CREATION OF A DIGITAL CORPUS OF BULGARIAN DIALECTS

Nikola Ikonomov, Milena Dobreva

Abstract: The paper presents our considerations related to the creation of a digital corpus of Bulgarian dialects.

The dialectological archive of Bulgarian language consists of more than 250 audio tapes. All tapes were recorded between 1955 and 1965 in the course of regular dialectological expeditions throughout the country. The records typically contain interviews with inhabitants of small villages in Bulgaria. The topics covered are usually related to such issues as birth, everyday life, marriage, family relationship, death, etc. Only a few tapes contain folk songs from different regions of the country.

Taking into account the progressive deterioration of the magnetic media and the realistic prospects of data loss, the Institute for Bulgarian Language at the Academy of Sciences launched in 1997 a project aiming at restoration and digital preservation of the dialectological archive. Within the framework of this project more than the half of the records was digitized, de-noised and stored on digital recording media. Since then restoration and digitization activities are done in the Institute on a regular basis. As a result a large collection of sound files has been gathered.

Our further efforts are aimed at the creation of a digital corpus of Bulgarian dialects, which will be made available for phonological and linguistic research. Such corpora typically include besides the sound files two basic elements: a transcription, aligned with the sound file, and a set of standardized metadata that defines the corpus. In our work we will present considerations on how these tasks could be realized in the case of the corpus of Bulgarian dialects. Our suggestions will be based on a comparative analysis of existing methods and techniques to build such corpora, and by selecting the ones that fit closer to the particular needs. Our experience can be used in similar institutions storing folklore archives, history related spoken records etc.

Keywords: phonology, corpus, corpus linguistics, audio archive, digitization, restoration, metadata, alignment, transcription, phonetics

Definition of a Corpus

In the broad sense *corpus* means a collection of data, either written texts or a transcription of recorded speech which can be used for linguistic description or language studies. According to this definition we still cannot speak about a digital corpus of Bulgarian dialects, because all we have on hand at the moment are… a pile of magnetic tapes and numerous audio files stored on digital media, which represent the dialectological archive of the Institute for Bulgarian language.

The Dialectological Archive

The dialectological archive of the Bulgarian language consists of over 250 audio tapes. All tapes were recorded between 1955 and 1965 in the course of regular dialectological expeditions throughout the country. Interviewers are dialectology researchers. The records typically contain interviews with aged people from small villages in Bulgaria. The topics covered are usually related to such issues as birth, everyday life, marriage, family relationships, death, etc. Only a few tapes contain folk songs from different region of the country.

The First Project

In 1997 the Institute for Bulgarian Language started a 2-years project with the support of the British council the following basic aims:

- to secure the further preservation of the audio tapes;
- to start digitization of the records and their storage on a digital recording media

In order to secure the preservation of the audio archive we re-considered following basic issues, which are relevant when handling and storing sound recordings:

- that they be kept free of any foreign matter deposits;
- that they be kept free of any pressure that might cause deformations; and
- that they be stored in a stable, controlled environment.

Till 1997 none of these requirements has been met. The results of the inspection have forced us to take urgent measures to ensure suitable storage conditions for the tapes. At the same time we started to digitize the archive records and to store the digital content on CD's. In the framework of the Project more than 30% of the records have been digitized.

The workflow included the following steps:

- Digitization (sampling frequency of 44.1 KHz,16 Bits, Stereo, using a professional sound card equipped with a high end ADC);
- Digital restoration (elimination of most frequently encountered disturbances: impulsive disturbances, wideband noise, and harmonic disturbances);
- Recording on a CD-R.

The current situation

Since 1999 digitization activities have been done in the Institute for Bulgarian Language on a regular basis Recently we changed the output format of the digital records from "wav" to "mp3" in order to save storage space. We also changed the recording media from CD-R to DVD for the same reasons. Due to the lack of financing the number of non-digitized tapes is still considerable (about 40%). The digitized records are not published electronically. Doing dialectological research under such circumstances is not easier than it was in the 50's.

What to do?

The solution is obvious – to create a digital dialectological corpus and make it available to the research community in a variety of formats:

- digitized sound (partly available);
- standard orthographic transcription;
- phonetic transcription;
- various levels of tagged text, all aligned.

Who will benefit from such an endeavor? First of all this will be the scientific community especially in such branches as:

- Arts and Humanities (cultural theory; history/geography and gender studies, linguistics, corpus linguistics, historical linguistics, speech recognition, text synthesis; dialectology);

- Sociology, social history and sociolinguistic research;
- Ethnography and cultural studies.

