TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation

Matúš Pleva, Jozef Juhár
Department of Electronics and Multimedia Communications,
Technical University of Košice,
Letná 9, 042 00 Košice, Slovakia
E-mail: Matus.Pleva@tuke.sk, Jozef.Juhar@tuke.sk

Abstract

This article presents an overview of the existing acoustical corpuses suitable for broadcast news automatic transcription task in the Slovak language. The TUKE-BNews-SK database created in our department was built to support the application development for automatic broadcast news processing and spontaneous speech recognition of the Slovak language. The audio corpus is composed of 479 Slovak TV broadcast news shows from public Slovak television called STV1 or "Jednotka" containing 265 hours of material and 186 hours of clean transcribed speech (4 hours subset extracted for testing purposes). The recordings were manually transcribed using Transcriber tool modified for Slovak annotators and automatic Slovak spell checking. The corpus design, acquisition, annotation scheme and pronunciation transcription is described together with corpus statistics and tools used. Finally the evaluation procedure using automatic speech recognition is presented on the broadcast news and parliamentary speeches test sets.

Keywords: broadcast news, Slovak language, spontaneous speech

1. Introduction

The Slovak language belongs to a group of Slavic languages, which are typical of inflection and free word order. These features make the Slovak automatic speech recognition task very complicated, and an extremely large amount of data is required for automatic large vocabulary spontaneous speech recognition. Different types of text and speech corpora are needed for complex applications such as automatic broadcast news (BN) processing or media monitoring. All focus conditions (Stern, 1997) should be distributed in the acoustical part of the speech corpus as well. The broadcast news monitoring and automatic speech transcription of BN shows are very popular issues nowadays, because the government regulation usually specifies the minimal amount of shows with hidden subtitles for hearing impaired spectators.

There are several BN corpuses already available in other languages. The *Czech* TV & Radio Broadcast News speech corpus contains 50 hours of recordings and *26 hours* of pure transcribed speech (Ircing et al., 2001). The *French* corpus of the ESTER Evaluation Campaign contains *100 hours* recorded from 6 French radio broadcasters using 16Khz/16bit quality (Galliano et al., 2006). The *French* ETAPE corpus consists of *30 hours* of TV and radio broadcasts, selected to cover a wide variety of topics and speaking styles, emphasizing spontaneous speech and multiple speaker areas (Gravier et al., 2012). The *Thai* Broadcast News Corpus contains about *17 hours* of speech data while the text corpus was transcribed from around 35 hours of television broadcast news (Jongtaveesataporn et al., 2008), but there is also an ongoing LOTUS-BN project with goal of collecting *100 hours* of the transcribed Thai BN shows (Chotimongkol et al., 2009). The RUNDKAST: *Norwegian* broadcast news speech corpus contains recordings of approximately *77 hours* of broadcast news shows from the Norwegian broadcasting company NRK (Amdal et al., 2008). The *Slovenian* BN database (SiBN) contains *29 hours* of the transcribed speech from public RTVSLO-1 TV station and 35 hours of recordings (Žibert & Mihelič, 2004). The *Iberian* KALAKA-2 BN corpus, created to support the Albayzin 2010 Language Recognition Evaluation, contains around *125 hours* of speech (Rodríguez-Fuentes et al., 2012). And of course the LDC Hub4 BN corpuses of *English* speech: *75 hours* in 1996 set and *72 hours* in 1997 set (Graff, 2002).

The Slovak language is a minor European language with approximately 5 million of native speakers. Despite that there are different types of speech corpora already available. For example, a large Slovak speech database was created as a part of SpeechDat-E (II) project (*100 hours* of speech over public switched telephone network A-law compression 8kHz sampling frequency, mainly simple commands, available as ELRA-S0095) (Pollak et al., 2000), a database named MobilDat (*100 hours*, similar corpus to SpeechDat but recorded over mobile GSM network from different environments, not publicly available) (Rusko et al., 2006), Parliament speech database (*136 hours* of annotated parliamentary speech from the Slovak parliament with 48kHz quality, contains mainly monologues, not publicly available) (Darjaa et al., 2011), APD project database (*250 hours* of read court proceedings, planned speech, contains only monologues, recorded in studio environment with 48kHz, not publicly available) (Rusko et al., 2011), etc. Unfortunately no Slovak annotated database consisting of different dialogs, spontaneous speech or live coverage with different background conditions is available for automatic broadcast news processing and spontaneous speech recognition task.

