THE DEVELOPMENT OF ASR FOR SLAVIC LANGUAGES IN THE MALACH PROJECT

Josef Psutka, Jan Hajič‡, William Byrne†

Dept. Cybernetics and Center for Computational Linguistics, Univ. of West Bohemia, Czech Rep.
UFAL and Center for Computational Linguistics, Charles Univ., Czech Rep. ‡
Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD USA †
psutka@kky.zcu.cz, hajic@ufal.mff.cuni.cz, byrne@jhu.edu

ABSTRACT

The development of acoustic training material for Slavic languages within the MALACH project is described. Initial experience with the variety of speakers and the difficulties encountered in transcribing Czech, Slovak, and Russian language oral history are described along with ASR recognition results intended to investigate the effectiveness of different transcription conventions that address language specific phenomena within the task domain.

1. INTRODUCTION

The goal of MALACH (Multilingual Access to Large Spoken Archives) (`www.clsp.jhu.edu/research/malach`) is to use automatic speech recognition and information retrieval 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). All aspects of ASR are challenging within this corpus. The speakers are usually elderly, their speech is often heavily accented and, due to the nature of the stories they relate, often highly emotional. The problem of developing ASR for this domain is challenging technically, partly because resources are not available for the languages within the domain of the collection. This paper focuses on experience gained in creating resources needed for acoustic modeling for the Slavic language testimonies in the collections [1, 2].

2. SPEECH TRANSCRIPTION CONVENTIONS

The audio files were divided into segments and annotated using the speech annotation software Transcriber [3]. Audio files are divided into segments roughly corresponding to sentences; we attempt to maintain 'linguistic segmentation'. The beginning of a segment is marked by `<b ti>`, with `ti` as the time in seconds. Speaker turns are marked by `<t ti> <<sp#, n, g>>`; by convention, `sp1` is the interviewer, and `sp2, sp3, ...` are other speakers, and `n` is the name and surname of the speaker (if known), and `g` is either `m` or `f` for the speaker gender. Incidents in which speakers speak simultaneously, are marked as:`<t ti> <<spk_1, n_1, g_1 + spk_2, n_2, g_2>>`, and `<unintelligible>` is used whenever any speaker cannot be understood. Everything spoken is transcribed with words; no numerals or punctuation is used. Sentences begin with lower case letters; only proper names and acronyms such as IBM and NATO are capitalized. If a word is spelled out, letters are capitalized and separated by white space.

If a speaker stammers, for example saying "thir thirty", the corresponding transcription is `thir- thirty`. The "-" indicates word fragments, which may be due to recording errors as well as to disfluencies. In such cases the "-" must be preceded or followed by white space, depending on whether only the end or the beginning of the word was spoken. If the "-" is neither preceded nor followed by any blank space it indicates a hyphenated word. In this corpus switching between languages within a testimony is very common; speech in languages other than the dominant or nominal language of the testimonies is enclosed in square brackers, i.e. `[]`. If transcribers are unsure about a portion of the transcription, it is set off by parentheses. Non-speech sounds are transcribed as `<click>, <cough>, <laugh>, <breath>, <inhale>, and <mouth>`. Background noise is marked according to the following rules: if no word overlaps with the background noise the mark `<noise>` is used; if a word or a part of an utterance overlaps with the noise, the mark `<noise_begin>` is used before the first affected word and the mark `<noise_end>` is used after the last affected word. Filled pauses are marked as `<UH>, <UM>, <UH-HUH>`, or `<UH-HUM>`. Distinct pauses and gaps in speech are marked with `<silence>`.

An example of the annotated file pronounced in Czech is shown in the following paragraph:

```
<t 26.800> <<spk2, f>>
  <mouth><inhale> to vám neřeknu data já si
  absolutně nepamatuju
<t 31.747> <<spk1, f + spk2, f>>
  SPEAKER1: aspoň roční období
  SPEAKER2: <mouth><inhale>
<t 33.372> <<spk2, f>>
  roční tož to mohlo být v třiaštyrc-
  dvaaštyrycet už třiaštyrycátém roce
<b 40.838>
  <noise_begin> protože to byl čas vždycky
  ten odstup <inhale><noise_end>
```

3. CZECH, RUSSIAN, AND SLOVAK PHONETICS

The basic **Czech phonetic alphabet** consists of 10 vowels, 29 consonants and 3 dipthongs. Vowels are either short ([i] (myš), [e] (les), [a] (pas), [o] (rok) and [u] (kus)) or long ([ii] (pít), [ee] (lék),

This work has been funded by the 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 MSM235200004 and LN00A063.

