# Models of Russian Text/Speech Interactive Databases for Supporting of Scientific, Practical and Cultural Researches

### Pavel Skrelin and Tatiana Sherstinova

Department of Phonetics, St. Petersburg State University Universitetskaja nab. 11, 199034, Russia paul@phonet.lang.pu.ru; tanya@ts4306.spb.edu

#### **Abstract**

The paper briefly describes the following databases: "Online Sound Archives from St. Petersburg Collections", "Regional Variants of the Russian Speech", and "Multimedia Dictionaries of the minor Languages of Russia", the principle feature of which is the built-in support for scientific, practical and cultural researches. Though these databases are addressed to researchers engaged mainly in Spoken Language Processing and because of that their main object is Sound, proposed database ideology and general approach to text/speech data representation and access may be further used for elaboration of various language resources containing text, audio and video data. Such approach requests for special representation of the database material. Thus, all text and sound files should be accompanied by information on their multi-level segmentation, which should allow the user to extract and analyze any segment of text or speech. Each significant segment of the database should be perceived as a potential object of investigation and should be supplied by tables of descriptive parameters, mirroring its various characteristics. The list of these parameters for all potential objects is open for further possible extension.

# 1. Introduction. Speech corpora and database requirements

The following speech corpora are currently being created and maintained by the Departments of Phonetics of St.Petersburg State University:

- 1. **Phonetic Fund of the Russian language.** Standard Russian speech corpus is used for description and analysis of modern Russian speech. It is intended for general phonetic analysis, teaching of phonetics and Russian speech technologies.
- 2. **Russian Speech Variability corpus** consists of records of the phonetically representative text, recorded by hundreds of speakers from different regions of Russian Federation and the Commonwealth of Independent States. It is designed primarily for research of regional phonetic variability of the Russian speech and forensic phonetics.
- 3. **Russian speech by foreigners.** The corpus borrows the structure and ideology from the Standard Russian speech corpora. It is intended for study of language interference and improvement of teaching methodologies on Russian pronunciation.
- 4. Phonetic Funds and Multimedia Dictionaries/ Lexicons of the minor Languages of Russia. Phonetic Fund of Buryat language is being created. A program to create a serial of multimedia lexicons of the endangering languages of Russia has been started. Nenets lexicon is made, and Nganasan lexicon is being prepared.
- 5. **Sound Collections of St. Petersburg Archives** (Sound Archive of the Institute for Russian Literature of the Russian Academy of Sciences the Pushkinsky Dom, and audio collections of different departments of St.Petersburg State University) embrace the recordings of speech and folklore of the peoples of the Russian Federation. Since we are phoneticians, our own professional objects of investigation are 1) Russian speech, 2) its variation inside and outside Russia, and 3) speech of other languages of the

Russian Federation. Nevertheless our speech corpora may be of considerable value to specialists of other scientific, social, cultural and applied disciplines (e.g., general linguists, ethnologists, musicologists, sociologists, folklorists, forensic phonetician and others). Therefore we see our task in elaborating the model of an interactive database, basing on material of our speech corpora and containing both sound recordings and correspondent multi-aspect descriptions, which may meet the requirements of specialists belonging to these quite different disciplines. That is to create a system that will allow the users of different scientific and cultural fields to perform their investigations of the recorded material both on textual and sound levels directly in the database interface.

Because of that a special representation of database material is necessary. Thus, all text and sound files should be accompanied by information on their multi-level segmentation, which should allow the user to extract and analyze any part (segment) of text or speech. Each significant segment of the database should be perceived as a potential object of investigation and should be supplied by tables of descriptive parameters, mirroring its various characteristics. The list of these parameters for each of the objects in concern should be open for further possible extension.

In addition to traditional off-line database interface we are going to elaborate similar on-line system, thus making database accessible in the World Wide Web, at least in its restricted abilities.

