Revised annotation conventions in Hungarian speech corpora

Katalin Mády[1], Anna Kohári[1], Tekla Etelka Gráczi[1], Péter Mihajlik[1,2]

[1] *HUN–REN Hungarian Research Centre for Linguistics*
[2] *Department of Telecommunications and Artificial Intelligence, Budapest University of Technology and Economics*

Abstract

This technical report presents the revised annotation conventions for one large and two smaller Hungarian speech corpora, the BEA Spoken Language Database, the Akaka Maptask Corpus, and the Budapest Games Corpus. Annotations relying on standard Hungarian orthography rather than actual and partly reduced phonetic realisations make it possible to run both linguistic and phonetic queries on a large amount of data. Since the vast majority of the recordings contain (semi-)spontaneous speech, non-lexical phenomena such as hesitations (filled pauses) and non-verbal events such as laughter are labelled. The frequency of the occurrences of these phenomena is demonstrated on the subset Release 1 of the BEA database on speech samples of 115 speakers. Unsurprisingly, laughter and communicative grunts were more frequent in spontaneous speech when expressed in relative numbers. Hesitations occurred more often in semi-spontaneous speech than in read and spontaneous speech showing that the task demanded a higher cognitive effort from speakers. The majority of questions were found in spontaneous speech since the reading tasks did not include interrogatives.

Keywords: speech database, annotation, read speech, spontaneous speech, discourse

1. Introduction

The development of speech databases has become essential during the last decades in order to assess data from a large number of speakers for various kinds of disciplines connected to speech research. The present study introduces the annotation conventions for one large and two smaller Hungarian speech corpora. The main emphasis lies on BEA –*BEszélt nyelvi Adatbázis*, 'Spoken Language Database' (Gósy, 2012) being the largest available speech database

Email addresses: mady.katalin@nytud.hun-ren.hu (Katalin Mády), kohari.anna@nytud.hun-ren.hu (Anna Kohári), graczi.tekla.etelka@nytud.hun-ren.hu (Tekla Etelka Gráczi), mihajlik.peter@nytud.hun-ren.hu (Péter Mihajlik)

for Hungarian at present. Two smaller task-oriented corpora, the Akaka Map-task Corpus (Molnár et al., 2023) and the Budapest Games Corpus (Mády et al., 2023) have been annotated according to the same guidelines. Audio and annotation files are available for research purposes for free via the website `phon.nytud.hu/voxbox/corpora/databases.html` that is being extended by new subsets of the corpora as annotations become available.

2. Basics on the BEA database

BEA is a large speech corpus containing Hungarian speech samples: repetition, spontaneous, semi-spontaneous, and read speech samples (Gósy, 2012). The recordings started in 2007 and lasted until 2017, purporting altogether 472 speakers' speech. The spontaneous speech tasks include an interview about the speaker's life (work/studies, education, hobbies, family, etc.), a quasi-monologue or a dialogue on their opinion about a current public topic, and a discourse with an additional discourse partner on a further current topic, see Table 1. The semi-spontaneous speech tasks include a summary of two short texts that were read aloud either by the experimenter or played to the participant from a recording. The read speech samples were collected by asking the speakers to read aloud 25 sentences of various lengths and a coherent text (title + 13 sentences). In the repetition task, the experimenter read 25 sentences separately, waiting for the speaker to repeat each sentence before moving on to the next one. The altogether eight modules were recorded with all speakers. The sequence of the speech tasks is shown in Table 1. For a detailed description of the recording protocol, see Gósy (2012).

The annotation of the speech samples started shortly after the first recordings were carried out. Annotations were prepared manually, i.e. from scratch, without the help of speech recognition tools. The time-consuming work was carried out with a large group of annotators after careful training. Neuberger et al. (2014) provide an overview on the various formats. The transcription guidelines have been described in several papers (Gósy, 2012; Neuberger et al., 2014). The

speech type	speech task	dynamics	typical order of recording
spontaneous speech	interview	quasi-monologue	2
	opinion on a topic	quasi-monologue or dialogue (depending on the speaker)	3
	discourse	trialogue	6
semi-spontaneous	summarisation of a heard text on plants	quasi-monologue	4
	summarisation of a heard historical anecdote	quasi-monologue	5
reading	reading aloud 25 sentences		7
	reading aloud a text on plants		8
repetition	repetition of 25 sentences		1

Table 1: The seven speech tasks of the BEA database

elaboration of the transcription and annotation methods, their processing and monitoring took place between 2007 and 2019 meaning the joint work of a large number of annotators and researchers.

