Problems and Solutions in the Creation of Czech and Slovak Lexica for Speech Technology Applications: General Experiences and LC-Star2 Lexica

P. Pollák *V. Hanžl*
Czech Technical University in Prague
Faculty of Electrical Engineering
Dept. of Circuit Theory
Praha, Technická 2, Czech Republic
pollak@fel.cvut.cz
hanzl@fel.cvut.cz

J. Černocký *P. Smrž*
Brno University of Technology
Faculty of Information Technology
Dept. of Computer Graphics and Multimedia
Brno, Božetěchova 2, Czech Republic
cernocky@fit.vutbr.cz
smrz@fit.vutbr.cz

Abstrakt — *This paper presents results of interdisciplinary research which is devoted to design and collection of lexica for speech technology applications. Such lexica are required by automated speech recognizers (ASR), text-to-speech synthesis systems (TTS), or translation systems. For the design and creation of such lexica, linguistics or phonetics solutions are sometimes constrained by the nature of ASR or TTS systems. Within this paper, we would like to present our general experiences in this field and also some experiences from creation of Czech and Slovak LC-Star2 Lexica.*

1 INTRODUCTION

Speech technology applications can be currently met in many different areas of ordinary human life, e.g. from the field of speech recognition, personal dictation systems are already widely used over the world. Other applications of continuous speech recognition are in the transcription of any spoken utterances in different situations as transcription of radio or TV programs, lectures, different meetings, court sessions, medical reports, etc. Next group are speech recognition applications in voice control of different systems or devices, in voice operated information systems which are based on recognition with relatively small vocabulary.

Further, text-to-speech conversion is another important speech technology application, i.e. speech synthesis. It is used especially in above mentioned control and information systems where more natural communication between human and machine is reasonable. Further application group can be in tools supporting the life of handicapped people where written documents as mails are converted by speech synthesizer into audio form. Finally, the speech technology tools can be met also in more general text processing as automated translation systems, spell checkers, text analyzers, etc.

All these mentioned applications have a common need of a lexicon though slightly different requirements are posed to these lexica for particular applications. For Automated Speech Recognition (ASR) and speech synthesis systems mainly the information about the pronunciation of particular words is the most important information. Especially, for translation systems and text processing systems the information about morphological form of given word can be useful.

Our motivation for this work is mainly to have a support for ASR, mainly on the level of the availability of large general lexicon with basic pronunciation variants of included words. Though for Czech (or Slovak) the conversion rules can be used for the generation of pronunciation, the lexicon is important also from the point of view of the vocabulary of words which should be recognized. Our activities in this field originate from our general activities within the research in the filed of LVCSR but also from the participation in European projects as LC-StarII [5], SpeechDat [2], SPEECON [8].

2 LEXICA IN SPEECH TECHNOLOGY

This session should summarize general requirements for lexica used in above mentioned speech technology applications. Of course, exact requirement is always specified by target application but when we suppose more general usage of collected lexica in several systems we can have slightly more extended and generalized content.

2.1 Amount and categories of entries

Firstly, needed size of collected lexicon is discussed. This is typically language dependent and it is not surprising that for the equivalent coverage of language vocabulary we need higher number of entries

for more inflective languages.

- *non-inflective languages* - approximately 100 000 entries give very good coverage,
- *inflective languages* - much more than 100 000 entries are necessary for successful usage in Large Vocabulary Continuous Speech Recognition system (LVCSR).

Consequently, the percentage of general words coverage is usually established as typical requirement for equivalent lexica in different languages. Within European project LC-Star the lexica of major European languages were collected and 95% coverage of common words was required in the specification. Typical number of entries are in the following table as it is mentioned in distribution catalogue of ELRA [1].

Language	Common words
US English	51 119
German	55 507
Spanish	55 854
Italian	56 420
Slovenian	64 521
Czech	**84 058**
Slovak	**96 622**
Finish	144 233

Table 1: Illustrative examples of number of common words in selected LC-Star lexica

Designing the content of the lexica, the entries are supposed to be from several different categories summarized in the following points:

General entries (common words) - This group of entries is always the core of each lexica as these words creates the basis of each collected language.

