Designing a Speech Corpus for the Development and Evaluation of Dictation Systems in Latvian

Mārcis Pinnis[1], Askars Salimbajevs[1], Ilze Auziņa[2]
[1]Tilde, Vienibas gatve 75a, Riga, Latvia
[2]Institute of Mathematics and Computer Science, University of Latvia, 29 Raina Blvd., Riga, Latvia
E-mail: marcis.pinnis@tilde.lv, askars.salimbajevs@tilde.lv, ilze.auzina@lumii.lv

Abstract

In this paper the authors present a speech corpus designed and created for the development and evaluation of dictation systems in Latvian. The corpus consists of over nine hours of orthographically annotated speech from 30 different speakers. The corpus features spoken commands that are common for dictation systems for text editors. The corpus is evaluated in an automatic speech recognition scenario. Evaluation results in an ASR dictation scenario show that the addition of the corpus to the acoustic model training data in combination with language model adaptation allows to decrease the WER by up to relative 41.36% (or 16.83% in absolute numbers) compared to a baseline system without language model adaptation. Contribution of acoustic data augmentation is at relative 12.57% (or 3.43% absolute).

Keywords: speech recognition corpus, dictation systems, Latvian

1. Introduction

Automatic speech recognition (ASR) technologies for Latvian have a relatively short history because even three years ago (i.e., in 2013 and before) there was no orthographically annotated speech corpus, which could be used for ASR purposes, available. However, there have been attempts to develop ASR systems for broadcast speech recognition (Oparin et al, 2013) in the Quaero project (Lamel, 2012) using acoustic model bootstrapping. Since the creation of the first orthographically and phonetically annotated speech corpus for Latvian (Pinnis et al., 2014) – the *Latvian Speech Recognition Corpus* (LSRC), ASR technologies for Latvian have been actively researched (e.g., Salimbajevs & Pinnis, 2014, Salimbajevs & Strigins, 2015b), and different application scenarios that have resulted in practical applications with ASR capabilities (e.g., Vīra & Vasiļjevs, 2014; Salimbajevs & Strigins, 2015a; Znotiņš & Dargis, 2014; Znotiņš et. al., 2015) have been investigated. However, the technology has yet to reach a level where it is applicable in dictation scenarios in text editors.

For the development of dictation systems, it was necessary to create a specific corpus that would: 1) better capture the speaking characteristics of speakers when dictating text to a computer and 2) contain spoken commands common to dictation scenarios (e.g., punctuation, formatting, special symbol, and action commands). Therefore, in this paper, we present the *Dictated Speech Corpus* (DSC) that has been created to address these requirements.

Speech corpora creation with spoken commands and speech recognition system development with spoken command support has been investigated also in related research. For instance, Paul and Baker (1992) and Bernstein and Danielson (1992) created speech corpora for English continuous speech recognition with both verbalised and non-verbalised punctuation marks. Digalakis et al. (2003) created a speech recognition system for Greek that could handle special symbol and formatting commands, however they were introduced only in the speech recognition system and were omitted from their speech corpus. Enravi (2012) in his thesis stressed the importance of spoken command support in dictation systems. Rusko et al. (2011) developed a dictation system for Slovak that supported spoken commands for punctuations. Although spoken commands are investigated in related research, little attention has been given to the actual annotation of spoken commands in speech corpora. The focus of this paper is the creation of a speech corpus with spoken commands for dictation system development.

The paper is further structured as follows: 1) section 2 describes the requirements of the corpus, 2) section 3 describes the main statistics of the created corpus, 3), section 4 provides a discussion about the challenges we faced when developing the DSC, 4) section 5 presents the evaluation of the corpus in ASR scenarios, and 4) the paper is concluded in section 6.

