Orthographic and Phonetic Annotation of Very Large Czech Corpora with Quality Assessment

Petr Pollák*, **Jan Černocký** †

*Czech Technical University in Prague, Faculty of Electrical Engineering
CVUT FEL K13131, Technická 2, 16627 Praha 6, Czech Republic
pollak@feld.cvut.cz

†Brno University of Technology, Faculty of Information Technology
Božetěchova 2, 612 66 Brno, Czech Republic
cernocky@fit.vutbr.cz

Abstract

The annotation is generally indivisible part of speech database. In this paper we are presenting common orthographic and phonetic annotation of large Czech databases. Phonetic annotation may be very important and gives more information than pronunciation lexicon with possible pronunciation variants. Moreover, for Czech language phonetic annotation means just small additional effort to standard ortographic transcription. The tool FTP-Trascriber developed for this purposes is also presented. In the second part we are presenting procedure of quality assessment applied to the annotation of large speech corpora collected at our laboratories. We are presenting semi-automated quality checks based on using several fully automated pre-checks decreasing necessarry additional manual effort.

1. Introduction

The traditional annotation of very large speech databases is usually based on orthographic transcription of spoken utterances, with some additional markers denoting special events as mispronunciations, word truncation, various types of non-speech events, etc. Phonetic transcription of each utterance is then generated using pronunciation dictionary.

The standard system of orthographic transcriptions with pronunciation dictionary seems to be problematic for the Czech. A standardized pronunciation lexicon is not easily available for our language. Moreover, Czech is written almost phonetically with a strong grapheme-to-phoneme correspondence, so the usage of rule-based conversion is advantageous. The tool *transc* performing this conversion was presented in (Pollák and Hanžl, 2002). Of course, there are many exceptions from regular pronunciation, typically for neologisms and foreign words. A special problem is the pronunciation of foreign proper names; their Czech pronunciation is rather random. Some exceptions are known and they may be incorporated into conversion rules as exceptions as it was also presented in (Pollák and Hanžl, 2002).

Nevertheless, many irregular pronunciations are not available in exception lexicon and also many unusual words are pronounced quite randomly without losing sense of the word. Concerning the generation of the pronunciation lexicon independently of utterance by expert in phonetics, we may find the situation that correctly generated pronunciation is not used uniformly or that all finally used pronunciations are not predictable. On the other hand, the marking such different pronunciation as mispronunciation does not seem to be good solution due to frequency of its appearance. That is the reason why we prefer on-line pronunciation check of each utterance during the annotation of collected speech data.

2. Annotation conventions for Czech

Concerning the reasons described above, it seems to be very convenient to annotate Czech speech databases both, orthographically and phonetically. Creation of phonetic annotation does not imply great additional effort. Having the rules for orthographic-to-phonetic transcription conversion we can obtain on-line prediction of phonetic transcription during the annotation. This is done by tool *transc*.

2.1. Tool '*transc*' and transcription syntax

This tool is using large list of hand-crafted context grammar rules which are applied in sequence, gradually converting orthography to pronunciation. Various assimilations take place during this conversion, most notable being interactions of voiced and voiceless phones. Many common sequences in words of foreign origin are also handled in this "regular pronunciation" stage. Other exceptions from regular pronunciation rules must be either included in external exception lexicon or marked by simple parenthesis convention, i.e. "(orthography/pronunciation)" in the input text. Additional special marks for non-speech events are possible according to database specific requirements.

Generated pronunciation is in simple proprietary phonetic alphabet (Pollák and Hanžl, 2002) so that the predicted pronunciation can be checked by medium trained person. The output can be also generated directly in SAMPA - the Czech part was after several years of discussions finally approved and placed at official SAMPA WEB-page (Wells, http://www.phon.ucl.ac.uk/home/sampa/home.htm).

More details about *transc* were presented at last LREC conference (Pollák and Hanžl, 2002).

2.2. Annotation post-processing

Having above described orthographic transcription with marked real pronunciation (if it is different from basic pronunciation rules), we have the input source with much more information and this transcription can be used for generation of different outputs according to different requirements. Bellow, the most important task are described.