2. TUKE-BNews-SK Corpus Design

During last years a *new broadcast news corpus TUKE-BNews-SK* for building acoustic and language models was created in our laboratory consisting of *265*

hours of recorded TV broadcast news shows and annotated using Transcriber tool (Barras et al., 2001). *178'152 speech utterances* extracted from the corpus suitable for continuous speech recognition acoustic model training cover around *186 hours of annotated corpus*. The recordings were made in MPEG2 format from digital broadcast of the Slovak public TV "Jednotka".

The textual part of the corpus brings important information also for spontaneous speech language model adaptation for future experiments, because the transcribed utterances in the shows contain not only planned but also a *32.7 hours* of spontaneous speech (F1 - condition in Table 1) which is a very challenging task. The distribution of all focus conditions and speaker gender is presented in the Table 1 and Table 2 below.

F0 – prepared speech in studio	94.38 h
F1 – spontaneous speech in studio	32.70 h
F2 – prepared telephone speech (reduced-bandwidth)	2.07 h
F3 – speech with music in background (SNR<10dB)	19.15 h
F4 – speech under degraded acoustical conditions	43.36 h
F5 – speech performed by a non-native speaker	1.24 h
FX – combination of the focus conditions listed above (F1-F5)	21.39 h

Table 1: Focus conditions distributions in Slovak BN Corpus (TUKE-BNews-SK).

Speaker gender	**Number of *utterances***	**Percent from all**
Female	88 941	47%
Male	99 882	53%
Speaker gender	**Number of *speakers***	**Percent from all**
Female	4 195	37%
Male	7 447	63%

Table 2: Gender distribution in Slovak BN Corpus (TUKE-BNews-SK) of all speech segments (it covers also utterances excluded from processing, because they contain malformed speech content).

The corpus contains 187'756 words in dictionary extracted from 1'691'122 tokens in 166'938 utterances from 11'345 speakers in the training set (statistics generated also using Nechala (2014) tool). The training set generation process includes filtering of inappropriate non-speech tags or speech errors (stammering speaker, words which even annotators could not understand, etc.).

3. The Annotation Scheme

The annotation scheme used in TUKE-BNews-SK was constructed from DARPA Hub4 evaluation campaign (Stern, 1997) and LDC corpus building instructions compiled together during COST-278 project and described in details by (Žgank et al., 2004b).

The annotation scheme was further extended for better description of frequent noise and non-speech events in our database. For example all noises from Transcriber were extended by their background alternative. The bell sound, overloading of the microphone input, applause and cheering was added because of frequent occurrence during outdoor or sports match reports.

More phonetic sets derived from Slovak SAMPA (Ivanecky & Nabelkova, 2002) were evaluated, because some phones have a rare occurrence and thanks to small training data they do not improve the whole recognition results. First of all, the SpeechDat based set was used as the main phonetic set with 57 phonemes named "SD" set. Next the reduction of the set was realized using only 45 most used phonemes named "SAV" set (no diphthongs, and different pronunciations of graphemes "v", "f", "r", "l" & "n"). And finally an extended version containing 51 phonemes (diphthongs - back again and "shva" phoneme introduced) was evaluated and named "SAVE" set.

Context dependent triphones were evaluated too and the state tying mechanism from MASPER initiative was compared with the results of the triphone mapping solution described in (Darjaa et al., 2011b).

4. The Pronunciation Transcription

The pronunciation dictionary was built using our Perl tool which uses reprogrammed & extended Ivanecky (2003) rules. The tool is generating mainly word level phonetic transcription as it was used in the standard MASPER training, but inter-word phone dependent transcription could improve the results for spontaneous speech. The inter-word transcription is difficult if there are noise tags or any other non-speech tags present, because the tags should be removed for phonetic transcription process and then restored in previous positions.

We plan to extend the phonetic transcription scripts to handle the tags in the sentence level processing and add all new pronunciation alternatives to the resulting phonetic dictionary automatically for the speech recognition task.