On the other hand, the experience which will be acquired throughout the project will serve the needs of various institutions with similar audio archives – folklore archives, history related records, etc. Last but not least the wide access to the data within the dialectological corpus will provide valuable information for lay persons, especially members of the local communities.

Coding and coding standards

According to the basic requirements each speech corpus designed for phonological research must as a minimum consist of the following:

- A sound file;
- An orthographic transcription aligned with the sound file;
- A set of standardized metadata that defines the corpus.

Sound files

For the completeness of the corpus all available tapes in the archive have to be further digitized and restored. It must be taken into account that the restoration and processing of the digital audio files are aimed only at making their content available; hence the playback quality is not a relevant parameter. In other words, cost reasons will specify the depth of the restoration efforts.

The transcriptions

Transcription is the conversion into written form, of a spoken language source. There are different types of transcription:

- Orthographic transcription, which is done according to the basic orthographic rules of a corresponding language;
- Phonetic or phonemic transcriptions, which is the process of matching the sounds of human speech to special written symbols (IPA and its ASCII equivalent, SAMPA for example) using a set of exact rules, so that these sounds can be reproduced later. Phonetic transcriptions present three well known problems. They are hugely time-consuming and subjective in the sense that different transcribers typically produce different representations for a given speech segment. As the size of the corpus grows, so does the difficulty of maintaining consistency of practice across the transcription.

For cost as well as reliability reasons, the basic transcription of the sound file must be orthographic. But even if orthographic transcriptions are less costly and more reliable, defining standards for consistent transcription of speech by means of standard orthography is not trivial, and must be addressed. Transcription and sound must be aligned, so that the sound corresponding to a specific part of the transcription can be easily accessed.

Depending on the goals of a specific project, other types of transcriptions, such as phonetic or phonemic transcription, may be added, but they should not supplant the orthographic transcription. Different projects will

have different needs for phonological tiers, depending on different kinds of use. The number of tiers is in principle limited, but a recommended list of relevant tiers might be useful.

A basic problem with all transcriptions is that they are products of interpretation. The result is that people do not trust each others transcriptions. Within a more long term perspective, the possibility of automatic transcriptions, which will make transcriptions at least more objective, (but not necessarily more correct), should be investigated.

The Metadata

Researchers need standards for coding of metadata in order to be able to work on each other's databases. Two basic questions have to be answered:

- What are the relevant metadata?
- How should they be coded?

In other words a specification of the relevant metadata is needed before we can decide how to code them. Here the question arises whether it is possible to define a set of metadata that is relevant for all projects, and whether project specific need to code additional metadata should be catered for by means of a set of general guidelines. Our research has shown that the IMDI[1] (ISLE[2] Meta Data Initiative) already offers a standard for different kinds of metadata.

How should metadata be coded? Up to now it has been a standard practice for corpus creators to define their own representational standards. The drawback of such an approach is that they are not easily portable. In response various standards have been proposed. The currently dominant one is the Text Encoding Initiative (TEI). However there are two basic problems with TEI for representation of phonetic /phonological corpora:

- The TEI recommendation for linguistics corpora is vestigial, and needs to be further developed if it is to be useful for any but the most basic representations.
- The overabundance of XML tags makes TEI-encoded corpora difficult to use directly, and requires development of XML-based analytical applications. Few of these exist currently.

Pending their appearance, we have accepted TEI as an archiving standard. We expect that TEI will be supplemented by provision of XSLT[3], tools which translate TEI representation into formats usable by existing non-XML-aware applications like relational databases.

As to the coding itself, XML should be recommended. It is flexible, and allows users to define their own tags. An important question is whether only standards for coding metadata should be recommended, or whether the coding standards should be extended to the linguistic content as well. The latter position implies that tags will reflect theoretical positions.

Digital Corpus of Bulgarian Dialects (DCBD)

The content of the Digital Corpus of Bulgarian Dialects (DCBD) will be provided in several types of representation:

[1] http://www.mpi.nl/IMDI/

[2] ISLE stands for International Standard for Language Engineering

[3] eXtensible Stylesheet Language Transformations

- Audio (partly available);
- orthographic transcription;
- part-of-speech tagged orthographic transcription;
- phonetic transcription.

The orthographic transcription of DCBD will contain a complete orthographic transcription of the audio recordings. The transcription process will consist of several (up to four) passes through the audio files. The first pass will produce a base text. The next passes (usually the second and the third) are correction passes aimed at improving the transcription accuracy. The last pass will be used for establishing uniformity of the transcription algorithm across the entire corpus. To avoid pre-judging discourse structure, capitalization and punctuation will not be used in the transcription. As a general principle, the DCBD will use the Standard Bulgarian orthography. In genuinely dialectal segments, it will use the Bulgarian dialect dictionary (in preparation).