- *pure orthographic transcription*

  Because the correct written form of the word always appears in the transcription, the pure orthographic transcription can be easily generated.

- *phonetic transcription*

  Similarly, the exact phonetic transcription of the utterance is available. It brings the advantage overcoming the necessity of choice between several pronunciation variants, especially, when these variants may differ just in several phonemes (which may be close). It allows for more precise training, mainly in the very beginning of the training procedure, when average initial models should be more precised for different phonemes.

- *context dependent phonetic transcription*

  The possibility of inter-word context dependent conversion between orthographic and phonetic transcription is one of the basic characteristics of tool *transc*. Such transcription may be also immediately generated and it should be quite precise, because irregular changes are already marked in input transcription and changes due to inter-word context are based on standard Czech pronunciation rules.

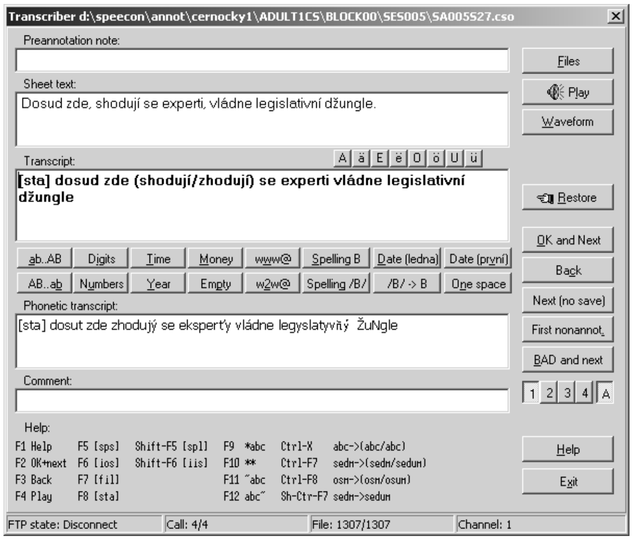
- *pronunciation lexicon*

  Many application work with pronunciation lexicon and it is also standard part of each very large corpora. It is clear that such lexicon can be also easily generated, including numbers of occurrences of pronunciation variants. Lexicon may be easily generated with context independent or context dependent entries.

3. FTP-Transcriber tool

For the purposes of above described annotation, the tool 'FTP-Transcriber' was developed (Boudy et al., 1999) and it was successfully used in the annotation of large Czech databases as "CISLOVKY" (Czech database of digits, ELRA catalog number S0077), Czech SpeechDat (ELRA catalog number S0094), and for Czech SPEECON (currently under validation procedure). It works under Windows 95/98/2000/NT/XP and it has an modular structure which allows easy configuration for different type of data formats, using different buttons and hot-keys, etc. The data can reside on a server and be accessed using the FTP protocol, but a stand-alone mode is also available.

The main difference from similar software, e.g. well known WWWTranscribe, see (Draxler, http://speechdat.phonetik.uni-muenchen.de/speechdat/WWWTranscribe.html), (Draxler, 2000) or others within SpeechDat projects (SpeechDat, http://www.speechdat.org), is that the annotator edits the field which is the input to above described tool for conversion between orthographic and phonetic transcription. During the annotation, the annotator has access only to the field 'Transcript'. Tool *transc*, built in *FTP-Transcriber*, converts this field to phonetic form each time a new character is entered. The conversion is very fast, so that the process is seamless to the user. The annotator checks it, and if he finds an incoherence, he uses the parenthesis convention to mark the correct pronunciation, see Fig 1.



Transcriber d:\speecon\annot\cernocky1\ADULT1CS\BLOCK00\SES005\SA005S27.cso

Preannotation note:

Sheet text:
Dosud zde, shodují se experti, vládne legislativní džungle.