5. Corpus Acquisition

The database was captured using Technisat AirStar PCI card of digital terrestrial broadcast (DVB-T) available in Kosice region. The audio data was mostly recorded in original transmitted stream of MPEG1 Audio Layer 2 coded stereo in 128kbit/s and 48kHz sampling rate quality. Audio data were converted to mono after extraction of the RAW waveform and down sampled to resultant 16kHz sampling rate format. The original audio is also available. The quality of the audio is affected by the compression algorithm used in DVB-T transmission. This format is a wide standard in the state-of-art digital broadcast systems,

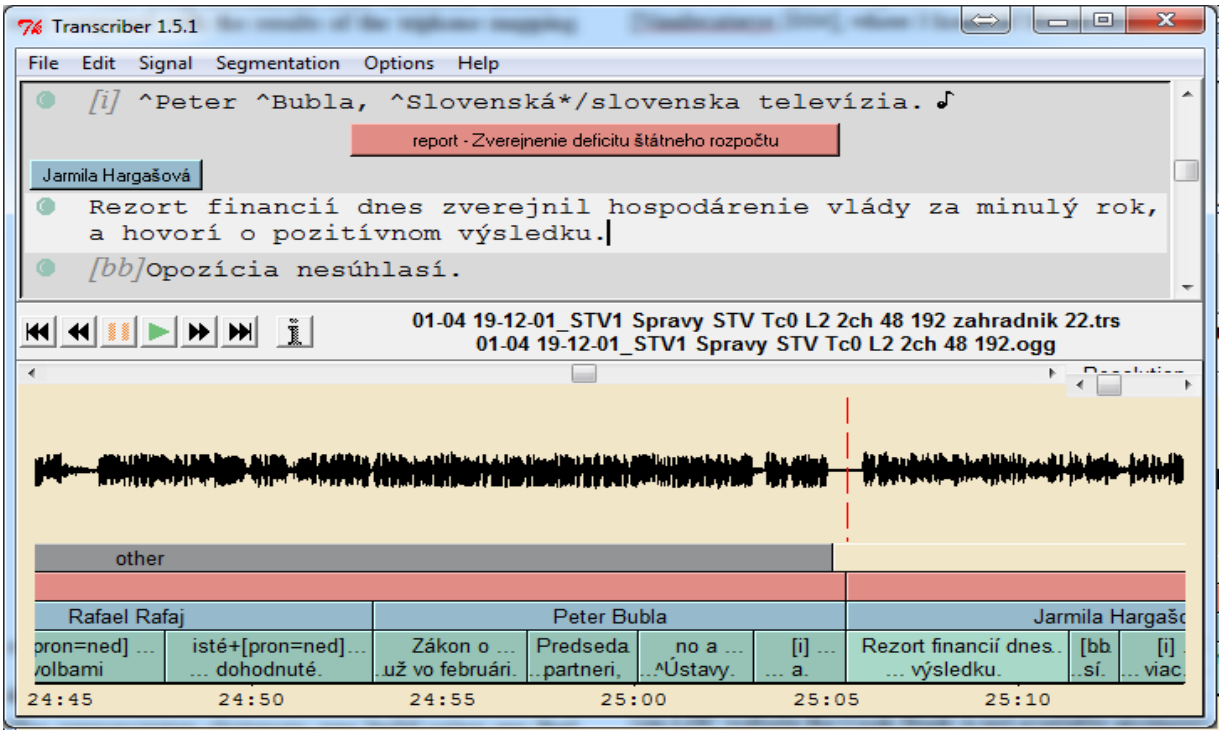

Figure 1: Example of the annotation in the chosen Transcriber tool

so the audio data will have the same characteristics in common BN automatic transcription system input.

The TUKE-BNews-SK database *was constructed in 3 phases* during 7 years of working on different topics.

In the *first phase* our department joined the COST-278 pan-European database initiative (Vandecatseye et al., 2004), where 3 hours of Slovak BN shows (private TA3 TV) were transcribed and segmentation and clustering algorithms were evaluated. In this phase the Hub4 LDC Corpus Cook Book transcription conventions (on LDC website the Cook Book is not available anymore) for annotation were used.

In the *second phase* the KEMT-BN1 database was constructed using previous experiences and consists of 48 hours of recordings and annotations (STV1 evening news). This database was used to train and evaluate the first Slovak BN acoustic models. Based on the results and experiences we have concluded that more language resources are needed to train acoustic models suitable for automatic continuous speech recognition of Slovak BN shows.

In the *third phase* the first Slovak automatic speech recognition system was built and next 210 hours of material was captured from STV1 (Jednotka) television, transcribed and evaluated (KEMT-BN2). An extended set (more detailed) of noise and non-speech tags was introduced for improving the third phase transcriptions and for future use of non-speech events processing during the language model evaluation.

