Issues in annotation of the Czech spontaneous speech corpus in the MALACH project

Josef Psutka*, Pavel Ircing*, Jan Hajič†, Vlasta Radová*
Josef V. Psutka*, William J. Byrne‡, Samuel Gustman♯

* Department of Cybernetics and Center for Computational Linguistics, University of West Bohemia, Plzeň, Czech Rep.
{psutka, ircing, radova, psutka_j}@kky.zcu.cz
† UFAL and Center for Computational Linguistics, Charles University, Praha, Czech Rep.
hajic@ufal.mff.cuni.cz
‡ Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
byrne@jhu.edu
♯ Survivors of the Shoah Visual History Foundation, Los Angeles, CA, USA
sam@vhf.org

Abstract

The paper present the issues encountered in processing spontaneous Czech speech in the MALACH project. Specific problems connected with a frequent occurrence of colloquial words in spontaneous Czech are analyzed; a partial solution is proposed and experimentally evaluated.

1. Introduction

The goal of MALACH (Multilingual Access to Large Spoken Archives) (`www.clsp.jhu.edu/research/malach`) is to use automatic speech recognition (ASR) and information retrieval (IR) techniques to provide improved access to the large multilingual spoken archives created by the Visual History Foundation (`www.vhf.org`). These archives contain approximately 52,000 interviews ("testimonies") in 32 languages of personal memories of survivors of the World War II Holocaust (116,000 hours of video).

Our research group concentrates on the ASR component for Slavic languages within this project. Although we have over five years of experience building large vocabulary continuous speech recognition systems for the Czech language, the Czech part of the MALACH project is the first task concerning unconstrained spontaneous speech. There are also other properties of the VHF corpus that are extremely challenging from the ASR point of view - the speakers are usually elderly, their speech is often heavily accented and, due to the nature of the stories they relate, often highly emotional. Consequently, we discovered several issues during the transcription of the Czech part of the archive that complicate both the transcription process and the consequent speech recognition and information retrieval procedures.

2. Goal of the paper

The aim of the paper is to present specific issues encountered in processing spontaneous Czech speech, with special emphasis on problems which can be solved or at least alleviated by appropriate handling of the speech data transcriptions. The main focus is put on the problem of discrepancy between standard (literary) form of the Czech language used in formal conversations, written documents, and - most importantly from our point of view - also the search engine queries, and colloquial (common) form used in spontaneous speech. We will describe patterns of colloquial word usage in spontaneous Czech speech and explain the impact on the ASR procedure. Then we will propose a partial solution of this problem and demonstrate its effect on the ASR performance.

3. Transcription procedure

Testimonies were delivered for further processing divided into 30-minute segments stored as the video file in MPEG-1 format. The interviewer and the interviewee have been recorded via lapel microphones on separate channels. Audio was extracted at 128 kb/sec in 16-bit stereo and 44kHz sampling rate.

In order to create the training data, we decided to transcribe a 15-minute segment from each of the training testimonies, 30 minutes into each testimony (i.e. at the beginning of the second segment), thus getting past the biographical questions and initial awkwardness.

The selected segments were divided (roughly) into sentences and transcribed using the speech annotation software Transcriber (Barras et al., 2000). Detailed description of the transcription format is given elsewhere (Psutka et al., 2004). Let us only mention here that in addition to lexical transcription, the transcribers also marked several non-speech events.

One of the most important issues which had to be decided before the transcription process started is the way of transcription of colloquial words. As is explained in detail in Section 5.1., colloquial Czech has its well-defined (orthographic) written form. The transcribers were instructed to use this orthographic transcription of colloquial words in order to keep transcriptions close to what was actually said. There were several reasons for this decision. Firstly, transcribing colloquial sentences using standard words is not an easy task, especially for transcribers without a solid linguistic background. Secondly, our automatic phonetic transcription algorithm, which is based on phonetic transcription rules, also requires orthographic transcription of

pronounced words and it would produce an incorrect phonetic baseforms for artificially "standardized" word forms. Thus it would be necessary to append information about the actual word form when using standard forms for transcribing colloquial speech. Both the above mentioned factors would cause a substantial slowing down of the transcription process if we chose to "standardize" colloquial word forms during the transcription. Using the orthographic transcription, the transcribers worked at a rate of fifteen times real time. Transcription inspection and verification requires additional effort at approximately twice real time.

