Performance of Czech Speech Recognition with Language Models Created from Public Resources

Vaclav PROCHAZKA[1], *Petr POLLAK*[1], *Jindrich ZDANSKY*[2], *Jan NOUZA*[2]

[1] Czech Technical University in Prague, Prague, Czech Republic
[2] Technical University in Liberec, Liberec, Czech Republic

{vaclav.prochazka, pollak}@fel.cvut.cz, {jindrich.zdansky, jan.nouza}@tul.cz

Abstract. *In this paper, we investigate the usability of publicly available n-gram corpora for the creation of language models (LM) applicable for Czech speech recognition systems. N-gram LMs with various parameters and settings were created from two publicly available sets, Czech Web 1T 5-gram corpus provided by Google and 5-gram corpus obtained from the Czech National Corpus Institute. For comparison, we tested also an LM made of a large private resource of newspaper and broadcast texts collected by a Czech media mining company. The LMs were analyzed and compared from the statistic point of view (mainly via their perplexity rates) and from the performance point of view when employed in large vocabulary continuous speech recognition systems. Our study shows that the Web1T-based LMs, even after intensive cleaning and normalization procedures, cannot compete with those made of smaller but more consistent corpora. The experiments done on large test data also illustrate the impact of Czech as highly inflective language on the perplexity, OOV, and recognition accuracy rates.*

Keywords

speech recognition, LVCSR, n-gram language models, public language resources.

1. Introduction

During the last two decades a lot of research activities have been focused on the development of large vocabulary continuous speech recognition (LVCSR) systems. Nowadays, well performing systems are available for almost all major languages. For many minor ones, including Czech, the research and development has made significant progress - see [1], [2], [3]. A LVCSR system is based on two principal components, a) an acoustic model (AM), which represents the acoustic-phonetic part of speech, and b) a language model (LM), which covers the linguistic level of spoken and written language. The latter is usually represented by so called n-grams, i.e. statistics on sequences of *n* adjacent words. The n-gram model, if properly built, has a considerable impact on the accuracy of the target LVCSR application. Although many procedures, algorithms and tools for n-gram computation have been developed, the process of building a well performing LM requires a lot of human effort spent by collecting a large enough text corpus, and its pre-processing, cleaning, balancing, etc. For highly inflectional languages, like Czech, a proper text corpus, namely its size and structure, is of great importance ([4], [5]), and its collection may take a long time. Therefore, a publicly available resource may be a good solution for a quick and efficient creation of a proper LM, at least for research purposes. Moreover, as there are publicly available toolkits, which provide easily usable instruments for a basic development of a LVCSR system (e.g. HTK [6] or Sphinx), it is reasonable to support the creation of LMs from public resources, too. So, the main motivation of this paper is to investigate and compare publicly available n-gram sources for Czech, discuss their properties, describe the procedures necessary for the creation of practical LMs, and eventually, test them in real LVCSR systems.

This work extends the experience with the publicly available WEB1T corpus distributed via Linguistic Data Consortium (LDC) [7], [8]. This type of corpus was used previously for several natural language processing tasks, like spell-checking and correction [9], word sense disambiguation and lexical substitution [10], information extraction [11], and also for speech recognition related tasks [12]. Several works have focused on more elementary aspects of this WEB corpus, like extending the size of available linguistic features, e.g. part-of-speech tagging and creating of tools for working with them [13], or proposing a method for memory and time efficient access to this huge amount of data [14]. It has also been shown that some well known smoothing methods, e.g. absolute discounting or modified Kneser-Ney, did not work well for corpora with fixed cutoff threshold, and hence, their modifications were proposed [15]. Some of these works also mentioned the risks associated with utilizing WEB-based n-grams, e.g. so called corpus noise, questionable text source or document selection [11], [13].

