Rules for automatic grapheme-to-allophone transcription in Slovene

Jerneja Gros, France Mihelič, Simon Dobrišek, Tomaž Erjavec, and Mario Žganec

University of Ljubljana, Faculty od Electrical Engineering,
Tržaška 25, SI-1000 Ljubljana, Slovenia
nejka@fe.uni-lj.si,
WWW home page: http://luz.fe.uni-lj.si

Abstract. The domain of spoken language technologies ranges from speech input and output systems to complex understanding and generation systems, including multi- modal systems of widely differing complexity (such as automatic dictation machines) and multilingual systems (for example automatic dialogue and translation systems). The definition of standards and evaluation methodologies for such systems involves the specification and development of highly specific spoken language corpus and lexicon resources, and measurement and evaluation tools [5]. This paper presents the MobiLuz spoken resources of the Slovene language, which will be made freely available for research purposes in speech technology and linguistics.

1 Introduction

The domain of spoken language technologies ranges from speech input and output systems to complex understanding and generation systems, including multimodal systems of widely differing complexity (such as automatic dictation machines) and multilingual systems (for example automatic dialogue and translation systems). The definition of standards and evaluation methodologies for such systems involves the specification and development of highly specific spoken language corpus and lexicon resources, and measurement and evaluation tools.

In the beginning, standards for these areas have been derived from the consensus within the spoken language community previously established in a number of European and national projects, with reference to important initiatives in the US and Japan. Primary among these have been the SAM projects (centered on component technology assessment and corpus creation), SQALE (for large vocabulary systems assessment) and both SUNDIAL and SUNSTAR (for multi-modal systems.)

Past and present projects with significant outputs in the domain of assessment and resources include ARS, RELATOR, ONOMASTICA and SPEECHDAT, as well as major national projects and programs of research such as VERBMOBIL in Germany. This has led to an initial documentation of existing practice which was relatively comprehensive but in many respects heterogeneous and widely dispersed. The lack of generic technologies and resources and the wide diversity of formats and specifications has hindered the effective reutilisation of existing resources. In 1993, the EAGLES (Expert Advisory Group on Language Engineering Standards) initiative was launched within the framework of the CEU's DGXIII Linguistic Research and Engineering (LRE) Programme, to accelerate the provision of standards for developing, exploiting and evaluating large-scale language resources. A special working group has been set up for this purpose, named the Spoken Language Working Group (SLWG). The project resulted in the publication of comprehensive guidelines documenting existing working practices in Europe and guidelines for spoken language resource creation and description [5].

2 Slovene Speech Corpora

For the Slovene language, several attempts in speech data collection were made in the past, resulting in various speech corpora: SNABI [12, 13], LUZ diphones [8], GOPOLIS [4] and SPEECHDAT-Slovene [14], distributed by ELRA.

The collected speech data mainly represented the domain of intended applications and are not available for distribution, except for the Slovenian SpeechDat(II) FDB- 1000 corpus containing phonetically rich sentences. The corpus consists of read and spontaneous speech and was recorded through an ISDN card (1.000 speakers). A phonetic lexicon with canonical transcriptions in SAMPA is also provided. However, due to its high cost (20.000 EUR) the Slovene SPEECHDAT corpus can hardly be used for research or even development purposes. We therefore decided to create a collection of various Slovene speech data, freely available for research purposes in speech technology and linguistics. Such speech resources are essential for building multi-lingual speech recognition and text-to-speech synthesis applications.

The idea is in the process of being finalized within the MobiLuz project funded by the Slovene Ministry of Science and Technology and the Slovene mobile telephony operator Mobitel d.d.

The MobiLuz project collects and integrates resources of several previous project, e.g. the EU Copernicus SQEL and MULTEXT-East and will provide three main deliverables:
1. MobiLuz Slovene speech corpus,
2. MobiLuz Slovene speech corpus annotations and lexicon and
3. MobiLuz speech tools.
The next three sections detail our work on these deliverables.

3 MobiLuz Slovene speech corpus

The MobiLuz Slovene speech corpus includes various collections of speech data pronounced by multiple speakers, either as read speech or spontaneous speech. In particular, the corpus consists of the following components:
- a fully updated and revised version of the 50 speaker GOPOLIS corpus of air

travel inquiries,
- isolated spoken commands (digits, common control commands etc.) and
- recordings of live dialogs.

3.1 Gopolis

The GOPOLIS corpus is a large multi-speaker speech database, derived from real situation dialogs concerning airline timetable information services. It was used as the Slovenian speech database within the SQEL project (Copernicus COP-94 contract No. 01634) for building a multi-lingual speech recognition and understanding dialog system [11], capable of passing information over the telephone line to a client in one of four European languages - German, Czech, Slovak and Slovenian. The name of the database has been derived from "GOvorjena POizvedovanja o Letalskih Informacijah v Slovenskem jeziku", meaning "Spoken Flight Information Queries in the Slovene Language".

