INFORMATICA, 2025, Vol. 36, No. 2, 369–384

DOI: https://doi.org/10.15388/25-INFOR591

Evaluation of Lithuanian Speech-to-Text Transcribers

Pijus KASPARAITIS

Institute of Computer Science, Faculty of Mathematics and Informatics, Vilnius University, Didlaukio str. 47, LT-08303 Vilnius, Lithuania
e-mail: pijus.kasparaitis@mif.vu.lt

Received: July 2024; accepted: April 2025

Abstract. For more than two decades, Lithuanian speech recognition has been researched solely in Lithuania due to the need for deep knowledge of Lithuanian. AI advancements now allow high-quality speech-to-text systems to be built without native knowledge, given sufficient annotated data is available. This study evaluated as many as 18 Lithuanian speech transcribers using a small piece of recording; 7 best ones were selected and evaluated using extensive data. The top system achieved a WER of 5.1% for Lithuanian words, with three others showing 8.7–9.2%. For other word-size tokens, such as numbers, speech disfluencies, abbreviations, foreign words, a classification adapted to the Lithuanian language was proposed. Different processing strategies for tokens of these classes were examined and it was assessed which transcribers tend to follow which strategies.

Key words: speech-to-text transcription, automatic speech recognition, word error rate, character error rate, Lithuanian.

1. Introduction

Automatic speech recognition (ASR) has been developed in Lithuania for over 20 years. A digit recognizer based on the Dynamic Time Warping (DTW) algorithm created by Lipeika *et al.* (2002) can be considered a pioneer in ASR research. Later, works in isolated word recognition using a Hidden Markov Model (HMM) followed, e.g. (Raškinis and Raškinienė, 2003). There have been many attempts to adapt recognizers of other languages to the Lithuanian language, e.g. (Maskeliunas *et al.*, 2009; Kasparaitis, 2008), to combine several recognizers (Rasymas and Rudžionis, 2014). Finally, deep neural networks have begun to be used for speech recognition (Pipiras *et al.*, 2019; Salimbajevs and Kapociute-Dzikiene, 2018). All the above-mentioned works are united by the fact that they were carried out in Lithuania, or at least by individuals who speak Lithuanian, as this required specific knowledge of the Lithuanian language. People used to work on recognizers of a single language (rarely several languages) and offered them to the local market.

The situation has changed dramatically in recent years since deep neural networks and other AI technologies have appeared. They enable the creation of Lithuanian speech recognizers even without knowing Lithuanian language. All that is needed is a sufficiently

large amount of Lithuanian speech data (voice recordings and their corresponding text). In addition, for a long time, speech recognition has been rarely used due to insufficient recognition accuracy. The new AI-based technologies not only enabled the creation of recognizers without knowing the corresponding language, but they also made it possible to increase recognition accuracy, to recognize not only isolated commands, but also to transcribe continuous speech into text. A new term has been adopted – Speech-to-Text Transcription (STT). Recognizers (transcribers) have become even more widely used in practice, and new applications have appeared, such as the automatic generation of video subtitles. Companies emerged that developed transcribers for many languages and offered them worldwide as a publicly available paid service. Below you can find some examples of transcribers that also support Lithuanian, their country of origin, and the declared number of languages supported: Sonix – San Francisco, USA, 49 languages; Voiser – Istanbul, Turkey, 71 languages and 135 variants; Scriptoman – Ashford, UK, 120 languages; Happy Scribe – Dublin, Ireland, more than 120 languages, dialects, and accents. Here and further in the text, see Table 1 for web addresses of the transcribers.

To create even better transcribers, even larger and more diverse speech corpora would be needed. For this purpose, in April 2024, the Lithuanian government announced a call for applications for the project "Creation of the Lithuanian speech corpus (for speech recognition purposes)"[1] whose goals are to create a 10 000-hour speech corpus annotated at the sentence level and a 500-hour speech corpus annotated at the phoneme level. Obviously, the creation of such corpora can be accelerated by automatic speech-to-text transcription tools. Speech recordings are collected and sent to an automatic transcriber. Since automatic transcribers can't ensure 100% accuracy, human annotators should review these transcriptions, compare them with the voice recordings, and make necessary corrections, resulting in final annotations of the recordings. The fewer errors the automatic transcriber makes, the less manual work is required.

Thus, the motivation for the present study arose from the high need to discover and analyse automatic transcription tools suitable for the Lithuanian language that would accelerate the creation of speech corpora. This work aimed:

- To collect and test as many publicly available Lithuanian speech transcribers as possible;
- To check whether they used the same or different recognition engines;
- To evaluate the recognition accuracy they achieved on Lithuanian words, to divide them into suitable and unsuitable ones;
- To propose a word-size token classification system adapted to the Lithuanian language;
- To examine different strategies for transcribing tokens of these classes;
- To assess which transcribers tend to follow which strategies.

Section 2 of the paper reviews similar scientific research in Lithuania and worldwide and describes the evaluation criteria. Section 3 is devoted to an experiment with many (18) transcribers and a small amount of data. In Section 4, the 7 selected transcribers are evaluated more thoroughly with a larger amount and more diverse data. Section 5 is devoted to handling other word-size tokens.

[1] https://www.esinvesticijos.lt/kvietimai/lietuviu-kalbos-garsyno-sukurimas-snekos-atpazinimo-tikslams

2. Related Works

2.1. *Evaluation Metrics*

Word Error Rate (WER) is the most widely used metric to evaluate ASR systems. It is derived from Levenshtein distance, or edit distance (McCowan *et al.*, 2004). When calculating WER, three types of errors are taken into account:

S – the number of substituted words in the automatic transcription;
D – the number of words in the reference that are deleted in the automatic transcription;
I – the number of words inserted in the automatic transcription that do not appear in the reference.