4. Characteristics of the transcribed data

4.1. Speech data

The above mentioned 15-minute segments from 336 speakers were transcribed for training purposes and the entire testimonies given by another 10 speakers were transcribed to create a test set. The ratio between males and females in terms of the number of speakers and the amount of transcribed speech is shown in Table 1.

	Training (336)		Test (10)	
	Male	Female	Male	Female
Speakers	145	191	5	5
Hours transcribed	36.25	47.75	13.15	9.7

Table 1: Transcribed speech data

4.2. Text data

The acoustic training set contains 43,702 different words and 565,517 tokens (running words). The distribution of words is considerably different in this corpus that in broadcast news or newspaper articles (Psutka et al., 2002). The test set contains 19,465 different words and 156,315 tokens. The OOV rate with a 43,702-word lexicon is 5.8%, which is an unacceptably high number. There are two basic ways of lowering the OOV rate. The first one involves transcribing more VHF data; this is not feasible since a rough extrapolation of the OOV curve indicated that we would need several hundreds of transcribed testimonies to reduce the OOV rate bellow 1% (Psutka et al., 2003), and that much Czech data is not even available in the VHF collection. The second option is to find external text data and use it to enhance both the lexicon and the text training corpus. The problem of finding the appropriate text data is also dealt with in (Psutka et al., 2003).

4.3. Features complicating transcription and ASR

The speech quality is often quite poor from the ASR point of view. There is frequent whispered or emotional speech along with many disfluences and non-speech events as crying, laughter, etc. Transcribers observed that the quality and fluency of speech was often affected by the age of speakers. The age of the oldest survivor was 94; the average age of all speakers was 75 years. The speaking rate was also quite variable, ranging from 64 to 173 words per minute, with an average rate of 113.

Other problems are related to several problematic word classes listed in Table 2 together with their frequencies by words (vocabulary types) and tokens.

Personal names	Place names	Foreign words	Colloq. words	Word fragments
5.0 / 0.7	4.7 / 1.6	4.2 / 0.5	8.9 / 6.8	4.3 / 1.1

Table 2: Problematic word classes - percentage by words/tokens.

The class of *personal names* contain first names and last names, including dialectical variants of the first names. *Geographical (place) names* cover the names of countries, cities, rivers and other places, as well as names of languages and nationalities. *Foreign words* class contains mostly Slovak and German words; English, Russian, Hebrew and Yiddish words are also quite frequent. Some of the foreign words appeared in isolation, but there were also stretches of continuous segments pronounced in a foreign language.

These three word classes cause problems in both the transcription process and the language modeling part of the ASR system. The transcription is difficult because many names and places are of foreign origin and it is not easy to determine their spelling; the transcribers often have to consult addition knowledge sources (dictionaries, WWW search engines) in order to spell the words correctly. The same naturally holds true for foreign words in general. The problems in language modeling are due to the fact that those word classes are underrepresented in a typical Czech text corpus used for estimation of language model parameters.

The class of *colloquial words* is by far the most frequent of the problematic word classes. It does not cause so much problems during the transcription (once we have decided to use the orthographic transcription of colloquial words, see Section 3.) but the language model is affected in a similar way - colloquial words also only rarely occur in the written text (see Section 5.1.). The problem of colloquial word usage in spontaneous Czech is the main focus of the rest of the paper.

Finally, the class of *word fragments* covers words whose transcription is not complete, either due to recording errors or speaker disfluences. Word fragments require special handling when building a language model since they induce discontinuities in the word stream. Currently, we simply exclude the word fragments from the language model training data, however, the development of a more sophisticated technique is highly desirable.

5. Colloquial words in spontaneous speech

5.1. Linguistic background

It can be successfully argued that a phenomenon called *diglossia* is present in today's Czech: *colloquial* Czech differs substantially from *standard* Czech (as defined by orthographic, morphological, lexical and syntactic rules by the Czech language normative bodies). Whereas standard Czech is used in most of Czech written materials [1] as well

[1] Standard Czech is taught in all schools as the only acceptable variant.

in official public speeches, such as TV news, in schools etc., colloquial[2] Czech can be heard in the homes and on the streets of most Czech cities and villages.[3]

The difference between just pronunciation variants (as found in English and many other languages) and the Czech case is that Czech spelling rules are phonetically based. Therefore, the colloquial Czech words have well-defined, but different spellings than their standard variants. In other words, colloquial Czech has its orthographic written form (and as this paper shows below, it is to our advantage to respect and use this fact).