Several tasks that originally required human resources have received substantial support by machine learning tools – transcribing speech to text was no exception (Bazillon et al., 2008) Due to the increase in computing capacities and the advances in Artificial Intelligence (AI) research, deep neural networks (DNN) have become far more efficient in automatic speech recognition (ASR) than the earlier ML (Maximum Likelihood) techniques (Hinton et al., 2012).

The technical improvement was a motivation to rearrange the transcription process for the speech material of 167 participants (35% of all speakers) for whom no manual annotations had been prepared previously. Along with the largely improved recognition rates (Mihajlik et al., 2022), word and phoneme segmentation tools became available for the team working on the BEA database. These tools required a slightly different input format of the annotated text than the MAUS segmentation system (Schiel, 2004) that had been used previously for the manually annotated recordings. Additionally, labels marking non-lexical vocalisations such as laughter or hesitation have been turned into English ones, since the corpus is meant to serve the international research community.

The current paper presents the conventions along which file names and annotation file structures are set up, and it gives a description of the annotation guidelines. Subsequently, a statistical overview is provided on a subset of the database with respect to the phenomena described in the next section.

3. Setup of sound and annotation files

Speech was recorded in mono wav files via a standing microphone for each discourse partner. File names have the following structure:
bea_461_m_24_readsent_stm
The underscores divide the following pieces of information: *bea* refers to the corpus, the 3-digits number is the speaker ID. The letters *f, m* refer to the speaker's gender (female, male), the 2-digit number of their age at the recording. The recording protocol included eight modules with spontaneous and non-spontaneous speech that are encoded in the file name accordingly, here for read sentences. The final unit refers to the fact that recordings were carried out with a standing microphone, which is the case for all BEA samples.

The current annotation files are in the textgrid format of the Praat software (Boersma & Weenink, 1999). First, signal detection was run on the wav files, by which chunks of spoken passages were marked as intervals in the previously created textgrids. Since the initial and final boundary of the utterance cannot

always be detected exactly, e.g. when starting with a voiceless stop containing an initial closure phase, utterances begin and end with a silent phase of maximal 300 ms. The attempt to separate and label speakers automatically was not successful because they were recorded on the same channel. This task was performed by a team of annotators during the manual corrections following ASR.

The target speaker is encoded as SPK on the first tier of the annotation file. EXP on the second tier refers to the experimenter, i.e. the researcher being present throughout the entire recording. In the discourse module, an additional discourse partner joined the conversation who is labelled as DP on tier 3. Finally, non-human noise, such as a creaking chair, is indicated on the last NOI tier. Apart from the discourse module of BEA, textgrids contain three tiers: SPK, EXP, and NOI. Overlapping speech is indicated by overlapping intervals on tiers devoted to different speakers. Very short overlaps, e.g. backchannelling signals from a different speaker are not always marked, since they are not recognised by ASR, and they do not sincerely affect acoustic analysis.

4. Annotation guidelines

Earlier annotation guidelines followed the principle that the text along with the wav files should follow audio events as exactly as possible. This approach was applied in order to enhance manual searches in the textgrid files, e.g. when selecting certain phone sequences in the material.

The revised annotation system has a different approach: instead of being close to the actual acoustic material, the annotated text follows standard Hungarian orthography with only few exceptions. The reason for choosing a broader transcription system is manifold. First, the orthographic forms enhance the searchability of the corpus when analysing certain lexical units for purposes other than the actual phonetic realisation. An example is the discourse particle *tehát* 'this means' that is often realised with reduction resulting in forms such as *tát, tet* etc. At the same time, a pragmatic analysis might not be interested

in the actual acoustic realisation but intends to deal with all occurrences of the lexical unit in various contexts. Another reason why the full lexical form is given is that modern tools for automatic speech segmentation can deal with varying grapheme-to-phoneme mappings, thus, it is not necessary for the annotator to remain close to the actual realisation. Besides, like in the example of the discourse particle *tehát*, a frequent realisation *tát* is identical to another lexical unit *tát* 'open/gape+3rd person singular' indicating a verb instead of a discourse particle. Third, using the lexical forms makes annotators' lives easier because they can rely on their understanding of the context, and they are not required to perform broad phonetic transcription which is in fact a different task.