If we define N_r as the total number of words in the reference transcription, then:

$$WER = \frac{S + D + I}{N_r}. \tag{1}$$

Metrics based on portions of information smaller than a word, such as Syllable Error Rate (SER) (Hui Jae *et al.*, 2023) or Character Error Rate (CER) (Silber-Varod *et al.*, 2021) can also be used. SER and CER are defined similarly to WER, i.e. the number of substitutions, deletions, and insertions divided by the length of the reference transcript. CER can be useful for the Lithuanian language, because often only the ending is mistakenly recognized, which does not change the word's meaning, but only in its grammatical form.

Other metrics found in the literature: Relative Information Loss (RIL), Word Information Loss (WIL), Weighted Keyword Error Rate (WKER) (Errattahi *et al.*, 2018), Match Error Rate (MER) (Silber-Varod *et al.*, 2021).

Despite its popularity, WER has some shortcomings:

- It does not distinguish between words that are important to the meaning of a sentence and those that are not.
- It does not take into account whether two words differ by just one character or completely.
- It does not account for the reason why errors might occur.

Many studies have examined the shortcomings of WER as an evaluation metric for ASR systems and many alternative metrics or improvements to the WER have been proposed. E.g.:

A hybrid evaluation metric Hybrid-SD (HSD) that takes into account both semantic correctness and error rate was proposed by Sasindran *et al.* (2023). To generate sentence dissimilarity scores (SD), a fast and lightweight transformer-based language model SNanoBERT was built. Evaluation metric Hybrid-SD is a weighted combination of SD score and non-keyword error rate.

Rugayan *et al.* (2023) analysed the Aligned Semantic Distance (ASD), which utilizes dynamic programming (DP) to find the optimal alignment between two sequences of token embeddings and calculates semantic closeness as the accumulated distance of the alignment.

In this work, WER and, to a lesser extent, CER were used.

2.2. *Works on Comparison and Evaluation of Transcribers*

In recent years, a number of articles have been published evaluating and comparing publicly available commercial speech transcribers. Most of them are for English language recognition, significantly less for other languages. Some of these works are presented below:

Georgila *et al.* (2020) evaluated the following publicly available ASR platforms: Amazon, Apple, Google, IBM, Kaldi, and Microsoft. Data collected from deployed spoken dialogue systems in 6 domains (in US-English) were used. Overall, Google cloud online video performed best, except for one domain where Apple cloud online had the lowest WER. A comparison with their previous evaluations from 2010 and 2013 on the same data sets showed great progress in ASR technology.

Fadel *et al.* (2023) evaluated the performance of five off-the-shelf speech recognition systems, namely Google speech-to-text API, VOSK API, QuartzNet, Wav2vec2.0, and CRDNN model, pre-trained on the Moroccan French corpus. The results indicated that the Google speech-to-text API had the lowest WER (38%). However, other ASR systems had relatively high error rates.

Silber-Varod *et al.* (2021) compared the performance of four ASR engines (Google Cloud, Google Search, IBM Watson, and WIT.ai) for recognizing American English, German, and Hebrew. The English ASR systems performed best, while the Hebrew and German ASR systems showed similar but lower performance. The best engines for American English were IBM Watson (for spontaneous speech) and Google Cloud (for lectures). The best engine for German and Hebrew was Google Cloud.

In their research, Siegert *et al.* (2020) focused on transcribing German spontaneous speech for three different types of applications using four speech APIs: Google Cloud Speech to Text API (GCS), Google Web Speech API (GWS), IBM Watson Speech to Text API (IBM), Wit.ai Speech to Text API (WIT). The result was that both Google speech APIs and WIT are recommendable for all three types of applications. Although IBM had the highest WER, it was able to transcribe almost all German interjections. Another conclusion was that none of the analysed ASR services improved over a period of eight months.

The cloud-based speech recognition Open APIs of seven domestic and foreign cloud companies (Kakao, ETRI, Naver, Microsoft, Google, IBM, Amazon) were compared by Yoo *et al.* (2021). The best results for Korean were shown by Kakao and Microsoft recognizers.

Kuligowska *et al.* (2023) selected three systems, Google ASR, Microsoft ASR and Techmo ASR, and compared their performance on a set of medical-related expressions spoken in Polish. Although all three recognizers showed similar results (the difference between the best and the worst being only 1.7%), the recognizers were ranked in descending order of performance as follows: Google ASR, Techmo ASR, and Microsoft ASR.

The transcription of spoken Ukrainian speech samples using three speech-to-text APIs was compared by Kobylyukh *et al.* (2023). It was concluded that Amazon Transcribe and Microsoft Azure Speech Services were more accurate than Google Cloud Speech-to-Text.

It can be noticed that the number of compared recognizers is not large, up to 7. Mostly these are recognizers of the most well-known companies. In addition, authors often complain about a lack of transcribers for their language in their articles published 3–5 years ago. E.g.:

Iancu (2019) stated that there were five major players offering ASR algorithms in the cloud, each with its own personal assistant: Google with the so-called Google Assistant, Apple with Siri, Microsoft with Cortana, Amazon with Alexa and IBM with Watson. Google was the only cloud provider that supported Romanian. Further in the article, the Google Cloud Speech-to-Text API was applied to various video e-learning resources in Romanian, achieving a WER of 30.96%.

Cumbal *et al.* (2021) stated that, as a lower-resource language, Swedish was not included in many prevalent ASR or Speech-To-Text (STT) systems (e.g. Amazon Transcribe, IBM Watson, Houndify, VOSK). They found three off-the-shelf ASR services that could process Swedish speech, but only two (Google Cloud and Microsoft Azure) had an API available. As a third recognizer they used an open-source Swedish ASR model available through the platform Huggingface. The focus was on the differences between native vs. non-native speakers and between read vs. spontaneous speech, rather than on the recognizers themselves.

