Is Spoken Hungarian Low-resource?: A Quantitative Survey of Hungarian Speech Data Sets

Péter Mihajlik∗,†, Katalin Mády∗, Anna Kohári∗, Fruzsina Sára Vargha∗, Gábor Kiss†, Tekla Etelka Gráczi∗, A. Seza Doğruöz‡

∗HUN-REN Hungarian Research Centre for Linguistics
mihajlik.peter, mady.katalin, kohari.anna, graczi.tekla.etelka@nytud.hun-ren.hu

†Budapest University of Technology and Economics
Departure of Telecommunications and Media Informatics
kiss.gabor@tmit.bme.hu

‡LT3, IDLab, Universiteit Gent
as.dogruoz@ugent.be

Introduction

- Hungarian is in the top 20 content languages of websites: https://w3techs.com/technologies/overview/content_language
- Still, finding Hungarian speech data sets with transcription, e.g., for Automatic Speech Recognition, is difficult.
- Key data set features: *total duration in hours / total number of speakers* are not typically available in Hub pages.
- No relevant foundational models for Hungarian – in opposite to text-based AI models like Hungarian BERT, LLM's, etc.

Related work

Even the most up-to-date and detailed **survey** on Hungarian language technology mention only a few Hungarian speech data sets without quantitative features:

- Kinga Jelencsik-Mátyus, Enikő Héja, Zsófia Varga, and Tamás Váradi. 2023. Language report Hungarian. In European Language Equality: A Strategic Agenda for Digital Language Equality, pages 155–158. Springer:

https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_18__Language_Report_Hungarian.pdf

Language data repositories:

- https://catalog.elra.info/
- https://catalog.ldc.upenn.edu/
- https://live.european-language-grid.eu/

Major issues: expired links, missing data sets, quantitative features not always available, etc.

About Hungarian

Characteristics of the Hungarian Language

- Hungarian belongs to the **Uralic language family** along with Finnish and Estonian.
- spoken by ~12.5 million native speakers.
- agglutinating language, rich morpho-phonology
- vowel harmony
- latin alphabet extended with diacritics (for vowel),
- word stress on first syllable

Spoken Hungarian Data Sets

With transcription

Methodology

- Search: Internet + Personal research network
- Criteria:
  - Hungarian Speech
  - Transcription available
  - Accessible for research
  - Published
  - (longer than 1 hour)

- Quantitative features (hours, speakers):
  - Based on data owners declaration
  - Only transcribed parts are considered
  - Conservative estimates (~)

Monolingual Data Sets

- Significant proportion spontaneous
- Most of them related for (socio)linguistic research and limited size
- Accessibility issues (e.g. BUSZI)
- Typically wide-band recordings

Name	[hours]	Speakers	Key Features
BUSZI	~600	250	Sociolinguistic Interviews recorded with magnetofon. Kontra and Váradi (1997)
Oasis Numbers	3	26	Domain-specific corpus of numbers, 5857 recorded words. Kocsor et al. (2000)
MTBA	~5	500	Mostly read telephone speech. Vicsi et al. (2002)
MRBA	~6	332	Phonetically balanced read text. Vicsi et al. (2004)
BME Broadcast	3	~10	TV broadcast news. Teleki et al. (2005)
Szeged Broadcast	28	n.a.	News broadcast from 8 tv channels. Gosztolya and Tóth (2010)
HuComTech	~50	112	Audioviusal, conversational and read speech of students. Pápay et al. (2011)
BEA Release 1	65	115	Studio, various speech types: conversational, monolog, read, repeated, etc. Gósy (2012)
Kivi	1	45	Short guided monologues. Kugler (2015)
SzöSzi	370	163	Sociolinguistic interviews with Szegedian speakers, partially transcribed. Kontra et al. (2016)
BEKK	20	56	Spontaneous conversations of students recorded in dormitories by smartphones. Bodó et al. (2017)
HuTongue	~500	15	Conversations of reality show characters recorded with head microphones. Szabó and Galántai (2017)
StaffTalk	101	20	Spontaneous conversations of teachers recorded through smartwatches. Szabó et al. (2021)
Akaka Maptask Corpus	5	46	Task-oriented dialogues recorded via head-mounted microphones. Molnár et al. (2023)
Budapest Games Corpus	9	12	36 task-oriented dialogues recorded via head-mounted microphones. Mády et al. (2023)
ForVoice120+	32	120	Various speech tasks for speaker identification. Sztahó and Fejes (2023)
Total approximately	**1798**	**1822**	

Table 1: Monolingual Hungarian Data Sets

Hungarian Data Sets as Parts of Multilingual Databases

Name	[hours]	Speakers	Key Features
BABEL	1	72	Clear speech, read text. Vicsi and Vig (1998)
SpeechDat(E)	65	1000	Read telephone speech, ELRA. Pollák et al. (2000)
SPEECON-1	~200	555	Read speech recorded through 4 microphone channels, ELRA. Speecon Consortium (2000)
CSLU: 22 Languages	~4	300	Prompted telephone speech, LDC. Lander (2005)
CSS10	10	1	Free audiobook. Park and Mulc (2019)
CommonVoice	151	1603	Read sentences, online collection. Ardila et al. (2019)
MaSS	~21	1-25	Sentence-aligned Spoken Utterances Extracted from the Bible. Boito et al. (2019)
VoxPopuli	63	143	EU parliamentary speech. Wang et al. (2021)
FLEURS	~12	n.a.	Read sentences from wikipedia, multi-purpose. Conneau et al. (2023)
HUN_ASR001_CN	286	254	Scripted speech data recorded via smartphones. Appen_China
Total approximately	**813**	**3928**	

