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

Abstract

Even though various speech data sets are available in Hungarian, there is a lack of a general overview about their types and sizes. To fill in this gap, we provide a survey of available data sets in spoken Hungarian in five categories (e.g., monolingual, Hungarian part of multilingual, pathological, child-related and dialectal collections). In total, the estimated size of available data is about 2800 hours (across 7500 speakers) and it represents a rich spoken language diversity. However, the distribution of the data and its alignment to real-life (e.g. speech recognition) tasks is far from optimal indicating the need for additional larger-scale natural language speech data sets. Our survey presents an overview of available data sets for Hungarian explaining their strengths and weaknesses which is useful for researchers working on Hungarian across disciplines. In addition, our survey serves as a starting point towards a unified foundational speech model specific to Hungarian.

Keywords: Hungarian, Speech Data, Low-resource Languages, Transcription

1. Introduction

Although (written) Hungarian is one of the most used 25 languages across websites[1], annotated spoken language data are relatively scarce. Hungarian speech data with time-aligned transcriptions is difficult to find in a quantity sufficient to train a contemporary Automatic Speech Recognition (ASR). Recent advancements in Self-Supervised Learning (SSL) (Baevski et al., 2020) allow for high accuracy speech models using limited supervised (transcribed) speech data (for fine-tuning), but state-of-the-art ASR systems still use several thousand hours of transcribed speech (Radford et al., 2022). In our survey, we provide a comprehensive quantitative overview of Hungarian labeled speech data sets accessible (at least) for research purposes. The amount of available data seems impressive both in terms of variability and quantity if compared to popular English ASR benchmark data sets such as (Godfrey and Holliman, 1993) or (Panayotov et al., 2015). However, unifying the annotations/transcriptions for a potential foundational speech model poses significant challenges. Nevertheless, we believe that such an overview is useful not only for the Speech Community and for an interdisciplinary LREC audience, but it will also pave the road to a unified Hungarian data set/speech model.

2. Related work

There is a wide range of data available on the Internet for English, but not much is known about the size and types of data sources available for low and mid resource languages. Therefore, there is a need for surveys and/or overview articles on less resourced languages (e.g., Eszter et al. (2012), Jelencsik-Mátyus et al. (2023) for Hungarian and Çöltekin et al. (2023) for Turkish, Rehm et al. (2014) for several languages). A quantitative overview of a given language's speech resources (e.g., the number of speakers, database length in hours) is uncommon even if it is highly appreciated by researchers/developers performing statistical analysis and/or model development. This also applies to Hungarian speech corpora. Although language resource catalogs like ELRA[2], LDC[3], ELG[4] are help-

[1]https://w3techs.com/technologies/overview/content_language

[2]https://catalog.elra.info/

[3]https://catalog.ldc.upenn.edu/

[4]https://live.european-language-grid.eu/

LREC-COLING 2024, pages 9382–9388
20-25 May, 2024.

ful for hosting the desired speech data, they do not provide a quantitative overview, partly because the data owners may not have provided it when submitting their data. Another issue is that the links for data sources often become invalid over time without any contact information (cf. (Çöltekin et al., 2023)).

3. Characteristics of the Hungarian Language

Hungarian belongs to the Uralic language family along with Finnish and Estonian, and it is spoken by approximately 12.5 million native speakers (mostly in Hungary, but also in neighboring countries, Europe and overseas (Kenesei et al., 2002)). It is an agglutinating language with rich morpho(phono)logy subject to vowel harmony (i.e. grammatical relations are expressed mostly by suffixes(Siptár and Törkenczy, 2000)). An extended version of the Latin alphabet is used in orthography, including diacritics for vowel quantity and quality (e.g., ó, ö, ő). It has relatively flexible word order (Kenesei et al., 2002), and sentences carry stronger prominence in their initial parts. Word stress is fixed to the first syllable (Reichel et al., 2018).

