Rapid Transcription Guidelines for Turkish Broadcast News[*]

Murat Saraçlar

May 1, 2012

Rapid transcription aims to generate a corpus for Turkish automatic speech recognition systems. The main goal is the accurate and quick transcription of speech by humans. Speech will be transcribed in terms of words used in written language, together with some non-speech information. For example, presence of noise, music or speech in the background will be noted. In addition, speech will be segmented into utterances when the speaker pauses. Finally, the speakers will be identified when possible.

1 Segmentation

The initial segmentations from an energy based automatic segmentation program will be corrected by the annotators. A few points that should be taken into account:

- Words should never be split into two.
- Each segment should contain speech from a single speaker and a single sentence.
- The segments should neither be too short nor too long
- A new segment should be started when there is a change in the acoustic environment (e.g. background music).

2 Transcription

Written Turkish language will be used for transcription. The transcription should use the written form and not the spoken form. If a word is not pronounced as it is written (e.g. a foreign word) this should be marked appropriately and the pronunciation should be provided in parentheses.

[*] Adapted from LDC HUB4 and Rapid Transcription Guidelines for English (`http://projects.ldc.upenn.edu/Transcription/quick-trans/index.html`). Translated from the original Turkish guidelines titled "Çabuk Yazılandırma Kılavuzu" by Murat Saraçlar.

Punctuation Punctuation, other than periods (.), question marks (?) and exclamations (!) indicating end of sentences, and apostrophes (') following proper nouns, need not be used.

Capitalization These should be used for beginning of sentences and proper nouns.

Truncation If the beginning or the end of a word is not heard, then these should be marked with a dash (-), e.g., -zartesi, pazarte-. In this case, if the transcriber is sure about what was said in the unheard portion, then this could be indicated in parentheses, e.g., (pa-)zartesi, pazarte(-si).

Special symbols These are not used. For example, "dolar" should be used instead of '$' and "yüzde" should be used instead of '%'. As indicated below, some of these symbols are used for other purposes.

Acronyms Acronyms pronounced as series of letters, should be marked using a tilde (˜) in the front. For example, ˜İETT, ˜TRT.
Acronyms pronounced as they are written should not be marked. For example, NATO.

Foreign words These are marked by a dollar sign ($) in the front. For example, $Washington. The correct foreign spelling should be used.
Foreign acronyms pronounced as series of letters should be marked using both a dollar and a tilde ($˜). For example, $˜ CIA.
For other foreign acronyms a dollar sign ($) is sufficient.
The actual pronunciation should be indicated in parentheses. For example, $Washington(Vaşington) or $Washington(Vaşingtın), $˜ CIA(si ay ey).

Letters Letters pronounced by themselves are marked by a tilde (˜). For example, '˜B' is pronounced as 'be'

Numbers Numbers are written as they are pronounced. For example, bin dokuz yüz doksan altı.

Hesitation and interjections These are marked with a percent (%). The standardized spellings are as follows:

- %ıı (duraksama - hesitation)
- %hı-hı (evet - yes)
- %ı-ıh (hayır - no)
- %aaa (şaşırma - surprise)
- %ooo
- %eee
- %ıı
- %mmm

- %off
- %ahh
- %vay

Noises These are marked with an asterisk (*):

- For human noises such as breaths, laughters, coughs, mouth noise etc., use *N
- For other noises such as door noises, squeaks, pops, clicks, use *G.

If these noises are not instantaneous and if they continue for a few words in the same segment, they are marked with a *G< at the beginning and a *G> at the end.

Unclear speech If speech is unclear, the best guess is put between double parantheses ((konuşma)). If the speech is unintelligible then empty double parentheses are (( )) used.

2.1 Please pay special attention to

- the use of apostrophes,
- the use of the conjunctions 'de' and 'da' and the suffixes '-de' and '-da'. The suffixes are part of the word whereas the conjunctions are separate words.
- the use of the conjunction 'ki' and the suffix 'ki'. Again, the suffix is part of the word whereas the conjunction is a separate word.
- the transcription of words and suffixes like 'ile', 'ise', 'iken', 'idi'. Here, the choice of the speaker should be followed. For example, annesiyle – annesi ile, öyleyse – öyle ise.
- the vowel alterations (a-ı, e-i). Here, the written form and not the spoken form should be followed. For example, instead of the spoken form "başlıyan", the written form "başlayan" should be used.

IMPORTANT The character encoding should be ISO-8859-9 (Latin5).