THE AUTOMATIC RUSSIAN TEXT TRANSCRIBER

E.B.Ovcharenko, J.V.Ipatov, S.B.Stepanova

Leningrad State University, USSR

ABSTRACT

This report gives a linguistic description of the program that automatically translates Russian orthographic text into transcription signs. This transcriber provides linguistic material for computer interface and allow complex automization for linguistic research, i.e. dictionary-making, text analyses, scientific apparatus compiling and besides to serve as a reliable base for different educating programs.

1.INTRODUCTION

The automatic transcriber (AT) is hoped an undispensable component of the Phonetic fund. Any language matherial (a dictionary or a text) may be transcribed with the help of the AT sufficiently providing the necessary phopnetic details.

Traditionally transcription is known as a system of signs and rules of using them for recording speech and its' sound structure. The main aim of a linguist when working out a system of automatic transcription is to determine and algorythmize the rules of correspondence between 33 Russian letters and speech sounds.

2.THEORY OF TRANSCRIPTION

Several automatic transcribers had been worked out by this time. All their differences can be reducec to 4 general items.

1) The choice of phonologic "ideology". There are several tendencies in modern soviet phonologic science which it their approach to the definition of the main minimum sound language unit - phoneme. Thus, one considers a phoneme to be a representative of a morpheme which is constant (eg, the final consonant of a word 'город' is /d/ as it is in 'города'); and other (Leningrad or Shcherba's phonological school) - to be an independent unit (eg, in the same word 'город' the final phoneme is /t/).

2) The form of presenting material. It may be detailed or simplified transcription (from general symbols C and V for consonants and vowels or phonemic transcription to the detailed indication of sound qualities).

3) Difference in choosing the object of transcription, i.e. that one may transcribe separate words, sentences or texts.

4) The choice of the pronunciation variant. For Russian there are two different standard variants - Moscow and Leningrad).

3.PHONEME TRANSCRIBER

The authors of the AT - members of the Department of Phonetics of Leningrad state University - tried to develope and improve all these aspects.

There is a phoneme block of the AT, responsible for forming the phoneme record of speech material according to Shcherba's phonologic theory.

In phoneme transcription the following phenomena of the Russian language system are considered.

1) In non-stressed Russian vocalism practacally there are no /o/ and /e/ vowels. Some frequent words coming from foreign languages, where this vowels still remain in non-stressed syllables, are included into a list of exceptions (eg, какао, радио, анданте etc.).

2) Voiced consonants are replaces with voiceless before a pause. (завод - /zavo"t/, мороз - /maro"s/, гроздь - /gro"s't'/ and so on.)

3) Regressive consonant assimilation in feature of voice and its' absence. (трубка - /tru"pka/, сделать - /z'd'e"lat'/ and so on.)

k4) Consonant assimilation in feature of softness and hardness. (шесть - /šе"s't'/, пенсионер - /p'in's'ian'e"r/ and so on.)

5) Consonant assimilation in the place of forming (сшить - /s͡šы"t'/, сжать - /ž͡ža"t'/, заказчик - /zaka"š':ik/, городской - /garacko"j/ and so on.

6) "Non-pronunced consonants" (честный - /č'e"snыj/, поздно - /po"zna/, счастливый - /š':isl'i"vыj/ and so on.)

7) "Double consonants". Formalisation of rules of pronuncing long or short consonants in place of two similar letters is rather complicated and in this case pronunciation dependes on the "double" letter position inside the word, on the neighbour letters and on morpheme borders. Special analyses of all Russian words containing double letter combinations helped to find the rules of their transcription (длинный - /dl'i"nnыj/, but: сделанный - /z'd'e"lanыj/, грамм - /gra"m/, ввод - /vvo"t/ and so on.)

The phoneme transcription is only one of the ways how to present the speech material with the help of the AT.

4.PHONETIC TRANSCRIBER

Any text can be simultaneously presented as a sequence of allo- phones - concrete realizations of phonemes determined by simple rules of correspondence and assimilation of sounds and reduction of unaccented syllables.

Examples:
сд"елка - [z'd'e"lkʌ],
изво"зчик - [izvo"š':ьk],
голова - [gъlʌva"].

There are hundreds of such elements. And at last, the description of speech flood as a sequence of phonetic elements giving quite detailed description of very subtile but important for perception of speech realization differences, eg, modification of vowels after labial consonants /b/, /p/, /v/, /f/: ба"л - [b°a"l], ва"за - [v°a"zʌ]; or after nasal consonants: ма"ма - [mã"mʌ̃], нёс - [n̰õs], different degrees of reduction of non-stressed vowels depending on their position to the stressed syllable,

different symbols for vowels after voiced and voiceless consonants, after soft and hard consonants. While transcribing consonants the following phenomena are conside- red: labial character of consonants before [o] and [u] (стук – [sˇ tˇ u"k], кот – [k° o"t]), appearance of faucal explosion in combina- tions дн, тн, бм, пм (дно – [d͡no"],обман – [ʌb͡mã"n], etc), lateral explosion in combinations тл, дл (подлый –[po"d͡lыj], etc), devoicing of sonorants in some positions (eg театр – [t'ia"tr̥], надсмотрщик– [nʌ̃cmõ"tr̥š':ьk] etc). In AT desinged for speech synthesis are considered changes of consonants at the end of words: affrication or aspi- ration of consonants /p/, /p'/, /t/, /t'/, /k/, /k'/ and nasals' implosiveness /m/, /m'/, /n/, /n'/ and so on.

The set of phonetic elements of this unit corresponds to the set of acoustic elements, which is sufficient for the automatic synthesis of distinctive and naturally sounding Russian speech. That's why the number of these elements is rather large – approximately 700 elements. Still, when presented in a graphic form, all the three types of transcription look compact and usual for phonetists and speech scientists. There are two systems (Roman and Cyrillic alphabets) of graphic representation of phonemes and those of the International Phonetic Alphabet.

The described AT allows to transcribe both separately taken words and sentences, in this case it realizes rules of words' bounds, eg, voicening of voiceless consonants having no pair: мех животного – [m'e" zыvo"tnъv^]; connection of nominative words and relying prepositions and particles into one phonetic word, eg, без огня – [b'izʌgn'̥a"], не знаю – [n'izna"̥u] and so on.

5.ORTHOEPIC STANDARD

The suggested transcriber is oriented to the modern literary Russian pronunciation standard. Two variants of orthoepic standard (Moscow and Leningrad) are generally acknowledged. But nowdays there is a definite tendency to eleminating of differences between variants,"to formation of some common pronunciation standard embracing features of both Moscow and Leningrad variants" (Л.А. Вербицкая, "Русская орфоэпия", 1976, p.115). The AT consideres all recent orthoepic researches.

6.STATISTICAL PROCESSING

Important advantage of the AT is its' ability to get information on statistical processing of the text, on distribution of letters, phonemes and allophones both in digital and graphic form. The program is written in algorythm language C for personal computers based on 8086-compatible processor.

7. SUMMARY

The AT described in the report worked out on the Shcherba's phonological princeples allows to get both phoneme and phonetic text transcription of two degrees of detalization. The AT takes into account sound modification inside a phonetic word and some phenomena taking place at the nominative words connections. The transcription is based on modern Russian language pronunciation standard. The program allows to make statistic processing of the text.