In the next sections, instructions for annotators will be presented. Guidelines reflecting Hungarian orthography are followed by those deviating from standard lexical forms. Section 4.2 specifies labels for non-lexical utterances of speakers, whereas section 4.3 describes an extended function of the NOI tier, originally reserved for non-human noise.

4.1. Relation to Hungarian orthography

The following punctuation marks occur in the annotations: .,?!. Since our current ASR system does not output punctuation marks, they are inserted manually by annotators during the manual correction process. Capital letters are reserved for proper names and abbreviated forms, therefore utterances start with lower-case letters. Simple hyphens are used according to Hungarian orthography, e.g. in multiple compounds longer than 6 syllables such as *teherautó-forgalom* 'van traffic'. Digits, however, are spelled out, i.e. *tizenhárom* for '13'.

4.1.1. Irregular orthography

Words with irregular orthography are given both with their actual phoneme sequence and with their lexical form in square brackets. This includes proper names with no direct grapheme-phoneme mapping, foreign names and words,

and words including digits. Another occurrence of the actual realisation and the proper written form is with mispronounced words. Some examples:

- kosut [Kossuth] (surname of the 19^{th} century politician Lajos Kossuth),
- váo [wow],
- cépluszplusz [C++] (the programming language),
- tévékettő [TV2] (television channel),
- szerklény [szekrény] (mispronounced wardrobe).

Colloquial pronunciations of lexical and morphological forms are usually given with the typical orthography in written text style or dictionary form. Most such non-standard forms include the deletion of certain segments such as word-final /r/, sometimes leading to vowel lengthening (e.g. *amikó* 'when' as a relative pronoun instead of standard *amikor*).

The same principle applies to the colloquial merger of the suffixes *-ban/ben* and *-ba/be*. If the former, the inessive (e.g. *in the house*) is replaced by the illative (e.g. *into the house*), as is frequently the case in colloquial speech, it is still written as *-ban/ben*, corresponding to the semantic context and to the usually written form. A slightly different case is the annotation of dialectal or non-standard word forms such as *köll* for standard *kell* 'needs to'. Since it is not a reduction, but an alternative word form, here the version *köll* is given.

This is handled differently with colloquial forms that are lexicalised, i.e. people would usually write them in informal style. A short list of such units contains *nemtom* for *nem tudom* 'I don't know', *asszem* for *azt hiszem* 'I think' etc.

Unfinished word fragments due to disfluencies, e.g. interruption or replanning by the speaker are marked with a double hyphen next to the incomplete word form, e.g. *ke-- kenyér* 'bread'. The label <**interr**> is used to signalise that the utterance is interrupted by the current speaker, and they continue the utterance within the same interpausal unit (IPU) with a different sentence

structure. This helps to filter out sentences from further analysis that rely on syntactically or prosodically complete phrase units.

4.1.2. Capital letters

Capital letters are primarily used for proper names. Additionally, they signalise spelled letters such as *T mint Tamás* 'T as in the proper name Tamás'. The same rule is applied to acronyms produced with individual letters such as *MTA*, the abbreviation for 'Magyar Tudományos Akadémia', Hungarian Academy of Sciences. Abbreviations for words whose letter sequence is produced as a word rather than spelled letters are annotated differently: although the university Eötvös Loránd Tudományegyetem is abbreviated as ELTE in Hungarian orthography, it is written as Elte here since it is pronounced as the sequence of the phonemes indicated by the letters. When a letter is not used as a letter, but as a mathematical symbol such as *x tengely* 'x-axis', the form is given in square brackets, preceded by the actual phoneme sequence: *iksz tengely [x tengely]*.

4.2. Non-lexical units and non-canonical speech

Spontaneous speech contains a number of non-lexical units that are verbal utterances of the speaker, but they are not directly connected to a lexical entry or even to existing phonemes of the language. The most frequent function of these units is hesitation marking by filled pauses (e.g. English *uh, umm*). In spontaneous settings with two or more interlocutors, such units are often used as communicative signals such as backchannels, expressions of emotional state or alike (e.g. English *m-hm, hmm)*. Ward (2006) describes the relationship between the form and the communicative function of such non-lexical units by the term *conversational grunts*. Filled pauses have language-dependent realisations (Horváth, 2020). Non-lexical forms with an intentional communicative meaning, often replacing lexical forms such as *yes, what?* have been studied less frequently. The latter category often contains nasal phoneme-like sounds in Hungarian (Reichel et al., 2023), unlike filled pauses which are most often realised as a schwa (Horváth, 2020).

