Corpus of Spoken Slovak Language

Milan Rusko[1] and Radovan Garabík[2]

[1] Department of Speech analysis and Synthesis, Institute of Informatics,
Slovak Academy of Sciences, Bratislava, Slovakia
milan.rusko@savba.sk
[2] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
http://korpus.juls.savba.sk/

Abstract. In this paper a short description of activities towards building a general speech corpus of spoken Slovak language is given. Different rôles and specific features of text corpus and speech corpus are investigated as well as the most frequent mistakes and misunderstandings of the concept of a speech corpus are mentioned. The concept of a big representative corpus of spoken language and its desired properties are presented. The paper gives an overview of the current state of the art in speech corpora all over the world. It explains the need for a national speech corpus and indicates some of the typical areas of research and applications taking advantage of the existence of such a corpus. The speech databases currently available in Slovakia are listed and the particularities of annotation structures of these databases are pointed out. The authors search for a general annotation structure suitable for the kind of speech corpus envisaged. Some of the basic concepts and technical solutions used in recording and computer aided annotation used for the existing speech corpora are described. The most significant problems standing in the way of building a big speech corpus are pointed out. Furthermore, a pilot version of a speech corpus is presented, containing several recordings and their orthographic transcription.

Keywords: *speech corpus, database, spoken speech, Slovak.*

1 Introduction

Speech corpora play an irreplaceable rôle in present-day automatic speech processing research and development. The information obtained from speech corpora and databases is used for building acoustic models for speech recognition, language models for natural language processing, dialogue models for dialogue management in human-machine interaction and many other purposes. Special speech databases are being built for "unit selection" or "corpus based" speech synthesizers. Every database is built for its particular purpose and is therefore application specific with regards to the choice of speech material and annotation aimed at covering the needs of the actual application.

It would certainly be helpful to have a general speech corpus available for the Slovak language that would allow for broad research in many scientific

areas ranging from linguistics, stylistic analysis, research of dialects, phonetics, phonology, from speech communication to extralinguistics, vocalics and speech acoustics. A pilot version of such a speech corpus, which could be considered as a statistically representative sample of the spoken speech communication in Slovakia is being prepared at the Slovak National Corpus department[1] of the Ľ. Štúr Institute of Linguistics, in collaboration with the Department of Speech analysis and Synthesis at the Institute of Informatics of the Slovak Academy of Sciences. The aim of the pilot version is to investigate the principal ways of building a spoken corpus, consider different possibilities for a transcription and query mechanism and prepare the way for a big, representative corpus. According to its expected volume and diversity of speech material the final corpus has to be collected with the mutual cooperation of several institutions. The benefit of having such a corpus available would be extraordinarily big not only for theoretical research, but also for commercial application development as well. The cultural consequences are not negligible either, since language represents a substantial part of national culture.

2 "Corpus" versus "database"

In principle, any collection of more than one text can be called a corpus – corpus being the Latin expression for "body", hence a corpus is any body of text. But the term "corpus" when used in the context of modern linguistics most frequently tends to have more specific connotations than this simple definition.

According to McEnery and Wilson [2]

> "the following list describes the four main characteristics of the modern corpus: sampling and representativeness, finite size, machine-readable form, and standard reference". Scientists are therefore interested in creating a corpus which is maximally representative of the variety under examination, that is, which provides them with an as accurate a picture as possible of the tendencies of that variety, as well as their proportions. The corpus should contain a broad range of speakers and genres which, when taken together, may be considered to "average out" and provide a reasonably accurate picture of the entire language population.
>
> The term "corpus" also implies a body of text of finite size, but this property does not have universal validity – it is possible to create a monitor corpus. This "collection of texts" is an open-ended entity – texts are constantly being added to it, so it gets bigger and bigger. The main advantages of monitor corpora are: dynamic nature – new texts can always be added, unlike the synchronic "snapshot" provided by finite corpora; and wider scope – they provide for a large and broad sample of language.

> Their main disadvantage is that they are not such a reliable source of quantitative data (as opposed to qualitative data) because they are constantly changing in size and are less rigorously sampled than finite corpora. [2]

(We prefer a national speech corpus to be open as to reflect the newest tendencies in Slovak speech communication.)