[aa] (rád), [oo] (móda), and [uu] (půl)). There is one “domestic” dipthong ([ow] (pouto)) and two dipthongs that appear in foreign words ([aw] (auto) and [ew] (euro)). Consonants may be divided into 8 plosives ([p] (pivo), [b] (bota), [t] (tón), [d] (dům), [tj] (tito), [dj] (děda), [k] (koš), and [g] (gól)), 4 affricates ([c] (cíl), [dz] (leckdo), [ch] (čas), and [dzh] (léčba), 11 fricatives ([f] (fotka), [v] (víno), [s] (sen), [z] (zub), [rsh] (tři), [rzh] (řád), [sh] (šaty), [zh] (žena), [j] (jaro), [x] (chléb) and [h] (had)), 2 liquids ([r] (růže) and [l] (led)), and 4 nasals ([m] (maso), [n] (nos), [ng] (banka), and [nj] (nic)). 11 consonants come in pairs, having the same manner and place of articulation and differing just by the unvoiced/voiced characteristic ([p]/[b], [t]/[d], [tj]/[dj], [k]/[g], [c]/[dz], [ch]/[dzh], [f]/[v], [s]/[z], [rsh]/[rzh], [sh]/[zh], and [x]/[h]). The other consonants (i.e. liquids, nasals, and glide [j], i.e. the sonorants) are always voiced. Additionally, the basic alphabet can be extended by allophones, as [mg] (tramvaj).

The **Russian phonetic alphabet** consists of six *vowels* ([a] (мама), [e] (бчера), [i] (путин), [o] (никто), [u] (каникулы), [y] (теплый)) and 36 *consonants*. Most consonants come in pairs, ”hard“ (non-palatalized) and ”soft“ (palatalised). Among plosives we put the non-palatalized [b] (обычной), [d] (туда), [g] (когда), [k] (собака), [p] (опыт), and [t] (защита) and their palatalized counterparts [B] (набирать), [D] (неделя), [G] (лагерь), [K] (банкир), [P] (опять), and [T] (интерес). Also *fricatives* ([f] (телефон)/[F] (кофе), [x] (вход)/[X] (тихий), [s] (русский)/[S] (спасибо), [w] (страшний)/[W] (будущий), [v] (свобода)/[V] (связь), and [z] (везде)/[Z] (озеро), *sonants* ([l] (около)/[L] (далеко), [m] (дома)/[M] (кроме), [n] (машина)/[N] (не), and [r] (первой)/[R] (пример)), and *affricates* ([c] (больница)/[C] (вечер)) appear in non-palatalized / palatalized pairs. There are 2 non-pair consonants: *fricative* [J] (также) and *sonant* [j] (русский). Native Russians living in Russia don’t use the phoneme “h” and usually replace it in foreign words, personal and geographical names by the phoneme “g” (in our Russian alphabet by [g] or [G]) (for instance in the name Harry) or “ch” (in our alphabet by [x] or [X]) (for instance in the surname Hussain). The native Russians living in the Ukraine territory as well as in Israel or USA learnt to pronounce “h” and they use this phoneme frequently in words in which it is currently used in local languages (for example in geographical names, personal names etc).

The **Slovak phonetic alphabet** comprises 11 vowels, 37 consonants, and 4 dipthongs. There are 6 short ([i] (pivo), [e] (meno), [a] (kapitola), [o] (noha), [u] (bubon), and [{] (mäso)) and 5 long vowels ([ii] (vít’az), [ee] (gén), [aa] (pohár), [oo] (katalóg), and [uu] (múr)), as well as dipthongs ([iˆa] (piatok), [iˆe] (mier), [iˆu] (paniu), and [uˆo] (kôň)). Of the consonants, there are 8 plosives ([p] (popol), [b] (žaba), [t] (vata), [d] (voda), [tj] (Mat’o), [dj] (hád’a), [k] (páka), and [g] (guma)), 4 affricates ([c] (cena), [dz] (medza), [ch] (oči), and [dzh] (džungla)), 8 fricatives ([f] (figa), [w] (vdova), [s] (osa), [z] (zima), [sh] (šek), [zh] (veža), [x] (chata), and [h] (hra)), and 16 sonorants ([r] (para), [r=] (vrch), [r:] (vŕba), [l] (skala), [l=] (vlk), [l:] (vĺča), [L] (l’ad), [m] (mama), [M] (amfiteáter), [n] (rana), [ng], (banka), [nj] (vaňa), [v] (slovo), [uˆ] (kov), [iˆ] (kraj), [j] (jama)). There are 10 unvoiced/voiced consonantal pairs in Slovak ([p]/[b], [t]/[d], [tj]/[dj], [k]/[g], [c]/[dz], [ch]/[dzh], [f]/[w], [s]/[z], [sh]/[zh], and [x]/[h]). Other consonants (i.e. sonorants) are always voiced.