Following Ward's terminology, our corpus annotations include two labels for non-lexical units:

- filled pauses signalising hesitation <**hes**>,
- communicative grunts in the function of backchannelling, assertion, question etc. <**hum**>.

Unlike in the earlier versions of BEA (see Section 4.4), annotators were not supposed to find a similar sequence of letters to indicate filled pauses. Instead, these are uniformly marked as <**hes**>. Disfluencies expressed by the lengthening of a segment or syllable, but without an independent non-lexical hesitation, are only marked word-medially if the word form is interrupted and contains a pause (see Section 4.1.1).

The label <**hum**> refers to the meaningful conversational grunts realised as *m-hm, m-m, hm* that are referred to as communicative grunts. They are frequent in informal communication in Hungarian, and their intonation is closely linked to their communicative function such as assertion, agreement, disagreement, question, surprise etc. Similarly to filled pauses, this kind of signal is not turned into phoneme sequences in the annotations. Given the promising results on the distinction between manually labelled hesitations and hummings using DNN (Reichel et al., 2023), we are currently expanding our model to automatically recognise these non-lexical signals along with speech sounds (Mihajlik et al., in print).

A special case of non-lexical speech is represented by passages in which the utterance is not intelligible even after careful listening. These units are marked as <**unint**> for 'unintelligible'.

A further group of meta-linguistic markers relates to either non-lexical or paraverbal phenomena that are either produced as separate units or realised on the top of (sequences of) words.

These are the following:

- laughter, either as a separate unit: <**laugh**>, or one or more words during

which the speaker laughs (speech-laugh): <***laugh***> *hát ez még soha nem jutott eszembe* </***laugh***> 'I have never thought of this';

- whispered speech: <***whisper***> *hogy magyarázzam el?* </***whisper***> 'how should I explain it?',
- singing, e.g. when the speaker refers to a song while singing a short passage of it <***sing***> *tovább, tovább, tovább* </***sing***> 'further, further, further' – cited from a well-known farewell song.

4.3. Extensions in two task-oriented corpora

Two further speech corpora have been annotated along the same guidelines: the Akaka Maptask Corpus (AMC) (Molnár et al., 2023) and the Budapest Games Corpus (Mády et al., 2023). These two smaller task-oriented corpora contain collaborative games with two interlocutors. In AMC, one participant received a map of a cave system, while the other was supposed to guide them on earth level to the appropriate exit. Thanks to the different perspectives of the two maps, participants were involved in intensive interaction. The Budapest Games Corpus is based on an object placing task with two participants, introduced in Gravano et al. (2007). Each speaker was seated in front of a laptop, divided by a board in order to prevent visual contact. Both participants saw a set of objects on their screen, in almost identical arrangement. One object was blinking on the first speakers' screen and was placed in the bottom panel for the second speaker. The first speaker was instructed to describe the exact position of the object with reference to the other objects, while the second speaker dragged the object to the suspected position using the mouse. The overlap of the target objects' position on the two screens was measured in percentage (amount of identical pixels). Speaker pairs were organised in groups and competed with the other groups, resulting in motivated and partly emotional exchange (joy, disappointment, surprise etc.). The experimenter did not participate in the task-oriented dialogues, but in some cases it was necessary to give support. If

the experimenter's speech is audible in the recording, the annotation is given on the NOI tier, that is otherwise used for non-human noise.

4.4. Deviations in the first subset of BEA

As mentioned in Section 2, the first subset of the BEA database was annotated manually by a large group of researchers, research assistants and students over years. Given the enormous input of human resources, these annotations form a specific subset of the database called Release 1. Since the annotation guidelines were different from the present ones, an attempt was made to turn these into the labels presented in the previous paragraphs. This could be done automatically for abbreviations of non-lexical vocalisations such as <laugh> instead of *NEV* (for Hungarian 'nevet') or filled pauses, i.e. <hes> instead of the phone sequence perceived in the signal (*ÖMM, öö* etc.).