Topic dependent entries - ASR systems are applied many times for the recognition of the content which may be strongly topic specific, e.g. medical reports, court protocols, broadcast news, weather forecast, etc. For these purposes special topic dependent lexica are created.

Proper names - As each language contains a lot of proper names which are sometimes overlapping between particular languages, the sublexica of names like persons, cities, streets, or companies are standardly collected.

Special application words - As we can meet many times the same applications of speech technology systems which are localized to different languages but which should use typically the same control commands, the lexica of these special application words are another very important and very typical part of lexica or other speech databases.

2.2 Annotation of lexical entries

Once we have particular lexical entries it is necessary to create detailed description of them. So called annotations of lexical entries are done on several levels: a) on *basic phonetic content* providing the information about word pronunciation and optionally including also word syllabification, b) on *morphological level* describing all morphological categories of words, c) on *general class categories* which may categorize words into meaning based subgroups as common words, persons, cities, streets, companies, special application words, etc.

3 COLLECTION OF LEXICAL DATA

3.1 Collection of text corpus

For the collection of general entries a big corpus contains texts of different topics must be created. Currently there is a lot of publicly available sources of different texts in electronic form.

- Mainly on the Internet we can find general text as newspaper articles, electronic books, etc. These sources are typically used for collection of topic independent corpora.
- Special documents for topic dependent lexica as medical reports, technical documents, parliament records, etc. can be find at the Internet in sufficient amount too.
- On the other hand for some more special applications private sources of text, which are not available from public sources, must be used.
- Proper names are collected from publicly available directories and official lists (government, institutional).

Finally, there are two special sources of Czech and Slovak text at the institutions of Czech (Slovak) National Corpus [3]. These corpora contain really huge amount of representative texts, on the other hand, there is not full open access to these data.

3.2 Tokenization procedure

Tokenization procedure is used to obtain particular lexical entries from source texts and it is performed typically in the following steps.

1. clearing of the text - i.e. HTML commands, figures, tables, formulae, and other special text parts are removed,
2. collection of sentences - however the separation of particular sentences is slightly redundant for lexical entries tokenization, it is important for language modelling,

3. analysis of word rates - the rates are computed basicly on the level of unigrams (but bigram or trigram rates are many times also analyzed) for the selection of the most frequent entries.

3.3 Irregular entries

After the above discussed procedure, a lot of irregular entries remains in the lexica. Typical examples of irregular entries are: numbers, abbreviations, spelling, foreign words, etc. It is suitable to solve the collection of these entries separately as much more manual actions is necessary to process these entries.

Collected lists may also contain a reasonable amount of errors. They appear due to the fact that majority of used sources contain non-revised text within which the errors might be more frequent. Consequently, manual revision of collected texts or lexica must be done. This check is done typically during the annotations commonly with completing further information to each lexical entry.

3.4 Automatization of the collection

Collecting of above mentioned entries, the compromise between automated and manual processing always must be done. Of course, the collection of more than 100 000 lexical entries must be performed mostly automatically, on the other hand, necessary manual check must be always provided due to rather big amount of incorrectness in available texts.

4 PHONETIC ANNOTATION

Phonetic content represents the most important information contained in the lexica for speech technology application. Both speech-to-text (ASR) and text-to-speech (synthesis) systems work with this information. So called orthoepic representation of words (i.e. regular pronunciation) is a required source for acoustic modelling in speech recognition systems or for the generation of proper sounds in speech synthesis.

4.1 Phonetic alphabets

Particular subword elements (phones) should be represented by exactly defined symbols. There are several phonetic alphabets which are used for this purposes. SAMPA alphabet [10] is used as an international standard. There is the official set of Czech SAMPA agreed by majority of labs dealing with phonetics in different research fields and it is available at the above mention WEB page. Such standard was not agreed for Slovak, so we are using so called Slovak SpeechDat SAMPA in our lexica.