6. Annotation Tools and Formats

All *manual* annotations (no texts was provided together with the recordings) were realized in the modified *Transcriber* tool (see Figure 1), where new noise and non-speech tags were introduced and the export to STM format was modified (to force all non-speech and noise tags to remain in the output text file, and to fix *UTF8* characters handling). An automatic Slovak grammar check was implemented and the Transcriber plugin modification was used during the third phase of the annotation process (also because of the faulty *UTF8* characters handling).

The native Transcriber xml files *.trs* (see Figure 2) are along with the original media files included in the final database.

```
<Event desc="i" type="noise" extent="instantaneous"/>
    Tí to však popierajú.
</Turn>
  <Turn speaker="spk4" mode="planned"
  fidelity="high" channel="studio"startTime="57.783"
  endTime="76.299">
<Sync time="57.783"/>
  V korupčnej kauze ide o nájomné byty v ^Košiciach
<Sync time="61.329"/>
  ktoré stavala firma ^Kame.
<Sync time="62.985"/>
```

Figure 2: Example of TRS native Transcriber XML format from the TUKE-BNews-SK corpus

The *STM* format transcriptions (the NIST Scoring toolkit Sclite format) were exported (see Figure 3) together with the *WAV* audio files that were used as the input for next processing of the corpus creation mechanism. The modified TCl/Tk Transcriber scripts are freely available

together with this submission through *LRE Map*. The database is distributed together with the original video files for speaker verification purposes. The annotators used the video files for identification of the real speaker names from headlines in the broadcast news.

```
stv1_hl_spravy_17 1 Jarmila_Hargašová 55.561 57.783
    <o,f0,female> [i] Tí to však popierajú.
stv1_hl_spravy_17 1 Katarína_Krajňáková 57.783
    61.329 <o,f0,female> V korupčnej kauze ide o
    nájomné byty v ^Košiciach
stv1_hl_spravy_17 1 Katarína_Krajňáková 61.329
    62.985 <o,f0,female> ktoré stavala firma ^Kame.
```

Figure 3: Example of the exported STM NIST Sclite format from the TUKE-BNews-SK corpus

The selection of the annotated data segmentation is also important. As you can see in the Figure 2/3 the silence inside a compound sentence shorter than 0.5 seconds was segmented in natural breakpoints (usually when the speaker makes a pause), so not a strict sentence level segmentation was chosen. Breakpoint in the middle of the silence part was inserted when the pause in the speech utterance is between 0.5 and 1.5 seconds (also in simple sentences). If the pause was longer than 1.5 seconds, a special silence segment was inserted. Foreign language utterances were marked with special tags, but the content was not annotated.

7. Evaluation of the Corpus

The acoustic model training for corpus evaluation process was realized using the extension of Refrec (Lindberg et al., 2000) and MASPER (Zgank et al., 2004) training scripts, which consist of algorithms for conversion of the databases in SpeechDat format (Pollak et al., 2000). Also the configuration script, which includes all possible combinations of configuration in one place was compiled and the mapping of noise and non-speech tags to different smaller sets was realized.

The training procedure was modified for continuous speech recognition and inter-word triphones creation. The unique triphone mapping algorithm (Darjaa et al., 2011 & 2011b) was implemented and parallel threads training modification for speeding up the evaluation was redesigned. Finally, the filtering scripts for improving the training utterances selection process were evaluated. For example: the sentences, where the forced alignment recognition algorithm failed during the MASPER training (Zgank et al., 2004), (so called outliers) were filtered out from next training purposes.

The resulting acoustic model was evaluated using language model built from different Slovak text corpora (approximately 10^9 tokens) in our department described in following papers (Hládek & Staš, 2010; Juhár et al., 2012; Zlacký et al., 2013) and the open source Julius recognition engine (Lee et al., 2009) was used for automatic speech recognition on broadcast news and parliamentary speech test sets. The 240 minutes (4h) subset of TUKE-BNews-SK corpus was extracted for this purpose containing 4343 sentences. The parliamentary testing set of 75 minutes contains 884 sentences from database compiled on UI SAV (Rusko et al., 2011). The results of the automatic transcription are presented in the Table 3. For comparing the impact of the acoustic similarity between testing and training set the acoustic model based on Parliamentary speech database (136h) was used for evaluation (Darjaa et al., 2011).

WER [%]	BN AM	Parliament AM
BN test set	*10.09*	13.59
Parliament test set	17.28	12.62

Table 3: Comparison of ASR test results of the acoustic model trained on Slovak BN Corpus (TUKE-BNews-SK) and acoustic model trained on Parliamentary speeches.