2. Requirements of the Speech Corpus

The DSC is the second speech corpus created specifically for speech recognition in Latvian. Therefore, as the basis for the requirements, we used the requirements designed for the LSRC. This means that the audio data format requirements and orthographic annotation requirements remain the same. In this paper, we describe the key differences that address the needs of the dictation corpus. The following requirements were set for the DSC corpus:

- the corpus had to contain approximately ten hours of orthographically annotated speech data (or 10% of the length of the LSRC corpus);
- to capture cleaner speech data than in the LSRC, the DSC had to feature speeches with background noise that is common to only office spaces and rooms with relatively low levels of noise;
- because we aimed at recording speakers of active working ages, in terms of the physical characteristics of speakers, the corpus had to contain speeches from speakers of two age groups (16-25 and 26-50). Furthermore, the corpus had to feature speakers from both genders in equal proportions;

- in terms of speech styles, the DSC had to feature both spontaneous speech and prepared (read) speech;
- to ensure that the collected corpus features spoken commands, in terms of content, the corpus had to contain speeches from possible dictation scenarios (e.g., formal and personal e-mail letters, formal applications and instructions, social network and SMT messages, short blog and news articles, etc.);
- The DSC had to contain four types of spoken commands – punctuation, special symbol, formatting, and action commands. Examples of possible commands are given in Table 1. The commands had to be annotated using specific spoken command tags (this is the only difference in data formats between the LSRC and the DSC). An example of a speech in the *Transcriber*[1] software (also depicting annotation of spoken commands) is given in Figure 1. The example shows usage of two spoken commands (two special symbol commands for an exclamation mark and the emoticon of a smiling face). An example of the XML-based orthographic annotation of spoken commands is given in Figure 2. The XML document contains the same speech fragment that is depicted in Figure 1.

Command type	Command	Example of pronunciation
Punctuation	full stop	*punkts*
	comma	*komats*
	new paragraph	*jauna rinda*
	exclamation mark	*izsaukuma zīme*
	euro sign	*eiro zīme*
Formatting	align left	*līdzināt pa kreisi*
	centre text	*centrēt tekstu*
	normal	*parasts teksts*
	bold	*treknraksts*
	italic	*slīpraksts*
Action	undo	*atsaukt tekstu*
	delete last word	*izdzēst pēdējo vārdu*