It is important to emphasise that the earlier versions of the database (sound and text files) that are in use by several research teams in and outside Hungary are set up according to the guidelines described in the earlier papers listed in Section 2. For example, earlier annotations included more non-lexical human noise such as coughing, breathing, clearing one's throat etc. However, these phenomena were not marked consistently by annotators which introduces problems when training ASR models. Therefore, these labels were deleted from the unified annotations of the BEA database.

The most important difference is that Release 1 does not contain punctuation marks because annotators were instructed not to use them. Interrogative forms are labelled as <**q**> ... </**q**>, similarly to passages spoken while laughing. Annotators were asked to set the interval boundaries exactly to the start and end of the utterance, thus the onset and offset of the speech signal is not preceded by silence, unlike in later textgrid versions. Thus, IPUs are generally shorter than in textgrids created later. The label <**interr**> was not used in Release 1 for marking incomplete syntactic structures.

Overlapping speech was often not annotated for any of the speakers but simply marked as overlap. In the current version, these units are labelled as

<**unint**> even if the speech is intelligible but not annotated for two or three speakers on the different tiers.

The second substantial difference is that the text originally contained exact representations of the actually spoken form of many words such as *miér, mér, mé* for *miért* 'why'. In some cases, both the actually spoken and the intended standard form are given as *mér [miért]*. In other cases, the text simply contains the reduced word forms. In order to find these occurrences, an automatic check of lexical forms according to Hungarian orthography was run on the texts. If words not contained in the lexicon were found, they were marked and manually checked. If the reduced form coincided with another existing word form such as *mér* 'measure 3rd person singular', these cases remained undetected and could not be turned into their standard orthographic form without further text processing. The same is true for reduced suffixes such as *-ba/be* instead of *-ban/ben*, a colloquial form for the inessive. Later revisions of Release 1 of the BEA database might contain adaptations to the annotation guidelines introduced above.

A further deviation from the aforementioned annotation principles regards words that were interrupted by the speaker. If the incomplete word form could be detected with the spell checking tool, the interruption was indicated by a double hyphen as in the current annotation conventions. When incompleteness was not marked, and the resulting form was an existing word, it remained unmarked in the final version.

Even if Release 1 does not fully rely on the annotation guidelines described above, most discrepancies could be detected and adapted to the current conventions automatically or manually.

5. BEA Release 1 in numbers

Release 1 of the BEA database contains 65 hours of recorded speech from 115 speakers, consisting of over 100,000 interpausal units. The total durations of the different speech styles are as follows: around 16 hours of read speech

and sentence repetition, 6 hours of semi-spontaneous speech and 43 hours of spontaneous speech.

label	read	semi-spontaneous	spontaneous	sum
<hum>	157	567	4188	4912
<hes>	436	2388	9035	11859
−−	721	486	3525	4732
<laugh>	238	206	4832	5276
</laugh>	2	2	62	66
<q>	108	61	947	1116

Table 2: Occurrences of various speech phenomena in the database

The numbers of occurrences of certain speech phenomena in the annotated material are shown in Table 2. As mentioned in the previous sections, we marked the occurrences of hesitations, word fragments, laughter, speech-laugh, humming and questions in the textgrids corresponding to the audio files. The occurrences of disfluencies (e.g. hesitations, word fragments) are numerous in both read and spontaneous speech, facilitating corpus-based statistical analyses.

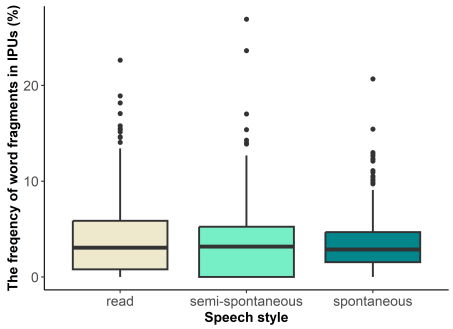

Figure 1: Relative frequency of word fragments in IPUs (%).