The sentence corpus was drawn from listening to recordings of real situation dialogs between anonymous clients' inquiries and telephone operators at the Adria Airways information center (15 hrs of speech stored on audiotapes).

The selected 300 typical sentences were compiled into the form of rewrite rules to obtain a generative sentence pattern grammar [7]. Using this grammar, we produced 22,500 different sentences for short introductory inquiries, long inquiries and short confirmations. 5,077 of them were selected to form the final sentence corpus.

Each of the total of 50 speakers (25 female and 25 male) read about 100 randomly selected unique corpus sentences and 71 sentences of welcome greetings, introductory phrases, short affirmations and farewell greetings, common to all of the speakers. Each session has a list of attributes with speaker and recording session descriptors.

3.2 Isolated Commands

A set of isolated commands spoken by various speakers has also been recorded. Included are the most common commands used in telecommunication applications and commands for navigating windows applications.

3.3 Recording conditions

The recording sessions were performed in a normal laboratory acoustic environment. Additional noise, such as background speech or slamming doors was avoided. The utterances were acquired by a close talking microphone and simultaneously by telephone. Thus an additional analysis of both audio devices was enabled.

A set of recording environment programs was developed for the HP 9000 workstation platform. The user interface program built for recording communicates with two audio servers over the network, displays and saves the acquired signals and displays the sentences that a speaker should utter. So the acoustic realizations are in form of read continuous speech. The program is also equipped with loudness detection and an acoustic messaging system. It takes care of the correct maximum loudness level, begin and end pauses and synchronizes the acquired telephone and microphone speech signals.

The audio servers use the HP 9000/735 common audio hardware components and the additional Gradient Technology DeskLab hardware with a full telephone interface (DeskLab is a data acquisition and play device, which communicates via SCSI with a workstation). The speaker has to press a space bar key to signal the start of the recording session and then again to finish it. The program requires at least half a second of silence at the beginning and at the end of the utterance. A sampling rate of 16 kHz for both microphone and telephone signals and a 16-bit data format with MSB-LSB byte order was chosen.

4 Corpus annotations and lexica

The GOPOLIS corpus is encoded in accordance with TEI recommendations [15], in particular, the base tagset for Transcriptions of Speech, the additional tagsets for Simple Analytic Mechanisms and Language Corpora, and some local modifications. The corpus contains the TEI header, giving the File, Encoding and Profile descriptions. Here general information about the corpus, including speaker descriptions is given. The body of the corpus consists of the 5,077 sentences (utterances), each marked with an ID and references to its speakers.

The utterances are segmented into words and punctuation marks, and each word is given in its orthographic form, as well as in the automatically derived phonetic transcription. Furthermore, the words were automatically tagged for their lemma and morphosyntactic description. Some sample data will contain also prosodic annotations.

Two lexica accompany the corpus: a pronunciation dictionary and a wordform lexicon with the morphosyntactic descriptions.

4.1 Transcription and tagging

The phonetic transcriptions in the corpus are based on the Slovene MRPA set [2, 17] containing machine-readable phonetic symbols equivalent to the Slovene IPA symbols [16].

The morphosyntactic descriptions and lexicon are based on the MULTEXT-East (Slovene) tagset and lexicon [3, 10]. The lexicon contains lemmas, their full inflectional paradigms and the morhosyntactic descriptions of the wordforms [6]. The descriptions have a feature-structure like format and encode informations such as part-of- speech, number, case, etc. These descriptions and lexicon are then used to automatically tag and lemmatise the corpus. The tagger used is TnT [1], which had been trained on the Slovene MULTEXT-East corpus.

4.2 Speech segmentation and alignment

The automatic segmentation is performed using the DTW based approach described in [4] where the speech material is automatically segmented and labelled using dynamic time warping alignment of a natural utterance with a synthesised speech signal. The synthesis of speech signals was achieved by simply concatenating labelled diphone speech signals using a simplified TD-PSOLA technique. The diphone inventory used was borrowed from the Slovenian text-to- speech system S5 [9]. The diphone inventory consists of a set of segmented and labelled diphone speech signals. Every diphone in the inventory is segmented into two separated phones and voiced phones were additionally marked with pitch period markers.

According to the phonetic transcription of the natural speech signal, a sequence of concatenated diphone speech signals was generated. A conventional DTW alignment of the utterance with the synthesised speech signal was performed with two sequences of feature vectors derived from both speech signals. This method has the advantage that we do not need labelled reference speech signals since we know the phone boundaries of the synthetic speech signal.