In this paper, we present a study focused on evaluating the performance of a Czech LVCSR system based on LMs obtained from publicly available resources. These are a) Czech WEB1T 5-gram corpus [7], and b) SYN2006PUB

5-gram corpus [16] obtained from the Czech National Corpus Institute. After employing them for building a series of scaled LMs, we compare their statistics, their perplexity and Out-Of-Vocabulary (OOV) rates, and eventually, we utilize them in prototype LVCSR systems created from the HTK toolkit. In order to see how far one can get with public resources (data and tools) we compare them to a system that has been developed on a professional level. It is a system for on-line broadcast news transcription designed at the Technical University in Liberec and supported by an LM based on very large collection of Czech texts provided by a private media mining company.

Our paper is structured as follows: In section 2, we give a detailed description of the two publicly available resources of Czech n-grams. Section 3 deals with the procedures that must be applied to the corpora before an LM can be computed. In section 4, we present a series of results from experimental evaluation done with large spoken datasets. After that we discuss the results, summarize the conclusions and mention the ongoing work.

2. Resources of Czech N-grams

The best resource for building an LM is a large and well balanced corpus of electronic texts in the given language. If it is not available (for various reasons), a file with n-gram statistics may be a good alternative source. This is the case we investigate in this paper. We have two different resources of n-grams and we want to see how good they are for building an LM for a LVCSR targeted on general spoken Czech. Let us present the two resources.

2.1 Czech Web 1T 5-gram Corpus

The Czech Web 1T 5-gram corpus (WEB1T) [7] is one of the 13 sets assembled from WEB pages by Google and published by the LDC. Short after the first 5-gram collection for English was published, similar data were issued also for 10 European languages, namely Czech, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Spanish and Swedish, and also for Japanese and Chinese. For Czech, the set represents the collection of about 136 billion of tokens that are available in a n-gram format. The most relevant characteristics of Czech WEB1T corpus are summarized in the following points:

- n-gram statistics are available for n = 1 to 5,
- original web sources were cleaned mainly by removing (X)HTML markup language tokens,
- the set contains token <UNK> used for numbers with more then 16 digits, words longer than 32 characters, non-European or invalid UTF8 characters, and rare words with occurrence lower than 40,
- tokens may contain both lower and upper case letters as they appear in WEB texts,
- cross-sentence context is not preserved,
- n-grams with occurrence lower than 40 were removed,
- text is encoded in UTF8, sentence start and end are marked by <S> and </S>.

Although some pre-processing has been done by the provider, the n-grams still contain a lot of invalid tokens such as foreign words, strings with a mix of alpha-numerical characters, or URLs. Also, some other tokens with marginal importance for speech recognition are present in the statistics, such as string differing only in case, or punctuation marks. The initial analysis also revealed the presence of n-grams that were evidently artifacts of original web-pages, like menus, headers or footers. It also contained items differing only in one or two tokens of the same type, e.g. a name or serial number. These n-grams have significantly increased some counts compared to those we would obtain from real web-page contents. As we have already shown in [17], further processing of this data is necessary.

2.2 SYN2006PUB 5-gram Corpus

SYN2006PUB is a synchronic corpus of written journalism created at the Institute of the Czech National Corpus [16]. It contains exclusively newspaper texts from November 1989 to the end of 2004 [18]. Unfortunately, the SYN2006PUB corpus is not available in full-text form for public use. Under a license agreement, one can get n-gram statistics, with the following properties and limitations:

- n-gram statistics are available for n = 1 to 5,
- tokens may contain both lower and upper case letters as they are used in standard Czech orthography,
- cross-sentence context is preserved,
- no word or n-gram were removed from the final set because of low occurrence rate (no cutoff),
- text is encoded in iso-8859-2 and for sentence boundary single token </s> is used.

In contrast to WEB1T 5-grams, SYN2006PUB 5-grams corpus (further called CNC 5-grams) is generally much cleaner, but additional filtering of invalid tokens, such as numeric expressions, foreign words, misspelled words or abbreviations, is still necessary, though on considerably smaller scale.