Table 1: Spoken command examples for dictation systems

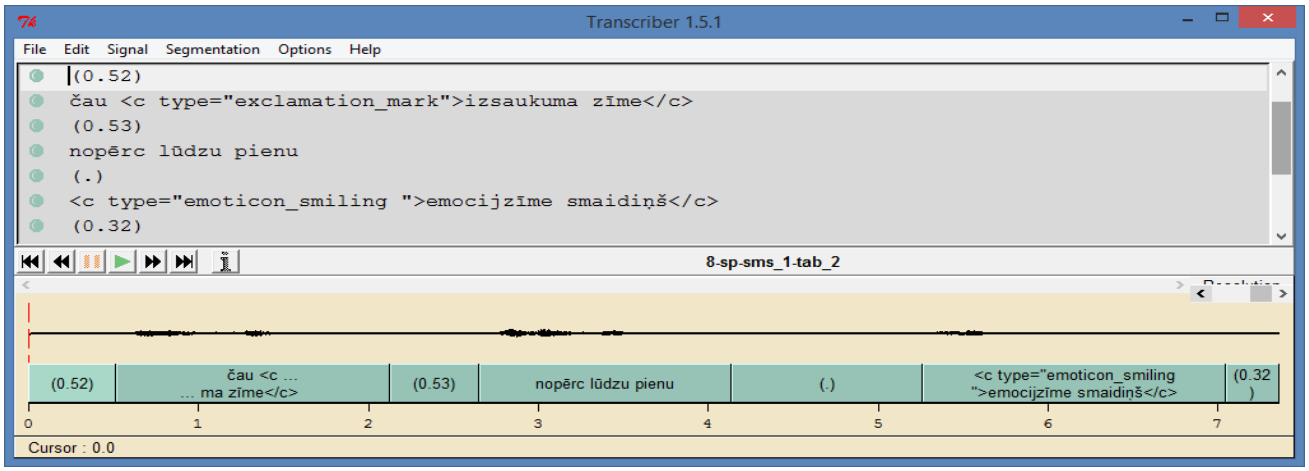

Figure 1: An example of the orthographic annotation of a spontaneously dictated SMS message in the speech transcription software *Transcriber*

```
<?xml version="1.0" encoding="UTF-8"?>
<files>
 <file name="[Recording Name]">
  <fragment place="2" recorder="2" speaker="8" type="2">
   <part length="0.52" audio_file="[File Path]" >(0.52)</part>
   <part length="1.611" audio_file="[File Path]" >čau &lt;c
    type="exclamation_mark"&gt; izsaukuma zīme &lt;/c&gt;</part>
   <part length="0.528" audio_file="[File Path]" >(.h)</part>
   <part length="1.482" audio_file="[File Path]" >nopērc lūdzu pienu</part>
   <part length="1.119" audio_file="[File Path]" >(.)</part>
   <part length="1.789" audio_file="[File Path]" >&lt;c
    type="emoticon_smiling"&gt; emocijzīme smaidiņš &lt;/c&gt;</part>
   <part length="0.316" audio file="[File Path]" >(0.32)</part>
  </fragment>
 </file>
</files>
```

Figure 2: An example excerpt of the orthographically annotated DSC depicting spoken command annotation

[1] Transcriber can be found online at: http://sourceforge.net/projects/trans/.

3. Statistics of the Speech Corpus

The speech corpus was annotated over a time period of two months. Each recording was annotated by one linguist and revised by a second linguist. After linguists completed their work, the annotation was semi-automatically validated and inconsistencies (e.g., annotations for words that were pronounced similarly, but annotated differently) were manually corrected. The total length of the DSC is 9 hours 19 minutes and 46 seconds. It consists of 287 speeches (22,763 running words) that are spoken by 30 speakers (15 men and 15 women; the data distribution with respect to the genders is given in Figure 3) of two age groups (see Figure 4 for more details). The corpus features speeches with both spontaneous and prepared (read) speech (see Figure 5).