According to Sinclair [3] a (text) corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. A computer corpus is a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance. A corpus can be divided into subcorpora. A subcorpus has all the properties of a corpus but happens to be part of a larger corpus. Corpora and subcorpora are divided into components. A component is not necessarily an adequate sample of a language and in that way is distinct from a corpus and a subcorpus. It is a collection of pieces of language that are selected and ordered according to a set of linguistic criteria that serve to characterize its linguistic homogeneity. While a corpus may illustrate heterogeneity, and also a subcorpus to some extent, the component illustrates a particular type of language.

The term annotated corpus is used for any corpus which includes codes that record extra information. (We think that according to this definition the existing Slovak speech databases can be considered as specialized satellite components of the future general speech corpus.)

Campbell has published a practical definition (coming out of several older definitions) explaining the difference between a database and a corpus [4] :

> A "database" is an organized collection of information, typically designed for ease of retrieval by computerized methods; a "corpus", on the other hand, is a collection of naturally-occurring spoken or written material in machine-readable form, that are in themselves more-or-less representative of a language for the systematic study of authentic examples of language in use. The important difference is that while both comprise an accumulation or assemblage of texts or recordings which can be considered as representative of a genre, the former is usually "constructed", and the latter "obtained". More specifically, a database is purpose-built; a store of information which is structured from the beginning, while a corpus is a body of information from which knowledge can be derived.

3 "Text corpus" versus "speech corpus"

In some countries the first attempt to build a general spoken language corpus was made by linguists who had experience in collecting and text corpora or by people from the speech processing community who had been involved in speech database construction. Therefore in some cases the speech corpus was treated very similarly to a text corpus supplemented with an "audio version" of the

text included in the corpus. The non-verbal cues or even prosody and other important information were omitted. The annotation then consists only of an orthographic transcription, some basic data about the identity of speaker and the situation when the speech was recorded.

Exaggerating a bit, one could say that a user of such a corpus finds himself in a position similar to that of patient with aprosodia – an inability to comprehend (or articulate) emotional voice tones and miss the affective or "feeling" content of speech. But the speech corpus offers a wide scale of information on different aspects of human communication, which should not be restricted to the textual and linguistic content.

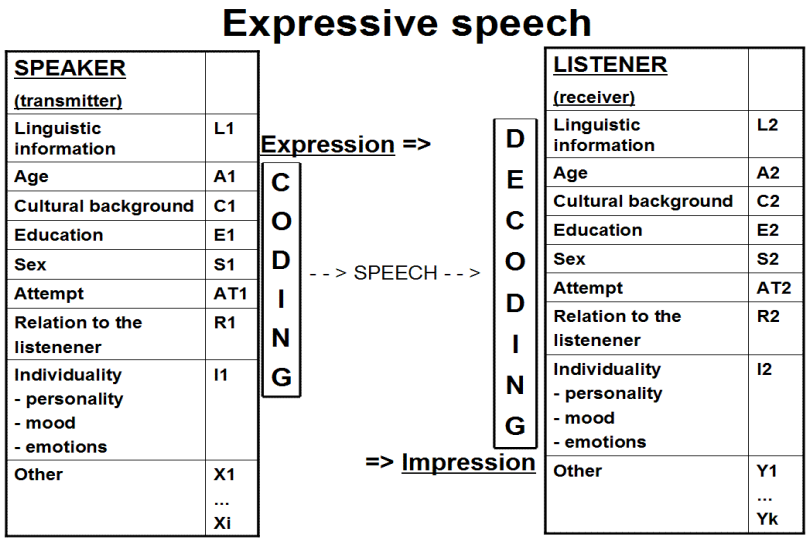

Fig. 1. Simplified scheme of transmission of various information from a speaker to a listener. Every part of the information carried by the speech signal can be an object of research and can be important for applications and should be therefore (at least partly) annotated in the speech corpus.

The corpus should be open to a broad scientific and public community, to allow for the novelty of previously unconsidered usage of the data. As Bird & Liberman say "Once created, a linguistic database may subsequently be used for a variety of unforeseen purposes, both inside and outside the community that created it." [5]

From an acoustical point of view, speech uses only several acoustic quantities (fundamental frequency, time duration of phonetic elements and pauses, intensity of acoustic pressure and frequency spectrum) to carry diverse information not only on the linguistic content, but also on the speaker and communication situation.