The number of occurrences of each phenomenon was divided by the total number of interpausal units (IPUs) and multiplied by 100 to get the percentage

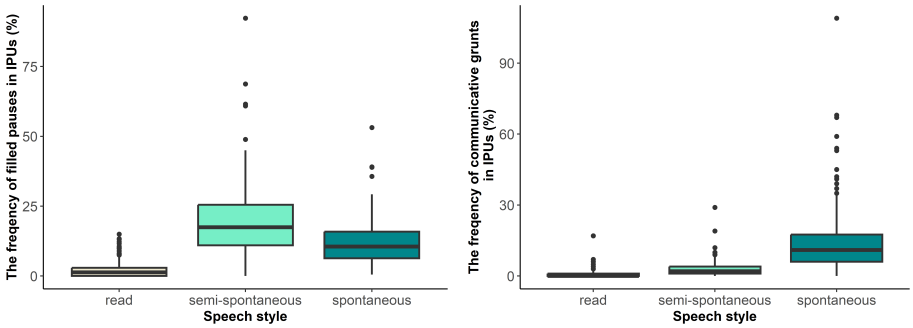

Figure 2: Relative frequency of hesitations and conversational grunts in IPUs (%).

values shown in the plots. Frequencies of various types of disfluencies are shown in Figures 1 and 2 for the different speech styles. Although drawing solid conclusions would require a more detailed analysis, the plots indicate that word fragments occur with a similar frequency in all speech styles, while hesitations can be encountered mostly in semi-spontaneous speech and are less typical in read speech. This is probably due to the fact that semi-spontaneous speech is task-oriented, i.e. speakers were supposed to retell a story instead of giving information about their lives or opinions they are more comfortable with. This example demonstrates the applicability of the database for annotated speech phenomena in a more general sense. The occurrences of tokens or phrases can be easily investigated with similar methods.

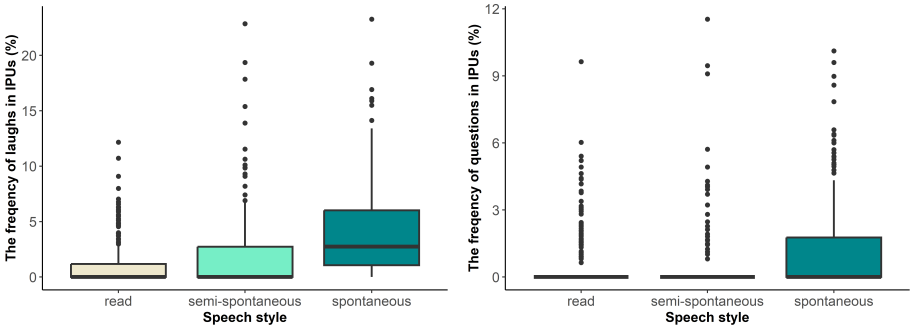

Figure 3: Relative frequency of laughters and questions in IPUs (%).

Laughters predominantly occur in spontaneous speech (Figure 3, left). However, they occur to a lesser extent in read and semi-spontaneous speech, as these tasks were also performed in the presence of a conversational partner. Laughters in these experiments seldom caused the speakers to laugh while saying one or more words. Communicative grunts are also found primarily in the spontaneous modules of the database, indicating consent in most cases (Figure 2, right). At the same time, in semi-spontaneous speech, experimenters often used communicative grunts as a feedback to encourage task continuation. Although questions only sporadically occur in the read and semi-spontaneous passages, their number is somewhat higher in spontaneous speech (Figure 3, right), which is relevant for the linguistic and phonetic investigation of various sentence types.

6. Final remarks

The revised annotation conventions will hopefully enhance linguistic and phonetic research on a large amount of annotated speech data. Both the read and (semi-) spontaneous parts of BEA are well suitable for segmental and prosodic analyses in speech research and for syntactic and semantic studies that require a large amount of data, along with the development and testing of tools in language and speech technology. The smaller Akaka Maptask Corpus and the Budapest Games Corpus are set up differently: The design was developed for both corpora in order to trigger intensive interaction between dialogue partners with varying semantic and pragmatic contents. The speech material contains various sentence types both in their canonical and non-canonical form, e.g. a large number of questions with speaker intentions other than information retrieval, e.g. surprise or uncertainty in form of self-directed questions.

The corpora (sound and annotation files) and the BEAST (BEA Speech Transcriber, Kádár et al., 2023) automatic speech recognition tool are available for research purposes for free. Researchers from countries within the EU and other countries that committed to the General Data Protection Regulation (GDPR) will receive access to the data after filling in the registration form and

signing a statement that they accept the GDPR guidelines. Research institutes from other countries need to sign an institutional contract that is in accordance with GDPR. When building your research on any of the corpora, please refer to the current paper or other relevant publications listed on our website `https://phon.nytud.hu/voxbox/bea/reg.html?lang=en`.