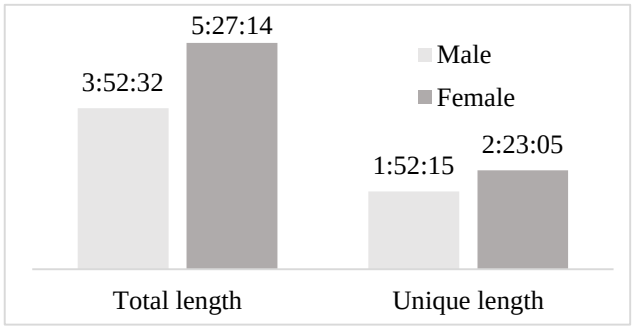

Figure 3: Statistics of the speech corpus with respect to the gender of speakers

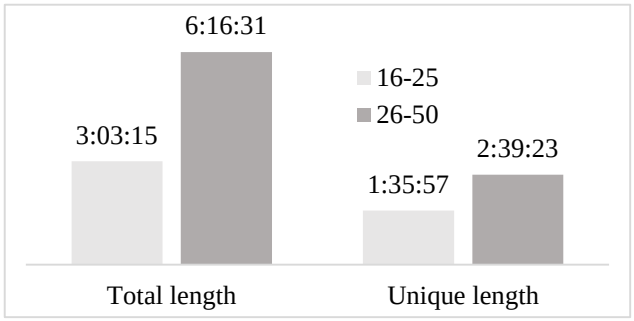

Figure 4: Statistics of the speech corpus with respect to the age of speakers

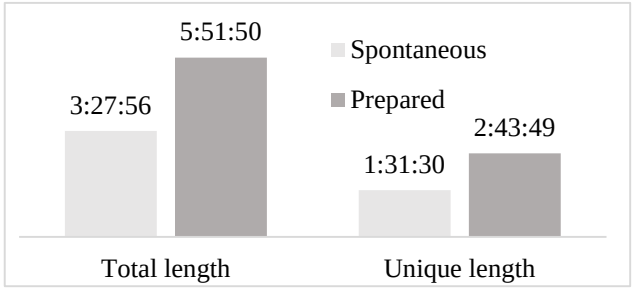

Figure 5: Statistics of the speech corpus with respect to speech styles

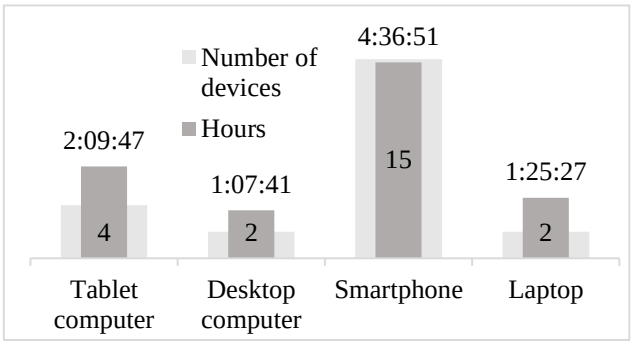

Figure 6: Statistics of the speech corpus with respect to different recording devices

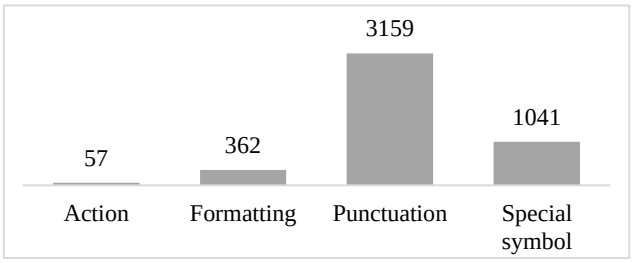

Figure 7: Occurrences of spoken commands in unique speeches

The speeches were recorded using 23 different recording devices (the statistics with respect to recording devices are given in Figure 6). Each speech was recorded simultaneously with one to four different devices (therefore, the tables and figures show the total length and the unique length). To perform the orthographic annotation only once per speech, the audio files representing the same speeches from different recording devices were manually cut to have similar start and end times.

The corpus features 36 different speech commands, which are used in dictation scenarios, and 93 different pronunciation variants of the commands. The most frequent commands are punctuation commands (over 68%), followed by special symbol commands (over 22%). The statistics of different command types are given in Figure 7. The corpus contains a total of 4,619 occurrences of spoken commands in unique speeches.

Analysis of the top 10 spoken commands (see Table 2) has shown that the three most frequent commands alone account for 62.5% of all spoken command occurrences. The statistics also show that the most frequent commands (e.g., comma, full stop, new paragraph) have one main pronunciation variant, whereas less frequent commands tend to have up to five (e.g., left parenthesis) and even six (e.g., left double quotation mark and delete last word) different pronunciation variants.

Command	Occurrences	% of all
comma	1095	23.7%
full stop	974	21.1%
new paragraph	821	17.8%
exclamation mark	289	6.3%
hyphen-minus	194	4.2%
colon	149	3.2%
normal text	124	2.7%
question mark	99	2.1%
right parenthesis	91	2.0%
italic text	85	1.8%
Total:	3921	84.8%

Command	Pronunciation	Occurrences	% of all
comma	*komats*	1095	23.7%
full stop	*punkts*	974	21.1%
new paragraph	*jauna rinda*	806	17.4%
exclamation mark	*izsaukuma zīme*	214	4.6%
hyphen-minus	*domu zīme*	189	4.1%
colon	*kols*	149	3.2%
normal text	*parasts teksts*	96	2.1%
question mark	*jautājuma zīme*	87	1.9%
italic text	*slīpraksts*	84	1.8%
exclamation mark	*izsaukumzīme*	74	1.6%
Total:		3768	81.5%

Table 2: Statistics of the top 10 most frequent spoken commands (left)
and the top 10 most frequent pronunciation variants (right) within the dictated speech corpus

4. Issues Identified During Corpus Creation and Lessons Learned

In this section, we briefly discuss issues that needed to be addressed in the speech data recording and orthographic annotation steps.