Phonetically, the length of vowels (a distinctive phenomenon in Czech) can be shortened or prolonged, depending on the particular word. It can happen in the root, ending or prefix; sometimes two such changes occur in one word (note that length is denoted by the accent ´, applicable to all vowels in Czech):

Coll. form	Std. form	*English gloss*
neni	není	*(he) is not*
jí	ji	*(I see) her (Acc.)*
novym	novým	*(by the) new (Instr.)*
vyjímka	výjimka	*exception*

However, the difference between colloquial Czech and standard Czech displays itself the most in morphology. Endings and prefixes are often changed in colloquial Czech:

Coll. form	Std. form	*English gloss*
novej	nový	*new (Nom. Sg. Masc.)*
nový	nové	*new (Acc. Pl. Masc.)*
vejtah	výtah	*elevator*
pracujem	pracujeme	*(we) work*
autama	auty	*(by the) cars*

Please note also that a form used in colloquial Czech in some context (e.g., a particular case, gender and number) may be equal to some standard form in a different context (cf. above the form *nový*). This makes automatic mapping from standard forms to colloquial and vice versa in general difficult if not impossible (cf. also below).

Lexical changes (i.e., changes to the root that apply to every form of the affected word) are quite common, too. They typically apply to the change of -ý- to -ej-, prefixing by v- (of words beginning in the standard Czech by o-), and the loss of initial j-:

Coll. form	Std. form	*English gloss*
bejt	být	*to be*
vokno	okno	*window*
sem	jsem	*(I) am*

Some of the colloquial forms are again, unfortunately, the same as some other standard forms (e.g. *sem* is *to here* in std. Czech), adding to the difficulty of mapping between colloquial and standard forms.

[2]In fact, the proper term for the colloquial Czech is *General Czech*, to emphasize that it is generally used.

[3]Except some areas in Moravia, the eastern part of the Czech Republic, including Brno, the largest Moravian city, where the standard prevails even in those settings.

Many borderline cases also behave similarly: for example, what was once non-standard spelling might become standard variant and vice versa (-z- and -s- in suffixes of words of foreign origin, -o- and -ó- in suffixes -on, etc.).

Differences in syntax are much less common; for example, in colloquial Czech the relative pronoun *which* (std. Czech: *který*) is often replaced by *what* (coll. Czech: *co*). However, in most spoken utterances, standard syntax is followed.

5.2. Impact on ASR

The automatic phonetic transcription and consequently also acoustic models clearly benefit from the orthographic transcription of colloquial words because it allows to capture the actual phonetic realization of a word quite precisely.

On the other hand, the influence of the abundance of colloquial words on the language model is rather negative. First of all, the orthographic transcription of colloquial words causes an unnecessary growth of the lexicon since a word is defined by its spelling in ASR and colloquial words not only differ in spelling from the corresponding standard form, but there are often several different colloquial variants of one standard word form. Consequently, the already sparse language model training data became even sparser.

Moreover, although the transcriptions of the testimonies proved to be by far the most suitable data for language model training in the MALACH project, some benefit can be gained through the combination of the language model trained on the transcriptions and the language model trained on the appropriately selected external data (Psutka et al., 2003). However, these additional data comprise mostly written text. Thus the frequency of colloquial words is very low and their statistics are not likely to be strengthened.

5.3. Proposed solution

In order to exploit both the advantage of close orthographic transcription of colloquial words in acoustic modeling and the benefit of standard word forms in language modeling, we decided to "standardize" the pronunciation lexicon. We went through the lexicon built from the original (orthographic) transcriptions and added a corresponding standard form to each colloquial word form, creating a new 3-column lexicon.

In order to illustrate that the number of colloquial forms for a single standard word form can be really high, we present a fragment from the "standardized" lexicon (columns contain standard form of the word, form observed in the transcriptions and phonetic baseform of the observed form, respectively):

```
ODJET  ODJET     o d j e t
ODJET  ODEJET    o d e j e t
ODJET  ODJEC     o d j e c
ODJET  ODJECT    o d j e c t
ODJET  VODJET    v o d j e t
ODJET  VODEJET   v o d e j e t
ODJET  VODJECT   v o d j e c t
ODJET  VODEJECT  v o d e j e c t
```

A new "standardized" text corpus was also generated by automatically replacing colloquial words in the origi-

LM training & decoding	AM Training	
	Colloq.	Standard
Colloq.	57.01%	56.46%
Standard	**58.85%**	58.06%

Table 3: Recognition accuracy using colloquial and standard forms in the testimony transcriptions

nal transcriptions with their standard counterparts using the above mentioned 3-column lexicon. Note that such procedure does not take into account word context and therefore the standardization process is far from perfect (see Section 5.1. for reasons). Nonetheless, even with this method we achieved some improvement in the recognition accuracy, as is presented in the next section.