2.3 Statistics of Available 5-gram Corpora

The n-gram counts for both the sets and all available n-gram orders are shown in Tab. 1. One can see much lower variability in WEB1T 5-grams. Although the number of unique unigrams is nearly 4 times higher for WEB1T corpus, with increasing n, the number of unique n-grams gets higher for the CNC set. For $n = 5$, it is even 3-times higher.

n-gram order	*WEB1T 5-grams*	*CNC 5-grams*
1	9 786 424	2 554 028
2	66 050 702	63 806 655
3	117 264 988	189 152 100
4	118 015 565	271 481 810
5	103 280 138	302 836 997

Tab. 1. Unique n-gram counts in original resource corpora.

3. Creating Language Models

As mentioned above, further cleaning and post-processing steps are necessary for both the n-gram sets, namely filtering and normalization. These were already proposed in [17]. Here we describe some additional improvements.

3.1 Token Filtering

The filtering procedure removes invalid words or word-forms in the n-gram set. The aim is to get more appropriate counts before we start LM computation and smoothing. The following two steps should be done:

- **removal of invalid strings** - tokens which do not form a proper Czech word (e.g. mixed alphanumeric strings or URLs) and all numbers are replaced by token <UNK>, for this purpose the following Czech alphabet letters were considered as legitimate: 'aábcčdďeéěfghiíjklmnňoópqrřsštťuúůvwxyýzž',
- **spell-checking** - only the words verified by the already established lexica or those accepted by Aspell [19] are accepted.

The second step is capable of removing almost all suspicious tokens but a proper spell-checking procedure is much slower than that based on regular expressions implemented in the first step. Therefore, we start with the first step, which is very fast and removes most invalid strings, and then the rest is cleaned by the second approach.

3.2 Normalization Steps

Because of slightly different characteristics of the two n-gram corpora, the following normalization steps were applied to get data with the same format and the same properties:

- **case unification** (both corpora) - all tokens were converted to lowercase form,
- **expansion of sentence boundary** (CNC only) - single sentence boundary token </s> was complemented by its counterpart <s> to match the WEB1T format, the n+1-grams created by this step were split into two n-grams,
- **partial cross-sentence context restoration** (WEB1T only) - new n-grams with cross-sentence context were added by shifting tokens and adding <UNK> token. E.g. from 4-gram 'w1 w2 w3 </s>', new 4-grams 'w2 w3 </s> <UNK>' and 'w3 </s> <UNK> <UNK>' were made. Similarly <UNK> was added from the left to n-grams beginning with <s>,
- **partial cutoff restoration** (WEB1T only) - if the occurrence count of an n-gram was higher than the sum of counts of n+1-grams with the same first *n* tokens, a new n+1-gram with appended token <UNK> was added, e.g. if there was bigram 'w1 w2 5' and there were trigrams 'w1 w2 w3 2' and 'w1 w2 w4 1', new trigram 'w1 w2 <UNK> 2' was added,
- **occurrence count rescaling** (WEB1T only) - n-gram counts were divided by the cutoff value, i.e. 40 in this case,
- **punctuation removal** (WEB1T only) - punctuation marks were discarded, which decreased the order of some n-grams.

The main reason for using the partial cutoff and cross-sentence context restoration is to create as many 5-grams as possible to guarantee that the other operations, e.g. the removal of punctuation marks, can work more efficiently. As the removal of punctuation marks reduces the n-gram order, filtering of WEB1T corpus was performed at the 5-gram level. In contrast, CNC n-gram filtering was performed at the 3-gram level, because this corpus does not contain punctuation marks and the n-gram order is not changed.

The WEB1T n-gram counts have been rescaled to get the smallest occurrence value equal to one. This is necessary for the next step, in which we want to apply n-gram smoothing techniques, like the Kneser-Ney or Witten-Bell ones. The rescaling is done by diving all the counts by the cutoff value, in our case number 40. This is a similar approach to that in [15], which gives a more appropriate probability to all unseen n-grams.

The final counts of unique n-grams after the filtering and normalization procedures are shown in Tab. 2. These counts include also the special tokens <UNK>, <s> or </s> up to the order of 3. If we compare the figures in Tab. 2 and Tab. 1, we can see a significant reduction of counts, namely in case of the WEB1T set. Its number of unigrams was reduced to one tenth, actually.

n-gram order	*WEB1T*	*CNC*
1	957 285	861 899
2	27 629 051	47 273 302
3	72 957 299	147 147 485

Tab. 2. Unique n-grams counts in both n-gram resources after filtering and normalization.