The Lithuanian Google speech recognizer, released for public use in 2015, was evaluated by Sipavičius and Maskeliunas (2016). The focus was on the impact of different types of noise and different signal-to-noise ratios on recognition accuracy. Using noise-free recordings, the average WER value for all speech recordings that were processed by the recognizer and produced results was 40.74%.

3. Experiments with a Small Amount of Data

To create a list of test transcribers, the keywords "speech to text Lithuanian" were entered into the search engine (https://www.google.com/), the first 200 results were reviewed, and 17 transcribers were selected to be used in further experiments. These were transcribers that:

1. Support the Lithuanian language;
2. Do not require any installation;
3. Can process recording from a file rather than from a microphone;
4. Have a demo mode that allows you to transcribe a small piece of a recording for free.

Due to its popularity, the 18th transcriber, Whisper-1, which requires installation, was added to the list. The list of transcribers in chronological order of testing is shown in Table 1.

A 4-minute long piece of audiobook was chosen as the test data. This recording was sent to all 18 transcribers and corresponding transcriptions were obtained. If any transcriber only accepted the shorter recording, an appropriately truncated recording was used. The text was taken from the book and adjusted to match the voice recording perfectly. The text contained 401 words, 3181 characters (including spaces). As soon as we

Table 1
List of transcribers tested.

No.	Name	Web site
1	Semantika	https://semantika.lt/Analysis/Transcriber
2	Amazon Transcribe	https://aws.amazon.com/pm/transcribe/
3	Tilde. Speech transcription	https://tilde.com/products-and-services/transcribe
4	Sonix	https://sonix.ai/languages/transcribe-lithuanian-audio
5	Notta	https://www.notta.ai/en/transcribe-lithuanian
6	Voiser	https://voiser.net/speech-to-text/lithuanian-lithuania-transcribe
7	Rask	https://www.rask.ai/tools/transcription/transcribe-lithuanian
8	Happy Scribe	https://www.happyscribe.com/transcribe-lithuanian
9	Vidby	https://vidby.com/transcription/lithuanian
10	Go Transcribe	https://go-transcribe.com/transcribe-Lithuanian-to-text
11	Scriptoman	https://scriptoman.ai/transcription/transcribe-lithuanian-audio
12	TurboScribe	https://turboscribe.ai/
13	Cockatoo	https://www.cockatoo.com/
14	Transkriptor	https://app.transkriptor.com/files
15	Intelektika	https://snekos-atpazinimas.lt
16	Google Cloud	https://cloud.google.com/speech-to-text?hl=en
17	NeuralSpace	https://www.neuralspace.ai/voiceai
18	Whisper-1	https://platform.openai.com/docs/guides/speech-to-text

Table 2
WER and CER of Lithuanian words, WER of non-Lithuanian words, when tested with a small amount of data.

No.	Name	WER Lithuanian		CER Lithuanian		WER non-Lithuanian	
1	Intelektika	5/364	1.37%	7/2975	0.24%	27/37	72.97%
2	Scriptoman	7/364	1.92%	9/2975	0.30%	25/37	67.58%
3	Go Transcribe	8/364	2.20%	11/2975	0.37%	24/37	64.86%
4	Sonix	8/364	2.20%	11/2975	0.37%	24/37	64.86%
5	NeuralSpace	10/364	2.75%	11/2975	0.37%	18/37	48.65%
6	Vidby	10/364	2.75%	12/2975	0.40%	19/37	51.35%
7	Voiser	10/364	2.75%	13/2975	0.44%	18/37	48.65%
8	Happy Scribe	10/364	2.75%	15/2975	0.50%	22/37	59.46%
9	Semantika	10/364	2.75%	16/2975	0.54%	30/37	81.08%
10	Tilde	6/280	2.14%	12/2232	0.54%	14/16	87.50%
11	Transkriptor	9/292	3.08%	13/2337	0.56%	13/25	52.00%
12	Amazon Transcribe	18/364	4.95%	30/2975	1.01%	6/37	16.22%
13	Rask	6/93	6.45%	10/767	1.30%	0/9	0.00%
14	TurboScribe	46/364	12.64%	67/2975	2.25%	0/37	0.00%
15	Whisper-1	50/364	13.74%	72/2975	2.42%	0/25	0.00%
16	Cockatoo	49/279	17.56%	71/2290	3.10%	0/25	0.00%
17	Google Cloud	27/156	17.31%	65/1261	5.15%	24/37	64.86%
18	Notta	26/50	52.00%	40/404	9.90%	9/37	24.32%

started comparing the texts generated by transcribers with the original text, we noticed that the transcribers showed significantly different abilities to transcribe non-Lithuanian words. It was decided that non-Lithuanian words should be examined separately. 37 such words were found (e.g. Android, PDF, Voice Dream Reader). The results (WER and CER of Lithuanian words, WER of non-Lithuanian words) arranged in ascending order of CER are shown in Table 2.

Table 3
List of errors and their corrections.

