The Slovene BNSI Broadcast News database and reference speech corpus GOS: Towards the uniform guidelines for future work

Andrej Žgank[1], Ana Zwitter Vitez[2], Darinka Verdonik[1]
[1] University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova ulica 17, SI-2000 Maribor, Slovenia
[2] Trojina, Institute for Applied Slovene Studies
Dunajska 16, SI-1000 Ljubljana, Slovenia
E-mails: andrej.zgank@um.si, ana.zwitter@guest.arnes.si, darinka.verdonik@um.si

Abstract

The aim of the paper is to search for common guidelines for the future development of speech databases for less resourced languages in order to make them the most useful for both main fields of their use, linguistic research and speech technologies. We compare two standards for creating speech databases, one followed when developing the Slovene speech database for automatic speech recognition – BNSI Broadcast News, the other followed when developing the Slovene reference speech corpus GOS, and outline possible common guidelines for future work. We also present an add-on for the GOS corpus, which enables its usage for automatic speech recognition.

Keywords: spoken language resources, transcription guidelines, less resourced languages

1. Introduction

In less resourced languages, the need for bigger speech databases is one of the more explicative gaps in the field of language resources, as it holds back the development of automatic speech recognition (ASR) as well as more thorough linguistic analyses of spoken language. The creation of speech databases is held back because of their time-consuming and expensive development.

Over previous years, standards have been developed for creating speech databases. In this paper, we compare two such standards, one followed when developing the Slovene speech database for automatic speech recognition – BNSI Broadcast News (Žgank et al. 2005), the other followed when developing the Slovene reference speech corpus GOS (Verdonik et al. 2013).

Based on experiences when creating both types of resources, the aim of the paper will be to search for common guidelines for the future development of speech databases for less resourced languages in order to make them the most useful for both main fields of their use, linguistic research and speech technologies. We will also present an add-on for the GOS corpus, which will enable its usage for automatic speech recognition, and critically examine the characteristics of the reference speech corpus GOS in the light of its use in ASR.

2. The BNSI Broadcast News Speech Database

The aim of the Slovene BNSI Broadcast News database was to produce a necessary language resource for Slovene large vocabulary continuous speech recognition within unrestricted domain. It comprises two subsets: the speech database with transcriptions for acoustic modeling, and the text corpus. In this paper, we will focus on the speech database only. It consists of two different types of news shows, *TV Dnevnik* and *Odmevi*, from the period 1999–2003. As these shows were taken from the archive, a broad spectrum of topics was covered. The manual transcriptions were done using the Transcriber tool (Barras et al. 2000). The transcription rules created by LDC, LIMSI and COST 278 BN SIG were taken as the baseline. Different language-dependent rules were added to this baseline set-up.

The final speech corpus consists of 42 shows covering a total length of 36 hours. The transcriptions cover 268k words, 37k of them being different. The number of speakers is 1565 (1069 males, 477 females). The Slovene BNSI Broadcast News database is available through ELRA/ELDA.

3. The GOS corpus

The reference corpus of spoken Slovene – GOS represents a balanced collection of authentic spoken discourse from all over Slovenia. Its main purpose was to enable broader linguistic research of the spoken language. The data consists of the most common speech situations: media, educational institutions, official settings and casual conversations.

The transcriptions follow most of the EAGLES (1996) and TEI (Burnard et al., 2014) recommendations for the transcriptions of speech in spoken corpora. They were done using the Transcriber tool. Two-level system of orthographic transcription was established: the pronunciation-based and the standardized. We found that such a system is very efficient and useful for languages with a great variety of dialects, such as Slovene. Along with the transcriptions some basic prosodic, textual and pragmatic marks are annotated.

The final corpus covers approx. 120 hours of recordings, resulting into over 1M words of transcriptions in total. It includes speech of 1555 speakers (859 male, 696 female).

The GOS corpus is available under the Creative

Commons license in three different modes or formats. In addition to online use within the online concordancing tool (www.korpus-gos.net), it is also available as a dataset intended for professional linguistic use and for language technologies. For the latter, two formats have been produced, the TEI P5 XML format (Erjavec, 2014) and a double set of Transcriber files. Audio files of public discourse are also available on the basis of individual license.

4. Differencies and similarities between both resources and future guidelines

Both of the two language resources, GOS and BNSI, need further growth and additional data, since their present scope is far from satisfactory. Both tasks demand a lot of manual effort, but they are also very similar. For minor languages like Slovene, where the finances for development of language resources are very limited, it is therefore rational to direct efforts into developing one single, huge speech database that would cover at the same time the needs of both ASR and linguistics as much as possible. In this spirit, we have therefore thoroughly examined both databases, outlining the main differences between them, and searched for solutions that would satisfy the needs of both fields of research.