Though such a standard is reasonable and is used especially for external exchange of data, for our internal purposes we are using our private alphabet which is very close to Czech orthography and which is due to this fact much more easily understandable (e.g. for orthography "*dědeček Ti poví pohádku*" we have in SAMPA "`Jedet_Sek ci povi: poh\a:tku`" and in our private alphabet "`ďedeček ťy pový pohátku`"). Moreover, each phone is represented here always by single character which yields to easier decoding of a word into particular phones.

4.2 Basic phonetic annotation

Problems which are met during the creation of phonetic annotations are mainly in differences between the regular and real pronunciation. These differences may cause severe failures in speech recognition accuracy, on the other hand, all possible (and many times rare) pronunciations cannot be included into such lexica. Consequently, the phonetic content of these lexica must be created respecting only the most important pronunciation variants.

We have dealt mainly with the creation of phonetic annotations of Czech and Slovak lexica which were realized in following steps.

1. automatically created orthoepic transcription is obtained by grapheme-to-phoneme conversion tool *transc* [7],
2. all entries are manually checked and possible differences are marked or more pronunciation variants are completed,
3. differences from regular pronunciation appears mainly for foreign words, names, numbers in numerical forms, abbreviations, etc.
4. orthoepic record is also obtained by *transc* tool in this case to guarantee transcription only by allowed characters representing particular Czech or Slovak phones,
5. pronunciation variants do not contain slang versions, regular context dependent changes of pronunciations, slight differences of pronunciation due to fast continuous speech.

The first versions of our phoneme-to-grapheme conversion rules were derived automatically on the basis of existing speech transcriptions (Slovak SpeechDat). These rules are continuously corrected and completed on the basis of reference knowledge, especially from [6] and [4], and also within the phonetic annotation of lexical entries or other speech databases.

```
frekventovanou  6       f r e k v e n t o v a n o\_u
frenštát        2       f r e n S t a: t
frenštátě       1       f r e n S t a: c e
freuda          4       f r o j d a           f r e\_u d a
freudovsky      10      f r o j d o f s k i   f r e\_u d o f s k i
freudu          1       f r e_u d u
frida           2       f r i d a
```

Table 2: Samples from Czech SpeechDat lexicon

4.3 Suprasegmental phonetic annotations

Mainly for the purposes of speech synthesis, the information about suprasegmental structure of words is interesting. That is the reason why lexica contain also the information about word syllables and about the position of the stress in the word. Especially, as the syllabification rules often are not exactly defined, possible decision about syllable boundaries is not simple and unambiguous.

Syllabic sounds in Czech and Slovak are all vowels and diphthongs. In some situations, we can find syllabic version of several consonants (e.g. "vlk", "trn", "stĺp", etc). According to additional rules, the borders between particular syllables in words are defined. The algorithm works with 100% accuracy on simple syllable structures, some inaccuracies may appear in setting of borders between syllables in some more complex cases.

This algorithm is well approximating similar acoustic items. Sometimes it may give results different from morphological based syllabification. On the other hand, as syllabification rules are not generally fixed (e.g. hyphenation rules are changing) and from the language engineering point of view, the results are coherent and useful for speech recognition and synthesis.

5 MORPHOLOGICAL ANNOTATION

The lexica contain frequently also morphological information. For highly inflected languages, the morphology can be of great use in language modelling, especially in class-based language models and their derivatives. This information is also of key importance for translation systems.

For purposes of speech technology following entries are the most important part of lexica

- *lemma* - basic form of the word, more lemmas of one word form are possible,
- all possible *POS tags* of given word form

Our procedure of morphological transcription was based on automated analysis by morphological analyzer AJKA [9] which was followed by manual checks. Manual correction is important mainly for the creation of morphological tags (POS tags) for proper names, words of foreign origin, word composites, neologisms, etc. where automated analysis is not working always correctly. Also simplified morphological annotation is used typically for multiword lexical entries which is based on morphology of the most important word of whole entry.

6 AVAILABLE LEXICA

Finally, we would like to present our participation on the creation of Czech and Slovak lexica of LC-Star standard and some other lexica available in our research labs.

6.1 Structure of target lexica

Different formats of particular lexica are naturally given by different requirements given by target system. Generally, we can create any structure of the lexica. On the other hand, there are many common projects using and defining internal standards (e.g. LC-Star or SpeechDat family formats) or formats of widely used tools (HTK) used by many labs over the world so keeping these standards is very reasonable. These two standards are presented now as examples, commonly with short data samples.