Name	Errors and their corrections
Intelektika	išbandytum(e te|ėte)reikia, tai(|p), vien(ą|a), veik(e|ia)
Scriptoman	aparatin(ės|ių), iš(|)bandytumėte, pasi(|)Bandykite, pa(s|k)lydus, Tai(|p), m(ąs|a)tančiam, (į |)klausimą,
Go Transcribe	aparatin(ės|ių), iš(|)bandytumėte, pasi(|)Bandykite, pa(s|k)lydus, Tai(|p), m(ąs|a)tančiam, (į |)klausimą, veik(ė|ia)
Sonix	aparatin(ės|ių), iš(|)bandytumėte, pasi(|)Bandykite, pa(s|k)lydus, Tai(|p), m(ąs|a)tančiam, (į |)klausimą, veik(ė|ia)
NeuralSpace	vie(|n), Liep(ą|a), j(ū|uo)s, Liep(ą|a), liep(ą|a), produkt(o|ų), viena(s|), (|j)į, kit(ą|a), Viena(s|)
Vidby	į(|)renginį, keturi(|s), j(ū|uo)s, paspaud(ė|ę), liep(ą|a), grį(š|ž)kite, tai(|p), neregi(ų|o), į(| jo), siūlom(o|u)s
Voiser	vie(|n), liep(ą|a), į(|)renginį, j(ū|uo)s, Liep(ą|a), liep(ą|a), vien(as|u), visuomenė(je|), viena(s|), (|j)į
Transcriptor	vie(|n), liep(ą|a), j(ū|uo)s, Liep(ą|a), liep(ą|a), vien(as|u), ne(|)reg(a|io), viena(s|), (|j)į
Happy Scribe	aparatin(ės|ių), j(ū|uo)s, iš(|)bandytumėte, pasi(|)Bandykite, paklyd(o su|us)rasti, Tai(p|), matan(tie|čia)m, neregi(ui|o), 2023(š|-)iais
Semantika	Nes(ių| jo)s, išbandytum(e te|ėte)reikia, (|į)kalb(a|ą), Liep(ą|a), (|ant)produktų, veik(ė|ia), (|į)kelti, Vien(u|a),
Tilde	nereg(ėtai|iams), charakte(te|)ristikų, Liep(ą|a), sintezatori(a|)us, balsu(su|), neregi(e|a)ms

Table 2 shows that the transcribers who took the last 6 places transcribed the Lithuanian voice unsatisfactorily. However, 4 of them transcribed non-Lithuanian words without errors. From the transcriptions generated by the first 11 transcribers, words with errors were selected, and their corrections were written. For this purpose, the following notations were adopted: the erroneous fragment of a word is enclosed in brackets, and its correct version is written after the symbol "|". For example, if "karstas" is recognized instead of "kartas", the deletion of the letter "s" is written as follows: "kar(s|)tas"; if "karas" is recognized instead of "kartas", the insertion of the letter "t" is written as follows: "kar(|t)as"; and finally, if "kardas" is recognized instead of "kartas", the substitution is written as follows: "kar(d|t)as". The words with errors and their corrections are shown in Table 3.

Table 3 suggests that transcribers 2–4 (Scriptoman, Go Transcribe, Sonix) use the same recognition engine, this will be tested in the next section. Other transcribers generally make mistakes in different places.

4. Testing with a Large Amount of Data

4.1. *Data*

For testing, 15 data files were taken from the Liepa project (Laurinčiukaitė *et al.*, 2018), and 2 more were taken from non-public sources. The files were selected to have approximately equal amounts of read and spontaneous speech, male and female voices. As for the recording sources, we tried to make them as diverse as possible: audiobook, dictaphone, phone, studio, and TV. Most speakers (9) were in the 26–60 age group; 2 were younger

Table 4
Distribution of token classes.

Class	Percentage
Regular Lithuanian words	94.0%
Numbers	1.9%
Short forms of Lithuanian words	1.1%
Speech disfluencies	1.3%
Others (non-Lithuanian words, abbreviations, etc.)	1.7%

than 12; 3 were 18–25; and 3 were over 60. All recordings were single-speaker, so there was no speech overlap, and no speaker identification was required.

The experiments were then carried out as follows: a 10-minute recording was cut from the beginning of each file and passed to the transcriber, and the generated transcription was downloaded. The resulting transcription and the original transcription taken from the corpus were tokenized into word-size chunks, arranged into two spreadsheet columns, and aligned at the word level.

Similar to the experiments with a small amount of data, where the WER of Lithuanian and non-Lithuanian words were calculated separately, this time, the tokens were divided into more classes: 1) regular Lithuanian words, 2) numbers, 3) short forms of Lithuanian words, 4) speech disfluencies, 5) others (non-Lithuanian words, abbreviations, etc.). The token class was indicated next to each token in the original transcription, allowing token counting for each class. See Table 4 for the distribution of token classes in the data examined.

As seen from Table 4, regular Lithuanian words make up as much as 94%, so their recognition essentially determines the accuracy of the transcriber. Later in this section, we will discuss this the most important characteristic of the transcriber, while classes 2–5 will be considered in the next section.

4.2. *Choosing Transcribers*

Since Table 3 showed that three transcribers (Go Transcribe, Scriptoman, and Sonix) make mistakes in the same words, it was decided to check this on a larger amount of data (6 files or 1 hour of recording), and if so, to continue experimenting with only one of them. Pairs of transcribers, Sonix-Go Transcribe and Sonix-Scriptoman, were taken, and the number of times both transcribers made a mistake in the same word, the number of times only the first one made a mistake, and the number of times only the second one made a mistake were counted. See the results in Table 5.

Table 5 shows that the transcribers Sonix and Go Transcribe make mistakes in the same words even in almost 97% of cases. Although the overlap is very large, these transcribers are not identical. When comparing the Sonix and Scriptoman transcribers, the match exceeds 90%. This is also a high degree of overlap. In addition, based on the same six files, WERs were calculated and obtained the following values: Sonix – 8.37%, Go Transcribe – 8.50%, and Scriptoman – 8.57%. Based on the slightly better results shown by Sonix and a high degree of overlap between these three transcribers, it was decided to use only Sonix in further experiments.

Table 5
Evaluation of error overlap between three transcribers.