The main differences between BNSI and GOS can be summarized on five levels:

4.1 Acoustic environment

The acoustic environment is one of the main characteristics which influence the development of acoustic models for an ASR system. The Slovene BNSI database has been compiled from the national broadcaster's archive and is comprised of professionally produced TV news shows where high quality equipment was used. The worst acoustic conditions concerned those scenarios where communication took place over narrow frequency band, for example in speech, spoken over the telephone, where field reports are given as part of news reports. Difficult acoustic environment for ASR is also the presence of background music.

The Slovene GOS corpus contains a much wider spectrum of various acoustic environments. One part of the corpus (approx. 41.2%), which includes informative and entertaining shows, has similar acoustic environments to those of the BNSI database. In addition to TV shows, which were the source in the case of the BNSI database, the GOS corpus also contains recordings from radio stations. The recordings from TV and radio stations originated mainly from broadcasters' archives, where MP3 encoding was applied to audio signal. All these environments are absolutely applicable as the domains of ASR and should be paid appropriate attention in the future.

The majority of the GOS corpus covers recordings of communication in personal contact or over the telephone. These two categories can be more challenging for ASR since they can be acoustically degraded due to the recording conditions and equipment used.

The main drawback of the GOS corpus recordings for ASR is that the need for low storage capacities of the files in order to make them easily transferable over the internet was made more important than presenting the whole spectrum of signal features, as the only intention was to keep the recordings understandable for humans, while the needs of ASR where not considered. In the future, the aim should be providing the highest possible quality of the signal on all recordings.

For the needs of ASR it is also important that the transcription includes information on acoustic environment. The BNSI therefore includes information on sound fidelity (e.g., channel noise, field, noisy...) for each turn while the GOS transcriptions do not. The future directions should be to cover this type of information in transcriptions.

4.2 Speech segmentation

It is not a trivial task to define an optimal speech segment when transcribing speech. The pauses in speech often do not correspond to sentences in the written text – other characteristics of speech need to be considered as well, such as the semantic and syntactic structures of the message, prosody, etc. For the linguistics, it is most important to segment units of speech into semantically, syntactically and prosodically coherent units. The exact borders of overlapping speech, for example, are less important than the coherence of segments, and were therefore not marked in the GOS corpus.

For ASR, on the other hand, it is very important to exactly separate the overlapping speech from speech of one individual speaker, to mark segments only where the pause in speech is long enough, and to prefer shorter segments than longer. While concentrating on such aspects of speech, the speech segmentation in the BNSI database when compared to the segmentation in the GOS corpus resulted in very different segment units. An example with overlapping speech is below:

Overlapping speech in the GOS corpus:
segment 1:
spk1: *lejte če lahko a lahko?* 'look if I may may I?'
spk2: *izvolite* 'please do'

Overlapping speech in the BNSI Broadcast News:
segment 1:
spk1: *in predvsem o strokovnih njihovih* 'and especially about their expert'
segment 2:
spk1: *pogledih , ki so zelo važni .* 'opinions which are very important'
spk2: *seveda . ampak , ampak ...* 'of course . but , but ...'

What we mark as two sections of speech in the BNSI would be just one single section of speech according to the GOS transcription standard.

We envisage the future work following the direction of the GOS corpus segmentation. It is very important for the corpus as a linguistic, not merely acoustic resource to

provide the basic transcription units, i.e., segments, as close to what we know as the clause and sentence unit in written texts as possible. However, the needs of keeping segments short where possible and to consider pauses in speech as the most appropriate potential place for the segment borders should also be applied in the guidelines.

4.3 Acoustic annotation

Annotation of acoustic events in the GOS corpus is very limited: we annotate only laughter, untranscribable speech sounds and pragmatically important sounds from outside world, like phone ringing. The BNSI, on the other hand, uses a very broad scope of different acoustic events wherever present, using the inbuilt Transcriber's list of events, since it is very important for acoustic training to carefully separate all acoustic events from the speech. Any kind of noise, even though very slight, such as are part of respiration, are carefully annotated in the BNSI. The background sounds, such as music, speech of other people, noises..., are carefully defined with time stamps.

While very careful annotation of each, even the slightest breathing can be a time-consuming task, the future work should nevertheless follow precise annotation of acoustic events and background noises in order to satisfy the needs of ASR. However, the scope of different events should be limited and generalized as much as possible.