Transcript:
[sta] dosud zde (shodují/zhodují) se experti vládne legislativní džungle

Phonetic transcript:
[sta] dosut zde zhodujý se eksperťy vládne legyslatyvňý ŽuNgle

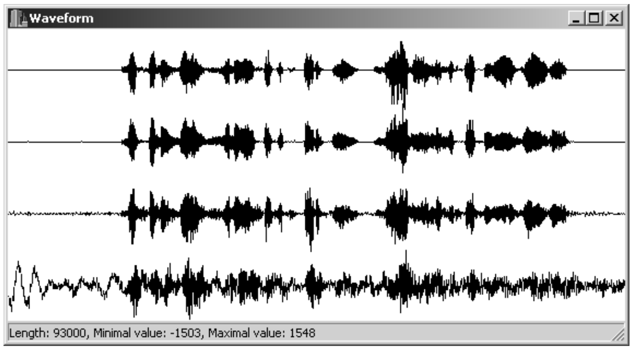
Comment:

Help:

```
F1 Help      F5 [sps]   Shift-F5 [spl]   F9  *abc   Ctrl-X      abc->(abc/abc)
F2 OK+next   F6 [ios]   Shift-F6 [iis]   F10 **     Ctrl-F7     sedm->(sedm/sedum)
F3 Back      F7 [fil]                    F11 ˜abc   Ctrl-F8     osm->(osm/osum)
F4 Play      F8 [sta]                    F12 abc˜   Sh-Ctr-F7   sedm->sedum
```

FTP state: Disconnect | Call: 4/4 | File: 1307/1307 | Channel: 1

Figure 1: Main window of FTP-Transcriber



Length: 93000, Minimal value: -1503, Maximal value: 1548

Figure 2: Additional window of FTP-Transcriber with signal waveforms

The annotator can also confront listen utterance with the signal waveform which can be displayed in separate window. For the annotation of databases with multi-channel signals, see Fig 2, it is possible to show all channels or the particular one according to the requirements.

Annotators providing such annotation must be trained, however, their training is not very difficult. Moreover, when CTU internal phonetic alphabet is used, the phonetic annotation is quite easy because for each phoneme has single character representation and whole phonetic transcription is very close to standard Czech orthography.

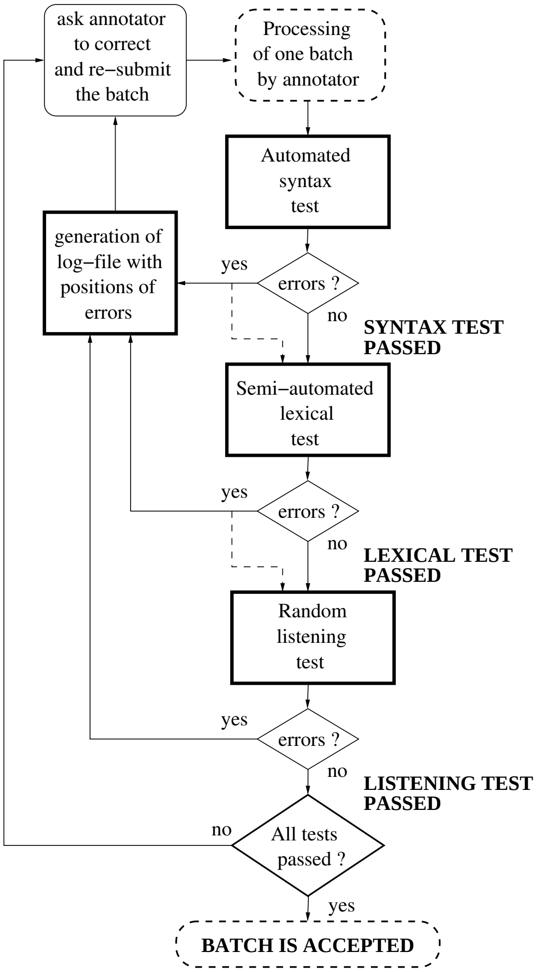


Figure 3: General flow graph of quality assessment

4. Annotation quality assessment

In the second part of this contribution, we would like to present our annotation processing procedure which is designed for achieving of maximal quality of annotations. We are providing several automated, semi-automated, and manual checks in the following steps, described also by flow graph on Fig 3.

