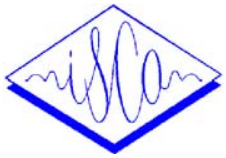

4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

BABEL: AN EASTERN EUROPEAN MULTI-LANGUAGE DATABASE

P.Roach[1], S.Arnfield[1], W.Barry[2],J.Baltova[3], M.Boldea[4], A.Fourcin[5], W.Gonet[6], R.Gubrynowicz[7], E.Hallum[1], L.Lamel[8], K.Marasek[9], A.Marchal[10], E.Meister[11], K.Vicsi[12]

[1] University of Reading, UK; [2] University of Saarbrucken, Germany; [3] Bulgarian Academy of Sciences; [4] Timisoara Technical University, Romania; [5] University College London, UK; [6] M.C.University, Lublin, Poland; [7] Polish Academy of Sciences, Warsaw, Poland; [8] LIMSI, Paris, France; [9] University of Stuttgart, Germany; [10] CNRS,France; [11] Institute of Cybernetics,Tallinn, Estonia; [12] Technical University of Budapest, Hungary

ABSTRACT

BABEL is a joint European project under the COPERNICUS scheme (Project #1304) comprising partners from five Eastern European countries and three Western ones. The project is producing a multi-language database of five of the most widely-differing Eastern European languages. The collection and formatting of the data conforms to the protocols established by the ESPRIT SAM project and the resulting EUROM databases.

1. INTRODUCTION

Standard formats and standard conditions are vital for the production of multi-language databases. Essential preliminary work in standardization for European languages has been carried out by the SAM project (Fourcin and Dolmazon [1]), and the standards established are now available for adding further languages to the list of those collected so far. The speech database resulting from ESPRIT-funded work so far is published as a CD-ROM database containing material recorded in Danish, Dutch, English, French, German, Italian, Norwegian, Swedish. Greek, Portuguese and Spanish. Details of the protocols used for this database, which is named EUROM1, are given in Chan [2] and can be examined via World Wide Web at the following addresses:
http://www.phon.ucl.ac.uk/resource/eurom1.html
ftp://pitch.phon.ucl.ac.uk/pub/eurom1/value_report/

2. THE BABEL PROJECT

In 1995, the Commission of the European Community awarded a grant under the COPERNICUS programme for scientific and technical collaboration with Central and Eastern Europe to a consortium which adopted the name BABEL. The membership of this consortium comprises the following:

- One coordinating partner: University of Reading, UK (P.Roach).
- Six partners in Central and Eastern Europe: Bulgarian Academy of Sciences and University of Sofia, Bulgaria (J.Baltova); Estonian Academy of Sciences, Estonia (E.Meister); Technical University of Budapest, Hungary (K.Vicsi); Technical University of Timisoara, Romania (M.Boldea), Marie Curie University, Lublin, Poland (W.Gonet), Institute of Fundamental Technical Research, Polish Academy of Sciences, Warsaw, Poland (R.Gubrynowicz).
- Five partners in Western Europe: University of Aix-en-Provence, France (A.Marchal); LIMSI, France (L.Lamel);
- University of Saarbrucken, Germany (W.Barry); University of Stuttgart, Germany (K.Marasek); University College, London, UK (A.Fourcin).

The project runs for three years (March 1995 to March 1998).

2.1 Database contents

The recording and analysis of the data is being done on identical speech workstations, again following SAM conventions. The workstation is PC-based, with the OROS AU-21 digital signal processing board. Recording is done direct to disk with a 20kHz sampling rate. Backup disks of the data are checked for technical quality by the Warsaw partner, and are archived at Reading University.

For each language, the recordings aim to follow the EUROM1 design, with the following ideal composition:

Many Talker Corpus (30 women, 30 men)
100 numbers
3 passages
5 sentences

Few Talker Corpus (5 women and 5 men)
100 numbers x 5
15 passages
25 sentences
C(C)VC(V) x 5

Very Few Talker Corpus (1 woman and 1 man)
C(C)VC(V) material embedded in 5 context phrases.
Context words x 5

In fact, within the EUROM1 recordings some differences in design between the lanaguges was unavaoidable, and some degree of latitute is similarly necessary in the case of BABEL.

2.2 Symbolic Transcription Conventions

It is intended that a substantial amount of the recorded data will be manually annotated at the phonemic level. The process of transcription of speech database material, using a speech workstation, differs somewhat from the traditional practice of "paper and pencil" transcription (Roach et al [3]). Only physically observable and segmentable entities are labelled. The labels may contain more or less phonetic detail by means of diacritic additions. The principles of the SAMPA transcription conventions used for the BABEL project were devised by J.C.Wells (see World Wide Web addresses given above). Each language presents its own problems for symbolization, and finalizing the conventions for each of the BABEL languages has been a major task for each of the partners involved. The details of these conventions will be published in full in due course, but they are summarised in their provisional form below:

- **Bulgarian**
  Vowels: i, e, a, 3, O, u
  Consonants: p, b, t, d, k, g, ts, dz, tS, dZ, f, v, s, z, S, Z, x, m, n, l, r
  (The diacritic ' is added to consonants which are palatal)

- **Estonian**
  Vowels: i, ii, e, ee, a, aa, A, AA, o, oo, u, uu, 7, 77, y, yy, 2, 22
  Consonants: p, pp, t, tt, k, kk, t', t't, f, ff, v, vv, s, ss, S, SS, h, hh, s', s's, m, mm, n, nn, n', n'n, l, ll, l', l'l, r, rr, j, jj

- **Hungarian**
  Vowels: i, i:, E, e:, O, A:, o, o:, 2, 2:, u, u:, y, y:
  Consonants: p, b, t, d, c, J\, k, g, f, v, s, S, Z, h, m, n, J, r, l, lj, t_s, d_z, t_S, d_Z

- **Romanian**
  Vowels: i, e, a, @, 1, o, u
  Consonants: p, b, t, d, k, g, f, v, s, z, S, Z, h, m, n, l, r, ts, tS, dZ

- **Polish**
  Vowels: i, e, a, o, u, 1, e~, o~
  Consonants: p, b, t, d, k, g, f, v, s, z, S, Z, s', z', h, m, n, n', N, t_s, d_z, t_S, d_Z, t_s', d_z', l, r, w, j

2.3 Distribution

The material of the corpus will be stored on CD-ROM and it is hoped to make it available to other researchers via ELSNET. Information about progress on the project can be read via World Wide Web at the following address:
http://midwich.rdg.ac.uk/research/speechlab/babel/

3. REFERENCES

1. Fourcin, A.J. and Dolmazon, J-M. "Speech knowledge, standards and assessment", *Proceedings of XII International Congress of Phonetic Sciences*, AIx-en-Provence, Vol.5, 430-433 (1991).

2. Chan, D., Fourcin, A. and others, "EUROM - A Spoken Language Resource for the EU", *Proceedings of Eurospeech '95*, Madrid, Vol.1, pp. 867-870, (1995).

3. Roach, P.J., Roach, H.N., Dew, A. and Rowlands, P. 'Phonetic analysis and the automatic segmentation and labelling of speech sounds', *Journal of the International Phonetic Association*, vol. 20.1, pp. 15-21, (1990)