# Torlak Transcription Guidelines

Teodora Vukovic

December 6, 2016

## 1 General principles

- Use symbols from Latin script for standard BCMS (with certain additions, see later). Write everything in lowercase (except accents, see later).
- Try to represent the language of the recordings as accurately as possible phonetically. Do not let the standard intervene!
- Try to be as precise as possible when it comes to marking time intervals for parts of transcriptions. There should not be empty/silent space at the beginning and at the end.
- Use as few non-alphabetical symbols as possible, other than the ones specified in the guidelines. Do not use the specified symbols for anything else.
- Do not use commas.
- Transcribe numbers with words.
- For word segmentation rely on standard Serbian. [1]
- Insert comments in an additional tier, if necessary.

## 2 File names

When naming transcript files, use the original recording file name. If necessary, specify if the transcript refers to video or audio file.

## 3 Speaker labels

When giving speaker labels in the transcript use the following scheme:
(Location - abbreviated to three letters)_(speaker number)
Example:
*first speaker from Selačka - SEL_1*

[1] Write post-positive articles and demonstratives as a part of the preceding word.

*third speaker from Debelica - DEB_3*

Example:
*RS_(speaker initials)*
Example:
*Biljana Sikimić - RS_BS*
*Svetlana Ćirković - RS_SĆ*

There is a document on Google Drive with speakers. List all the speakers from the same place together. (Add a new row when you add a new speaker.)

# 4 Segmentation

Segmentation into utterances - meaningful wholes, determined intuitively. Determine boundaries of text chunks according to pauses, intonation or overall structure and meaning.

# 5 Transcription

## 5.1 Phonology

- Phoneme elisions - no special marking, i.e. no reconstruction of full forms
- Accent - mark only the place of the accent. The accented vowel is marked with uppercase: e.g. 'danAs', 'kUća mi bilA dOle'
- Vowels
  - Semivowel - marked with a schwa:[2]
  - Diphthongs - 'uograda'
  - Other
    - Open 'e' - sounds like 'a', therefore transcribe as 'a': sometimes word 'selo' is pronounced as 'salo'
- Consonants
  - Palatalized *k* (one possibility: in etymological positions ć, elsewhere ḱ)[3]
  - Be careful with 'j', it is not always pronounced: e.g. 'edno' istead of 'jedno'

[2]Speed tip: use '**w**'.
[3]Speed tip: use '**q**'

  - Be careful with 'h/x', it is usually replaced or not pronounced at all: e.g. 'sarana' istead of 'sahrana', 'leb' istead of 'hleb', 'mej' istead of 'meh', 'kožuv' istead of 'kožuh'.
  - Pay attention to the cases where there is a 'v' instead of 'f' in standard Serbian: e.g. 'višek' istead of 'fišek', 'vuruna' istead of 'furuna'

Pay attention to other similar cases that might appear or phonological characteristics in general. Try not to let the standard intervene. (see General remark)

## 5.2 Prosody

- Pauses
  - Unfilled (silent) pauses - mark using middle-dot from HIAT keyboard panel in Exmaralda. Where relevant, differentiate between long and short pauses (e.g. ·, ··, ···; or '((3s))')
  - Filled pauses, i.e. hesitations: ah, aah, hmm , etc. When it occurs within a word mark it using ((?)): e.g. 'nee((?)) znam', 'mislim((?)) da je tako',
- long sounds - double or tripple letter, depending on the length: e.g. 'jaao', 'staari zavet', 'miliceee'
- intonation - high/low, rising/falling (only when relevant): 'a onda se ((on)high) pojavio'
- tempo, speed, volume, etc. (only when prominent) - ((fast)I had no idea about that!), ((loudly)what was that?), or labels such as *forte, piano*, etc.

## 5.3 Other

- Overlapping - mark the overlapping segments with square brackets.

  Example:
  A: i onda sam ja išla tamo [da vidim šta ima]
  B: [da baš tamo] a tamo nije bilo ničega.
- non-linguistic information[4] - ((laughter)), ((cough)), ambient sounds: ((wind)); circumstances, additional necessary information, etc.
- way of speaking - ((laughing) that is so funny), ((unsure) I don't know), ((loudly) zamisli)!
- Interruptions, self-corrections, reformulations or sudden stops - mark with a slash. e.g. 'Onda me je pit/ onda mi je rekao...'

[4]Use english for non linguistic comments within text.

- Unintelligible passages - mark using ((XXX)). For longer segments, specify duration ((XXX)4s)
- Uncelar word or a sequence - put word or sequence between hash-tags: #dobro to tamo bilo#, #nisam#

**General remark:** make notes in a shared Google document of any uncertainties or characteristics neglected in the guidelines. With every note give the name of the transcript and approximate (or exact) time on the recording. In cases of new transcription items that needs to be marked, use some special characters, such as *, #, %, once for phoneme-level units, and before and after the passage for words or longer parts of text, so that they could be easily found and automatically replaced later (e.g. #word#). Use only one for one category and use them systematically and consistently. Make notes about their use and ideally give a few examples.

# 6 Meta-data

## 6.1 Speaker Meta-data

Speaker meta-data should contain several fixed categories and some additional optional ones, depending on the speaker and the situation. Within some of the categories certain fixed values could be established, but the arbitrary ones should also be allowed. Lacking information could perhaps be reconstructed or guessed, another option could be to try to contact the informants and ask, or marked as 'not available'.

Use the following attribute schema:

- sex - male, female
- age
- Year of birth
- Residence (place where the person lives)
- Origin (place of birth)
- education - none, 4 grades, primary school, high school, technical school, university
- occupation

## 6.2 Recording Meta-data

Rescording meta-data should contain several fixed categories and some additional optional ones, depending on the recording and the situation. Within some of the categories certain fixed values could be established, but the arbitrary ones

should also be allowed. Lacking information could perhaps be reconstructed or guessed, another option could be to try to contact the informants and ask, or marked as 'not available'.

Use the following attribute schema:

- location - name of the place where the interview was recorded [5]
- geo-coordinates - longitude and latitude copied from Google maps.
- sub-region - predefined sub-regions within the area (Zaplanje, Budzak, Stara Planina, etc.)
- location 2 (setting) - a particular place where the interview was recorded, e.g. family house, restaurant, market, etc.
- text type / narrative genre - oral history, personal experience narrative, recipes, instructions,
- topics - a set of predetermined labels [6] (arbitrary ones are allowed), e.g. Christmas, wedding, childbirth, etc. in form of list that corresponds to the whole recording / transcript file

[5] In certain situations this can be the name of the place where the informant lives.
[6] Use the questionnaire.