RO-SAM EUROM

Author(s): Babel project

Institute: University “Politehnica” of Timisoara

Address: Vasile Parvan 2, 1900 Timisoara, Romania

Email:boldea@cs.utt.ro

Date: 1997-09-29 (created) 2004-05-10 (updated)

Version: 3

1. INTRODUCTION

This is a small portion of the Romanian speech data built within the framework of the Copernicus project BABEL. The XML encoding of the speech transcription has been achieved within the „Multext-East” Copernicus project. The entire speech corpus can be aquired from ELRA (ELRA-S0170). The ELRA description says: “The BABEL Romanian Database is a speech database that was produced by a research consortium funded by the European Union under the COPERNICUS programme (COPERNICUS Project 1304). The project began in March 1995 and was completed in December 1998. The objective was to create a database of languages of Central and Eastern Europe in parallel to the EUROM1 databases produced by the SAM Project (funded by the ESPRIT programme).

The BABEL consortium included six partners from Central and Eastern Europe (who had the major responsibility of planning and carrying out the recording and labelling) and six from Western Europe (whose role was mainly to advise and in some cases to act as host to BABEL researchers). The five databases collected within the project concern the Bulgarian, Estonian, Hungarian, Polish, and Romanian languages.

The Romanian database consists of the basic "common" set which is:

* The Many Talker Set: 50 males, 50 females; each to read 4 connected passages, 1 block of 2-3 "filler" sentences, 4 phonemically compact sentences, 3-7 individual sentences, and 26 numbers.
* The Few Talker Set: 5 males, 5 females from the Many Talker Set; each to read additionally 3 blocks of syllables and, in 4 supplemental sessions, 16 connected passages, 4 blocks of 2-3 "filler" sentences, 4 repetitions of the 26 numbers.

* The Very Few Talker Set: 1 male, 1 female from the Few Talker Set; each to read additionally 5 pairs of context words and the syllables in these 5 contexts.”

Part of the translation of the BABEL texts from English into Romanian was carried on at RACAI. Below is the description of the sample we contribute to MetaShare. The XML encoding, done within the Multext-East Project (1995-1998) is due to Tomaz Erjavec of Josef Stefan Institute.

SPEECH FILE FORMATS

The audio data is stored in separate WAVE files, which are PCM encoded at 22Khz, 16 bits per sample, mono.

1.1 DIRECTORY STRUCTURE

All files are stored in a single Directory.

1.2 FILE NAMING CONVENTIONS

Each filename starts with the string „spch„ followed by the two digit file number and the string „-ro.wav” (eg. „spch01-ro.wav”). The file number is 0-based.

1.3 LABEL FILES

The file „spch-ro.xml” is an XML file containing the TEI header (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-teiHeader.html) for this resource and separate descriptions for each audio file.
An example of one audio file description is explained below:

```
   <text id="mtes-ro." lang="ro">
     <body lang="ro">
       <div id="sro.1" n="O0" type="block">
         <head>*BLOCK: O0</head>
         <p id="sro.1.2">
<s id="sro.1.2.1"> Sentence 1 </s>
<s id="sro.1.2.2"> Sentence 2 </s>
<s id="sro.1.2.3"> Sentence 3 </s>
<s id="sro.1.2.4"> Sentence 4 </s>
<s id="sro.1.2.5"> Sentence 5 </s>
         </p>

         <ab>[<xref url="spch00-ro.wav">speech file</xref>]</ab>
       </div>
.
.
.
     </body>
   </text>
```

Each tag refers to one audio file and contains the following tags: <head>, <p> and <ab>. The <p> tag has an <s> tag for each sentence in the audio file (each audio file contains 5 sentences). The <ab> tag is the reference to the audio file on the disk (filename).

2. DATA

The database is split into 40 audio files, each file containing a number of 5 sentences, with a total number of 200 sentences.

2.1 PHONETIC TRANSCRIPTION (YES/NO)

No phonetic transcription was provided and the file transcription.txt was automatically generated with our phonetic transcription tool (no text normalization present).

3. SEGMENTATION

3.1 SEGMENTATION TYPE

The audio files are segmented at sentence level.

3.2 TOKEN COUNT

The database is composed of 200 sentences with a total number 2203 words (1026 distinct).
The occurance count of each allophone in the corpus is as follows:

Allophone	Count	Allophone	Count
@	5	l	380
ch	155	m	320
o@	58	n	615

z	102	o	345
e@	120	a@	190
p	337	a	1491
r	707	b	92
s	604	d	334
t	770	e	1013
u	516	f	117
v	143	g	89
w	43	je	5
h	14	ij	32
i	742	pau	444
j	183	zh	27
k	387	dz	21

The database contains the following isolated digits:

DIGIT	COUNT
1	1
2	5
3	9
4	3
5	5
6	6
7	2
8	3
9	2

No utterances of digit “0” where found.

The database contains the following natural numbers:

Number	Count	Number	Count
10	3	60	2
13	1	62	1
14	1	63	1
15	3	80	1
16	3	84	1
17	1	89	1
20	3	500	1
23	1	584	1
30	4	700	1
36	1	762	1
38	2	900	1
40	2	989	1
46	1	1000	1

		1989	1

The database contains one spontanous date: 6 March 1989 and three occurances of spontanous time strings where found: 06:15, 05:30 and 10:30

The total number of diphones is 10212 with 518 distinct values and the total number of triphones is 10023 with 2705 distinct values.

4. LEXICON

A lexicon file (LEXICON.TBL) was automatically generated. Each line contains one word and its occurence count separated by <tab>:

word<tab>count

5. SPEAKER DEMOGRAPHIC INFORMATION

Unknown.

5.1 ACCENT/REGIONS

Unknown.

5.2 SPEAKER AGES

Unknown.

5.3 SPEAKER OVERLAP

Unknown.

6. RECORDING CONDITIONS

6.1 SOFTWARE

Unknown.

6.2 HARDWARE

Unknown.

7. TEST MATERIAL

None provided