5.4. Experimental results

We performed several ASR experiments to support the statements about the role of standard/colloquial transcriptions in acoustic model (AM) and language model (LM) training and decoding, which we made in the previous sections.

The baseline system is a conventional HTK cross-word triphone mixture Gaussian system trained on 84 hours of speech, with approx. 6K states and 97K Gaussians (Psutka et al., 2003) along with a bigram language model estimated from the transcriptions.

We investigated all possible combinations of standard/colloquial transcription usage in AM/LM training (see Table 3). The column "AM training - Colloq." denotes acoustic training using the original transcriptions (produced by transcribers) and the second and third column of the 3-column lexicon, whereas the column "AM training - Standard" denotes training with the "standardized" transcriptions and the first and third column of the "standardized" lexicon (note that in such case the colloquial forms are regarded as pronunciation variants of the standard forms; the proper variants are chosen via forced alignment during the training). Similarly, the "LM training & decoding - Colloq." row describes the situation when the second and third column of the lexicon are used in the decoding process and original transcriptions are used for the language model estimation and the "LM training & decoding - Standard" row denotes the usage of the first and third column of the lexicon (together with appropriate weights representing the relative frequency of the given colloquial form) in the decoding and the usage of the "standardized" transcriptions for the language model building.

As can be seen from Table 3, the best performance is obtained by using the colloquial forms during acoustic model training while restricting the language model to formal forms both in the lexicon and in the LM estimation process. Such outcome is fully compatible with our previously stated expectations.

Having achieved a decent improvement using a "standardized" lexicon, we decided to re-create the experiment from (Psutka et al., 2003) and interpolate our best (i.e. standardized) language model with the language model built on the data selected from the Czech National Corpus (CNC).

λ	AM Training	
	Colloq.	Standard
0.00	55.79%	55.17%
0.25	59.84%	58.87%
0.50	**61.43%**	59.80%
0.75	61.38%	**60.61%**
1.00	59.06%	58.11%

Table 4: Recognition accuracy for interpolated language model

Results for both acoustic model training scenarios and various interpolation coefficients λ are summarized in Table 4. The improvement of 2.58% ("AM training - Colloq.") is approximately 0.6% absolute bigger than the improvement achieved when interpolating the language model trained on the CNC selection with the language model estimated using the original colloquial transcripts (see (Psutka et al., 2003)). Thus our "standardized" lexicon also better matches written Czech text resources.

6. Conclusion

We have presented a method for handling colloquial words in the Czech testimonies that allows to switch arbitrarily between the standard and colloquial output of the decoder. This method not only proved to be beneficial for the ASR performance, but it will also have important consequences in the upcoming IR stage of the MALACH project. The reason is that although colloquial words are very frequent in Czech spontaneous speech, they will rarely or never be typed into a search engine, the ultimate goal of our project. Therefore the decoder ability to deliver sentences in the standard form may be very useful.

7. Acknowledgements

Support for this work was provided by NSF (U.S.A.) under the Information Technology Research (ITR) program, NSF IIS Award No. 0122466 and by the Ministry of Education of the Czech Republic, projects No. MSM235200004 and No. LN00A063.

8. References

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 2000. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33.

J. Psutka, P. Ircing, J. V. Psutka, V. Radová, W. J. Byrne, J. Hajič, S. Gustman, and B. Ramabhadran. 2002. Automatic transcription of Czech language oral history in the MALACH project: Resources and initial experiments. In *Proceeding of TSD 2002*.

J. Psutka, P. Ircing, J. V. Psutka, V. Radová, W. J. Byrne, J. Hajič, J. Mírovský, and S. Gustman. 2003. Large vocabulary ASR for spontaneous Czech in the MALACH project. In *Proceeding of Eurospeech 2003*.

J. Psutka, J. Hajič, and W. J. Byrne. 2004. The development of ASR for Slavic languages in the MALACH project. In *Proceeding of ICASSP 2004*.