Pointing out bad practices in speech corpora building Campbell says [4] "when designing speech databases, care is usually taken to exclude all inarticulate prosody, since it is associated with "ill-formed" speech". (We agree, that the speech is not ill-formed, but our knowledge is still insufficient and the models we have developed are not able to model the natural speech communication correctly.)

A segment in spoken language is an individual consonant, vowel, tone, or stress that makes up a word. An utterance is made up of both segments and supra-segmental features. These are broadly divided up into prosody and paralinguistics. Prosody refers to pitch, loudness, duration, intonation and tempo. Paralinguistics, which is much more difficult to measure, refers to the expression of speaker characteristics, individuality (personality, mood and emotion) – the speaker's attempt and his relationship to the listener. These nonverbal or suprasegmental elements of a speech utterance constitute a significant part of its meaning. The nonverbal cues of the voice are the object of study of vocalics.

The speech corpus should therefore contain different information and various levels of annotation, such as:

- sound file properties (name, description, format, recording conditions, copyright, etc.)
- linguistic information (various transcriptions, linguistic annotation – morphological tags, part of speech tags, syntax, semantic annotation, prosody annotation, etc.)
- extralinguistic information (dialogue and communicative acts annotation, voice quality, pauses, fillers, disfluences, elements specifying background noise and signal quality etc.)

4 General and representative corpus of spoken language

Several attempts have been made to design a relatively general and representative corpus for many terrestrial (and even extraterrestrial[6]) languages – mainly for the "big ones", like English, American, Chinese, Japanese, Spanish, French, Korean, but also for Polish, Irish, Scottish (Gaelic), Czech, Croatian and others. For illustration we will mention some details on some of them.

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20^{th} century, both spoken and written. The spoken part includes a large amount of unscripted informal conversation, recorded by volunteers selected from different age, region and social classes in a demographically balanced way, together with spoken language collected in all kinds of different contexts, ranging from formal business or government meetings to radio shows and phone-ins to broadcast news and conversational telephone speech [7].

There are two parts to the 10-million word spoken corpus: a demographic part and a context-governed part.

The Demographic part of the Spoken Corpus was recorded by 124 volunteers from different social groups. They were male and female volunteers from a wide range of ages, and they lived at 38 different locations across the UK. Recruits used a personal stereo to record all their conversations unobtrusively over two or three days, and logged details of each conversation in a special notebook. Those who took part in the recordings were asked after the conversation to give permission for their speech to be included in the corpus. Information about the participants, such as age, sex, accent and occupation, was recorded when available.

The Context-Governed part of the Spoken Corpus was created with the intention to collect roughly equal quantities of speech recorded in each of the following four broad categories of social context:

- Educational and informative events (lectures, news broadcasts, classroom discussion, tutorials)
- Business events (sales demonstrations, trades union meetings, consultations, interviews)
- Institutional and public events (sermons, political speeches, council meetings, parliamentary proceedings)
- Leisure events (sports commentaries, after-dinner speeches, club meetings, radio phone-ins.)

The Spoken Language Corpus of Swedish at Göteborg University, which is general and covers the whole of Sweden (although it is not called "national"), is an incrementally growing corpus of spoken language samples from several languages which presently consists of 1.26 million words from about 25 different social activities. Because spoken language varies considerably in different social activities with regard to pronunciation, vocabulary, grammar and communicative functions, the goal of the corpus is to include spoken language from as many social activities as possible in order to facilitate research that will provide a more complete understanding of the rôle of language and communication in human social life [8].

The recording facilities covered are: auctions, bus driver/passenger conversation, court, dinner, discussion, factory conversation, formal meeting, hotel, informal conversation, information, service (phone), interview, lecture, market, medical consultation, religious service, retelling of article, rôle play, shop, task-oriented dialogue, therapy, trade fair, travel agency.

The Czech National Corpus has several projects of spoken corpora available [9] – the Prague Spoken Corpus (PMK), the Brno Spoken Corpus (BMK) and ORAL2006.