Table 2: Hungarian Data Sets as Parts of Multilingual Collections

- Mostly read speech
- Considerable amount of telephone speech
- Open data sets available

Name	[hours]	Speakers	Key Features
HPSDB	~16	308	Several types of speech, Parkinson's disease and healthy control. Kiss et al. (2021)
LAPASDA	~7	~400	Read text and sustained sounds; dysphonia and healthy control, RBH scores. Sztahó et al. (2021)
DEPISDA	~17	400	Read text, spontaneous; depression and healthy control, BDI-II scores. Hajduska-Dér et al. (2022)
HuMenDisCo	n.a.	90	Spontaneous; bipolar, schizophrenia and schizoaffective disorder and healthy control. Szabó et al. (2023)
Total approximately	**40**	**1198**	

Table 3: Hungarian Pathological Speech Data Sets

Data collected to monitor/diagnose menthal health disorders.

- Wideband recordings
- High SNR
- Healthy control included

Pathological Speech

Child(-related) Speech

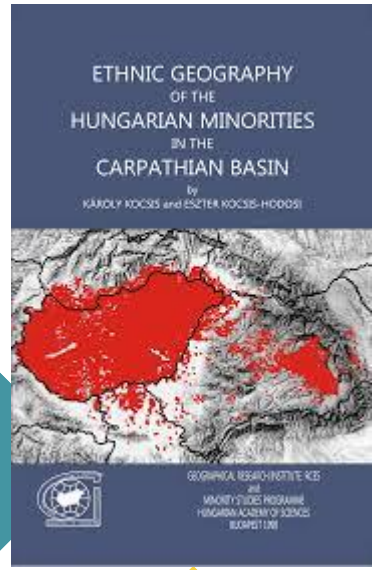
Name	[hours]	Speakers	Key Features
SPECO	~15	72	5-10 years old children, read speech and sustained segments. Csatári et al. (1999)
SPEECON-2	~10	50	8 to 15-year-old children, home environment, read sentences, toy commands, etc. Iskra et al. (2002)
CHILDES	~13	1	A boy between the ages 1;11 and 2;11, freeplay, home environment. Réger (2004)
Monyek (Eng. HUKILC)	~20	62	4.5–5.5 ages old children, 20–30 minutes of talks. Orosz and Mátyus (2014)
TiniBea	~13	18	8 female, 10 male speakers of 16–18 years, similar protocol with BEA. Gyarmathy and Neuberger (2015)
GABI	~50	~100	3-18 years old speakers, various speech types. Bóna et al. (2019)
HIDS	~9	68	Infant-directed speech, longitudinal; read text, semi-spontaneous storytelling. Kohári and Mády (2023)
Total approximately	**117**	**353**	

Table 4: Hungarian Child-Related Data Sets

- Mostly child speech
- Some infant-directed speech (from mother)
- Teenagers included

Dialectical Speech

- Mostly from the Carpathian basin
- Hungarian minorities from
  - Slovakia
  - Romania



https://www.mtafki.hu/konyvtar/kiadv/Ethnic_geography.pdf

Name	[hours]	Speakers	Key Features
MNyHk	13	~110	Interviews from 79 locations in and around Hungary. Vargha (2007)
VIHk	1	18	Interviews from 17 locations in Transylvania (Romania). Fazakas (Gál) (2013)
SzMNyHk	3	34	Interviews in Hungarian dialects from Slovakia. Presinszky (2020)
MoMa	18	~100	Interviews in Moldavian (Romanian) Hungarian (Csángó). Eris et al. (2023)
Total approximately	**35**	**262**	

Table 5: Hungarian Dialectal Speech Data Sets

Overview & Conclusions

- Total duration [hours]: 2.800
- Total number of speakers: 7.500
- Versatile and diverse collection
- Applicable for Hungarian foundational speech model
- Unification of transcriptions poses a significant challenge
- Data type distribution not balanced
- More (accessible) data is needed…

Duration distributions Speaker distributions

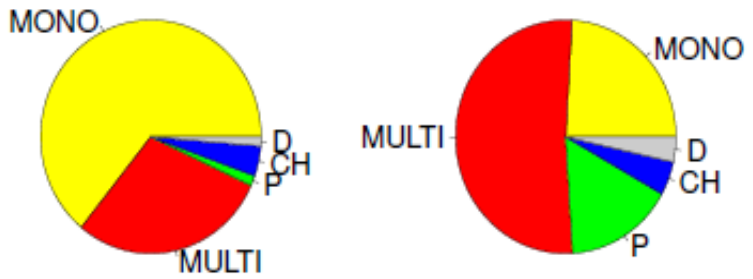


Figure 1: Types of Hungarian data sets. Left: Length in hours, Right: Number of speakers. MONO: monolingual, MULTI: Hungarian from multilingual collections, P: pathological, CH: child-related, D: dialectal.

Thank you

Questions:

mihajlik.peter@nytud.hu

Hungarian researchers were supported by by the NKFIH K143075, K135038 and NKFIH-828-2/2021(MILAB) projects of the NRDI Fund.