For all three languages we developed rule-based phonetic transduction, which are used to automatically transform the majority of the words in the transcriptions to their phonetic forms. These rules have been described in detail in previous work (see citations). Not surprisingly, this task is quite complex. For example, the phonetic transcription of the Czech word “oběd” (Engl. “lunch”) has two variants: [o b j e t] and [o b j e d], and it also happens that [o b j e t] and [o b j e d] are two phonetic variants of the standard word “oběd”. Owing to cross-word (voice) assimilation this word can be pronounced in either way depending on the following word in the utterance and the manner of speech. If “oběd” is followed by a pause or if the next word starts with an unvoiced consonant or with some of 10 Czech vowels or with [m], [n], [nj], [l], [r], [j] then the pronunciation will be [o b j e t] (for example: “oběd měl” [o b j e t m nj e l] (Engl.“lunch had”). The variant [o b j e d] will otherwise appear in the remaining cases. The phonetic pronunciations are also rich in alternatives: for example the Czech word “poněvadž” (Engl. “since”) has the following 4 correct phonetic variants: [p o nj e v a ch], [p o nj e v a d zh], [p o nj e v a t sh], and [p o nj e v a dzh].

3.1. Exceptions to Pronunciation Rules

There are many words which must be treated as exceptions to the pronunciation rules and those words are transcribed and corrected manually. For example, the pronunciation of “automatizace” (Engl. “automation”) is an exception to standard Czech and also Slovak phonetic rules. Using the standard rules this word would be phonetically transcribed as [aw t o m a tj i z a c e] (Czech) and [a uˆ t o m a tj i z a c iˆa] (Slovak). But these would be incorrect due to the word’s foreign origin. The correct pronunciations are therefore manually added to each dictionary : [aw t o m a t i z a c e] (Czech) and [a uˆ t o m a t i z a c iˆa] (Slovak). Notably, however, in the case of Russian, the pronunciation rules do not produce exceptions for “автоматика” (Engl. “automation”).

Generally speaking, most exceptions to the rules of phonetic transcription in Czech and Slovak are connected with words containing sequences of characters: -ti-, -di-, and -ni- which, are pronounced in words of Czech origin as [tj i], [dj i], and [nj i] and as [t i], [d i], and [n i] in words of foreign origin. The majority of exceptions to the Russian phonetic rules can be found among words containing the character -o-. If the position of this character in the word is before the stress then “o” is actually read as “a”. Example: “Москва” (Engl. “Moscow”) [m a s k v a] (because the stress is on the “a” “Москвá”), “море” (Engl. “sea”) [m o R e] (because the stress is on the “o” “мóре”). The difficulty in applying rules automatically is partly due to the difficulty in determining stress within words. We rely on the native Russians involved in the transcription effort to verify automatically produced pronunciation and to correct them as needed.

4. ANNOTATION OF SPONTANEOUS SPEECH

All manual annotations were in the orthographic form of the words. This means that the eventual colloquial words were neither transformed to standard (formal, non-colloquial) forms nor written phonetically.

Colloquial Czech: Czech colloquial words are usually not considered to be phonetic variants of standard Czech words in that they can be properly written in their colloquial orthographic form. But these orthographically written colloquial words do not frequently appear in formal or semi-formal text (novels, newspapers, letters, e-mail, etc.). Colloquial speech (in Czech) is currently used in non-professional but also in professional life (most university lectures may partly use colloquial words). But in TV Broadcasts, in newspapers, and in the official speech of the Czech officials, standard words and their pronunciations are used. For example, the standard Czech word “oběd” (Engl. “lunch”) has pronunciations [o b j e t] and [o b j e d]. If we wish to write this word phonetically, then we obtain “objet”, but this form is used neither in standard nor in colloquial Czech. But, there does exist the standard Czech word “objet” (Engl. “to go round”). Similarly, the word “oběd” has also a colloquial variant, “voběd” with the two pronunciations [v o b j e t] and [v o b j e d].

