny *e* (the whole word being *atque*), but later Sysana T. Pedro wrote:

> Actually, I looked closer at the first image you sent — because I never saw a small e taking the place of a UE sign — and I revised my opinion. It is a UE sign, not an e. The ink on the print is blurred, so the lower half of the sign looks like a blob instead of a downward curve, clockwise. The full sign should look like a 3. (originally it was a semicolon; cursive evolution afterwards caused the ligature between the top dot and the lower comma, hence the 3-shaped sign)

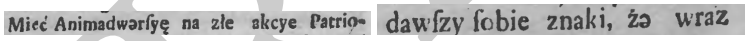
She also answered my question whether encoding it as LATIN SMALL LETTER Q LIGATED WITH FINAL ET is evidently wrong and/or misleading:

No, not really. MUFI guidelines do call it LATIN SMALL LETTER ET.
The comment adds: “used as suspension marks in a number of contexts, e.g. for ‘et’ in ‘videlicet’, for ‘us’ in ‘quibus’, ‘omnibus’, for ‘ue’ in ‘neque’, ‘cumque’, for ‘m’ in ‘nam’, ‘omnem’, for ‘est‘ in ‘prodest’, ‘interest’, etc. In the paleographical literature, it is usually described as “a sign looking like the number 3”. It is difficult to find a wholly suitable name; the name proposed here takes the abbreviation of ‘et’ in ‘videlicet’ to be prototypical. ”
Because of the blur, I can’t say whether is is ligated or not. A clean example would help but I understand you have only this one.

2.2 Multiple occurences

1. 2 occurences of the sequence LATIN SMALL LETTER I (0x000069) and COMBINING GRAVE ACCENT (0x000300) changed to LATIN SMALL LETTER I WITH GRAVE (U+00EC), namely encoding «*koniowí*» changed to «*koniowì*» and «*przynìoſt*» by «*przynìoſt*», cf.

2. 2 occurences of the character LATIN SMALL LETTER SCHWA (0x000259) changed to LATIN SMALL LETTER TURNED E (U+01DD) representing printing error, cf.



3. 2 occurences of the Private Use Area character 0x00F514 (LATIN SMALL LETTER W WITH BREVE) replaced by the sequence of the characters LATIN SMALL LETTER W (U+0077) and COMBINING BREVE (U+0306).

4. 4 occurences of the Private Use Area character 0x00F511 (LATIN SMALL LETTER S WITH MACRON ABOVE) replaced by the sequence of the characters LATIN SMALL LETTER S (0x000073) and COMBINING MACRON (0x000304).

5. 6 occurences of the character SPACE (0x000020) preceding COMBINING LATIN SMALL LETTER O (0x000366) replaced by NO-BREAK SPACE (00A0) to facilitate proper text segmentation.

6. 10 occurences of the Private Use Area character 0x00F501 (LATIN SMALL LETTER C WITH MACRON ABOVE) replaced by the sequence of the characters LATIN SMALL LETTER C (0x000063) and COMBINING MACRON (0x000304).

7. 10 occurences of the Private Use Area character 0x00F50B (LATIN SMALL LETTER L WITH APOSTROPHE) replaced by the sequence of the characters LATIN SMALL LETTER L (0x00006C) and APOSTROPHE (U+0027).

8. 11 occurences of the character BULLET (0x002022) replaced by the character MIDDLE DOT (U+00B7).

9. 11 occurences of the character BULLET OPERATOR (0x002219) replaced by the character MIDDLE DOT (U+00B7).

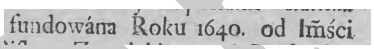
10. 14 occurences of the Private Use Area character 0x00F50A (LATIN SMALL LETTER D WITH APOSTROPHE) replaced by the sequence of the characters LATIN SMALL LETTER D (0x000064) and APOSTROPHE (U+0027).

11. 22 occurences of the Private Use Area character 0x00F50E (LATIN SMALL LETTER Q WITH ACUTE ACCENT) replaced by the sequence of the characters LATIN SMALL LETTER Q (0x000071) and COMBINING ACUTE ACCENT (U+0301).

12. 24 occurences of the Private Use Area character 0x00F50D (LATIN SMALL LETTER Q LIGATED WITH FINAL ET AND ACUTE ACCENT) replaced by the sequence of the MUFI characters LATIN SMALL LETTER Q LIGATED WITH FINAL ET (E8BF) and COMBINING ACUTE ACCENT (U+0301), cf. e.g

(the word in the example is *Actisque*).

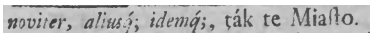
13. 29 occurences of the Private Use Area character 0x00F519 (LATIN SMALL LETTER M WITH TILDE) replaced by the sequence of LATIN SMALL LETTER M (0x00006D) and COMBINING TILDE (U+0303), cf. e.g



14. 72 occurences of the Private Use Area character 0x00F516 (LATIN SMALL LETTER Z WITH TILDE) replaced by the sequence of LATIN SMALL LETTER Z (0x00007A) and COMBINING TILDE (U+0303).

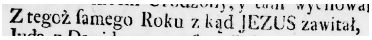
15. 120 occurences of the Private Use Area character 0x00F504 (LATIN SMALL LETTER G WITH RING ABOVE) replaced by the sequence of LATIN SMALL LETTER G (0x000067) and COMBINING LATIN SMALL LETTER O (U+0366), cf. e.g

16. 177 occurences of the Private Use Area character 0x00F50C (LATIN SMALL LETTER Q WITH ACUTE ACCENT ABOVE AND SEMICOLON ON THE RIGHT) replaced by the sequence of LATIN SMALL LETTER Q (0x000071), COMBINING ACUTE ACCENT (0x000301) and the MUFI character LATIN ABBREVIATION SIGN SEMICOLON (F1AC), cf. e.g



In the example the first occurence of the semicolon is LATIN ABBREVIATION SIGN SEMICOLON (the whole word being *aliusque*), but the original encoding as SEMICOLON (0x00003B) is not changed. Disambiguation of the semicolon glyphs should be done at a later stage with appropriate tools.

17. 664 occurences of the Private Use Area character 0x00F51D (LATIN SMALL LETTER Z WITH HOOK ABOVE) replaced by the sequence of LATIN SMALL LETTER Z (0x00007A) and COMBINING HOOK ABOVE (U+0309), cf. e.g



3 License

Acknowledgment

Besides Sebastian Kempgen and Susana Tavares Pedro already mentioned above I would like to thank also Izabela Wiencek for the verification of Latin abbreviations.