4. Hungarian Speech Data Sets

In this section, we focus on various Hungarian speech corpora by types and consider only Hungarian audio data (accompanied at least by text transcriptions in Hungarian). In Tables 1–5 the key features (e.g., size in terms of hours and speakers) of the data sets are summarized

4.1. Monolingual Hungarian Data Sets

Table 1 provides an overview of the recordings with healthy adult speakers of Standard Hungarian. MTBA was recorded via a public switched telephone network with reduced bandwidth, a simplified, SpeechDat(E)-like (Pollák et al. (2000)) corpus. A larger data set is BEA (Spoken Hungarian Database) (Gósy (2012); Neuberger et al. (2014)) with 472 recorded speakers, for which annotations are still in progress. The subset BEA-Base Mihajlik et al. (2022) is devoted to benchmarking ASR systems. BUSZI (Budapest Sociolinguistic Interviews) has limited access (i.e., written data can be searched online, but the sound files are available only on-site).

4.2. Hungarian Data Sets as Parts of Multilingual Databases

Hungarian subsets of multilingual data sets are summarized in Table 2. Only data with a minimum

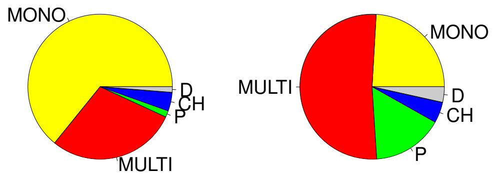

Figure 1: Types of Hungarian data sets. Left: Length in hours, Right: Number of speakers. MONO: monolingual, MULTI: Hungarian from multilingual collections, P: pathological, CH: child-related, D: dialectal.

length of one hour are listed. This collection is diverse in terms of content. CommonVoice Hungarian showed a rapid growth in 2023 from 20 to 153 hours. SPEECON and SpeechDat(E) are important spoken language resources, but they lack spontaneous speech. The final data set is called HUN_ASR001_CN which looks promising (in terms of size and speakers), but no further details are available.

4.3. Pathological Speech

Table 3 collects data sets created to support research and diagnostics related to neurological, physiological or psychiatric disorders affecting speech. However, the speech tasks to be performed by the speakers (introduction, readings, sustained vowels etc.) are partly different across data sets which makes it difficult to examine the disorders comparatively.

4.4. Child-related Speech

Table 4 outlines data sets with child or child-directed speech (by adults). GABI (Child Language and Speech Database and Information Repository), the largest corpus containing recordings with 422 children/adolescents is currently under development.

4.5. Hungarian Dialects

All data sets contain time-aligned narrow phonetic transcriptions (see Table 5). MNYHK (Hungarian Dialectal Talking Book Series) contains transcribed interviews or interview segments recorded between 1960-1964.VIHk (István Vőő Talking Book) includes short interviews mostly from Transylvania (part of in Romania) in the 1960's. SzMNyHk (Talking Books of Hungarian Dialects in Slovakia) consists of interviews recorded recently in and around Nitra. MoMa (Hungarian in Moldova) is composed of transcribed interviews with Moldavian Csángós starting from 1990. The Hegedűs Archive Horváth (2007) is the first collection of Hungarian dialectal speech, recorded around the 1950s and digitized in 2005.

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

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

5. Discussion and Conclusions

Although none of the data sets in the previous section is comparable in size and in variability (e.g. with LibriSpeech, Panayotov et al. (2015) or AISHELL-2, Du et al. (2018)), summing all the Hungarian speech corpora, we still reach remarkable quantities of 2.800 hours and 7.500 speakers in our survey. According to Tables 1–5 and Figure 1, this hypothetical super data set is highly diverse and