In the audio data recording step, the most difficult aspect was the recording of spontaneous speech with spoken commands. It was difficult for the speakers to think of a topic during the recording sessions that would contain the spoken commands that we had envisioned. Therefore, we instructed speakers prior to the recording about possible topic areas (e.g., commenting a news article, answering to a tweet, composing an informal SMS message, composing an e-mail message that they would have to send during the day, etc.) and provided a list of possible spoken commands that they could use during the recording session. However, speakers were instructed to not limit themselves to the list if they thought that a different spoken command was necessary.

Contrary to read speech where speakers during recording sessions pronounced almost all spoken commands, which were included in the texts, we observed that during spontaneous speech recording sessions speakers often (unintentionally) skipped spoken commands. For instance, commas that separate subordinate clauses were often not pronounced. This is because the speakers were focused on how to express their thoughts, while forgetting about punctuation marks that would need to be pronounced.

During the annotation process and after, the DSC was automatically validated for machine readability. This allowed us to identify issues in the annotations that would possibly corrupt the data in our speech recognition system training workflows (e.g., incorrect number of words in the orthographic annotation that a spelling correction is linked to, unclosed tags, overlapping tags, etc.).

After the automatic validation, all commands and spelling corrections were manually reviewed for correctness and consistency. This step was necessary, because we identified that the orthographic transcriptions and spelling corrections of foreign words, abbreviations, acronyms, and named entities were not always consistent (also equally sounding words were in some cases annotated inconsistently). For named entities, the capitalisation of words was not always according to the annotation guidelines (i.e., only proper nouns that are always written with an upper-case letter should have been written with an upper-case letter). There were also instances where common words (e.g., currency names) were annotated as commands although the words are frequently found in texts as normal words.

Several examples of mistakes and corrected annotations are given in Table 3. All these issues allowed us to identify weak points in our annotation guidelines. The guidelines were, therefore, clarified so that annotation inconsistencies in further (and also future) annotation efforts would be mitigated.

5. Evaluation

The baseline ASR system was trained on the 100 hour long Latvian Speech Recognition Corpus (Pinnis et al., 2014) using the Kaldi ASR toolkit (Povey et al., 2011).

The following training procedure was used for Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) training:

- First, monophone models were trained on a small subset of the shortest utterances in the training corpus.
- Then, more complex models (triphone models, models with Linear Discriminant Analysis (LDA) transform, and models with feature space Maximum Likelihood Linear Regression (fMLLR) speaker adaptation) were trained sequentially, increasing the amount of training data each time.
- In the last step, speaker adaptive training (SAT) was performed on all available training data.

Next, this final HMM-GMM model was used to create phoneme alignments for all 100 hours of training data. The phoneme alignments were then used as inputs and outputs for deep neural network (HMM-DNN) acoustic model training.

Initial annotation	Corrected annotation	Reason
deviņdesmit **<c type="euro_sign">** eiro **</c>**	deviņdesmit eiro	Currency names are not spoken commands
<c type="**undo**"> atsaukt iepriekšējo vārdu </c>	<c type="**ctrl_backspace**"> atsaukt iepriekšējo vārdu </c>	Wrong command type selected in the annotation
<c type="at_sign"> **et** zīme </c> <c type="at_sign"> **at** zīme </c>	<c type="at_sign"> **et** zīme </c> <c type="at_sign"> **et** zīme </c>	Inconsistent annotation of equally pronounced spoken commands
siā [**3** , SIA]	siā [**1** , SIA]	Incorrect number of words in the orthographic annotation that the spelling correction is linked to
poverpoint [1 , en , **p**ower**p**oint]	poverpoint [1 , en , **P**ower**P**oint]	Incorrect capitalisation of foreign language named entities
Laba **D**aba	**l**aba **d**aba	Incorrect capitalisation of words frequently occurring in lowercase forms in Latvian named entities
ekolain [1 , en , **Ecoline**]	ekolain [1 , en , **Ecolines**]	Incorrect spelling of (foreign language) named entities
seši diena **[2 , 6Diena]** dē divi ā er **[4 , D2AR]** el tē vē viens **[4 , LTV1]**	seši diena dē **[1 , D]** divi ā er **[2 , AR]** el tē vē **[3 , LTV]** viens	Abbreviations or named entities with numerals are annotated as separate words