SpeechDat standard see [2] and sample in tab. 2
- lexica related to collected speech databases
- containing only the pronunciation
- close to widely used standard for HTK tools
- our Czech SpeechDat lexicon - 19 313 entries
- our Czech SPEECON lexicon - 25 562 entries

LC-Star standard see [5] and sample in tab. 3
- XML format - different categories of annotation
- DTD files - check of structure correctness
- same coverage requirements for lexical entries
- common set of application words for all languages
- Czech LC-Star - 132 574 entries
- Slovak LC-Star - 145 313 entries

Finally, our labs at CTU or VUT uses private lexica for LVCSR systems of different topic. These lexica are not publicly available. The contents of these lexica are usually limited to pronunciation and for suitable usage within general LVCSR system we use the lexica with approximately 500 000 entries.

```
<ENTRYGROUP orthography="abbého">
   <ENTRY>
      <NOM class="common" gender="masculine" animate="yes"
           number="singular" case="genitive" />
       <LEMMA>abbé</LEMMA>
       <PHONETIC>" a - b e: - h\ o</PHONETIC>
   </ENTRY>
   <ENTRY>
       <NOM class="common" gender="masculine" animate="yes"
            number="singular" case="accusative" />
       <LEMMA>abbé</LEMMA>
       <PHONETIC>" a - b e: - h\ o</PHONETIC>
    </ENTRY>
</ENTRYGROUP>
<ENTRYGROUP orthography="nabereme">
   <ENTRY>
      <VER person="1" number="plural" gender="not_specified" animate="not_specified"
              tense="present" voice="active" aspect="perfect" mood="indicative" />
      <LEMMA>nabrat</LEMMA>
      <PHONETIC>" n a - b e - r e - m e</PHONETIC>
   </ENTRY>
</ENTRYGROUP>
```

Table 3: Samples from Czech LC-Star2 lexicon

7 CONCLUSIONS

General procedure and experiences of the creation of lexica for application within speech technology systems were described. We have presented mainly the experiences with the collection of several Czech and Slovak lexica, commonly with the description of typical content of such lexica. Publicly available lexica via ELRA contain usually about 50 000 lexical entries. As it was presented, for successful usage within LVCSR system the size of the lexicon for Czech must be increased, so we are continuously working on extension our internal Czech lexica.

ACKNOWLEDGEMENT

The activities of this work were supported by grant GACR 102/08/0707. The general part of the work of the first author is also part of research activities within MSM 6840770014 "Perspective Informative and Communications Technicalities Research".

The creation of LC-StarII Czech and Slovak lexica was supported by Harman/Becker, Ulm, Germany which is also owner of the collected lexica.

References

[1] European languange resources association (ELRA) - home page. http://www.elra.info.

[2] Pages of all SpeechDat projects. http://www.speechdat.org.

[3] Český národní korpus. Domovská stránka http://ucnk.ff.cuni.cz.

[4] Á. Král. *Pravidlá slovenskej výslovnosti. Systematika a ortoepický slovník.* Matica slovenská, 2008.

[5] A. Moreno. LC-StarII. Lexica and corpora for speech-to-speech translation components. http://www.lc-star.org.

[6] Z. Palková. *Fonetika a fonologie češtiny.* Univerzita Karlova, vydavatelství Karolinum, 1994.

[7] P. Pollák and V. Hanžl. Tool for Czech pronunciation generation combining fixed rules with pronunciation lexicon and lexicon management tool. In *Proc. of LREC'02*, May 2002.

[8] P. Pollák and J. Černocký. Czech SPEECON adult database. Technical report, Nov 2003. http://www.speechdat.org/speecon.

[9] R. Sedláček and P. Smrž. A new Czech morphological analyzer *ajka.* In *Proc. of TSD 2001.*, Springer-Verlag, 2001.

[10] J. C. Wells and et al. Sampa - computer readable phonetic alphabet. WEB page http://www.phon.ucl.ac.uk/home/sampa.