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

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

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)
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

includes read and conversational speech, broadcast news, monologues, phone calls across diverse speakers (e.g., healthy adults, children, pathological and dialectal speakers). In that sense, spoken Hungarian may appear to be well-resourced. However, developing a speech recognition model based on the conglomerate of the previous data sets would be a challenging task. First, many of the speech corpora are difficult to access. Next, transcription format, style and methodologies are highly diverse. Transcription data is sometimes stored in a separate document without time-alignment, and some words (or entire texts) are annotated in their phonetic form instead of standard orthography. Annotation systems differ in how foreign words are handled, and abbreviations, numbers, hesitations, noises and punctuation marks vary across data sets without consistency even within the same corpus. Despite the impressive agglomerate numbers, the distribution of "speech hours" and speakers is rather uneven. One of the largest corpora (HuTongue) is collected only from 15 speakers, whereas corpora with high numbers of speakers are often tiny (MTBA, MRBA). This leads to issues about representativeness of the collected speech corpora in comparison to the diversity in the general population (Doğruöz et al., 2023). In some data sets audio recording is not controlled sufficiently (e.g. BEKK, StaffTalk), or it is of studio quality (e.g., BEA, HuComTech). Read speech is overrepresented, and conversational speech is less easily accessible because of GDPR (General Data Protection Regulation)/privacy issues. All in all, potential training data does not align well with the foreseeable ASR applications (e.g., meeting transcription, e-mail dictation). Thus, spoken Hungarian can still be considered underresourced for Speech Technology. More specifically, larger scale data sets (e.g., SwitchBoard Godfrey and Holliman (1993)), meetings or lectures are missing. Furthermore, the diversity of data sets/annotation approaches pose significant challenges if a unified foundational speech model was to be trained. Although new data sets arise and older ones are being processed further, the demand for larger, more aligned and more accessible transcribed speech data is growing rapidly. However, having an overview of existing resources is the first step for the unification of these data sets. We also hope that our survey will be beneficial for researchers who conduct research about Hungarian across disciplines.

6. Acknowledgement

The authors in Hungary were supported partially by the NKFIH K143075, K135038 and NKFIH-828-2/2021(MILAB) projects of the NRDI Fund.

7. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

A. Seza Doğruöz, Sunayana Sitaram, and Zheng Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5751–5767, Singapore. Association for Computational Linguistics.

Simon Eszter, Lendvai Piroska, Németh Géza, and Olaszy Gábor. 2012. *The Hungarian Language in the Digital Age*. Springer.

Viktória Horváth. 2007. A hegedűs-archívum (1942-1962) feldolgozásának alapelvei. *A Magyar Nyelvtudományi Társaság Kiadványai*, (228):270–276.

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.

István Kenesei, Robert M Vago, and Anna Fenyvesi. 2002. *Hungarian*. Routledge.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Georg Rehm, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, and Tamás Váradi. 2014. An update and extension of the meta-net study "europe's languages in the digital age".

Uwe D. Reichel, Katalin Mády, and Stefan Benus. 2018. Acoustic profiles for prosodic headedness and constituency. In *Proceedings of the 9th International Conference on Speech Prosody*, pages 699–703, Poznań, Poland. International Speech Communication Association (ISCA).

Péter Siptár and Miklós Törkenczy. 2000. *The phonology of Hungarian*. OUP Oxford.

Ç. Çöltekin, A.S. Doğruöz, and Ö. Çetinoğlu. 2023. Resources for turkish natural language processing: A critical survey. *Language Resources and Evaluation*, 57:449–488.

8. Language Resource References

Appen_China. Hungarian asr001 cn.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus.

C. Bodó, Z. Kocsis, and F. S. Vargha. 2017. A budapesti egyetemi kollégiumi korpusz: Elméleti és módszertani kérdések. In *Élőnyelvi kutatások és a dialektológia*, pages 169–177.

M.Z. Boito, W.N. Havard, M. Garnerin, É.L. Ferrand, and L. Besacier. 2019. MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible.

Judit Bóna, Tímea Vakula, and Viola Váradi. 2019. Gabi–hungarian child language and speech database and information repository. *The Phonetician*, 116:41–52.

A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Ferenc Csatári, Zsolt Bakcsi, and Klára Vicsi. 1999. A hungarian child database for speech processing applications. In *EUROSPEECH-99*.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. AISHELL-2: transforming mandarin ASR research into industrial scale. *CoRR*, abs/1808.10583.