The part-of-speech tagged transcription is a morphological - syntactic annotation. It represents the basic linguistic analysis. It will be done automatically using software tools called taggers. The tagger for Bulgarian texts is called "GrammLab"and is distributed freely by BACL (Bulgarian Association of Computer Linguistics).

The phonetic transcription is in fact discretization of the analog speech signal into phonetic segment sequences. DBCD will contain phonetic transcriptions of all the interviews. The process will include following basic steps:

- Selection of transcription scheme, that is, a set of symbols each of which represents a single phonetic segment (for example IPA)
- Partition of the linguistically-relevant parts of the analog audio stream such that each partition is assigned a phonetic symbol.
- The result will be a set of symbol strings each of which will represent the corresponding interview phonetically. These strings can then be compared and processed.

The usefulness of the DBCD would be enhanced by provision of an alignment mechanism to relate the representational types to one another, so that corresponding segments in the various types can be conveniently identified and simultaneously displayed. The first task in this process is the necessity to define how large the alignment segments should be - phonetic segment by phonetic segment, word-by-word, sentence by sentence, or utterance by utterance? The answer has to take into account two basic factors: the research utility, and the feasibility in terms of cost.

All interviews consist of a sequence of "interviewer-question, interviewee-answer" pairs in which the utterance boundaries are generally clear-cut; rarely there is some degree of overlap on account of interruption and third-party intervention. The format of the interviews makes alignment at the granularity of utterance the natural choice.

In practice the alignment process has to take into account, that time is a meaningful parameter only for the audio level of representation in the corpus, and that text has no temporal dimension.

A time interval t is selected, and the audio level is partitioned into some number n of length-t audio segments s, $s(t \times 1), s(t \times 2)...s(t \times n)$, '$\times$' denotes multiplication.

Corresponding markers are inserted into the other levels of representation such that they demarcate substrings corresponding to the audio segments. In XML such marker could be the `<anchor>` tag), where, the `'id'` attribute will specify a real-time offset from the start of the audio file.

Future cooperation

The future cooperation in the field will be achieved in the frames of the network *European Corpus Phonology Group (CorPho)*. It will assemble researchers and research teams interested in combining insights from theoretical phonology, both diachronic and synchronic, linguistic variation studies, phonetics, and corpus linguistics.

Bibliography

[Clua, 2006] Clua Esteve; Lloret, Maria-Rosa "New tendencies in geographical dialectology: The Catalan Corpus Oral Dialectal (COD)". In: Montreuil, Jean-Pierre (ed), *New Perspectives on Romance Linguistics. Vol. 2: Phonetics, phonology, and dialectology.* Amsterdam / Philadelphia: John Benjamins. (Available at: http://www.uv.es/foncat.)

[Ihalainen, 1990] Ihalainen, Ossi "A source of data for the study of English dialect syntax: The Helsinki Corpus." In Aarts, Jan & Willem Meijs (eds) *Theory and Practice in Corpus Linguistics.* Amsterdam: Rodopi. 83-103.

[Lloret, 2002] Lloret Maria-Rosa; Perea, M. Pilar "A Report on Corpus Oral Dialectal del Català Actual (COD)". *Dialectologia et Geolinguistica* 10: 1-18.

[Meurman-Solin, 2001] Meurman-Solin, Anneli. "Structured Text Corpora in the Study of Language Variation and Change", *Literary and Linguistic Computing*, Vol. 16, No. 1, 5-27.

[Moisl, 2005] Moisl H., Jones V., "Cluster analysis of the Newcastle Electronic Corpus of Tyneside English: a comparison of methods", Literary and Linguistic Computing 20, 125-46.

[Moisl, 2006] Moisl H., Maguire W, Allen W., "Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English". In: F. Hinskens, ed. Language Variation. European Perspectives. Amsterdam: Meertens Institute.

Web resources

[NECTE] A linguistic time-capsule: the Newcastle electronic corpus of Tyneside English (available on http://www.ncl.ac.uk/necte/).

Authors' Information

Nikola Ikonomov– Head of Laboratory on Phonetics and Speech Communication, Institute for Bulgarian Language, BAS, Shipchenski prohod 52, Sofia-1113, Bulgaria, Institute for Mathematics and Informatics, e-mail: nikonomov@ibl.bas.bg.

Milena Dobreva – Head of Dept. on Digitization of Scientific Heritage, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev St., bl. 8, Sofia-1113, Bulgaria, e-mail: dobreva@math.bas.bg