3.3 Created Language Models

Target unigram, bigram, and trigram LMs were created from the cleaned data using two LM smoothing techniques: Witten-Bell and Kneser-Ney discounting. Because

unigram	*bigram counts*			*trigram counts*		
(vocab-size)	*WEB1T*	*CNC*	*TUL*	*WEB1T*	*CNC*	*TUL*
60 000	8 820 828	27 305 979	-	19 404 319	17 976 344	-
120 000	11 004 727	35 231 568	-	22 023 489	19 501 433	-
180 000	11 988 619	39 047 471	-	23 014 945	20 028 748	-
240 000	12 529 836	41 223 572	-	23 500 813	20 268 094	-
340 000	13 005 789	43 210 293	130 362 668	23 885 402	20 443 214	133 363 851

Tab. 3. Unique bigram and trigram counts in analyzed LMs.

both the methods yield very similar results (the former being a little bit worse), in the next text we will mention only the performance of the latter. To investigate the impact of the vocabulary size on the perplexity, OOV, and later also on the recognition score, vocabularies with the following sizes were created: 60 K, 120 K, 180 K, 240 K, and 340 K words. All of them and their corresponding LMs were created using the SRILM toolkit [20]. These LMs were compared with bigram and trigram language models computed for a 340 K word lexicon at the Technical University in Liberec from internal resources provided by a private Czech media mining company [21]. These resources covered 12 GB of full texts of all major Czech newspapers from 1989 to 2010 as well as verbatim transcriptions of many broadcast (TV and radio) programs from the last decade. This lexicon and LM is employed in several practical applications (see e.g. [23]), and in this work it served mainly as a reference denoted as TUL LM.

The counts of unique n-grams for n = 1,2 and 3 and all the three LMs are shown in Tab. 3. It is interesting to see that the WEB1T models contain about three times less bigrams than the CNC ones, but about 10 % more trigrams. The trigram components of the LMs were created with cutoff 2, which is the reason why the CNC trigram counts are lower compared to the bigram ones. This is not true for the WEB1T LMs, as its n-grams were created from significantly larger source corpus. Therefore, after the rescaling step mentioned in Section 3.2, there was a large number of trigrams that appeared at least twice. These facts must be taken into account when the three LMs are compared from the absolute numbers point of view. Their basic parameters are summarized in Tab. 4.

parameter	*WEB1T & CNC*	*TUL*
orders	1,2,3	2,3
vocab-sizes [K]	60,120,180, 240,340	340
vocab-selection	most freq. words	most freq. words
<UNK> token	skipped	skipped
<s> token	kept	kept
</s> token	removed	removed
discounting	Kneser-Ney	Kneser-Ney
cutoff 1,2-gram	1	1
cutoff 3-gram	2	2

Tab. 4. Parameters of all compared LMs.

4. Evaluation and Experimental Part

The created LMs were evaluated in terms of perplexity, OOV and recognition accuracy in LVCSR systems. The evaluation was done with two large databases of spoken Czech.

4.1 Description of Test Data

Each of the test sets contained audio recordings and their transcriptions. The latter were manually checked and normalized if necessary (e.g. all numbers and some abbreviations were converted into full text forms). The transcriptions were used mainly for the LM evaluation on the linguistic level (perplexity and OOV rates), while the audio recordings served as the input for the LVCSR systems.