8. Conclusion

Our goal was to develop a big Broadcast News speech database for Slovak BN and spontaneous speech which will be available through ELRA/ELDA association. We are working hard to acquire the broadcaster agreement of using the captured multimedia content and annotations outside of our laboratory, so the database is not freely available language resource in the time of the submission. Unfortunately the negotiation procedure could take more time and effort than expected during the corpus construction.

Finally we are working intensively on the web online (bn.kemt.fei.tuke.sk) automatic multimedia indexing database which will be available for the public, where any new media file could be uploaded and after automatic transcription process the subtitles for the corresponding media will be available. The resulting audio or video file could be played together with subtitle in optional karaoke format and edited afterwards. Also an audio query search engine will be included based on Gubka (2013).

9. Acknowledgements

The research presented in this paper was supported by Research and Development Operational Program funded by the ERDF under the project numbers ITMS-26220220141 (50%), ITMS-26220220182 (25%) & ITMS-26220220155 (25%).

10. References

Amdal, I., Strand, O. M., Almberg, J. and Svendsen, T. (2008). RUNDKAST: an Annotated Norwegian Broadcast News Speech Corpus. In *Proceeding of LREC 2008,* Marrakech, Morocco, pp. 1907- 1913.

Barras, C., Geoffrois, E., Wu, Z. and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. In: *Speech Communication. Special issue on Speech Annotation and Corpus Tools*, vol. 33(1-2), pp. 5-22.

Chotimongkol, A., Saykhum, K., Chootrakool, P., Thatphithakkul, N. and Wutiwiwatchai, C. (2009).

LOTUS-BN: A Thai broadcast news corpus and its research applications. In *International Conference on Speech Database and Assessments, 2009 Oriental COCOSDA,* IEEE, Nat. Electron. & Comput. Technol. Center (NECTEC), Pathumthani, Thailand, pp. 44-50.

Darjaa, S., Cerňak, M., Beňuš, Š.,Rusko, M., Sabo, R. and Trnka, M. (2011). Rule-based triphone mapping for acoustic modeling in automatic speech recognition, *Text Speech and Dialogue* 2011, Pilsen, Springer LNAI series, vol. 6836, pp. 268-275.

Darjaa, S., Cerňak, M., Trnka, M., Rusko, M. and Sabo, R. (2011b). Effective Triphone Mapping for Acoustic Modeling in Speech Recognition, *Proceedings of Interspeech 2011*, Florence, Italy, pp. 1717-1720.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J. F., Mostefa, D. and Choukri, K. (2006). Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. *Proc. of LREC 2006,* Vol. 6, Genoa, Italy, pp. 315-320.

Graff, D. (2002). An overview of Broadcast News corpora. *Speech Communication*, vol.*37* (1), pp. 15-26.

Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A. and Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *International Conference on Language Resources, Evaluation and Corpora*. *LREC 2012,* Istanbul, Turkey, pp. 114-118.

Gubka, R., Kuba, M. and Jarina, R. (2013). Universal approach for sequential audio pattern search. *Federated Conference on Computer Science and Information Systems, FedCSIS 2013*, art. no. 6644057, pp. 565-569.

Hládek, D. and Staš, J. (2010). Text mining and processing for corpora creation in Slovak language. *Journal of Computer Science and Control Systems*, Vol. 3 (1), ISSN 1844-6043, pp. 65–68.

Ircing, P., Krbec, P., Hajic, J., Khudanpur, S., Jelinek, F., Psutka, J. and Byrne, W. (2001). On large vocabulary continuous speech recognition of highly inflectional language—Czech. In *Proc. 7th European Conf. Speech Communication and Technology, Aalborg (Denmark), EUROSPEECH / INTERSPEECH,* pp. 487-489.

Ivanecky, J. and Nabelkova, M. (2002). Phonetic transcription SAMPA and Slovak language (Foneticka transkripcia SAMPA a slovencina), *Jazykovedny casopis*, vol. 53, pp. 81-95 (in Slovak).

Ivanecky, J. (2003): Automatic speech phonetic transcription and segmentation (Automatická transkripcia a segmentácia reči). PhD thesis, Technical university of Kosice, KKUI FEI, (in Slovak).

Jongtaveesataporn, M., Wutiwiwatchai, C., Iwano, K. and Furui, S. (2008). Thai Broadcast News Corpus Construction and Evaluation. In *Proceedings of LREC 2008*. Marrakech, Morocco, pp. 1249- 1254.