No.	Sonix vs. go transcribe	Sonix vs. scriptoman
1	79-1-2	77-3-6
2	157-0-10	155-2-6
3	144-7-4	143-8-11
4	130-0-1	129-1-6
5	108-0-0	107-1-1
6	197-1-0	177-21-20
Sum:	815-9-17	788-36-50
Percentage:	96.91%	90.16%

The 5 transcribers that showed the best results with a small amount of data were selected for further examination (see the top of Table 2): Happy Scribe, Intelektika, Sonix, Vidby, Voiser.

4.3. *Recognition of Lithuanian Words*

All available 17 files (2 hours and 50 minutes of recording) were submitted to the transcribers, and WER values were calculated for Lithuanian words. Although Vidby outperformed Voiser and Happy Scribe in an experiment with a small amount of data, it performed extremely poorly this time. WER values ranged from 15.1% to 50.9% (average 30.9%); hence, Vidby will not be considered further.

For the other four transcribers, the number of errors obtained in each file and the WER estimate are shown in Table 6. The best result for the file is written in bold. The total number of errors and the total WER are shown at the bottom of the table. Intelektika showed the lowest WER (total 5.1%, 1.0–17.2% for individual files, depending on the recording quality) with a significant gap from others. There was only one file (No. 11) where Happy Scribe outperformed him and one (No.16) where Voiser outperformed him. The other three transcribers showed similar reasonably low WER: Happy Scribe – 8.7%, Voiser – 8.9%, Sonix – 9.2%.

Paired t-tests were performed with the results in Table 6, which showed that differences between Intelektika and the remaining three transcribers are statistically significant. A comparison of the remaining pairs showed that differences are not statistically significant.

4.4. *Comparing Public vs. Nonpublic, Read vs. Spontaneous Speech*

As mentioned earlier, most data was from a public source – the Liepa project. This is good, considering that other testers will be able to use it, but the bad thing is that it is not known whether this public source was used to train any transcribers. Therefore, two more files were taken from a non-public source, and WER was calculated using only the public and non-public data. The results are in the upper part of Table 7. Unfortunately, other features of the recordings, such as speaker or text complexity, do not allow us to predict whether

Table 6
Number of errors and WER of the top four transcribers when transcribing Lithuanian words.

No.	Total words	Happy scribe		Intelektika		Sonix		Voiser	
1	1268	73	5.8%	**23**	**1.8%**	74	5.8%	65	5.1%
2	1344	134	10.0%	**43**	**3.2%**	145	10.8%	163	12.1%
3	944	109	11.5%	**53**	**5.6%**	119	12.6%	106	11.2%
4	1007	115	11.4%	**88**	**8.7%**	126	12.5%	137	13.6%
5	908	81	8.9%	**49**	**5.4%**	60	6.6%	74	8.1%
6	885	46	5.2%	**9**	**1.0%**	72	8.1%	46	5.2%
7	730	42	5.8%	**24**	**3.3%**	42	5.8%	28	3.8%
8	1355	93	6.9%	**45**	**3.3%**	99	7.3%	94	6.9%
9	1233	88	7.1%	**12**	**1.0%**	90	7.3%	55	4.5%
10	883	98	11.1%	**57**	**6.5%**	96	10.9%	102	11.6%
11	1217	**171**	**14.1%**	209	17.2%	179	14.7%	280	23.0%
12	994	109	11.0%	**81**	**8.1%**	132	13.3%	143	14.4%
13	984	97	9.9%	**32**	**3.3%**	105	10.7%	101	10.3%
14	1041	71	6.8%	**48**	**4.6%**	78	7.5%	53	5.1%
15	960	66	6.9%	**36**	**3.8%**	66	6.9%	45	4.7%
16	1282	136	10.6%	84	6.6%	136	10.6%	**82**	**6.4%**
17	1282	65	5.1%	**37**	**2.9%**	65	5.1%	48	3.7%
Sum:	18317	1594		930		1684		1622	
WER:			8.7%		5.1%		9.2%		8.9%

Table 7
Comparison of public/non-public sources and read/spontaneous speech.

Type of speech	Total words	Happy scribe		Intelektika		Sonix		Voiser	
Public sources	15753	1393	8.8%	809	5.1%	1483	9.4%	1492	9.5%
Non-public sources	2564	201	7.8%	121	4.7%	201	7.8%	130	5.1%
Read speech	10222	972	9.5%	555	5.4%	1039	10.2%	1109	10.8%
Spontaneous speech	8095	622	7.7%	375	4.6%	645	8.0%	513	6.3%

the public recordings of the Liepa project were used to train any of the transcribers unless it can be stated that the recognition leader remained the same, i.e. Intelektika.

In addition, the accuracy of recognition of read and spontaneous speech was calculated. The results are in the lower part of Table 7. Better results were obtained for spontaneous speech, although it was expected to be the opposite, showing again that other recording characteristics are more important than the type of speech. We can only state that the transcriber Voiser is more oriented towards spontaneous speech.

5. Transcribing Tokens of Other Classes

The processing of tokens that are not regular Lithuanian words may also depend on the purpose for which we intend to use the transcription. Two cases will be considered: when it is more important that the text closely matches the acoustic representation, e.g. when creating a phoneme-level annotated speech corpus, and when grammatically correct and

Table 8
Recognition of numbers by the top four transcribers.

Recognized as	Happy scribe		Intelektika		Sonix		Voiser	
Text, correct	224	59.3%	224	59.3%	215	56.9%	44	11.6%
Text, error	24	6.3%	13	3.4%	17	4.5%	56	14.8%
Number, unambiguous	73	19.3%	77	20.4%	84	22.2%	97	25.7%
Number, ambiguous	42	11.1%	52	13.8%	50	13.2%	130	34.4%
Number, error	15	4.0%	12	3.2%	12	3.2%	51	13.5%

easy-to-read text is required, e.g. for video subtitles. The processing of such tokens will be discussed in more detail below.