The first source of test data was Czech SPEECON database [24]. It had been created as a part of a large international project whose goal was to provide speech community with spoken data in many languages. Its structure and content was designed so that the audio recordings would serve primarily for the training of acoustic models suitable for speech recognition in various tasks and under diverse conditions. For our experiments we have chosen that part of Czech SPEECON DB which contains recordings of fluent sentences read by several tens of speakers. The read sentences are slightly specific as they were tailored to be phonetically rich and balanced. This means that sometimes they may not represent common spoken utterances. On the other side, the database offers a lot of data suitable both for LVCSR training as well as for independent testing. For the evaluation of the linguistic issues we have taken all 5183 sentences available in the SPEECON database and denoted this set as SPEECON_SENT. Its smaller subset, denoted as SPEECON1, was chosen for time intensive speech recognition experiments performed with HTK tools. (The remaining part of recordings was used for training the acoustic model of HTK-based LVCSR system, see Section 4.3.1.) A larger subset, containing 1000 sentences, was utilized in speech recognition experiments performed with the TUL system, whose acoustic model was trained on TUL's own databases.

The second source of evaluation data was a test set prepared at TUL. The TUL_TRANS set covers 271 minutes of utterances taken from broadcast news and talk shows recorded in 2010. Obviously, this data was kept separate from those used in AM and LM training. The main parameters of all the test sets are summarized in Tabs. 5 and 6.

audio data			
test corpus	*words*	*utter.*	*t[min]*
SPEECON1	4 975	577	57
SPEECON2	8 694	1 000	88
TUL	36 097	436	271

Tab. 5. Word and utterance counts in speech test corpora.

text data		
test corpus	*words*	*sent.*
SPEECON_SENT	41 045	5 183
TUL_TRANS	32 185	405

Tab. 6. Word and sentence counts in text test corpora.

4.2 OOV Rate and Perplexity

The results from the linguistic analysis are presented in Tabs. 7 and 8. We can see how the vocabulary size influences the OOV rates and n-gram LM perplexities for both the text sets. It is evident that the CNC LMs demonstrate significantly lower perplexities when compared to the WEB1T ones, even if the size of the original WEB1T corpus was considerable larger. Also we may notice that the SPEECON_SENT set shows higher OOV and perplexity values than the TUL_TRANS one, which can be explained by the fact that the sentences of the former set used some less frequent (but phonetically rich) words.

LM	*OOV [%]*	*unigram*	*bigram*	*trigram*
WEB1T 60	7.24	5 449	1 768	954
WEB1T 120	3.89	6 867	2 155	1 137
WEB1T 180	2.66	7 591	2 340	1 236
WEB1T 240	2.02	8 039	2 493	1 314
WEB1T 340	1.60	8 368	2 615	1 375
CNC 60	5.18	3 482	809	630
CNC 120	3.02	4 141	918	714
CNC 180	2.20	4 456	964	749
CNC 240	1.76	4 640	1 000	779
CNC 340	1.48	4 781	1 026	800
TUL 340	1.05	15 007	1 917	1 428

Tab. 7. Perplexities of analyzed LMs for TUL_TRANS corpus.

LM	*OOV [%]*	*unigram*	*bigram*	*trigram*
WEB1T 60	15.32	6 517	3 033	1 775
WEB1T 120	9.01	10 304	4 753	2 640
WEB1T 180	5.97	13 158	6 053	3 288
WEB1T 240	4.17	15 413	7 157	3 842
WEB1T 340	2.63	17 735	8 302	4 428
CNC 60	12.46	4 357	1 528	1 228
CNC 120	7.15	6 316	1 986	1 560
CNC 180	4.56	7 641	2 309	1 804
CNC 240	3.24	8 460	2 514	1 962
CNC 340	1.98	13 850	2 768	2 161
TUL 340	3.02	21 477	4 232	3 294

Tab. 8. Perplexities of analyzed LMs for SPEECON_SENT corpus.

4.3 LVCSR Performance

From the practical point of view, the most relevant results came from speech recognition experiments. These were performed with two LVCSR systems: a prototype modular system based on HTK tools [6], and the broadcast news transcription system developed at TUL [5]. For evaluation, we used the standard word accuracy measure defined as

$$ACC = \frac{N - S - D - I}{N} \cdot 100\ [\%] \qquad (1)$$

where N is total number of words in testing subset, S, D and I are numbers of substituted, deleted, and inserted words.