I. Each annotator is working on given block of the data. This block should have a reasonable size to do a compromise between efficiency of provided checks and feedback to the annotator.

II. The first step in quality assessment is based on *syntax test*. The most evident errors should be found here, typically usage of allowed characters, correct usage of special marks, detection of missing files, empty annotation fields, etc.

III. As the second quality check, the *semi-automated lexical test* is provided. This test is described by flow chart on Fig‘4 and it contains the following principal sub-steps:

1) Mini-lexicon is generated from finished annotations.

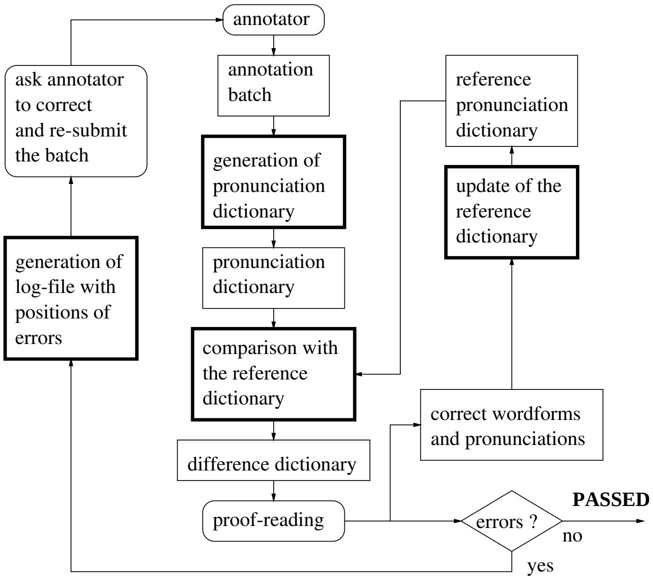

annotator
ask annotator to correct and re-submit the batch
annotation batch
generation of pronunciation dictionary
pronunciation dictionary
reference pronunciation dictionary
update of the reference dictionary
generation of log-file with positions of errors
comparison with the reference dictionary
difference dictionary
correct wordforms and pronunciations
proof-reading
errors ?
PASSED
no
yes

Figure 4: Flow graph of semi-automated lexical test

2) Generated mini-lexicon is compared with the reference lexicon with already checked and approved entries.

3) Experts manually check unknown entries, and probable typographic errors and pronunciation transcription errors are marked. Known entries are supposed to be correct.

4) Listening of utterances with marked strange entries is provided and block of the data with commented errors is returned to the annotator for correction.

5) Correct new entries are added to reference lexicon.

Checks described above are repeated till the annotation is completely accepted. Approved entries are added to the reference lexicon, so that only a small fraction of entries need to be reviewed and listened in the next iteration.

IV. Finally, *random listening test* must be done. Several utterances are selected from defined categories, and correctness of the transcription is checked at all.

V. The annotation package is *accepted* only if all three above described test are successfully passed.

This annotation procedure was successfully used for annotation of above-mentioned databases and we hope that reached quality of the annotations is very high.

This assumption was confirmed by independent validation provided by SPEX (Nijmeghen, Netherlands). A Czech native speaker has performed the check of two large databases transcription, i.e. Czech SpeechDat - ELRA catalog number S0094 (Černocký et al., 2000) and Czech SPEECON (not available yet). Errors were found in the following percentage from checked items:

- speech transcription errors:
3.4% - long utterances in Czech SpeechDat,
2.1% - short utterances in Czech SpeechDat,
1.8% - long utterances in adult SPEECON DB,
0.5% - short utterances in adult SPEECON DB,
(allowed limit was 5%),
- non-speech transcription errors:
0.9% - long utterances in Czech SpeechDat,
1.1% - short utterances in Czech SpeechDat,
1.1% - long utterances in Czech adult SPEECON DB,
1.2% - short utterances in Czech adult SPEECON DB,
(allowed limit was 20%).

The last two figures show, how the numbers of words which had to be read decreased during the processing of SPEECON annotations (Fig 5), commonly with generally decreasing errors made by one group of involved annotators (Fig 6).