Table 3: Examples of inconsistencies identified during the validation of the orthographic annotation

The language model (LM) was a 3-gram model trained on a 22 million sentence text corpus, which was collected from Latvian web news portals. The text corpus was specially processed to adapt it to the dictation task:

- Formatting and action commands were artificially added as bigram and trigram counts.
- Punctuation and special symbols were replaced with their respective pronunciations.

"New line" commands were appended after every second sentence in the text corpus.

A roughly one hour long held-out data set from the DSC was used as an evaluation data set. The remaining data was used for augmenting training data after the last step of HMM-GMM training. This means that only the SAT was repeated.

First, the existing general transcription system (without LM adaptation and trained on the 100 hour long LSRC data set) was evaluated on the held-out set from the DSC. The resulting high word error rate (WER) of 40.69% indicates that there is a significant difference between dictation and transcription tasks.

Next, HMM-GMM models were trained and evaluated. The results in Table 4 show that by augmenting training data with 8 hours of dictated speech it was possible to reduce the WER by relative 17% (or 7.36% absolute). The effect of the special text corpus processing was also evaluated. Without this adaptation, it is difficult for both systems to correctly recognise commands and the resulting WER is very high. Even in the case when the adapted acoustic model is used, the WER still is 35.71%. By using the special text corpus processing alone, the WER can be improved by relative 30.69% (or 13.22% in absolute numbers). However, in this case, the improvement from using DSC in acoustic training is much smaller; the WER is reduced only by 6.53% (or 1.95% absolute).

Next, HMM-DNN models were trained using maximum likelihood criteria and a similar comparison was performed (see Table 5). As in previous case, special text corpus processing for LM training gives significant improvement – 32.93% relative (or absolute 13.4%). By augmenting training data with dictated speech, the WER is further improved by 12.57% (or 3.43% absolute).

Training set	Language model	WER
Baseline (100 hours)	Not adapted	43.07%
	Adapted	29.85%
Augmented (108 hours)	Not adapted	35.71%
	Adapted	27.90%

Table 4: Evaluation of HMM-GMM models

ASR system	WER
Baseline with non-adapted LM (100 hours)	40.69%
Baseline with adapted LM (100 hours)	27.29%
Augmented (108 hours) with adapted LM	**23.86%**

Table 5: Evaluation of HMM-DNN models

If this result is compared to a non-adapted general domain system that is trained on the 100 hour long LSRC data set without language model adaptation, then the overall

improvement from both language model adaptation and acoustic data augmentation is at relative 41.36% (or 16.83% in absolute numbers).

It is noticeable that there is a relatively small difference between baseline HMM-DNN and baseline HMM-GMM systems without LM adaptation. This is the result of a mismatch between training and testing conditions. However, when both LM adaptation and additional training data is used, the difference between HMM-DNN and HMM-GMM becomes large – 14.48% relative improvement (or 4.04% absolute).

We performed also experiments with discriminative training, however we did not observe any improvement. On the contrary, the WER degraded by approximately 10% when compared to a non-discriminatively trained system. Discriminative training had no effect in the case of general transcription system. This is probably caused by the fact that the size of DSC is relatively small in comparison with the remaining training data (the DSC contribution is approximately 7.5%). Therefore, the acoustic models get more adapted to the 100 hours of non-dictation speech data. Analysis of this issue is beyond the scope of this paper.