5.1. *Numbers*

When it comes to transcribing numbers, the results produced by the transcriber can be divided into the following five groups: 1) numbers transcribed as text, e.g. “vienas” (one); 2) an attempt was made to transcribe a number as text, but it was transcribed incorrectly; 3) transcribed as a number and this number can be unambiguously converted to text, i.e. the number is in the nominative case or the number is given an ending that determines its grammatical form, e.g. “12-aisiais” (in twelfths). The part of the number that indicates thousands, millions, or billions can be indicated as an abbreviation, e.g. “3 mln.” (3 million); 4) recognized as a number, but based on this single number, the grammatical form cannot be unambiguously recognized, but a person could do so based on a broader context; 5) an attempt was made to recognize it as a number, but it was incorrect, or the structure of the recognized number does not correspond to the usual notation, e.g. “šimtas 18” instead of “118”, “3 1000” instead of “3000”). If we are talking about a task focused on acoustic representation, we can consider only groups 1 and 3 as suitable results. If the text is intended for a human reader, then group 4 is also appropriate.

There were 378 numbers in the analysed records. If a number consisted of several words, it was treated as several words. See Table 8 for how the transcribers deal with numbers.

5.2. *Short Forms of Lithuanian Words*

In spontaneous spoken Lithuanian, some word forms can be shortened. See Table 9 for the most common short word forms. Short forms are not the norm of the Lithuanian language; therefore, long forms should be used in written language. However, it is worth keeping the short forms if you want the text to be closer to the acoustic representation.

211 short forms were found in the analysed records. See Table 10 for how the transcribers handle them.

As can be seen from Table 10, Intelektika preserves the short form in less than 5% of cases, and changes it to the long form in 85% of cases. Happy Scribe and Sonix behave almost identically: they preserve the short forms in about 20% of cases and change it to the long form in 60% of cases. Voiser preserves the short form in almost 50% of cases, and changes it to the long form in only 35% of cases.

Table 9
Most common short word forms in Lithuanian.

Grammatical form	Examples of long forms	Examples of short forms	Translation to English
Dative case plural of a noun, adjective, numeral, pronoun, participle.	Visiems trims dainuojantiems linksmiems berniukams.	Visiem trim dainuojantiem linksmiem berniukam.	To all three singing cheerful boys.
Instrumental case of a noun, adjective, numeral, pronoun, participle.	Su visomis trimis dainuojančiomis linksmomis mergaitėmis.	Su visom trim dainuojančiom linksmom mergaitėm	With all three singing cheerful girls.
Locative case of a noun, adjective, numeral, pronoun, participle.	Visuose trijuose dainuojančiuose linksmuose berniukuose.	Visuos trijuos dainuojančiuos linksmuos berniukuos.	In all three singing cheerful boys.
Verb plural I and II person all tenses.	Einame, einate, ėjome, ėjote, eidavome, eidavote, eisime, eisite.	Einam, einat, ėjom, ėjot, eidavom, eidavot, eisim, eisit.	We go, you go, we went, you went, we used to go, you used to go, we will go, you will go.
Verb infinitive	Eiti	Eit	To go

Table 10
Processing of short word forms by the top four transcribers.

Recognized as	Happy Scribe		Intelektika		Sonix		Voiser	
Short form	47	22.3%	9	4.3%	42	19.9%	105	49.8%
Long form	128	60.7%	181	85.8%	128	60.7%	74	35.1%
Error	36	17.1%	21	10.0%	41	19.4%	32	15.2%

Table 11
Processing of speech disfluencies by the top four transcribers.

Result	Happy Scribe		Intelektika		Sonix		Voiser	
Recognized	56	22.8%	100	40.7%	41	16.7%	113	45.9%
Removed	137	55.7%	78	31.7%	148	60.2%	76	30.9%
Error	53	21.5%	68	27.6%	57	23.2%	57	23.2%

5.3. *Speech Disfluencies*

The voice recordings contained various speech disfluencies, such as filled pauses, repetitions, false starts, restarts, and incomplete words. If we want the transcriber's output to be as close to the acoustic representation as possible, such disfluencies should be preserved, and if clean text is needed, it is better to remove them. In total, 246 segments were marked as disfluencies in the recordings. See Table 11 for how different transcribers handle them.

Happy Scribe and Sonix eliminate most of the disfluencies (55% and 60%, respectively), while Intelektika and Voiser eliminate only about 30%.

5.4. *Remaining Token Classes*

After reviewing the results generated by the transcribers, four more token classes were identified. The first class is Lithuanianized foreign words, usually names and surnames,

Table 12
Processing of remaining token classes by the top four transcribers.

Result	Happy scribe		Intelektika		Sonix		Voiser	
Lithuanianized foreign words								
Correct	44	20.4%	78	36.1%	43	19.9%	89	41.2%
Error	172	79.6%	138	63.9%	173	80.1%	127	58.8%
English words and English abbreviations, recognized as:								
English word	21	35.0%	5	8.3%	21	35.0%	22	36.7%
Lithuanian transcription	1	1.7%	4	6.7%	1	1.7%	2	3.3%
Error	38	63.3%	51	85.0%	38	63.3%	36	60.0%
Lithuanian abbreviations								
Correct	19	86.4%	16	72.7%	16	72.7%	17	77.3%
Error	3	13.6%	6	27.3%	6	27.3%	5	22.7%
Lithuanian words, that can be abbreviated, recognized as:								
Word	22	53.7%	26	63.4%	29	70.7%	19	46.3%
Abbreviation	18	43.9%	14	34.1%	9	22.0%	19	46.3%
Error	1	2.4%	1	2.4%	3	7.3%	3	7.3%

e.g. “Maikas”, “Plautila”, “Vinė”, “Nastasė”, “Štefanas”. It is generally expected that the transcriber will recognize them as Lithuanian words. 216 such words were found; see the results in the upper division of Table 12. Unfortunately, the transcribers encountered difficulties recognizing such words since they are not common Lithuanian words. Voiser did the best – about 41% correct, Intelektika was 5% behind, and Happy Scribe and Sonix recognized only about 20%.