The PMK was collected during the years 1988–1996 and was the first available corpus of spoken Czech language. The audio recordings were taken in the city of Prague and surroundings, and the corpus was designed to contain four main sociolinguistic variables – speaker's sex, age, education and discourse type, and for simplicity all divided into two sets (man/woman; under 35/over 35 years; less than university/university education; formal speech/informal speech). The corpus contains 674 992 words and is available only in the form of transcribed text. The BMK was collected during the years 1994–1999 in the city of Brno, following the same structure as the PMK.

The most recent ORAL2006 tries to get recordings from the whole area of Bohemia, divided into four main regions. The sociolinguistic distribution of the recordings is kept balanced according to the speaker's age, sex and education, and less to the region of origin. The corpus contains recordings of 754 persons, amounting to 1 312 282 tokens of transcribed text.

4.1 Available speech databases in Slovak

The first professional speech database in Slovak was *SpeechDat-E SK* [10], following the SpeechDat specification [11] and having recordings from 1000 speakers. In spite of the fact that this database is specialized for training and testing speech recognizers in teleservices, it contains phonetically rich sentences which can be used for some purposes in speech research [12].

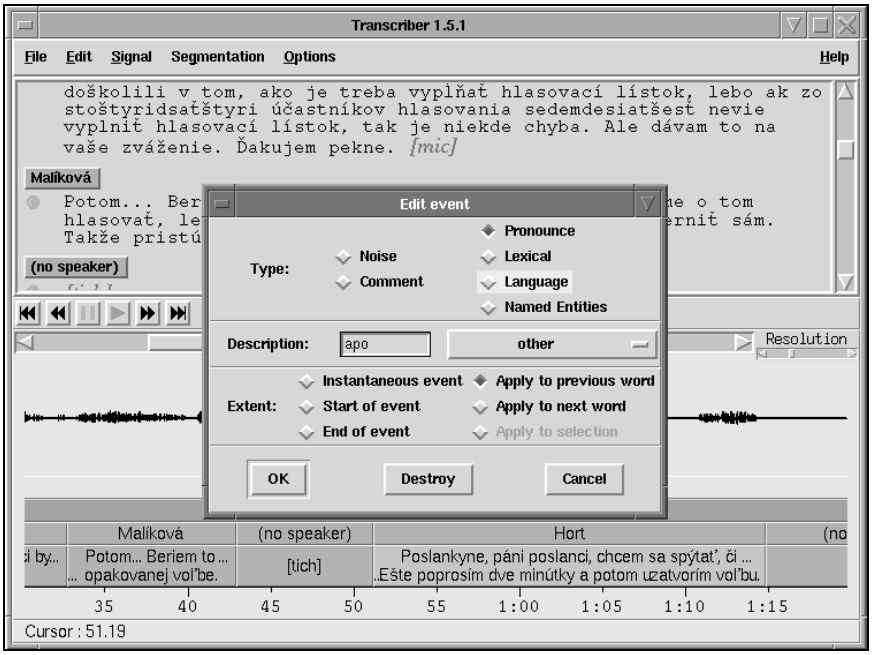

Fig. 2. Annotation of parliament proceedings in Transcriber [15].

MobilDat-SK, which was developed in a frame of the IRKR project [13] is a mobile phone counterpart of SpeechDat with 1100 speakers. Moreover this database contains an unprompted item, where every speaker answers to one of a set of simple questions (How do you get from your house to the closest supermarket? How do you cook scrambled eggs? etc.)

The *TV news audiovisual database* is being built at Technical University Košice for the purpose of experiments in speech recognition, which should have an application in automatic TV news subtitling [14].

The *TV debates* (e.g. "Pod lampou") *audiovisual database* is being built at Technical University Košice for the purpose of experiments in dialogue modeling and expressive speech recognition, which should have an application in automatic TV program subtitling.

The *Parliament proceedings audiovisual database* is being built at the Institute of Informatics, Slovak Academy of Sciences for the purpose of experiments in speech recognition, which should have an application in automatic Parliament proceedings transcription.

SyntDat – a speech synthesis database designed for unit selection speech synthesis (used in Kempelen 2.0 to 2.2 synthesizers) [16].

