Development of a Large Spontaneous Speech Database of Agglutinative Hungarian Language

Tilda Neuberger, Dorottya Gyarmathy, Tekla Etelka Gráczi, Viktória Horváth, Mária Gósy, and András Beke

Research Institute for Linguistics of the Hungarian Academy of Sciences
Departement of Phonetics, Benczúr 33, 1068 Budapest, Hungary
{neuberger.tilda,gyarmathy.dorottya,graczi.tekla,
horvath.viktoria,gosy.maria,beke.andras}@nytud.mta.hu

Abstract. In this paper, a large Hungarian spoken language database is introduced. This phonetically-based multi-purpose database contains various types of spontaneous and read speech from 333 monolingual speakers (about 50 minutes of speech sample per speaker). This study presents the background and motivation of the development of the BEA Hungarian database, describes its protocol and the transcription procedure, and also presents existing and proposed research using this database. Due to its recording protocol and the transcription it provides a challenging material for various comparisons of segmental structures of speech also across languages.

Keywords: database, spontaneous speech, multi-level annotation.

1 Introduction

Nowadays the application of corpus-based and statistical approaches in various fields of speech research is a challenging task. Linguistic analyses have become increasingly data-driven, creating a need for reliable and large spoken language databases. In our study, we aim to introduce the Hungarian database named BEA that provides a useful material for various segmental-level comparisons of speech also across languages. Hungarian, unlike English and other Germanic languages, is an agglutinating language with diverse inflectional characteristics and a very rich morphology. This language is characterized by a relatively free word order. There are a few spoken language databases for highly agglutinating languages, for example Turkish [1], Finnish [2]. Language modeling of agglutinating languages needs to be different than modeling of languages like English [3]. There are corpora of various sizes, different numbers of speakers and diverse levels of transcription. TIMIT Acoustic-Phonetic Continuous Speech Corpus was created for training speaker-independent speech recognizers. This database consists of sentence reading from 630 American English speakers; includes time-aligned orthographic, phonetic and word transcriptions [4]. The Verbmobil database (of 885 speakers) was developed also in the 90's with speech technological purposes [5]. The spoken part of the British National Corpus (100 million words) [6] consists of informal dialogues that were collected in different contexts, ranging from formal business or government meetings to radio shows. The London–Lund Corpus contains 100 texts

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 424–431, 2014.

of spoken British English. The basic prosodic features, simultaneous talk, contextual comment (laughs, coughs, telephone rings, etc.) were marked in the annotation [7]. The Switchboard corpus [8] includes 2,400 telephone dialogues of 543 American English speakers. It was developed mostly for the applications in speaker identification and speech recognition. There are also some corpora of audio and transcripts of conversational speech, such as HCRC map task corpus [9] or Buckeye corpus [10], and natural meetings, such as ICSI (International Computer Science Institute) Meeting Corpus [11] or AMI (Augmented Multi-party Interaction) Meeting Corpus [12]. Although the earliest databases had consisted of written and spoken English texts, new corpora were developed also in other languages in the past decades (e.g, the German Kiel Corpus [13], Danish spoken corpus [14]. The CSJ (Corpus of Spontaneous Japanese) is one of the largest databases; it contains 661 hours of speech by 1,395 speakers including 7.2 million words [15]. EUROM1 [16] and BABEL [17] are multilingual databases, containing samples of various languages giving possibility to compare the phonetic structures of these languages using similar materials and recording protocols in all languages. Recordings of spoken Hungarian were first compiled at the beginning of the twentieth century; unfortunately, this material was destroyed. Various types of dialectical speech materials were recorded in the 1940s; these recordings were archived in the late nineties and are available for studying at the Research Institute for Linguistics of the Hungarian Academy of Sciences, RIL. The Budapest Sociolinguistic Interview contains tape recorded interviews with 250 speakers (2–3 hours each) made in the late eighties [18]. The Hungarian telephone speech database (MTBA) is a speech corpus containing read speech recorded via phone by 500 subjects. It was designed to support research and developments in the fields of speech technology [19]. The HuComTech Multimodal Database contains audio-visual recordings (about 60 hours) of 121 young adult speakers that represent North-East Hungary [20]. The developing of the largest Hungarian spontaneous speech database, BEA (the abbreviation stands for the letters of the original name of the database: BEszélt nyelvi Adatbázis ‘Speech Database ‘Speech Database’) started at the Phonetics Department of RIL in 2007. This database involves a great number of speakers who speak relatively long, contains various styles of speech materials, and has various levels of transcriptions.

2 Database Specification

At the moment of writing this paper, the total recorded material of BEA comprises 333 recordings, meaning 300 hours of speech material (approximately 4,500,000 words). The shortest recording lasts 24 minutes and 27 seconds, the duration of the longest is 2 hours, 24 minutes and 47 seconds; the average length is 51 minutes (SD: 15.8). The majority of them appear between 40 and 60 minutes. Speech materials from 184 female and 149 male speakers are available at the moment. For each recording, the following data are documented: the participant’s age, schooling, job, height, weight, whether s/he is a smoker. The youngest participant is 19 years old, while the oldest one is 90 years old. The mean age of speakers is 39 years (SD: 18.8). The majority of the participants are in their twenties and thirties.