# 2. Standard Speech Database for the Phonetic Fund of Russian Language

Our first attempt to create phonetic database for linguistic research took place while elaborating Russian Standard Speech database for the Phonetic Fund of the Russian Language (Bondarko et al., 1992). Database consists of corpus of sound files and their descriptions. Recordings of the Phonetically Representative Texts pronounced

by standard Russian speakers form database sound material. The detailed description was made for each syllable of the database. The structure of descriptions was specially developed by the phoneticians so that to present in details the real characteristics of Russian speech. It comprises in particular the "real" transcription, "ideal" phonemic and phonetic transcriptions, attributes of the sound, duration of signal segments, acoustic features, and phonetic comments. The peculiarity of Standard Speech Database consists in detailed phonetic description of all syllables. Database may be used in phonetic and related studies, for building of speech algorithms, and evaluation of speech resources, speech technologies and products (Sherstinova, 1997; Kuznetsov & Sherstinova, 1999).

# 3. Multimedia Dictionaries of the minor Languages of Russia

Modern computer technologies place at researches' disposal the proper instruments for preserving and studying small and endangering languages. We started our work in this direction with creation of Nenets Digital Multimedia Lexicon, which is made in cooperation with Saint-Petersburg Institute of Linguistic Researches of the Russian Academy of Science. Nenets is the small people living along Arctic Ocean between Kola Peninsula and the mouth of river Yenisej. In spite of huge territory, original population counted to 35000 persons in 1989. Nenets is the native language for 77,1% of them. Comparing to other small Northern languages of Russia this one is a rather safe one. The main aim of Multimedia Nenets Lexicon is to preserve the specialties of Nenets phonetics and to allow the researcher to conduct the phonetic investigation of his own.

Nenets Lexicon was elaborated in the form of the relational database; audio data are stored in raw-format. The structure of Lexicon is based on the Nenets-Russian Dictionary composed by N.M.Tereshchenko where from it borrows the completeness of translation, text examples, linguistic remarks and comments. The number of words in the core database is currently limited to 4000. Variants of pronunciation are given for every word likewise in Jurak-Samojedisches Worterbuch, composed by T.Lehtisalo. Along with recordings of standard word pronunciation, another dialect variants of pronunciation may be presented, and additional phonetic information (syllables number, accent) is added.

The "Phonetic form" of the database currently includes:
1) sound button for standard pronunciation of the word;
2) speaker's name; 3) transcription of the word; 4) syllable number; 5) syllable accent; 5) word accent information; 6) sound buttons for another variants of pronunciation and speakers names. Nenets Multimedia Lexicon may be used to entirely present and study Nenets phonetics and grammar. It may be enlarged by other speakers' recordings, by new words, by new phrase examples. We are going to enrich the Lexicon by recordings of texts of the different types. The interactivity of system lies in user's possibility to expand the database by her/his own transcription and commentaries. Our Computer Lexicon is the first attempt to preserve and reflect the peculiarities of Nenets actual

speech. We consider it to be a model for future description of other minority languages of Russia (Ljublinskaja & Sherstinova, 1999).

Thus, our next project concerns the creation of similar system for Nganasan (former Tawgy - Samojed) language, belonging to Northern-Samojedic subgroup of Samojedic Group of Uralic languages Family. Nganasans are the most Northern people in Eurasia living on Tajmyr peninsula in Russia. The number of Nganasans in 1989 year was 1278 persons according to Total Russian Census. Because of that it is extremely important to record their present pronouncing.

# 4. Online Sound Archives from St.Petersburg Collections

Since 1998 the project on cataloguing and webrepresentation of speech and folklore sound collections of St. Petersburg is carried out in the Department of Phonetics of St.Petersburg State University. Our task here is to preserve the unique audio archives by restoration the old recordings, transforming analogue recordings into digital form (from tapes into CD-ROMs), creation of complete electronic catalogues on archive materials, development of web-sites dedicated to correspondent sound collections and including search system on their funds, and web-publication of the most interesting sound recordings or their fragments. The regional and national importance of this project consists in preservation of the national heritage of Russia, unification of sound archives of St. Petersburg in one uniform regional system, standardization of methods and concepts of sound database Internet representation, and the development of proper user interface.