E. Eris, A. L. Huszár, Á. Kalivoda, B. Sass, N. Vadász, and F. S. Vargha. 2023. Moldvai magyar korpusz – részletek tánczos vilmos gyűjtéséből.

Noémi Fazakas (Gál). 2013. *Elmélet és empíria a szociolingvisztikában*, chapter Az l kiesése mint nyelvi változó a Vöő István-féle hangoskönyvben. Budapest, Gondolat Kiadó.

John J. Godfrey and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62. Web Download.

Mária Gósy. 2012. BEA – a multifunctional Hungarian spoken language database. *The Phonetician*, 105–106:51–62.

G. Gosztolya and L. Tóth. 2010. Keyword spotting experiments on broadcast news data using phone-based technologies (in hungarian). In *Proceedings of MSZNY*, pages 224–235.

Dorottya Gyarmathy and Tilda Neuberger. 2015. Egy hiánypótló adatbázis a tini bea. *Beszédkutatás*, pages 209–221.

B. Hajduska-Dér, G. Kiss, D. Sztahó, K. Vicsi, and L. Simon. 2022. The applicability of the beck depression inventory and hamilton depression scale in the automatic recognition of depression based on speech signal processing. *Frontiers in Psychiatry*, 13:879896.

Dorota Iskra, Beate Grosskopf, Krzysztof Marasek, van den Heuvel Heuvel, Frank Diehl, and Andreas Kiessling. 2002. Speecon-speech databases for consumer devices: Database specification and validation. In *Proceedings LREC 2002*, pages 329–333, Paris, France. ELRA, European Language Resources Association.

G. Kiss, M. G. Tulics, A. Z. Jenei, and D. Sztahó. 2021. Analysis of cross disorder severity prediction problems based on speech features. *PROCEEDINGS E REPORT*, 71.

A. Kocsor, L. Toth, A. Jr. Kuba, K. Kovacs, M. Jelasity, T. Gyimothy, and J. Csirik. 2000. A comparative study of several feature space transformation and learning methods for phoneme classification. *International Journal of Speech Technology*, 3(3/4):263–276.

Anna Kohári and Katalin Mády. 2023. A longitudinal study of pauses, interpauseal units and clauses in infant-directed speech. In *Proceedings of the 20th International Congress of Phonetic Sciences*, pages 2364–2368, Prague, Czech Republic. Guarant International.

M. Kontra and T. Váradi. 1997. *The Budapest Sociolinguistic Interview: Version 3*. Number 2 in Working Papers in Hungarian Sociolinguistics. MTA Nyelvtudományi Intézet, Budapest.

Miklós Kontra, Miklós Németh, and Balázs Sinkovics. 2016. *Szeged Nyelve a 21. Század Elején*. Gondolat Kiadó, Budapest.

Nóra Kugler. 2015. *Megfigyelés és Következtetés a Nyelvi Tevékenységben*. Tinta Kiadó, Budapest.

T. Lander. 2005. Cslu: 22 languages corpus.

Katalin Mády, Anna Kohári, Uwe D. Reichel, Ádám Szalontai, and Péter Mihajlik. 2023. The budapest games corpus. In *Beszédkutatás – Speech Research Conference*, pages 75–77, Budapest.

Péter Mihajlik, András Balog, Tekla Etelka Gráczi, Anna Kohári, Balázs Tarján, and Katalin Mády. 2022. BEA-Base: A benchmark for ASR of spontaneous Hungarian. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1970–1977, Marseille, France. European Language Resources Association.

Cecília Sarolta Molnár, Katalin Mády, Péter Mihajlik, and Beáta Gyuris. 2023. The Akaka Maptask Corpus. In *Beszédkutatás – Speech Research Conference*, pages 81–83, Budapest.