4.3.1 Experiments with HTK tools

Unigram, bigram, and trigram language models were tested using a recognizer based on the HTK Toolkit with the following setup: signal parameterization based on 39-MFCC feature vector (12 cepstral coefficient + log energy + 1st and 2nd derivatives). The acoustic model was speaker independent, made of tied-state cross-word triphones with 32 mixtures per state, in total 106 432 Gaussians. It was trained on 52 hours of speech from Czech SPEECON database (office subset). The decoding was performed by HDecode, which enabled us to employ large lexicons up to 340 K words and LMs up to trigrams. The decoder was set up to use word insertion penalty equal to -10, LM weight factor equal to 10, and the pruning beam width was set to 200 [6], [22]. Most experiments took a significant portion of time as the real-time factor was between 5 (for the 60 K-word lexicon and unigrams) up to 20 (340 K + trigram) on a Linux machine with Intel Core i3 550 @ 3.20 GHz and 3 GB RAM.

The results from the experiments are summarized in Tab. 9. As expected there is strong correlation between the accuracy and the perplexity values presented in Tab. 8. Again, we can see that the LMs computed from the smaller CNC corpus significantly outperform those based on WEB1T. In absolute measure, the CNC LMs were about 4-5 % better than the latter ones. The relative improvement in accuracy values between trigram and bigram LMs is 2-3 % and seems to be slightly higher for the WEB1T LMs. The main conclusion from this series of experiments is that both the publicly available n-gram resources are suitable at least for basic research work in speech recognition of Czech.

LM	*unigram*	*bigram*	*trigram*
WEB1T 60	39.40	50.83	53.57
WEB1T 120	46.19	58.39	60.72
WEB1T 180	48.44	61.25	64.08
WEB1T 240	49.59	63.30	66.11
WEB1T 340	50.47	64.42	67.48
CNC 60	44.28	56.18	58.15
CNC 120	50.55	63.84	65.79
CNC 180	52.80	66.73	68.82
CNC 240	53.97	67.82	70.03
CNC 340	55.20	69.19	71.32

Tab. 9. Recognition accuracy of HTK-based LVCSR for SPEECON1 corpus.

4.3.2 Tests on TUL transcription system

The final series of experiments was performed with the LVCSR system developed at TUL. Its engine was designed to support primarily on-line speech recognition tasks, such as voice dictation and on-line subtitling of broadcast programs. Hence, it is not as flexible as the HTK tools and its current version accepts only bigram LMs, not the trigram ones. On the other side, it is fast and it can run in real time even for the largest 340 K lexicon. For the experiments, we used the following setup: 39 MFCC features, floating cepstral mean subtraction (with 1 s window), speaker independent acoustic model based on tied-state triphones with 3400 physical states and 32 Gaussians per state. The AM was trained on 120 hours of speech (a mix of read speech and broadcast speech from major Czech TV and radio stations). The Viterbi decoder is optimized for speed and memory usage. The 340 K lexicon together with the corresponding LM (with 130 M unique bigrams) occupy only 190 MB in memory. More details about the LVCSR system can be found in [5].

On this system we could compare the performance of the LMs based on the two publicly available n-gram sets with that made from private resources. The full-text corpus provided by the media mining company is probably the largest corpus of Czech electronic texts and includes also a large amount of transcriptions of spoken communication. This feature makes it suitable especially for real speech recognition tasks as it can be observed from the experiments with the TUL_TRANS test set in Tab. 10. The difference in the performance between the TUL 340 LM and the WEB1T 340 LM is more than 10 %. The results achieved with both the TUL_TRANS and SPEECON2 test sets prove again that the CNC-based models are significantly better than the WEB1T-based ones.

LM	*TUL_TRANS*	*SPEECON2*
WEB1T 340	72.45	65.94
CNC 340	79.63	76.63
TUL 340	82.60	77.93

Tab. 10. Recognition accuracy of real-time LVCSR with bigram LMs tested on TUL_TRANS and SPEECON2 corpora.

5. Conclusions

The goal of this work was to investigate the potential of publicly available linguistic resources for speech recognition in Czech language. We show that if one cannot get free access to a representative and large enough collection of full texts in Czech, there is still a possibility to utilize n-gram statistics provided either by Google in its WEB1T corpus or by the Institute of the Czech National Corpus in its SYN2006PUB corpus. Both allow for creating a language model that makes large-vocabulary speech recognition of Czech possible, at least at a basic level. This can be a good starting point for further improvements.