Currently the following collections are the subjects of cataloguing: Sound Archive of the Institute for Russian Literature of the Russian Academy of Sciences (the Pushkinsky Dom) and two collections of St.Petersburg State university (archive of the Department of Phonetics and collection of the Folklore Study). All sound materials are of unique character: they represent archival records of national folklore, dialects and small languages of Russia, earliest of which are dated the beginning of leaving century.

The Web site "Sound Archives from St.Petersburg Collections" may be found at: http://www.speech.nw.ru/phonetics/homepage.html

Sound Archive of the Institute for Russian Literature is the largest folklore archive in Russia, whose recordings number 100000 hours of sounding. General information on its archive collections is given at: http://www.pushkinhouse.spb.ru/structure/unit11.shtml

On the basis of Sound Archive card index the computer database is created, which is further converted into HTML-format for Internet representation. For web-publication the samples of records belonging to different genres, dialects and languages, which are interesting from either phonetic, linguistic or folklore point of view, are selected. Further they are deciphered, segmented onto phrases, syntagmes and words, transcribed and exposed to the expert phonetic analysis. The results of this analysis as well as recordings themselves become accessible to the network users by means of online interface.

The homepage of the site contains four basic hyperlinks: 1) technical requirements; 2) general information on the project; 3) textual catalogue of sound recordings; 4) acoustic database of collection samples and their fragments.

Textual catalogue of sound recordings represents the table derived from relational database supported by Microsoft Access 97/2000, which is filled by the employees of Pushkinsky Dom basing on archive card index. Lines (records) of the table contain description of concrete sound recordings, while fields reflect parameters of description.

The main description parameters are the following: 1) number of a collection; 2) serial number; 3) archival code; 4) record title; 5) initial line of record; 6) genre; 7) quantity of speakers/performers; 8) principle performer (up to 12 names); 9) date of record; 10) place of record (city, village); 11) geographic region; 12) nationality of speakers/performers; 13) language; 14) quality of record; 15) mono/stereo record; 16) speed of record; 17) general comment. Internet-catalogue at the given moment contains the following fields: 1) serial number of record in query results (N); 2) collection number (HK); 3) archival code; 4) record title; 5) initial line of record; 6) genre; 7) quality of record; 8) principle performer (the first one); 9) date of record; 10) place of record (city, village); 11) geographic region; 12) comment.

Textual catalogue of sound recordings on the first page contains the list of collections and their sections accessible via Internet. The description of collections N 002, 004, 005, 009 (records executed in the Arkhangelsk and Leningrad regions of Russia) is now completed. By means of clicking on correspondent hyperlink the user evokes the collection search system. Nowadays catalogue search system in online mode provides search on genres (prose, epos, lamentation, song, calendar song), place of record (region) and on a symbolical line, which should be included into the name of sound recording or its initial line.

#### Acoustic database

To get access to acoustic database of recording samples and their fragments it is necessary that the user fills in on-line registration form or enters his personal password. Thereafter the page similar to basic textual catalogue is opened. As distinct from the textual catalogue, each line of the table presented here has a hyperlink marked by small CD-ROM icon, which allows user's access to correspondent audio files and to the whole list of documents accompanying this particular recording.

The appearance of acoustic database pages is rather traditional for the Internet: they consist of two vertical frames: auxiliary left frame carries out functions of the menu, while the right one contains the actual information. By default here appears the orthography of recording. The menu allows to receive recordings transcription (either in symbol codes or as a graphic file).

Sound recordings may be listened to in the following modes: the whole recording, by selected syntagmes or by selected isolated words. Last two modes provide the most dynamical work with the material. For listening the user specifies in the menu a required mode of reproduction, whereupon in the principle frame appears orthography of

the text with the appropriate hyperlink breakdown. The user clicks the interesting fragment (word, syntagme, etc.), and the correspondent sound file is generated, downloading and listened by means of browser (or default) audio-player.