5 Some controversies

Discourse markers, that have more or less generally accepted transcription in English e.g. sounds representing backchannels and minimal positive feedback (yes, yeah, yah, okay, mhm, hm, aha, uhu), negative minimal feedback (no, n-n, uh-uh), hesitation (er, erm), exclamations – joy/enthusiasm (yay, yippee, whoohoo, mm:), questioning/doubt/disbelief (haeh), astonishment/surprise (a:h, o:h. wow, poah), apology (oops), disregard/dismissal/contempt (ts, pf), exhaustion (ooph), pain (ouch, ow), requesting silence (sh, psh), anticipating trouble (oh-o:h) etc. are still waiting to be get a standardised transcription in Slovak.

We have no experience with transcription of onomatopoeic noises. Intonation modelling needs a generally accepted annotation scheme which still does not exist although the first attempt towards the definition of Slovak ToBI has already been made [17].

We have no annotation scheme for many supralinguistic and extralinguistic phenomena (e.g. emphasis, voice quality and many others).

If we accept a grapheme to be the smallest element of written text, it would be reasonable to define a phoneme to be the smallest element annotated in a speech corpus. This means that speech recognition technology in Slovak capable of finding phoneme boundaries with acceptable reliability would be needed. For pitch contour and voice quality measurement we often need pitch

marks. Their determination is not language dependent, but reliable pitch marking is still a difficult task.

6 Obtaining Slovak speech recordings

Apart from recordings originating in the specialized databases mentioned earlier, a large part of our proposed corpus will consist of recordings obtained on purpose. The main sociolinguistic data observed will be speaker's sex, age, education, discourse type, conformance with the standard language and region of origin (inspired by the Czech spoken corpora). Although there is a huge potential in spoken corpora for dialect studies, our corpus will focus (at least from the beginning) on standard Slovak. Therefore the recordings will be made primarily in urban areas.

7 Corpus manager

There are several requirements for the corpus query possibilities, each targeting a different end-user group. On one hand, we want a powerful tool for working on the transcribed text, for statistical analysis on the various aspects of the data. This is easily achieved by a standard corpus manager interface, offering all the usual functions for the transcribed text. However, the existing text corpus managers offer no easy possibilities of linking with the specific sound data – this is not necessarily an insurmountable disadvantage per se, because any serious research on the acoustic level will be supposedly performed with rather specialized tools and for specific purposes, and it is not quite feasible trying to accommodate all the possible uses.

The corpus also has to be usable for casual users, without the need to install specialized client software and to study the (often complicated) program controls. Following the ubiquity of web applications, it is obvious that the corpus should be accessible through a simple WWW interface, with a possibility for the user to directly access the relevant sound sample. These two approaches are not exclusive, there is no reason not to provide both possibilities (in fact, a similar system was deployed in the (text) Slovak National Corpus).

