Transcription

The transcription is done in an html editor (e.g. MS Word). The transcriber can use basic formatting (underline and strikethrough). Certain things are encoded using special codes.

The resulting html text is then converted to the feat xml format by transcription supervisors while they review the transcription (the conversion is part of the feat editor or can be run independently).

Basic codes

- priv – text removed for privacy reasons (town, name, email, ...)
  I was born in {town}<priv>
- dt – pre-printed text
  {Once upon a time, there was}<dt> a boy who
- co – comment by the transcriber (to a single word, paragraph or the whole document)
  some tteeext{really hard to read}<co>

  {the whole document is really hard to read}<co doc>
- img – pictures in the text (with a short comment)
  My <img dog> is very nice.
- XXX<gr> – text in foreign script
  I was born in XXX<gr>.

Correction (corrections made by the author of the text)

- in – insertion
  Are {we}<in> there yet?

  Are we th{e}<in>re yet?
- {ab -> cd} ab was changed to cd
  {We are -> Are we} there yet?
- strike-through (deleted text) is transcribed as strike-through
  Are ~~woo~~ we there yet?
- || – splitting words
  Are||we there yet?
- tr – long distance word-order changes
  {Yet}<tr-from-1> are we there <tr-here-1>?

Uncertainty in transcription

- {v|w}e – either *ve* or *we*
  Are {v|w}e there yet?

  Are {ve|we} there yet?
- { } – uncertain space
  Are{ }we there yet?

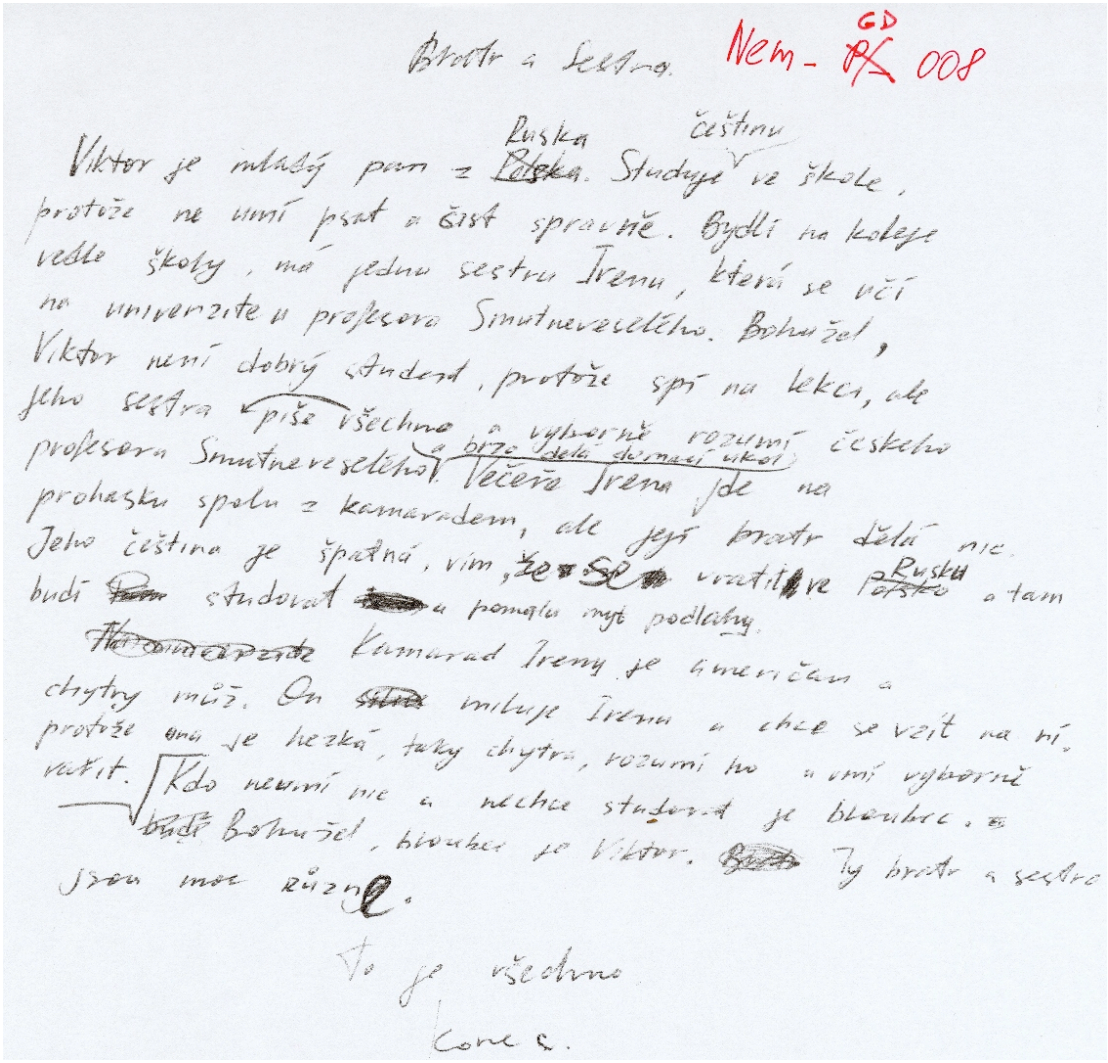
Encoding unusual diacritics

Texts written by foreigners often contain unusual diacritics. All diacritics can be written directly if the transcriber knows how. If not the following codes can be used:

[a“] – umlaut ä
[a:] – long umlaut
[c,] – cedilla
[c;] – ogonek
[z.] – dot above
[a^] – circumflex
[n~] – tilde
[L/] – crossed L
[e,] – accent grave
[m-] – line above
[ao] – ring above

The accents can be combined [n,o].

Example of a text with corrections



Bratr a Sestra.

Viktor je mladý pan z ~~Polska~~ Ruska. Studuje {češtinu}<in> ve škole, protože ne umí psat a čist spravně. Bydlí na koleje vedle školy, má jednu sestru Irenu, která se učí na univerzite u profesora Smutneveselého. Bohužel, Viktor není dobrý student, protože spí na lekci, ale jeho sestra {piše všechno -> všechno piše} a vyborně rozumí českeho profesora Smutneveseleho {a brzo delá domací ukol}<in>. Večeře Irena jde na prohasku spolu z kamaradem, ale její bratr dělá nic. Jeho čeština je špatná, vím, že se vratit ve ~~Polsko~~ Rusk~~o~~u a tam budí studovat u pomalu myt podlahy.

Kamarad Ireny je {A|a}meričan a chytry můž. On miluje Irenu a chce se vzít na ní. protože ona je hezká, taky chytra, rozumí ho a umí vyborně vařit.

Kdo neumí nic a nechce studovat je bloubec. ~~budi~~ Bohužel, bloubec je Viktor. Ty bratr a sestra jsou moc různ~~y~~e.

To je všechno.

Konec