Particular item of the menu allows to open "phonetic commentary", in which pronunciation features of recording are stated and analyzed. Besides that, general tendencies, characteristic of language, regional dialect, genre or temporary period may be described. On the basis of this material research of coexistence and interference of Russian, Nenets and Komi languages in territory of the Arkhangelsk region of Russia has already began (Bondarko et al., 1999).

### 5. Regional Variants of the Russian Speech

Our new project - the online database "Regional variants of the Russian speech recordings" - aims at developing of the interactive expert/scientific database for online information retrieval and multidimensional investigations. Database is designed primarily for research of regional phonetic variability of the Russian speech. Its sound corpus consists of recordings of the phonetically representative text, recorded by hundreds of speakers from different regions of Russian Federation and the Commonwealth of Independent States. The collection of this corpus had been performed by researches of the Department of Phonetics during last 30 years. The recorded speech is being segmented into words, syntagmes and sentences. All information about segmentation is automatically processed and converted into auxiliary database, which is to maintain speech retrieval on the user's request (Skrelin & Sherstinova, 1999).

Special tables of phonetic peculiarities for each regional pronunciation variant were created by experts/phoneticians after the detailed analysis of this speech corpus. This phonetic information is also presented in the database form, allowing dynamic on-line access on user's request. Each level of segmentation reveals its own phonetic peculiarities, describing particular regional pronunciation (e.g., phonetic realization of the word or individual melody of the sentence). Database interface organizes this information as the hypertextual tables providing audio access to the corresponding sound samples (words, syntagmes, sentences).

The logical components of the database are the following: speech material, information on its segmentation, auxiliary database of descriptive parameters and phonetic peculiarities, permanent and dynamic comments and descriptions, and user's interface. Each significant segment of the database (word, phrase, etc.) is perceived as a potential object of investigation and is supplied by tables of descriptive parameters, mirroring its various characteristics. The list of these parameters for each of the objects in concern is open for further possible extension.

The online user observing the orthography of the recorded text may select any its fragment and receive (or listen to) the corresponding audio file, which is automatically generated by the user's selection. If his selection coincides with the object previously described in the database (word, syntagme or sentence) he may receive all information from the database. The user may view the tables of phonetic peculiarities of segments for each regional

pronunciation variant and listen to audio samples. The search engine comprehending all logical component of the database is being created. In the future it is planned to allow all users the possibility to create their own comments and descriptions, which are to be included into general database.

Auxiliary database of segments parameters may be expanded so that to include new aspects of description necessary to support the needs of users, belonging to different scientific and cultural fields (at present beside phoneticians the system is used by folklorists and dialectologists). The results of online investigation become accessible via WWW, and the database itself may be considered as the model of interactive language database for humanitarian, NLP and SLP researches.

The online system is currently in the stage of development. At the present moment technical decision for realization of separate logic database components is determined (speech corpora; auxiliary database of the information on segmentation; database of phonetic attributes; online user interface and the serving CGI-programs) as well as their integration in the united system; the preparation of speech material and development of system modules is made. Web-publication of the site is planned to the end of 2000 (information about URL-address of the site will be announced at: http://www.speech.nw.ru/

### 6. Technical Features

Internet catalogues and on-line systems, as well as majority of accompanying information (texts, comment, descriptions) are now made in Russian, because original catalogues of sound recordings are in Russian and describe for the first part the Russian material. The basic coding of sites is CP1251. The sites use standard Cyrillic fonts with one exception: the installation of a special font Times Trn3 is necessary for transcribing symbols reproduction. This font the user can receive via network and establish on her/his computer. (The alternative possibility is to work with graphic files). The information about it is given on special page of "Technical Requirements".

Samples of recordings are stored now in raw-format. Therefore for their listening through a network the user need a computer with a sound card like SoundBlaster and appropriate browser Plug-in or other system media player. In a near future we plan the conversion of sound material from the raw-format into RealAudio one, which allows streaming audition of sound signals through the web in a real time mode.