We analyzed and compared the two corpora from linguistic and speech recognition points of view. Both the perplexity figures as well as the results from the recognition experiments demonstrated that the CNC corpus is more appropriate for statistic language modeling than the 10 times larger WEB1T set. It just proves the fact that the recently very popular large-scale web harvesting has its limits. Yet, these automatically gathered statistics can find an appropriate usage, e.g. for complementing existing language models, especially when properly weighted and mixed. This is one of the topics we want to investigate in future.

We have also presented several techniques used to clean and normalize n-gram corpora, which are necessary for language model preparation. Without these operations, the resulting models would be less efficient. This has been proven also during the TUL language model building, which outperformed the other models not only because of the large source data but also due to very intensive pre-processing and cleaning phases [5].

Acknowledgements

The activities presented within this work were supported by grants GA ČR 102/08/0707 “Speech Recognition under Real-World Conditions”, TA ČR TA01011204 “Living archives”, and by the research activity MSM 6840770014 “Perspective Informative and Communications Technicalities Research”.

References

[1] NOUZA, J., ZDANSKY, J., DAVID, P. Fully automated approach to broadcast news transcription in Czech language. *Text, Speech and Dialogue: Lecture Notes in Computer Science*, 2004, vol. 3206/2004, p. 401 - 408.

[2] VANEK, J., PSUTKA, J. Gender-dependent acoustic models fusion developed for automatic subtitling of parliament meetings broadcasted by the Czech TV. *Text, Speech and Dialog: Lecture Notes in Computer Science*, 2010, vol. 6231/2010, p. 431- 438.

[3] PSUTKA, J., PSUTKA, J., IRCING, P., HOIDEKR, J. Recognition of spontaneously pronounced TV ice-hockey commentary. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokyo (Japan), 2003, p. 83 - 86.

[4] IRCING, P., KRBEC, P., HAJIC, J., PSUTKA, J., KHUDANPUR, S., JELINEK, F., BYRNE, W. On large vocabulary continuous speech recognition of highly inflectional language - Czech. In *Proceedings of INTERSPEECH*. Aalborg (Denmark), 2001, p. 487 - 490.

[5] NOUZA, J., ZDANSKY, J., CERVA, P., KOLORENC, J. A System for information retrieval from large records of Czech spoken data. Text, Speech and Dialogue: Lecture Notes in Computer Science, 2006, vol. 4188/2006, p. 485 - 492.

[6] *The Hidden Markov Model Toolkit (HTK)*. Version 3.4.1. 2009. [Online] Available at: `http://htk.eng.cam.ac.uk`.

[7] BRANTS, T., FRANZ, A. *Web 1T 5-gram, 10 European languages, version 1*. Philadelphia: Linguistic Data Consortium, 2009. [Online] Available at: http://www.ldc.upenn.edu.

[8] BRANTS, T., FRANZ, A. *Web 1T 5-gram, English, version 1*. Philadelphia: Linguistic Data Consortium, 2006. [Online] Available at: http://www.ldc.upenn.edu.

[9] ISLAM, A., INKPEN, D. Real-word spelling correction using Google Web 1T n-gram with backoff. In *International Conference on Natural Language Processing and Knowledge Engineering NLP-KE 2009*. Dalian (China), 2009, p. 1- 8.

[10] NULTY, P., COSTELLO, F. Using lexical patterns in the Google Web 1T corpus to deduce semantic relations between nouns. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions SEW 2009*. Boulder (USA), 2009, p. 58 - 63.

[11] TANDON, N., DE MELO, G. Information extraction from web-scale n-gram data. In *Web N-gram Workshop : Workshop of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Geneva (Switzerland), 2010, p. 8 - 15.

[12] ZWARTS, S., JOHNSON, M. The impact of language models and loss functions on repair disfluency detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*. Portland (USA), 2011, p. 703 - 711.