5 Speech Tools

The MobiLuz speech corpus will be accompanied by the Sigmark software developed in cooperation with Masterpoint, a user-friendly program interface which allows manual editing, viewing and marker corrections of speech signals on various levels (phone, word and phrase levels).

Markers can also be copied from one level to the other which proves to be very useful for synchronising marker sets on different levels.

6 Conclusion

The MobiLuz project is aiming to set up an infrastructure of Slovene spoken resources necessary for building various multi-lingual speech recognition and synthesis applications. For example, the GOPOLIS speech corpus, which will become available in the scope of MobiLuz, has already been successfully used for training of the HMM-based speech recognition engine in the multilingual automatic dialog system SQEL [11].

7 Acknowledgements

The development of MobiLuz speech corpus was supported by the Slovene mobile telephony operator Mobitel d.d. and grant of the Slovene Ministry of Science and Technology. The development of the resources contained in the corpus have also been supported by the EU Copernicus projects SQEL and MULTEXT-East.

References

1. Brants, T. (2000): TnT-A Statistical Part-of-Speech Tagger. Proceedings of the ANLP-NAACL, in print, Seattle.
2. Dobrišek S., Kačič Z., Gros J., Horvat B. and Mihelič F., (1996): An Initiative for Standardisation of Phonetic Transcription of Slovenian Speech, Proceedings of the Fifth Electrotechnical and Computer Science Conference ER K'96, pp. 247–250, Portorož, Slovenia, 1996.
3. Dimitrova, L., Erjavec, T. Ide, N. Kaalep, H.J., Petkevič, V. and Tufis, D. (1998): Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. COLING-ACL '98 Proceedings, pp. 315-319.
4. Dobrišek S., Gros J., Mihelič F. and Pavešić N. (1998): Recording and Labelling of the GOPOLIS Slovenian Speech Database, Proceedings of the First International Conference on Language Resources and Evaluation, pp. 1089-1096. Granada, Spain.
5. EAGLES Handbook (1997): Handbook of Standards and Resources for Spoken Language Systems. Editors D. Gibbon, Roger Moore and Richard Winski. Berlin: Mouton de Gruyter.
6. Erjavec T. (1998): The MULTEXT-East Slovene Lexicon. Proceedings of the ERK'98 Conference, Portorož, Slovenia, pp. 189-192.
7. Gros J., Mihelič F. and Pavešić N., (1995): Sentence Hypothesisation Using Ng-Gram Models, In Proceedings of the the Fourth European Conference On Speech Communication and Technology, pp. 1759-1762, Madrid, Spain.
8. Gros J., Ipšić I., Mihelič F. and Pavešić N. (1996): Segmentation and labelling of Slovenian diphone inventories, COLING'96, pp. 298-303, Copenhagen, Denmark.
9. Gros, J., Pavešić, N. and Mihelič, F. (1997): Text-to-speech synthesis: a complete system for the Slovenian language. Journal of Computing and Information Technology. 5(1). pp. 11-19.
10. Ide, N., Tufis, D. and Erjavec, T. (1998): Development and Assessment of Common Lexical Specifications for Six Central and Eastern European Languages. Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98, Granada, pp. 233- 240.
11. Ipšć I., Mihelič F., Dobrišek S., Gros J. and Pavešić N. (1998): An overview of the spoken queries in European languages : the Slovenian spoken dialog system. Proceedings of the scientific conference Artificial Intelligence in Industry from Theory to Practice and 3rd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs, High Tatras, Slovakia, pp. 431-438.
12. Kačič Z. and Horvat B. and Derlič R. (1994): Zasnova baze izgovorjav slovenskega jezika SNABI. Proceedings of the ERK'94. Portorož, Slovenia.
13. Kačič Z. and Horvat B. (1998): Izgradnja infrastrukture, potrebne za razvoj govorne tehnologije za slovenski jezik. Proceedings of the Conference on Language Technologies for the Slovene Language. Ljubljana. pp. 100-104.
14. Kaiser J. and Kačič Z. (1998): Development of Slovenian SpeechDat Database. Proceedings of the Workshop On Speech Database Development for Central and Eastern European Languages, Granada, Spain, 1998.
15. Sperberg-McQueen, C.M., and Burnard, L., eds. (1994): Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford.
16. Šuštaršič R., Komar S. and Petek B. (1998): Slovene IPA Symbols, Illustrations of the IPA.
17. Zemljak M., Kačič Z., Dobrišek S. and Gros J. (2000): A Machine-readable Phonetic Transcription of the Slovene Speech, in preparation.