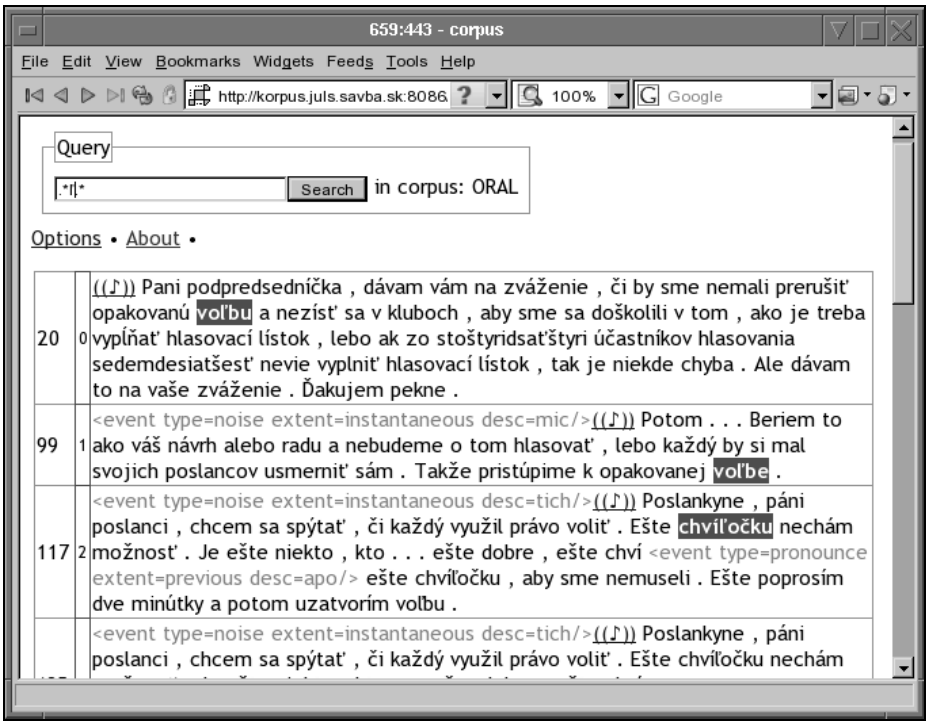

Fig. 3. Speech corpus web interface

8 File formats

For transcription, we are using the *transcriber* [15] software, which allows the annotators to define speakers' identities, define various types of extralingual events and speech phenomena and seamlessly integrate the audio and text data. The transcriber stores its data in a native XML format with links to the audio files and timestamps at synchronization points. We take advantage of this format and use the corpus manager to index the (postprocessed) XML files directly.

There are two conflicting requirements for the audio file format – the first is to maximize the sound quality, the second is to minimize the file size. Given the expected longevity of the spoken corpus and the ever-decreasing cost of storage media, sacrificing quality for the sake of saved disk space is not applicable anymore. This holds even in the case of inaudible quality degradation (using a high bitrate lossy compression protocol). Therefore it is desirable to

archive the original audio data in either an original format, or by using only lossless compression. On the other hand, there are uses for the corpus requiring only access to speech without very noticeable distortion, e.g. demonstration to casual users or as a part of a foreign language instruction process. For web-based services, the size of the transmitted files is important, as well as use of a common multi-platform format, not requiring installation of specialised software.

For the original format, we decided to use the FLAC lossless codec[18], giving a compression ratio of about 50 per cent compared to uncompressed PCM data (more for stereo input). Unfortunately, most modern budget dictaphones use proprietary WMA[1] or DSS formats, which are already lossy compressed. Therefore we expect some of the audio records in the corpus obtained from external sources to be in the WMA format (there is a lack of relevant software and tools needed for DSS format processing and conversion), which can potentially preclude the usage of the data for some specialised purposes, since the sound is already mapped to a psychoacoustic model – primarily, the corpus would not be usable for the development of new psychoacoustic models. However, when keeping the quality at a sufficiently high level, even frequency analysis as required by phoneticians is applicable.

For the format presented to users, we decided to use the lowest compression quality (bitrate) that gives only slight perception of quality distortion.

We used primarily the SPEEX codec[19]. SPEEX was designed specifically for speech encoding at lower bitrates, and gives an excellent compression ratio. Another advantage is a special decoder mode enhancing perceived sound quality (we found that sometimes the SPEEX encoded data sound subjectively better than the original). Before encoding, the sound samples were downmixed to one mono channel and downsampled to ultra-wideband frequency (32 000 Hz, one of the recommended sampling rates for the SPEEX codec). The files were encoded using variable bitrate encoding, encoding complexity 10, at quality 6, which gives an average bitrate of 23 kb/s.

Because of a rather lesser SPEEX penetration to the usual desktop PC systems, we decided to offer Ogg/Vorbis[20] as an alternative (downmixing to single channel, but without resampling, since the Vorbis codec does not have strict recommendation as per the sampling frequency, and resampling often makes the audio sound subjectively worse compared to SPEEX). We used the experimental aoTuV encoder[21] optimized for lower bitrates. Encoding was done at quality -1, giving an average bitrate of 40 kb/s.

Users can therefore choose between SPEEX, Ogg/Vorbis and original (or FLAC) format. There is also a Java applet available, playing SPEEX format for users unable or unwilling to install the required codecs.

---

1 We are using the general name WMA here, although technically WMA can mean several different incompatible codecs (WMA, WMA Pro, WMA Lossless, or WMA Voice).