Juhár, J., Staš, J. and Hládek, D. (2012). Recent Progress in Development of Language Model for Slovak Large Vocabulary Continuous Speech Recognition. In: *New Technologies – Trends, Innovations and Research*, C. Volosencu (Ed.), InTech Open Access, Rijeka, Croatia, ISBN 978-953-51-0480-3, pp. 261-276.

Lee, A. and Kawahara, T. (2009). Recent Development of Open-Source Speech Recognition Engine Julius. *Proceedings of the Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, APSIPA ASC 2009*, Sapporo, Japan, pp. 131-137.

Lindberg, B. et al. (2000). A Noise Robust Multilingual Reference Recogniser Based on Speechdat (II), *Proceedings of Interspeech 2000*, Beijing, China, October 16-20, 2000, pp. 370-373.

Nechala, M (2014) Corpus of speech recordings in Slovak language (in Slovak) [Diploma thesis] University of Matej Bel in Banska Bystrica Slovakia, Faculty of Natural Sciences. Banská Bystrica 2014 (in press).

Pleva, M. and Juhár, J. (2013). Building of Broadcast News Database for Evaluation of the Automated Subtitling Service. *Communications (Komunikacie)*, vol. 15 (2A), ŽU EDIS, ISSN: 1335-4205, pp. 124-128.

Pollak,P., Černocky, J., Choukri, K., Heuvel, H., Vicsi, K., Virag, A., Siemund, R., Majewski, W., Sadowski, J., Stzaroniewicz, P., Tropf, H., Ostrouchov, J., Rusko, M. and Trnka, M. (2000). SpeechDat (E) - Eastern speech databases. In: *Proceedings of LREC`2000*. Satellite workshop XLDB - Very large Telephone Speech Databases. - Athens, Greece, 2000. pp. 20-25.

Rodríguez-Fuentes, L. J., Penagarikano, M., Varona, A., Diez, M. and Bordel, G. (2012). KALAKA-2: a TV Broadcast Speech Database for the Recognition of Iberian Languages in Clean and Noisy Environments. In: *Proceedings of LREC 2012,* Istanbul, pp. 99-105.

Rusko, M., Trnka, M. and Daržagín, S. (2006). MobilDat-SK - a Mobile Telephone Extension to the SpeechDat-E SK Telephone Speech Database in Slovak. In: *Proceedings of XI International Conference Speech and Computer, SPECOM 2006*, Sankt Peterburg, Russia, ISBN 5-7452-0074-X, pp. 485-488.

Rusko, M., Juhár, J., Trnka, M., Stas, J., Darjaa, S., Hládek, D., Cerňák, M., Papco, M., Sabo, R., Pleva, M., Ritomský, M. and Lojka, M. (2011). Slovak automatic transcription and dictation system for the judicial domain. In: *Proc. of the 5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, pp. 365-369.

Stern, R. M. (1997). Specification of the 1996 Hub 4 broadcast news evaluation. In: *Proceedings of the 1997 DARPA Speech Recognition Workshop*.

Vandecatseye, A. et al. (2004). The COST278 pan-European Broadcast News Database, *Proceedings of LREC 2004*, Vol. 6, May 2004, Lisbon, pp. 873-876.

Zgank, A. et al. (2004): The COST 278 Initiative – Crosslingual Speech Recognition with Large Telephone Database, *Proceedings of LREC 2004*, Lisbon, May 26-28, May 2004, pp. 2107–2110.

Žgank, A., Rotovnik, T., Maučec, M. S., Verdonik, D., Kitak, J., Vlaj, D., Hozjan, V., Kačič, Z. and Horvat, B. (2004b). Acquisition and Annotation of Slovenian Broadcast News Database. In *Proceedings of the 4th International Conference on Language Resources and Evaluation – LREC 2014*. Lisbon, Portugal, May 26-28, pp. 2103 - 2106.

Žibert, J. and Mihelič, F. (2004). Development, evaluation and automatic segmentation of Slovenian broadcast news speech database. *Proceedings of LREC 2004*, Lisbon, May 26-28, pp. 2095-2098.

Zlacký, D., Staš, J. and Čižmár A. (2013). Supervised Text Document Clustering Algorithm with Keywords in Slovak. In: *Proceedings of Redžúr 2013: 7th International Workshop on Multimedia and Signal Processing,* May 1, Smolenice, Slovakia, STU Bratislava, pp. 31-34.