4.4 Speech transcription

The BNSI speech transcription is orthographic standard transcription (i.e., the same as in written texts). However, as Slovene speech is known for its many reductions, annotations were added where the pronunciations deviated from the standard ones, which can be very common in spontaneous speech. Example:
ne bi glasovali 'we would not vote'<Event desc="*****" type="pronounce" extent="previous"/>

An asterisk sign, added to the word *glasovali* 'vote', means that this word was not pronounced as standard g l a – s O – v /a: – l i,[1] but reduced, in this case g l a s – v /a: – l i. Such transcription system was applicable to the very formal speech in media information shows, covered in the BNSI database. However, when we transcribe the informal speech of all the different dialects (Slovene is a language known for a great variety of dialects) in casual conversation, as was the case in the GOS corpus, the standardized transcription becomes a very challenging task. An example is given in Table 1.

Utterance	Standard pron.	Actual pron.
mislim 'I mean'	m /i: – s l i m	m /i: – s m
tako 'so'	t a – k /o:	t k /o:
koliko 'how much'	k /o: – l i – k O	k /o: k
si 'did you'	s i	s
rabil 'need'	r /a: – b i U	r /a: – b u

Table 1: Example of GOS speech transcriptions.

[1] All phonetic trancriptions in this paper follow the Slovene MRPA, based on SAMPA (Zemljak et al., 2002).

In this utterance, each word is reduced. Such vocal reduction is a very spread phenomenon in speech of many Slovene regions and marking words with asterisk makes no meaning as half of the words or more would be marked.

Additionally, problems in defining the standard transcription are in some cases challenging. For example, dialectical words appear in the data that do not have a parallel in standard language, e.g., *spedenan* 'tidy' (standard *urejen*). Such non-standard words may have diverse pronunciations in different regions: s p E – d /e: – n a n vs. s p E – d /i: – n a n.

In other cases, it is a challenging task to define whether the pronounced form is just a reduction or should be considered as a new word, e.g., the regionally specific word *ka* is used in positions where the standard language uses such diverse set of words as *kaj* 'what', *ker* 'because', *da* 'that', *ki* 'which', *ko* 'when', *kar* 'what'. As such examples are not known in advance, but should be discussed along with the transcription process, it is extremely hard to keep the transcription uniform with many transcribers.

For all these reasons, the two-level transcription system was developed for the GOS corpus: the basic transcription was pronunciation-based, following the realized acoustic forms of words as faithfully as possible, i.e., approaching the phonetic transcription, however, it was done with the characters of the Slovene orthographic system, not indicating neither the stress position, nor the length or quality of the vowel. For the sentence in Table 1, such transcription is: *mism tko kok s rabu*. The standardized transcription was added to this basic transcription by an expert. The pronunciation-based transcription represents a valuable starting-point for phonetic transcription, while the standardized transcription is a basis for lemmatization, POS annotation, parsing, language modeling, etc. We estimate that it takes on general additional 10 minutes of work for 1 minute of recording to make a double transcription instead of a single standard transcription. The extra effort needed pays off by more uniform standard transcription, more evident and enriched representation of spoken language for linguistic users and saves a lot of effort to phonetic transcription.

Such a solution is absolutely applicable to the needs of ASR, and we have found it to be a very efficient for transcribing Slovene speech in the general domain.

4.5 Types of speech

One of the main goals of the GOS reference speech corpus was to collect prevailingly spontaneous speech. Written-to-be-read speech was systematically avoided, therefore the corpus can be considered as a corpus of spontaneous speech.

The BNSI database, on the other hand, consists of news shows, where written-to-be-read speech is frequently present (69.26%) and only some parts of news shows contain spontaneous speech (30.74%). The higher part of spontaneous speech is more challenging for ASR development since it contains linguistic characteristics (restarts, duplicated words, fillers, etc.) that are absent in the newspaper text, which is usually used as a main resource for language modeling.

The speakers' demographic criteria in the BNSI database and the GOS corpus also differ, due to the different corpora development strategies. While the timelines and

availabilities of news shows in the broadcaster's archive had a very important impact on selection of data for the BNSI database, the corpus GOS data was carefully balanced also according to demographic criteria. Therefore in private discourse section of the GOS corpus, the speakers are balanced according to gender, age, education, region, first language and country of residence. In the BNSI database we have no such balance, for example the number of male speakers exceeds the number of female speakers.

The types of speech to be covered in the future should be defined in a way to balance the needs of linguistics for the representative data, i.e., covering all types of discourse and all channels of spoken communication, and the needs of ASR for the data from the domains where the technology seeks its main applicability. Among these are most certainly all media domains as well as lectures and classes.