The second class consists of English words, including English abbreviations, e.g. “Brexit”, “startup”, “blockchain”, “AI” (Artificial intelligence). These words should be recognized as English words, although Lithuanian transcription would also be acceptable. 60 English words and abbreviations were found. The results are shown in Table 12, division 2. As was already seen from experiments with a small amount of data (Table 2), transcribers who recognize Lithuanian well, recognize English poorly. The WER of the transcriber Intelektika is as high as 85%, and the WERs of the remaining three transcribers are about 60%.

The third tiny class (only 22 cases) is Lithuanian abbreviations, e.g. “JAV” (USA), “PVM” (GDP, Gross domestic product). See the results in Table 12, division 3. They were recognized well, the WER was 13% to 27%, but there was too little data for more accurate conclusions.

The last fourth class includes normal Lithuanian words that can be replaced with abbreviations, and some transcribers did so, e.g. “procentų” – “proc.” or “%”, “valandą” – “val.” (hour), “litrų” – “l.” (liter), “prieš Kristų” – “pr. Kr.” (BC, Before Christ). The resulting abbreviation may no longer reflect the grammatical form of the word used, so such a change is suitable for generating text intended for human reading but not very suitable if a transcription corresponding to the acoustic representation is required. 41 cases were found, see the results in the lower division of Table 12. The fewest words were replaced with abbreviations by Sonix (22%), followed by Intelektika (34%), Happy Scribe (almost 44%), and Voiser (over 46%).

Table 13
Comparison of transcribers in terms of acoustics/text orientation.

Feature	Acoustics oriented	Text oriented
Transcribing numbers as text/numbers	Happy Scr., Intelektika, Sonix	Voiser
Short/long word forms	Voiser	Happy Scr., Sonix, Intelektika
Disfluencies retained/removed	Intelektika, Voiser	Happy Scribe, Sonix
Lithuanian words retained/abbreviated	Sonix, Intelektika	Happy Scribe, Voiser

Finally, we conclude this section with an analysis of which transcribers' behaviour is more focused on acoustic matching, which is more focused on text readability, and what each transcriber is more focused on. As shown in Table 13, all transcribers have features of both orientations. Intelektika concentrates more on acoustics, while Happy Scribe focuses more on text readability.

6. Conclusions

Recently, we have been experiencing a real boom of speech-to-text transcribers. As many as 18 publicly available commercial transcribers capable of converting Lithuanian voice recordings into text were found on the Internet. After testing with a small piece of high-quality audiobook recording, the transcribers were ranked by quality. The top 7 were subjected to more in-depth testing with many recordings of varying quality.

The analysis of errors showed that three transcribers (Go Transcribe, Scriptoman, and Sonix) made the same errors when tested with a small amount of data. This led to the assumption that they use the same recognition engine, but the tests with a large amount of data showed this is not true. So, it can be concluded that all 18 transcribers are different.

All texts produced by transcribers were tokenized, and tokens were divided into Lithuanian words (94% of tokens) and other tokens. The recognition accuracy of Lithuanian words was evaluated. The transcriber Intelektika showed the best results (the lowest WER, equal to 5.1%). Very similar WERs were provided by three other transcribers: Happy Scribe, Voiser, and Sonix (8.7–9.2%). Go Transcribe and Scriptoman could also be added to this group, whose errors are similar to those of Sonix. The accuracy of these 5 transcribers is also sufficiently high, although it is far behind the accuracy of Intelektika, and this difference is statistically significant.

The following classification, adapted to the Lithuanian language, was proposed for other tokens: numbers, short forms of Lithuanian words, speech disfluencies, Lithuanianized foreign words, English words and abbreviations, Lithuanian abbreviations, Lithuanian words abbreviated by the transcriber. Their processing depends on the goal, whether it is more important that the text reflects as accurately as possible what was actually said, or that the text is grammatically correct and easy to read. For the first purpose, numbers should be transcribed as text, short forms of words should be preserved (not replaced with long ones), speech disfluencies should be left in place (not removed), Lithuanian words should not be abbreviated, and for the second purpose – vice versa. All four top transcribers have some features that focus on the first goal and some that focus on the second.

Intelektika is slightly more focused on the first goal, while Happy Scribe focuses slightly more on the second.

The present work should be helpful for evaluating new transcribers of the Lithuanian speech that will appear in the future, as well as for tracking the quality progress of transcribers that have already been evaluated.

References

Cumbal, R., Moell, B., Lopes, J., Engwall, O. (2021). "You don't understand me!": comparing ASR results for L1 and L2 speakers of Swedish. In: *INTERSPEECH 2021*, pp. 4463–4467. https://doi.org/10.21437/Interspeech.2021-2140.

Errattahi, R., El Hannani, A., Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: a review. *Procedia Computer Science*, 128, 32–37. https://doi.org/10.1016/j.procs.2018.03.005.

Fadel, W., Toumi, B., Buvet, P.-A., Bourja, O. (2023). Adapting off-the-shelf speech recognition systems for novel words. *Information (Switzerland)*, 14, 179. https://doi.org/10.3390/info14030179.