Acknowledgements

This work was funded by the National Research, Development and Innovation Fund (NKFIH), grants K 135038, K 143075 and FK 128814.

We would like to thank András Balog for his indispensable help with the standardisation of previous annotations and his extensive contribution by the application of ASR tools. Uwe Reichel participated in creating the set of non-canonical labels in an earlier, more detailed annotation system for the Budapest Games Corpus. We are grateful to our annotators: Lili Cziáky, Péter Csényi, Éva Alíz Ernhöffer, Gergő Zsolt Gila, Flóra Hegyi, Boglárka Kaposvári, Sára Kovács, Boglárka Mákos, Katalin Pirsel, Henrietta Pokk, Luca Pollak and Szilárd Tóth for the manual correction of the ASR-based annotations.

References

Bazillon, T., Estéve, Y., & Luzzati, D. (2008). Manual vs. assisted transcription of prepared and spontaneous speech. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). URL: `http://www.lrec-conf.org/proceedings/lrec2008/pdf/277_paper.pdf`.

Boersma, P., & Weenink, D. (1999). *PRAAT, a system for doing phonetics by computer*. Technical Report Institute of Phonetic Sciences of the University of Amsterdam. 132–182.

Gósy, M. (2012). BEA – a multifunctional Hungarian spoken language database. *The Phonetician*, *105–106*, 51–62.

Gravano, A., Beňuš, v., Chávez, H., Hirschberg, J., & Wilcox, L. (2007). On the role of context and prosody in the interpretation of 'okay'. In *Proc. 45th Annual Meeting of Association of Computational Linguistics* (pp. 800–807). Prague.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*, 82–97. doi:10.1109/MSP.2012.2205597.

Horváth, V. (2020). Filled pauses in hungarian: their phonetic form and function. *Acta Linguistica Hungarica*, *57*, 288–306.

Kádár, M., Dobsinszky, G., Mády, K., & Mihajlik, P. (2023). "Feeding the beast" – A BEA Speech Transcriber továbbfejlesztése és integrálása neurális nyelvmodellel. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY-2023* (pp. 135–145). Szeged.

Mády, K., Kohári, A., Reichel, U. D., Szalontai, A., & Mihajlik, P. (2023). The budapest games corpus. In T. E. Gráczi, V. Horváth, K. Juhász, A. Kohári, V. Krepsz, & K. Mády (Eds.), *Beszédkutatás – Speech Research Conference* (pp. 75–77). Budapest.

Mihajlik, P., Balog, A., Gráczi, T. E., Kohári, A., Tarján, B., & Mády, K. (2022). BEA-Base: A benchmark for ASR of spontaneous Hungarian. In *Proceedings of the Language Resources and Evaluation Conference* (pp. 1970–1977). Marseille, France: European Language Resources Association. URL: https://aclanthology.org/2022.lrec-1.211.

Mihajlik, P., Meng, Y., Kádár, M. S., Linke, J., Schuppler, B., & Mády, K. (2024). On disfluency and non-lexical sound labeling for end-to-end automatic speech recognition. In *Interspeech 2024, Kos, Greece* (pp. 1270–1274). doi:10.21437/Interspeech.2024-2157.

Molnár, C. S., Mády, K., Mihajlik, P., & Gyuris, B. (2023). The Akaka Maptask Corpus. In *Beszédkutatás – Speech Research Conference* (pp. 81–83). Budapest.

Neuberger, T., Gyarmathy, D., Gráczi, T. E., Horváth, V., Gósy, M., & Beke, A. (2014). Development of a large spontaneous speech database of agglutinative Hungarian language. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech and Dialogue. TSD 2014. Lecture Notes in Computer Science* (pp. 424–431). Springer.

Reichel, U. D., Kohári, A., & Mády, K. (2023). Acoustics and prediction of non-lexical speech in the Budapest Games Corpus. In *Beszédkutatás – Speech Research Conference*.

Schiel, F. (2004). MAUS goes iterative. In *Proceedings of the 4. International Conference on Language Resources and Evaluation* (pp. 1015–1018). Lisbon, Portugal: European Language Resources Association.

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics and Cognition*, *14*, 113–184.