Spontaneous Regional Russian: There are some problems with regional variants of pronunciation of many words. The main differences appear in different pronunciation of one or more characters in the word in comparison with standard Russian. Example: The Russian word “когда” has the standard pronunciation according to the phonetic transcription [k a g d a] but many times this word was pronounced as [k o g d a]. The native Russian transcribers assessed these words not as colloquial words and/or only accented speech but rather as a speech of Russians whose pronunciation of many words is partly modified by a non-Russian environment (Ukraine, Israel, etc.) where may have lived for a long time. These instances marked by placing the incorrectly pronounced portion of words between asterisks, as in “к*о*гда”, and the region in question was transcribed phonetically.

Spontaneous Slovak: In this collection, we observed a relatively small number of words pronounced in dialectal or colloquial form. Generally speaking, we conclude from this collection that Slovaks rarely use colloquial speech. There are only several colloquial variants of standard Slovak words. Slovak colloquial words can appear in words which contain character “ä” or “ô”. In colloquial words, “ä” can be realized as “e” (in urban areas or towns) or as an “a” (spoken typically by older people in rural areas); “ô” can be pronounced as “o” or “ó”. For example, the Slovak word “devät”’ has a colloquial variant “devet”’, and another standard pronunciation “môj” has the variant “moj”.

5. ANALYSIS OF THE TRANSCRIPTIONS

It is interesting to note that the most frequent words in all three processed Slavic languages are very similar. (see correspondences: “sem” v. “я” v. “som” and “sme” v. “мы” v. “sme”), which correspond to “I did” (in English); “dělal sem” (in Czech, “sem” is a colloquial variant of the standard word “jsem”); “я работал” (in Russian); “robil som” (in Slovak, “som” is a standard - grammatically correct Slovak word); “we did” (in English); “dělali sme” (in Czech, “sme” is a colloquial variant of the standard word “jsme”); “мы работали” (in Russian); “robili sme” (in Slovak, “sme” is a standard - grammaticaly correct Slovak word).

Personal Names contain first names and last names, including dialectal variants of first names. This class contains roughly an equal number of first and last names, however, it is to be expected that the number of the last names will grow far more rapidly than the number of first names as the size of the corpus increases. It can be estimated from the numbers given in Table 2, that nearly each hundredth running word is a personal name.

Geographical (Place) Names cover the names of countries, cities, rivers and other places, as well as names of languages and nationalities.

Foreign Words are mostly German and Slovak words in Czech testimonies, German and Czech words in Slovak testimonies, both also English, Russian, Hebrew, Yiddish words. Some of the foreign words appeared in isolation, but there were also continuous segments in German and Czech/Slovak, for example. In Russian testimonies we can found words of German, English, Ukraine, Hebrew and Yiddish origin.

	Word	*English*	Relative Freq.
1	a	*and*	0.041
2	to	*it, that*	0.034
3	se	aux. word	0.023
4	sem*	*am*	0.019
5	že	*that*	0.019
6	v	*in*	0.016
7	sme*	*are*	0.016
8	tak	*so*	0.016
9	tam	*there*	0.016
10	na	*on*	0.013
1	и	*and*	0.033
2	в	*in*	0.022
3	не	*not*	0.018
4	я	*I*	0.016
5	это	*it, that*	0.012
6	мы	*we*	0.012
7	что	*what, that*	0.011
8	а	*but*	0.011
9	на	*on*	0.011
10	там	*there*	0.008
1	a	*and*	0.033
2	to	*it, that*	0.022
3	sa	aux. word	0.018
4	sme	*are*	0.016
5	som	*am*	0.012
6	že	*that*	0.012
7	v	*in*	0.011
8	tal	*so*	0.011
9	na	*on*	0.011
10	tam	*there*	0.008

Table 1. Ten Most Frequent Words in Czech, Russian and Slovak (from top to bottom). The asterisk denotes a colloquial word form.

Disfluencies In manual transcripts and then also in the formed lexicon there are many stammered words. The number of such words in individual testimony usually depends on how comfortable the survivors are in being filmed.

Example: .... `<inhale><UH> ta tehdejší <UH> an- manda- vládou <UH>`
Typically 1% of the words in running speech are disfluent (in each language) and these are usually word-initial fragments. These are not regular; most fragments occur no more than once.

