DEVELOPMENT AND USE OF THE ESTONIAN L2 CORPUS

Einar Meister, Lya Meister

Institute of Cybernetics at Tallinn University of Technology, Estonia
einar@ioc.ee, lya@phon.ioc.ee

Keywords: Estonian, L2 speech, corpus design, quantity opposition

1. INTRODUCTION

Corpus-based approach in L2 speech research is an increasing trend facilitated by technological progress enabling cheap storage of large amounts of speech data and by availability of free software tools for manual or (semi)automatic annotation, and analysis of annotated corpora. An ideal corpus of L2 speech should comprise a large variety of speakers with different native language backgrounds, different speaking styles, extensive annotations in terms of both segmental and suprasegmental phonology, and more than one target language [2].

The paper introduces the Estonian Foreign Accent Corpus (EFAC) [8] which aims to provide high-quality L2 speech data for studies of L2 phonology and for language technology developments. The corpus collection was initiated at the Institute of Cybernetics, Tallinn University of Technology in 2006. Currently, it includes speech recordings of 180 L2 speakers of Estonian representing 18 different native languages and the reference group of 20 native Estonian speakers. The corpus is annotated on word and segmental levels.

The research on Estonian L2 speech based on this corpus has been so far focused on the acquisition of Estonian vowel categories, and on L2 prosody, specifically on the acquisition of short/long categories and quantity contrasts.

2. CORPUS DESIGN

The corpus is designed for studies of L2 acquisition of the main categories of the Estonian phonological system – vowels, consonants, diphthongs, consonant clusters, and quantity oppositions.

The text corpus involves 130 neutral sentences including all Estonian vowels and frequent diphthongs as well as all consonants and frequent consonant clusters in two-syllable target words representing the Estonian quantity oppositions. The target words are embedded in short meaningful sentences of similar structure, e.g.:

- *Kaotasin **sada** krooni raha.* Q1: *sada* /sata/ 'hundred', nom.sg.
- *Palun **saada** talle artikli koopia.* Q2: *saada* /saata/ 'to send', sg.imperat.
- *Soovin **saada** kolme kuu aruannet.* Q3: *saada* /saa:ta/ 'to get'
- *Kaunis **lugu** kõlas raadiost.* Q1: *lugu* /luku/ 'story', nom.sg.
- *Uue **luku** paigaldus maksab sada krooni.* Q2: *luku* /lukku/ 'lock', gen.sg.
- *Vana **lukku** pole mõtet parandada.* Q3: *lukku* /luk:ku/ 'lock', part.sg.

The constructed sentences are rather simple in order to be easily comprehensible and readable for learners with intermediate knowledge of Estonian. In addition, the text corpus involves eight questions, two passages, and some prompts to elicit spontaneous speech (self-introduction, description of three pictures).

3. SPEAKER RECRUITMENT

The main criteria for L2 speaker selection were:

- proficiency level of Estonian (at least "intermediate" or higher, foreign accent must be perceived by native listeners),
- native language,
- age of learning Estonian (adult learners were preferred),
- no hearing and speaking disorder.

Ideally, it would be good to achieve balance by sex and age, but in reality it is rather impossible. A group of native speakers has also been recruited – 10 male and 10 female subjects from monolingual Estonian-speaking families living in the capital area.

Different recruitment schemes have been used – invitations have been distributed in local universities and newspapers, teachers of Estonian giving language courses for adults and teachers of Estonian working at different foreign universities have been approached, etc. All recruited speakers have filled a questionnaire containing questions about their age, native language, age of learning Estonian, where and how they have learnt the language and how often they use it, as well as self-assessment of their knowledge of Estonian ("elementary", "intermediate", "ad-

vanced" or "proficient"). All participants were paid a small amount of money.

The number of speakers in the L2 subject groups is: Russian – 50, Finnish – 30, Latvian – 20, German – 15, Lithuanian – 13, French – 13, Japanese – 6, Swedish – 6, Spanish – 5, English – 5, and Italian – 5; other language (Dutch, Slovak, Polish, Portuguese, Hindi, Azeri, and Irish) are represented by 1-2 subjects only. Attempts will be made to recruit more subjects so that all L2 subject groups will be represented by at least 10 speakers.

Among the L2 subjects ca. 60% are females and 40% males. Age of the L2 subjects ranges from 16 to 67 years (mean 29.4, median 26.1). Most L2 subjects started to learn Estonian at the age 18-30 years, some subjects at the age 7, and some at the age over 40. All subjects have studied several foreign language, mostly English, German, or French.

4. DATA COLLECTION

Majority of the speech recordings have been done in our own recording studio; several L2 speakers have been recorded at their home universities in Finland (Oulu, Helsinki and Turku), France (Paris), Austria (Vienna), Latvia (Riga), Lithuania (Vilnius). By now we have recorded 180 non-native speakers of Estonian representing 18 different language backgrounds and the reference group of 20 native Estonian speakers. Currently, the total duration of speech data is approximately 80 hours.