#### 7. Access Peculiarities

Textual catalogues of collections are exposed in the Internet with the easy access. As the collections of Sound Archive of the Institute for Russian Literature of the Russian Academy of Sciences (the Pushkinsky Dom) represent the National property of Russia, the access to their sound recordings can be carried out by means of the personal password given personally to the frequent visitors of the site, or after filling the standard user's registration form. The presented samples may be used in scientific and cultural purposes exclusively. The non-authorized copying, duplicating

and any use for commercial purposes are pursued under the

#### 8. Conclusion

The implementation of the described projects aims at decision of scientific, educational, and cultural tasks. Our developments are orientated to linguists, phoneticians, experts in the field of library business, speech technologies, multimedia developers, ethnographers, specialists in folklore, psychologists, sociologists and all users of the world wide web interested by the speech aspect of language. In addition to web-sites, which are currently created ("Sound Archives from St. Petersburg Collections" and "Regional Variants of the Russian Speech"), the preparation for webpublication the other speech corpora and audio collections is planned (in particular, "Phonetic Fund of the Russian language. Standard Russian speech corpora (Moscow and Petersburg standard variants)", "Multimedia Nenets Dictionary", "Fairy Tales of Russian North"; "Ritual Poetry of the Russian North: Lamentations", "Folklore of Volga region Germans", etc.).

Though our databases are addressed to researchers engaged mainly in Spoken Language Processing (in particular: speaker recognition, forensic phonetics, identification of speaker's origin, linguistic and phonetic interference studies) and because of that its main object is Sound (and its parameters), proposed database ideology and general approach to text/speech data representation and access may be further used for elaboration of various language resources containing text, audio and video data. Thus, text, speech and video files may be stored in the database segmented into meaningful units and supplied by corresponding descriptive information, which provides content-based access on user's request.

### 9. Acknowledgements

- 1. Web-site "Sound Archives from St. Petersburg Collection" is created in the framework of international project "Sound Archives on the World Wide Web with Sound Recordings from St. Petersburg collections", supported by INTAS foundation (Work programme 1999-2001, project Nr. 1705).
- 2. "Regional variants of the Russian speech in the Internet" project is supported by the Russian Foundation of Humanitarian Research (project N 99.04-12015v).

#### 10. References

Bondarko L.V., Svetozarova N.D., Skrelin P.A. 1992. Phonetic Fund of the Russian language" - the research program of the Department of Phonetics of Leningrad State University. In: Bulletin of the Phonetic Fund of the Russian Language. Issue 4. (pp. 5–16) Bochum (Germany) - St. Petersburg (Russia).

Bondarko L.V., Skrelin P.A., Volskaja N.B., Sherstinova T.Y., Kastrov A.Y. 1999. Concept and Structure of the Web-site Sound Catalogue of Saint-Petersburg Collections. In: Proceedings of the First All-Russia Scientific Conference "Electronic Libraries: Prospective Methods and Techniques, Electronic Collections" (pp. 216-218)

- "Digital Libraries'99". St. Petersburg State University, October 19-22 1999.
- Kuznetsov, V.I. & Sherstinova, T.Y. 1999. Phonetic Database of the Russian Speech Variability. In: Proceedings of the XIVth International Congress of Phonetic Sciences (pp. 1705–1708). San Francisco, CA. August 1-7, 1999. ICPhS99. Vol. III.
- Ljublinskaja, M.D. & Sherstinova T.Y. 1999. Sounding Dictionary of the Nenets Language. In: Proceedings of the XXVIII High-School Scientific and Methodology Conference of University Teachers and Post-graduate Students. St. Petersburg State University.
- Skrelin, P.A. & Sherstinova, T.Y. 1999. On-line Database "Regional Variants of the Russian Speech Recordings" as the Model of Interactive Database for Philological Researches. In: Proceedings of the Second All-Russia Scientific Conference "Internet and Modern Society" (pp. 55-57). St. Petersburg State University. November 29 - December 3, 1999.
- Sherstinova, T.Y. 1997. Database FON2 of Russian Normative Speech. In: Proceedings of the Second International Workshop "Speech and computer" SPECOM'97 (pp. 81–84) 27-30 October 1997, Cluj-Napoka Romania.