9 Levels of transcription

Different levels of transcription are possible, each of them putting different strain on the annotation process. In our corpus project, we selected three different levels – orthographic, phonetic/phonemic and suprasegmental transcription.

9.1 Orthographic transcription

Orthographic transcription is the most straightforward, and the basic type of annotation that distinguishes a simple collection of recordings from a speech corpus. We decided the orthographic transcription in our corpus should follow the standard Slovak orthography, transcribing only the differences from standard Slovak pronunciation as an additional word attribute. This both makes the transcription easier to read as well as allowing us to deploy usual NLP tools (e.g. morphology analysis, lemmatization). In some areas, we follow standard Slovak pronunciation, as opposed to the prescribed official one. In particular, the pronunciation of letter *ä* as /ɛ/ does not warrant specific transcription, but its pronunciation as /æ/ does. Similarly, pronouncing the syllables *le*, *li* and *lí* as /lɛ/, /li/ and /li:/ is not marked, but palatalized pronunciation /ʎɛ/, /ʎi/ and /ʎi:/ is. Even though officially correct, it has for all practical purposes disappeared from standard Slovak.

Although tempting, we have chosen not to use the standard punctuation symbols to denote extralingual information (such as pauses and hiatus in speech), since human annotators are prone to unconsciously deploying such marks where orthography rules require, not where the phenomena really occur. We are using specific annotation software possibilities instead, with usual punctuation marks (comma, colon, exclamation mark etc.) being at the annotator's discretion. For the same reasons, we are not using capital letters for any special purpose, the annotators can capitalize words as they feel natural. We recommend putting the dot at the end of sentences as dictated by the logical flow of the document (not by pauses in discourse), the sole purpose of this is to help the automatized analysis tools (in particular morphology analyzer), where marking the end of sentences sometimes improves the processing accuracy.

9.2 Phonemic/phonetic transcription

Phonetic transcription is useful for speech recognition, speech synthesis and basic linguistic research. However, making a correct phonetic transcription requires trained annotators with a good knowledge of language phonetics and is rather time consuming and sometimes controversial. Therefore we decided to include phonemic transcription, with just some phonetic features

(distinguishing several most frequent allophones). This requires designing a general model of phonemic analysis of the Slovak language usable for the transcription process – to our knowledge, no such analysis universally accepted among Slovak linguists exists so far. Only a part of the corpus will be manually transcribed phonetically (in addition to the orthographic transcription). For he rest of the corpus, an automatic grapheme-to-phoneme conversion will be available.

9.3 Suprasegmental annotation

A suprasegmental annotation scheme must provide a mechanism for indicating suprasegmental structure such as word/syllable boundaries and stress markings. The specification may address other types of suprasegmental structure.
A different phonological intonation annotation scheme is needed for every particular language. Inspired by the successful ToBI (Tones and Break Indices) for American English [22] the intonation annotation scheme Sk-ToBI was introduced for Slovak [17]. ToBI annotation by hand is extremely time consuming, therefore only a limited part of the corpus will be annotated manually. This can later serve for training automatic annotation algorithms.

10 Copyright issues

It can be argued that recorded "natural" speech is not protected by the Slovak Republic copyright law (the law is not very clear about the issue). However, the recordings cannot be distributed without consent from the author, as long as there are any data from which the author's identity can be inferred, and according to the current laws it is nearly impossible to legally record somebody without informing him in advance. This means that we are unable to get recordings of really natural speech, and the representative part of the corpus has to be recorded in other ways – e.g. masking the recording as sociological research or public opinion poll, so that the recorded subjects are not aware of the linguistic nature of the recordings. Even so, we cannot expect to obtain spontaneous natural speech.

11 Conclusion

In spite of the fact that we aware of the complexity and resource cost of building a general and representative speech corpus in Slovak we believe that Slovak linguists and speech researchers will proceed in a common effort towards a Slovak speech corpus that could be included in the Slovak National Corpus, as it is common in the leading corpora in the world.