6. Conclusion

In the paper, we have presented the Dictated Speech Corpus that was created for dictation system development and evaluation in Latvian. We discussed the requirements and statistics of the DSC. We also provided a discussion about challenging issues during the creation of the corpus and also about lessons learned for future corpus creation efforts. The corpus has been evaluated in an ASR scenario, and the results show that DSC in combination with language model adaptation allows to decrease the WER in a dictation scenario by up to relative 41.36% (or 16.83% in absolute numbers) compared to a baseline system trained on the LSRC data set without language model adaptation. The improvement from augmenting acoustic model training data is at relative 12.57% (or 3.43% in absolute numbers).

7. Acknowledgements

The research leading to these results has received funding from the research project "Information and Communication Technology Competence Centre" of EU Structural funds, contract Nº. L-KC-11-0003 signed between the ICT Competence Centre and the Investment and Development Agency of Latvia, Research No. 2.4 "Speech recognition technologies".

8. References

Bernstein, J., & Danielson, D. (1992). Spontaneous Speech Collection for the CSR Corpus. In *Proceedings of the Workshop on Speech and Natural Language* (pp. 373–378). Association for Computational Linguistics.

Digalakis, V., Oikonomidis, D., Pratsolis, D., Tsourakis, N., Vosnidis, C., Chatzichrisafis, N., & Diakoloukas, V. (2003). Large Vocabulary Continuous Speech Recognition in Greek: Corpus and an Automatic Dictation System. In *Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003)* (pp. 1565–1568).

Enarvi, S. (2012). *Finnish Language Speech Recognition for Dental Health Care*. Aalto University.

Lamel, L. (2012). Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data. In *The Fifth International Conference: Human Language Technologies-The Baltic Perspective*.

Oparin, I., Lamel, L., & Gauvain, J. (2013). Rapid Development of a Latvian Speech-to-text System. In ICASSP'13 (pp. 2–6). Vancouver, Canada.

Paul, D. B., & Baker, J. M. (1992). The Design for the Wall Street Journal-based CSR Corpus. In *Proceedings of the Workshop on Speech and Natural Language* (pp. 357–362). Association for Computational Linguistics.

Pinnis, M., Auziņa, I., & Goba, K. (2014). Designing the Latvian Speech Recognition Corpus. In *Proceedings of LREC 2014*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Rusko, M., Juhár, J., Trnka, M., Staš, J., Darjaa, S., Hládek, D., … Lojka, M. (2011). Slovak Automatic Transcription and Dictation System for the Judicial Domain. In *Human Language Technologies as a Challenge for Computer Science and Linguistics: 5th Language & Technology Conference* (pp. 365–369). doi:10.1007/978-3-319-08958-4_2

Salimbajevs, A., & Pinnis, M. (2014). Towards Large Vocabulary Automatic Speech Recognition for Latvian. In *Human Language Technologies – The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014* (Vol. 268, pp. 236–243).

Salimbajevs, A., & Strigins, J. (2015a). Latvian Speech-to-Text Transcription Service. In *Proceedings of Interspeech 2015* (pp. 722–723).

Salimbajevs, A., & Strigins, J. (2015b). Using sub-word n-gram models for dealing with OOV in large vocabulary speech recognition for Latvian. In *Proceedings of NODALIDA 2015* (pp. 281–285).

Vīra, I., & Vasiļjevs, A. (2014). The Development of Conversational Agent Based Interface. In *Human Language Technologies - The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014* (Vol. 268, p. 46).

Znotiņš, A., & Darģis, R. (2014) Baseline for Keyword Spotting in Latvian Broadcast Speech. In *Human Language Technologies - The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014* (Vol. 268, pp. 75-82).

Znotiņš, A., Polis, K., Darģis, R. (2015) Media Monitoring System for Latvian Radio and TV Broadcasts. In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association (Interspeech 2015)* (pp. 732-733).