Georgila, K., Leuski, A., Yanov, V., Traum, D. (2020). Evaluation of off-the-shelf speech recognizers across diverse dialogue domains. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6469–6476. https://aclanthology.org/2020.lrec-1.797/.

Hui Jae, Y., Oh, E.-B., Kim, J.-M. (2023). Comparison of automatic speech recognition system for school-aged children's narratives: naver clova speech and google speech-to-text. *Communication Sciences & Disorders*, 28, 30–38. https://doi.org/10.12963/csd.23952.

Iancu, B. (2019). Evaluating google speech-to-text API's performance for Romanian e-learning resources. *Informatica Economica*, 23, 17–25. https://doi.org/10.12948/issn14531305/23.1.2019.02.

Kasparaitis, P. (2008). Lithuanian speech recognition using the English recognizer. *Informatica*, 19(4), 505–516. https://doi.org/10.15388/Informatica.2008.227.

Kobylyukh, L., Rybchak, Z., Basystiuk, O. (2023). Analyzing the accuracy of speech-to-text APIs in transcribing the Ukrainian language. In: *CEUR Workshop Proceedings*, Vol. 3396, pp. 217–227. https://ceur-ws.org/Vol-3396/paper18.pdf.

Kuligowska, K., Stanusch, M., Koniew, M. (2023). Challenges of automatic speech recognition for medical interviews – research for Polish language. *Procedia Computer Science*, 225, 1134–1141. https://doi.org/10.1016/j.procs.2023.10.101.

Laurinčiukaitė, S., Telksnys, L., Kasparaitis, P., Kliukienė, R., Paukštytė, V. (2018). Lithuanian speech corpus Liepa for development of human-computer interfaces working in voice recognition and synthesis mode. *Informatica*, 29(3), 487–498. https://doi.org/10.15388/Informatica.2018.177.

Lipeika, A., Lipeikienė, J., Telksnys, L. (2002). Development of isolated word speech recognition system. *Informatica*, 13(1), 37–46. https://doi.org/10.3233/INF-2002-13103.

Maskeliunas, R., Rudzionis, A., Ratkevicius, K., Rudzionis, V. (2009). Investigation of foreign languages models for Lithuanian speech recognition. *Elektronika ir Elektrotechnika*, 91(3), 15–20. https://eejournal.ktu.lt/index.php/elt/article/view/10271.

McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., Bourlard, H. (2004). *On the Use of Information Retrieval Measures for Speech Recognition Evaluation*. IDIAP Research Report 04-73. IDIAP Research Institute. https://publications.idiap.ch/downloads/reports/2004/rr04-73.pdf.

Pipiras, L., Maskeliunas, R., Damaševičius, R. (2019). Lithuanian speech recognition using purely phonetic deep learning. *Computers*, 8(4), 76. https://doi.org/10.3390/computers8040076.

Rasymas, T., Rudžionis, V. (2014). Combining multiple foreign language speech recognizers by using neural networks. In: *Human Language Technologies–The Baltic Perspective*, Vol. 268, pp. 33–39. https://doi.org/10.3233/978-1-61499-442-8-33.

Raškinis, G., Raškinienė, D. (2003). Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system. *Informatica*, 14(1), 75–84. https://doi.org/10.15388/Informatica.2003.005.

Rugayan, J., Salvi, G., Svendsen, T. (2023). Perceptual and task-oriented assessment of a semantic metric for ASR evaluation. In: *Proceedings of the INTERSPEECH 2023*, pp. 2158–2162. https://doi.org/10.21437/Interspeech.2023-1778.

Salimbajevs, A., Kapociute-Dzikiene, J. (2018). General-purpose Lithuanian automatic speech recognition system. In: *Proceedings of the 8th International Conference, Baltic HLT*, pp. 150–157.

Sasindran, Z., Yelchuri, H., Rao, S., Prabhakar, T. (2023). H_eval: a new hybrid evaluation metric for automatic speech recognition tasks. https://doi.org/10.48550/arXiv.2211.01722.

Siegert, I., Sinha, Y., Jokisch, O., Wendemuth, A. (2020). Recognition performance of selected speech recognition APIs – a longitudinal study. In: *Speech and Computer: 22nd International Conference, SPECOM 2020*. Springer-Verlag, pp. 520–529. 978-3-030-60275-8. https://doi.org/10.1007/978-3-030-60276-5_50.

Silber-Varod, V., Siegert, I., Jokisch, O., Sinha, Y., Geri, N. (2021). A cross-language study of speech recognition systems for English, German, and Hebrew. *Online Journal of Applied Knowledge Management*, 9(1), 1–15. https://doi.org/10.36965/OJAKM.2021.9(1)1-15.

Sipavičius, D., Maskeliunas, R. (2016). "Google" Lithuanian speech recognition efficiency evaluation research. In: Dregvaite, G., Damasevicius, R. (Eds.), *Information and Software Technologies*. Springer International Publishing, Cham, pp. 602–612. 978-3-319-46253-0. https://doi.org/10.1007/978-3-319-46254-7_49.

Yoo, H., Seo, S., Im, S., Gim, G. (2021). The performance evaluation of continuous speech recognition based on Korean phonological rules of cloud-based speech recognition open API. *International Journal of Networked and Distributed Computing*, 9(1), 10–18. https://doi.org/10.2991/ijndc.k.201218.005.

P. Kasparaitis (born in 1967) graduated from Vilnius University (Faculty of Mathematics) in 1991. In 2001, he defended his PhD thesis "Lithuanian Text-to-Speech Synthesis". Presently, he is an associate professor at Vilnius University. His current research interests include text-to-speech synthesis, speech recognition, and other areas of computer linguistics.