Tilda Neuberger, Dorottya Gyarmathy, Tekla Etelka Gráczi, Viktória Horváth, Mária Gósy, and András Beke. 2014. Development of a large spontaneous speech database of agglutinative Hungarian language. In P. Sojka, A. Horák, I. Kopeček, and K. Pala, editors, *Text, Speech and Dialogue. TSD 2014. Lecture Notes in Computer Science*, pages 424–431. Springer.

György Orosz and Kinga Mátyus. 2014. An mlu estimation method for hungarian transcripts. In *Text, Speech and Dialogue: 17th International Conference*, pages 173–180, Brno, Czech Republic. Springer.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR Corpus Based on Public Domain Audiobooks. In *ICASSP*.

K. Park and T. Mulc. 2019. Css10: A collection of single speaker speech datasets for 10 languages.

P. Pollák, J. Cernocky, J. Boudy, K. Choukri, H. Heuvel, K. Vicsi, A. Virag, R. Siemund, W. Majewski, J. Sadowski, and P. Staroniewicz. 2000. Speechdat (e)-eastern european telephone speech databases.

Károly Presinszky. 2020. Digital methods in researching slovakia hungarian regional dialects. *Hungarian Studies: a Journal of the International Association for Hungarian Studies and Balassi Institute*, 34(1):98–107.

K. Pápay, S. Szeghalmy, and I. Szekrényes. 2011. Hucomtech multimodal corpus annotation. *Argumentum*, 7:330–347.

Zita Réger. 2004. The hungarian réger corpus. *CHILDES Database*.

Speecon Consortium. 2000. *ungarian Speecon database*. SpeeCon. Speecon Project, distributed via ELRA, ID: ELRA-S0297, Speecon resources, 1.0, ISLRN 517-412-534-623-4.

M. K. Szabó, V. Vincze, C. Guba, B. Dam, A. Solymos, A. Bagi, and I. Szendi. 2023. Humendisco: A hungarian speech corpus of schizophrenia, schizoaffective and bipolar disorders.

Martina K. Szabó and Júlia Galántai. 2017. Egy magyar nyelvű spontán beszélt nyelvi korpusz (hutongue) létrehozásának tapasztalatai. In *XXVI. MANYE Kongresszus Konferenciakötete*, Pécs.

Martina Katalin Szabó, Veronika Vincze, Orsolya Ring, István Üveges, Eszter Vit, Flóra Samu, Attila Elemér Gulyás, Júlia Galántai, Zsuzsanna Szvetelszky, Eliza Hajnalka Bodor-Eranus, and Károly Takács. 2021. Stafftalk: Magyar nyelvű spontán beszélgetések korpusza. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged.

Sztahó, D. and Fejes, A. 2023. *Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings*.

D. Sztahó, G. Kiss, and M. G. Tulics. 2021. Deep learning solution for pathological voice detection using lstm-based autoencoder hybrid with multitask learning. In *Biosignals*, pages 135–141.

C. Teleki, V. Szabolcs, T. S. Levente, and V. Klára. 2005. Development and evaluation of a hungarian broadcast news database. In *Forum Acusticum*.

Fruzsina S. Vargha. 2007. A magyar nyelvjárási hangoskönyv és elemzési lehetőségei. In *Nyelv, területiség, társadalom : A 14. Élőnyelvi Konferencia (Bük, 2006. október 9–11.) előadásai*, pages 429–443. Budapest, Magyar Nyelvtudományi Társaság (MNYT).

Vicsi, Kocsor, Csaba Teleki, and László Tóth. 2004. Beszédadatbázis irodai számítógép-felhasználói környezetben. In *Second Conference on Hungarian Computational Linguistics (MSZNY 2004)*, page 315, Szeged.

K. Vicsi, L. Tóth, G. Kocsor, G. Gordos, and J. Csirik. 2002. Mtba - magyar nyelvű telefonbeszédadatbázis. *Híradástechnika*, 8:35–39.

K. Vicsi and A. Vig. 1998. First hungarian speech database. In *Beszédkutatás '98*, pages 163–177.

C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation.