ILZE AUZIŅA,
ROBERTS DARĢIS,
GUNA RĀBANTE-BUŠA

LU Matemātikas un informātikas institūts
ilze.auzina@lumii.lv, roberts.dargis@lumii.lv, g.rabante@gmail.com

Fonētiski marķēts latviešu valodas runas korpuss

2013. gadā IT kompetences centra ERAF līdzfinansētā projekta "Informācijas un komunikāciju tehnoloģiju kompetences centrs" (finansēšanas līgums L-KC-11-0003) nozares pētījuma Nr. 2.9. "Runas korpusa izveide, principi, metodes un realizācija" laikā tika izveidots fonētiski marķēts latviešu valodas runas korpuss (Pinnis, Auzina & Goba 2014). Korpusa dati iegūti no ortogrāfiski marķētā latviešu valodas runas korpusa (apjoms — 100 stundas). Korpusā ietverti 67 runātāju balss ieraksti. Fonētiski transkribēti aptuveni 4 stundu ilgi ieraksti — galvenokārt ziņu raidījumi, publiskas diskusijas, intervijas un šovi.

Fonētiskajai transkribēšanai no ortogrāfiski marķētā latviešu valodas runas korpusa atlasīti īsākie pieejamie audioierakstu vienumi — frāzes. Iekļaušanai fonētiski marķētajā korpusā pirmajā posmā izvēlētas tikai tās frāzes, kurās nav fona trokšņu un izolētu trokšņu. Tas nozīmē, ka frāze (iespēju robežās) neietver pauzes, kas garākas par 0,3 sekundēm, fizioloģiskus trokšņus, pārteikšanos vai neskaidras runas daļas. Lai korpusā būtu pēc iespējas vairāk runātāju datu, no viena runātāja audiodatiem tika atlasīts minimālais skaits frāžu tā, lai tās kopumā ietvertu visus iespējamos fonēmu pārus.

Dati ir pierakstīti mašīnlasāmajā fonētiskajā transkripcijā, norādot fonēmas, to robežas un atsevišķus fonēmu variantus. Fonēmu apzīmēšanai izmantots latviešu valodas mašīnlasāmais fonētiskais alfabēts, kurš izstrādāts, izmatojot SAMPA (Speech Assessment Methods Phonetic Alphabet) mašīnlasāmo fonētisko alfabētu.

Sākotnēji dati marķēti manuāli skaņu apstrādes programmā WaveSurfer, bet, izmantojot jau fonēmu līmenī nomarķētos datus, lietota automātiska runas fonētiskās transkribēšanas programma. Automātiski iegūtā transkripcija manuāli pārbaudīta un labota.

Fonēmu varianti, kas atspoguļoti fonētiskajā transkripcijā, galvenokārt parāda: 1) pagarinātus vai garus līdzskaņus, 2) reducētus patskaņus, 3) nezilbiskus patskaņus.

Atsauces

Pinnis, M., Auziņa, I., & Goba, K. (2014). Designing the Latvian Speech Recognition Corpus. In Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14).

SAMPA computer readable phonetic alphabet [tiešsaiste]. Pieejams: http://www.phon.ucl.ac.uk/home/sampa/home.htm [skat. 2015. g. 31. maijā].