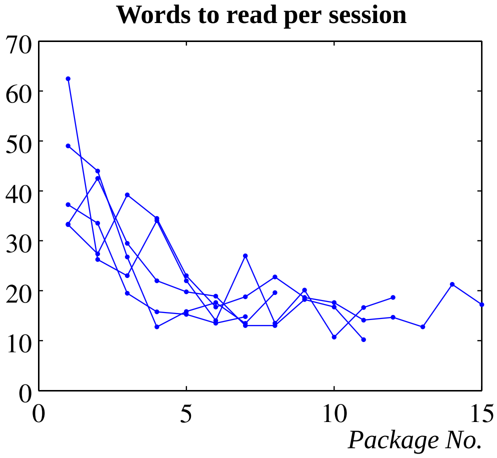


Figure 5: Numbers of words to read per session during SPEECON annotation

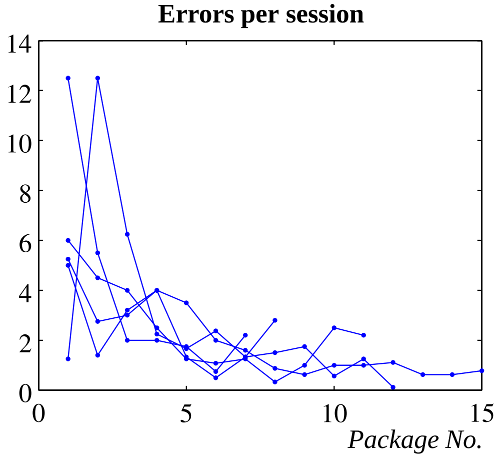


Figure 6: Numbers of errors per session during SPEECON annotation

5. Conclusions

In this paper we are presenting common orthographic and phonetic annotation which seems to be very convenient for the annotation of large Czech speech databases. The most important results of this work could be summarized in following points:

- Principal rules for common orthographic and phonetic transcription of speech utterances were defined.
- The tool FTP-Transcriber were created for the purposes of such common transcription.
- The procedure for annotation quality assessment was tested during the annotation of several large databases.

6. Acknowledgments

This research activity was supported at CTU by grant “Voice Technologies for Support of Information Society”, No. GACR-102/02/0124 (2002-2004), and internal CTU research project “Research in Information Technologies and Communications”, No. 34/99143/331.

Jan Černocký was supported by post-doctoral grant of Grant Agency of Czech Republic, No. GA102/02/D108.

The implementations of FTP-Transcriber were supported partly by EC projects SpeechDat(E), No. INCO-Copernicus 977017, and SPEECON, No. IST-1999-10003.

7. References

Boudy, J., J. Kochanina, M. Rusko, J. Černocký, P. Pollák, P. Staroniewicz, and A. Virag, 1999. Specification and adoption of annotation tools. Technical report, SpeechDat(E). Deliverable ED3.1, workpackage WP3.

Draxler, Ch., 2000. WWWTranscribe Tutorial. In *LREC 2000 - Second Internaltional Conference on Language Resources and Evaluation, XLDB - Very Large Telephone Speech Databases*. Athens.

Draxler, Ch., http://speechdat.phonetik.uni-muenchen.de/speechdat/WWWTranscribe.html. Transcription via the www.

Pollák, P. and V. Hanžl, 2002. Tool for Czech pronunciation generation combining fixed rules with pronunciation lexicon and lexicon management tool. In *Proc. of LREC’02, Third International Conference on Language Resources and Evaluation*. Las Palmas, Canary Islands - Spain.

SpeechDat, http://www.speechdat.org. Pages of all SpeechDat projects.

Černocký, J., P. Pollák, and V. Hanžl, 2000. Czech recordings and annotations on CD’s - Documentation on the Czech database and database access. Technical report, SpeechDat(E). Deliverable ED2.3.2, workpackage WP2.

Wells, J. C., http://www.phon.ucl.ac.uk/home/sampa/home.htm. Sampa home page.