5. GOS as a speech recognition resource

A language resource applied for speech recognition development, needs to be organized into three subsets: training, development and evaluation. The training set is used for developing acoustic and language modes, whilst the evaluation set is needed for testing the accuracy and performance. The development subset is needed to accordingly optimize the speech recogniser's parameters, such as beam pruning, language model interpolation weights, word insertion penalties, etc.

The part of the GOS corpus with public discourse has 68 hours of recorded and transcribed speech. The size of the ASR development and evaluation subset was set to 3 hours each. The subset's recordings were manually selected from the GOS corpus with the goal to diversify the recordings time span and to prevent the overlap between subsets. Only the recordings from broadcast shows (TV and radio) were used for the development and evaluation subset, as they are acoustically compatible with the BNSI database. The GOS ASR development subset has 6 different recordings from years 2008 and 2009, and the GOS ASR evaluation subset has 5 different recordings from the same two years. The remaining recordings can be used for training the acoustic models in a solo mode or in combination with any other suitable Slovene spoken resource. The comparison of GOS ASR and BNSI evaluation subset characteristics in the role of a speech recognition resource is given in Table 2.

	GOS ASR eval	**BNSI eval**
Number of words	27900	22744
Vocabulary size	5424	7501
OOV rate(%)	9.32	4.22
Number of speakers	44	215

Table 2: GOS ASR and BNSI evaluation subsets characteristics.

The GOS ASR evaluation subset has smaller vocabulary size than the BNSI eval, but the out-of-vocabulary rate, calculated on a typical 64k words Slovene broadcast news speech recognition system (Maučec, et al. 2013) vocabulary, is significantly higher. This is probably caused by the spontaneous and entertaining type of discourse, which is present in the GOS ASR subset. The main difference between GOS and BNSI from the speech recogniser's point of view is that GOS comprises only those parts of broadcasts, where spontaneous speech was present. This is also reflected in the smaller number of various speakers in the GOS ASR subset. Thus, we can expect significantly lower speech recogniser's word accuracy on the GOS ASR evaluation subset. Accordingly to some preliminary experiments, the decrease of speech recognition accuracy for Slovene spontaneous speech can be greater than 15% (Maučec et al. 2013).

6. Conclusions

As a result of all presented characteristics, the GOS speech corpus could be used as a valuable resource for improved acoustic modeling of ASR system, since it extends the BNSI database. The GOS corpus public discourse audio files are available on the basis of an individual license. Thus, these GOS development and evaluation subsets can be used in future as the first openly-available resources for benchmarking Slovene large vocabulary continuous speech recognition systems and algorithms. In future databases it seems reasonable to perform certain adjustments on the levels of segmentation, acoustic annotation and types of speech to be covered.

7. Acknowledgements

This research work was partially funded by the Slovene agency ARRS under the contract number P2-0069, and the European Social Fund and the Slovene Ministry of Education, Science, Culture and Sport.

8. References

Barras, C., Geoffrois, E., Wu, Z., Liberman, M. (2000). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33/1-2, 5–22.

Burnard, L., Bauman, S. (2014). P5: Guidelines for electronic text encoding and interchange: 8 Transcriptions of speech. TEI—text encoding iniciative. URL:http://www.tei-c.org/Vault/P5/1.7.0/doc/tei-p5-doc/en/html/TS.html.

EAGLES. (1996). Preliminary recommendations on spoken texts. EAGLES Document EAGTCWG-STP/P.

Erjavec, T. (2014). TEI Schema for GOS speech corpus of Slovene. URL:http://nl.ijs.si/ssj/gos/schema/tei_gos_doc.pdf.

Maučec, M.S., Kačič, Z., Žgank, A. (2013). Speech recognition for interaction with a robot in noisy environment. *Electrical Review*, 5/2013, pp. 162-166.

Verdonik, D., et al., (2013). Compilation, transcription and usage of a reference speech corpus : the case of the Slovene corpus GOS. *Language resources and evaluation*, doi: 10.1007/s10579-013-9216-5.

Zemljak Jontes, M., Kačič, Z., Dobrišek, S., Žganec Gros, J., Weiss, P. (2002). Računalniški simbolni fonetični zapis slovenskega govora. *Slavistična revija,* 50/2, 159-169.

Žgank, A., et al (2005). BNSI Slovenian broadcast news database - speech and text corpus. *Interspeech 2005*, September, 4-8, Lisbon, Portugal.