Problem Words - Russian Generally speaking, our initial experience with a processing of Russian testimonies suggests that besides the variable word order of Russian language (Russian together with Czech and some other Central and East European languages belong to a family of Slavic languages which share this phenomenon) the dominant challenges are due to accented speech. Unlike English, where the accent is due to the native tongues of the speakers, much of the Russian accents are due to regional differences in spoken Russian (regional variants of pronunciation). We

Personal Names	Place Names	Foreign Words	Problem Words	Disfluency
5.0 / 0.7	4.7 / 1.6	4.2 / 0.5	8.9 / 6.8	4.3 / 1.1
3.5 / 0.7	5.5 / 1.8	0.6 / 0.3	20.3 /5.3	2.2 / 0.8
4.5 / 0.8	5.0 / 1.9	2.0 / 0.4	2.3 / 0.4	3.5 / 1.0

Table 2. Frequencies of Word Classes by Word Type and Token (type/ token - percent) in Czech, Russian and Slovak.

found out that this accent is usually caused by the territory where the survivors are now living and where they were interviewed.

Studying the demographic information provided by the VHF, we found out that from about 7 thousand of Russian testimonies stored in the VHFs digital archives nearly one half (3500) were provided in Ukraine, about 1500 in Israel, 900 in U.S.A., and only one tenth (700) in Russia. The native Russians living outside Russia often adopted local non-Russian words and used them in their personal vocabulary. As observed earlier, this may account for the unusual phonology observed in speech otherwise expected to be 'Russian'.

In Table you can see very high number of problematic words: 20.3% items in vocabulary and 5.3% of running words in testimonies. Analyzing these words we found that the most of them are regional variants of standard words differing from standard forms by different pronunciation of one or more characters in the word. This pronunciation is often influenced by the incorrect placing of a stress. The number of typical colloquial variants of standard words in Russian testimonies is very low in comparison with Czech testimonies. Examples: Russian word “сегодня” has a standard phonetic transcription [S e v o d N a] but the phonetic transcription of a regional variant of this word is [S e g o d N a]. Standard Russian word “сейчас” with the phonetic transcription [S e j C a s] has its colloquial variant “щас” with the phonetic transcription [W a s].

Problematic Words - Slovak: The number of problematic words in Slovak testimonies is very low. The troublesome words that were observed appear to have been created by merging Slovak and Czech words, so that some of these are new formulations (native Slovak transcribers could not explain them).

6. COLLOQUIAL FORMS IN ACOUSTIC AND LANGUAGE MODEL TRAINING

We performed several Czech ASR experiments to determine the role of standard (formal, non-colloquial)/colloquial transcriptions in acoustic and language training and decoding.

The objective was to determine what benefit, if any, may be had by using the standard forms in acoustic and language model training. We went through the pronunciation lexicon built from the original (orthographic, therefore generally colloquial) transcriptions and added a corresponding standard form to each colloquial word form.

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 [1] along with a bigram language model estimated from the transcriptions.

We investigated several training and decoding scenarios where the colloquial training set transcriptions (produced by annotators) were replaced by their standard (formal, non-colloquial) forms; these transcriptions were used both in acoustic and language model training. In order to be able to use the 'standard' transcriptions in acoustic model training and decoding, the colloquial versions were allowed as pronunciation variants of the standard forms, and the variants were chosen via forced alignment.

Lexicon & Language Model	Acoustic Training Transcriptions	
	colloq.	standard
colloq.	57.01%	56.46%
standard	58.85%	58.06%

Table 3. Recognition Accuracy with Colloquial and Standard Forms in Acoustic and Language Model Training Transcriptions

These transcriptions were used in acoustic and language model training with the results shown in Table 3. We find that 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. Even though the differences seem to be relatively minor, the “standardization” of the data has important consequences in the upcoming IR stage of the project - the colloquial forms of the words will rarely or never be typed into a search engine and thus a standard output of the decoder is highly desirable.

7. CONCLUSION

We have presented an overview of the transcription procedures for Slavic languages in the MALACH project along with a summary and analysis of the phenomena encountered in these languages within the domain of spontaneous oral history.

8. REFERENCES

[1] J. Psutka, P. Ircing, J. V. Psutka, V. Radová, W. Byrne, J. Hajič, J. Mírovský, and S. Gustman, “Large Vocabulary ASR for Spontaneous Czech in the MALACH Project,” in *EUROSPEECH*, 2003.

[2] J. Psutka, I. Iljuchin, J. Psutka, V. Trejbal, W. Byrne, J. Hajič, and S. Gustman, “ Building LVCSR System for Spontaneously Pronounced Russian testimonies in the MALACH project: Initial Steps and First Results,” in *Proceedings of the Text, Speech, and Dialog Workshop*, 2003.

[3] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, “Transcriber: development and use of a tool for assisting speech corpora production,” in *Speech Communication*, January 2000, vol. 33.