[13] DEKANG LIN, et al. New tools for Web-scale n-grams. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation LREC'10*. Valetta (Malta), 2010, p. 19 - 21.

[14] GUTHRIE, D., HEPPLE, M. Storing the web in memory: space efficient language models with constant time retrieval. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing EMNLP 10*. Cambridge (USA), 2010, p. 262 - 272.

[15] YURET, D. Smoothing a tera-word language model. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Singapore, 2008, p. 141 - 144.

[16] *Czech National Corpus - SYN2006PUB*. 2006. [Online] Available at: http://ucnk.ff.cuni.cz/english/syn2006pub.php.

[17] PROCHAZKA, V., POLLAK, P. Analysis of Czech Web 1T 5-gram corpus and its comparison with czech national corpus data. *Text, Speech and Dialogue: Lecture Notes in Computer Science*, 2010, vol. 6231/2010, p. 181 - 188.

[18] *Institute of the Czech National Corpus*. [Online] 2010. Available at: http://www.korpus.cz.

[19] *GNU Aspell*. [Online] Cited 2010-11-17. Available at: http://aspell.net/.

[20] STOLCKE, A. SRILM – An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing ICSLP 2002*. Denver (USA), 2002, p. 901 - 904.

[21] *Newton Media*. [Online] 2010. Available at: http://www.newtonmedia.cz.

[22] NOVAK, J. R., DIXON, P. R., FURURI, S. An empirical comparison of the T3, Juicer, HDecode and Sphinx3 decoders. In *Proceedings of 11th Annual Conference of the International Speech Communication Association INTERSPEECH 2010*. Makuhari (Japan), 2010, p. 1890 - 1893.

[23] NOUZA, J., ZDANSKY, J., CERVA, P., SILOVSKY, J. Challenges in speech processing of Slavic languages (case studies in speech recognition of Czech and Slovak). *Lecture Notes in Computer Science*, 2010, vol. 5967/2010, p. 225 - 241.

[24] POLLAK, P., CERNOCKY, J. *Czech SPEECON Adult Database (technical report)*. 2004.

About Authors...

Vaclav PROCHAZKA was born in 1982. He received the Master degree in Computer Science from CTU Prague in 2008. Currently, he is a Ph.D. student on Czech Technical University, Faculty of Electrical Engineering. His research focuses on creation and adaptation of language models from Internet resources for use in automated speech recognition systems.

Petr POLLAK was born in 1966 in Usti nad Orlici, Czechoslovakia. After the graduation (Ing. 1989) he joined the Czech Technical University in Prague where he has also received his further degrees (CSc. 1994, Doc. 2003). He works as teacher and researcher in the Speech Processing Group at the Faculty of Electrical Engineering. His most important activities are in robust speech recognition, speech enhancement, speech database collection, and other related activities. He was the responsible person for several EC projects aiming at speech database collection realized in co-operation with leading European industrial partners (Speech-Dat, SPEECON, LC-StarII, and others). He is responsible person for the grant GACR 102/08/007 "Speech Recognition under Real-World Conditions" and he leads the "Signal Processing Team" in Research activity MSM 6840770014 "Perspective Informative and Communications Technicalities Research".

Jindrich ZDANSKY was born in 1978. He received the Master degree in Electronics and Electronic Systems from CTU Prague and the Ph.D. degree in Technical Cybernetics from TU Liberec, in 2002 and 2005, respectively. He is currently an assistant professor at the Institute of Information Technology and Electronics TUL. His research interests are speaker-change detection and speech recognition.

Jan NOUZA was born in 1957. He received his M.Sc. and Ph.D. degrees at the Czech Technical University (Faculty of Electrical Engineering) in Prague in 1981 and 1986, respectively. Since 1987 he has been teaching and doing research at the Technical University in Liberec. In 1999 he became full professor. His research focuses mainly on speech recognition and voice technology applications (voice-to-text conversion, dictation, broadcast speech processing and design of voice-operated tools for handicapped persons). He is the head of SpeechLab group at the Institute of Information Technology and Electronics.