References

1. Slovak National Corpus. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2006. Available from WWW: http://korpus.juls.savba.sk/
2. McEnery, T., Wilson, A., "Part Two: What is a Corpus, and what is in it?" Web pages to be used to supplement the book "Corpus Linguistics" published by Edinburgh University Press, ISBN: 0-7486-0808-7. http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fra1.htm
3. Sinclair, J., School of English, University of Birmingham http://www.ilc.cnr.it/EAGLES96/corpustyp/node5.html
4. Campbell N.: Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language, Journal Language Resources and Evaluation, Issue Volume 39, Number 1 / February, 2005, Publisher Springer Netherlands, pp. 109-118.
5. Bird St. and Liberman M., A Formal Framework for Linguistic Annotation. Speech Communication, 33 (1,2), pp 23–60, 2001.
6. Corpus of Spoken Martian. http://www.elsnet.org/nps/0014.html
7. British National Corpus, http://www.natcorp.ox.ac.uk/
8. Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E. and Ottesjö, C. (2000, December). The Spoken Language Corpus at the Department of Linguistics, Göteborg University [55 paragraphs]. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [Online Journal], 1(3). Available at: http://www.qualitative-research.net/fqs-texte/3-00/3-00allwoodetal-e.htm [Date of Access: 20th of May 2007].
9. Waclawičová, M.: Mluvené korpusy v ČNK: několik poznámek k mluveným projevům a polyfunkčním výrazům. In: Korpusová lingvistika: Stav a modelové přístupy. Studie z korpusové lingvistiky, sv. 1. Eds. F. Čermák, R. Blatná. NLN a ÚČNK, Praha, 2006. P. 347–358.
10. Heuvel, H., Boudy, J., Bakcsi, Z., Černocký, J., Galunov, V., Kochanina, J., Majewski, W., Pollak, P., Rusko, M., Sadowski, J., Staroniewicz, P., Tropf, H. S.: Five Eastern Europen Speech Databases for Voice-Operated Teleservices Completed. In: Eurospeech 2001 – Aalborg, Denmark, 2001.
11. http://www.speechdat.org
12. Rusko, M., Daržagín, S., Trnka, M.: SpeechDat-E telephone speech database as an important source for basic acoustic-phonetic research in Slovak. In: Proceedings of the International Congress on Acoustics, ICA 2004, Kyoto, Japan, part I. p. II-1676 – II-1682. ISBN 4-9901915-6-0.
13. Juhár J., Ondáš S., Čizmár A., Rusko M., Rozinaj G., Jarina R.: "Development of Slovak GALAXY/VoiceXML Based Spoken Language Dialogue System to Retrieve Information from the Internet", Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh, Pennsylvania, USA, 2006. ISSN 1990-9772, pp. 485–488.

14. Pleva,M., Juhár, J., Čižmár, A.: Vývoj a evaluácia multilingválnej databázy pre systémy automatickej transkripcie správ elektronických médií. About development and evaluation of multilingual database for automatic broadcast news transcription systems. Acta Electrotechnica et Informatica, Vol.4, No.2, 2004, ISSN 1335-8243, pp.56-59.
15. Barras, C., Geoffrois, E., Wu, Z., Liberman, M., 1998. Transcriber: a free Tool for Segmenting, Labeling and Transcribing Speech. In: Proc. First Int. Conf. on Language Resources and Evaluation (LREC 98), Granada, Spain, pp.1373–1376.
16. Rusko, M., Daržagín, S., Trnka, M., Cerňak M.: Slovak Speech Database for Experiments and Application Building in Unit-Selection Speech Synthesis. In: Proceedings of Text, Speech and Dialogue, TSD 2004, Brno, Czech Republic, pp. 457 – 464.
17. Rusko, M., Sabo, R., Dzúr, M.,: Sk.ToBI Scheme for Phonological Prosody Annotation in Slovak, in: Lecture Notes in Artificial Intelligence 4629, Springer Verlag, 2007, pp. 334–341, ISBN 978-3-540-74627-0.
18. http://flac.sourceforge.net/
19. http://www.speex.org/
20. http://xiph.org/vorbis/
21. http://www.geocities.jp/aoyoume/aotuv/
22. Silverman, M. et al.: ToBI: A standard for labeling English prosody, Proceedings of the 2nd International Conference of Spoken Language Processing, Banff (1992), 867–870.