Technically, the recordings are mostly of high quality (sampling frequency 44.1 kHz, 16 bit, wav-format); most speakers were recorded with two condenser microphones (a close-talking mic and a desktop mic). During the recording prompts were displayed on a screen one by one, correctness of reading was monitored by a recording operator. On average, a recording session lasted 25-30 minutes.

5. ANNOTATION

The corpus is annotated in different ways: the whole corpus is segmented and labelled automatically on word and phone levels (by using an Estonian text-to-speech aligner), and the target words representing quantity oppositions are segmented manually (currently only for native Estonian speakers and for L2 speakers with Russian, Finnish, Latvian and Japanese backgrounds) using Praat software [1]. We proceed with the manual annotation since the automatic text-to-speech aligner trained on Estonian native speech does not provide satisfactory results in the case of L2 speech.

6. CORPUS-BASED STUDIES

The corpus constitutes a valuable resource for the studies of L2 acquisition of Estonian phonological categories. Estonian is a quantity language exploiting the duration cue for manifesting phonological quantity oppositions. The quantity oppositions can occur in vowels and diphthongs, and also in consonants and consonant clusters. The corpus allows us to study the acquisition of quantity contrasts by L2 subjects with different phonological relevance of duration in their L1, e.g:

- Russian and Spanish: none,
- English and German: only as a secondary cue for vowels,
- Swedish: a primary cue for most vowels and consonants,
- French: only some vowels and consonants,
- Italian: only consonants,
- Latvian and Lithuanian: only vowels,
- Finnish and Japanese: consonants and vowels.

So far, we have reported the results on the acquisition of short/long categories by Russian subjects [6] and quantity contrasts by native speakers of Russian, Finnish, and Latvian [9], [4], [3], [5], as compared to L1 speakers. The results reveal differences in L1 and L2 groups due to the different role of duration in subjects' L1.

In addition, we have studied the acquisition of Estonian vowel categories by Russian and Japanese subjects [7], [10].

In our future research we plan to study:

- how does native language type (quantity vs. non-quantity language) affect the production and perception of quantity contrasts,
- to what extent do L2 subjects use spectral and other prosodic cues (such as F0 and intensity) as secondary cues in the production and perception of Estonian quantity contrasts,
- the relationships between L2 production and perception,
- the role of orthography in the production of Estonian quantity oppositions,
- L2 speech rhythm.

As an ultimate goal, we hope to propose a new (or at least to extend an existing) L2 theoretical model that is adequately able to address prosodic features, especially duration.

7. AVAILABILITY

The corpus will be available via the Center of Estonian Language Resources (http://keeleressursid.ee/eng/) and via EU CLARIN infrastructure (http://www.clarin.eu/).

8. REFERENCES

[1] Boersma, P., Weenink, D. Praat: doing phonetics by computer [computer program]. Version 5.4.08, retrieved 25 March 2015 from http://www.praat.org/.

[2] Gut, U. 2009. *Non-native Speech. A Corpus-based Analysis of Phonological and Phonetic Properties of L2 english and German*. Peter Lang.

[3] Meister, E., Meister, L. 2013. Native and non-native production of Estonian quantity degrees: comparison of Estonian, Finnish and Russian subjects. Asu, E. L., Lippus, P., (eds), *Nordic Prosody: Proceedings of the XIth conference, Tartu 2012* Frankfurt am Main. Peter Lang Verlag 235–243.

[4] Meister, E., Meister, L. 2013. Production of Estonian quantity contrasts by native speakers of Finnish. *Interspeech 2013: 14th Annual Conference of the International Speech Communication Association* Lyon, France. 330–334.

[5] Meister, E., Meister, L. 2014. Estonian quantity degrees produced by Latvian subjects. *Linguistica Lettica* 22, 85–106.

[6] Meister, L., Meister, E. 2011. Perception of the short vs. long phonological category in Estonian by native and non-native listeners. *Journal of Phonetics* 39(2), 212–224.

[7] Meister, L., Meister, E. August 2011. Production and perception of Estonian vowels by native and non-native speakers. *Interspeech 2011* Florence, Italy. 1145–1148.

[8] Meister, L., Meister, E. 2012. Aktsendikorpus ja võõrkeele aktsendi uurimine. *Keel ja Kirjandus* 55(8–9), 696–714.

[9] Meister, L., Meister, E. 2012. The production and perception of Estonian quantity degrees by native and non-native speakers. *Interspeech 2012: 13th Annual Conference of the International Speech Communication Association* Portland, Oregon. 886–889.

[10] Nemoto, R., Meister, E., Meister, L. 2015. Production of Estonian vowels by Japanese subjects. *18th International Congress of Phonetic Sciences* Glasgow, UK. xx–xx.