Abstract

One of the major tasks of the working group "Enabling Technology and Research" in the ESPRIT-project 2589 (SAM) is the production of standard European speech databases. Within SAM, standard recording protocols and several tools for recording and annotion of speech databases on the standard SAM workstation (SESAM) have been developed. The whole EUROM.1 database covers eight European languages (Danish, Dutch, British English, French, German, Italian, Norwegian and Swedish).

Introduction

In scientific and practical application terms the SAM project has been outstandingly successful. Within its funded lifetime it was given a rare Flagship Award by the European Commission and many hundreds of research papers and reports and working documents were produced which have subsequently been widely used and referenced, multi- and poly-language corpora were initiated and complete complementary sets of interlinked software tools made available to other workers. The project served as a model for similar activities in Japan, USA, Australia and China. It has led to the formation of the very influential COCOSDA network, at the world level of activity, and to the authoritative EAGLES overview of work in Spoken Language Systems in Europe itself. These achievements have only been possible as the result of a substantial degree of dedication, and some hardship, on the part of a core of workers.

The first tangible outcome of the SAM group's collaboration took place in June 1986 when a complete proposal was made to the European Commission for the definition and establishment in Spoken Language Engineering of a European foundation for multi- and poly-lingual standards in performance assessment and evaluation. Although the proposal was well-received by external experts, European Community difficulties made it impossible to obtain funding. The initiation of the SAM Definition Phase ESPRIT Project 1541 (2II87 - 31III88) followed a further full proposal, and it was during this period that a theoretical consensus was established and the main practical approaches to multi-language cooperation were agreed. The work packages were designed by the industrial and academic partners so that practice and theory were cooperatively addressed. Continuing longer term funding uncertainty gave the need for a further interim proposal which led to the Extension Phase of 1541 (1IV88 - 28III89) during which the first CDROM corpus, EUROM0, was made, phonotypically described and practically applied in conjunction with the special phonetic, linguistic and software tools which were in course of development and being implemented on SESAM the reference workstation. This was followed by a proposal for a five year period of sustained, and coherently supported, work within ESPRIT II. Although the new workplan was well-received as a scheme of

multinational research and development it proved once more impossible, but this time for local administrative reasons within the Commission, to obtain stable support. The five year programme was cut to three years (with the elimination of the means for the production of EUROM1 as one economy) and then the three year period was reduced to two separately negotiated eighteen month projects. In the event once initiated the first of these projects remained unfunded by the EC for six months and some younger colleagues were lost - as was their work. EUROPEC (European Program d'Enregistrement de Corpus), for example, was in consequence only finally functional well after the planned period of data acquisition and a major part of the work on EUROM1 has been done without funding long after the end of the main project. The main SAM project was then fragmented by the EC into two artificially separate "research" and "application" groups, SAM_R and SAM-A. The three year SAM-A Project 6819 in ESPRIT II (coordinated by Vocalis) would have provided a basis for the completion of EUROM1 but the EC announced its termination before the end of its first year of existence. A somewhat similar administrative fate has led to the stillbirth of the SpeechBase Spoken Language Resources Project which was conceived by the European Speech Community as a vital resource component of the Framework IV Programme. The present EUROM1 activity has derived some benefit from being associated with the narrowly application oriented SpeechDat Project (LRE 66314) but the corpus has finally been produced with VALUE support mostly as the result of the continuing dedication of largely unfunded SAM people.

A brief acknowledgement follows to the workers and institutions primarily responsible for the final completion of EUROM1.

Borge Lindberg	Danish	Aalborg University
Cor in't Veld	Dutch	SPEX, Leidschendam
Dominic Chan	English	UCL, London
Steve Nevard		
Tim Sherwood		NPL, Teddington, UK
Jerome Zeiliger	French	ICP, Grenoble
Daffyd Gibbon		
Gunter Braun	German	Universität Bielefeld
Giuseppe Castagneri	Italian	CSELT, Turin
Torbjorn Svendsen	Norwegian	Trondheim
Knut Kvale		
Lennart Nord	Swedish	KTH, Stockholm
Roger Lindell		

Individual centres validated their own corpora. The major work involved in overall compilation and correction was done at NPL by Tim Sherwood and at UCL by Dominic Sai Fan Chan, with inputs from Steve Nevard and Adrian Fourcin - who wrote this note.

List of Contractors and Staff Participating in the SAM 2589 Project, Grouped by Country

Denmark
JT
Sven Danielsen
Vagn Wissing

IES
Paul Dalsgaard
Børge Lindberg
Ove Andersen

France
CNRS/GRECO
Jean-Marc Dolmazon, ICP

CNRS workers on the project belong to the following laboratories:
Aix
Mario Rossi
Denis Autesserre
Chaslav Pavlovic
Daniel Hirst

CERFIA
Guy Perennou
Nadine Vigouroux

CRIN
Jean Paul Haton
Anne Boyer
Christine Bourjot
Dominique Fohr
Jacques Noël

ENST
Gerard Chollet
Jean Pierre Tubach

ICP
Christian Benoît
Louis Jean Boë
Geneviève Caelen
Jean François Sérignat
Jean Claude Caërou
Jérome Zeiliger

LIMSI
Jean Luc Gauvain
François Neel
Maxine Eskénazi

Germany
AEG
Helmut Mangold

Bielefeld
Dafydd Gibbon
Gunter Braun

Bochum
Jens Blauert
Ute Jekosch
Dieter Michel

Italy
CSELT
Giuseppe Castagneri
Lucia Vacchetta
Alberto Pacchiotti
Mario Oreglia

CNRS
Kiki Vagges
Piero Cosi

FUB
Andrea Paoloni
Andrea di Carlo

The Netherlands
TNO
Louis Pols
Jan Verhave
Tammo Houtgast
Herman Steeneken
Jeroen van Velden

PTT-RNL
Lou Boves
Jan Hendriks
Bert van Heugten

Norway
ELAB
University of Trondheim
Torbjørn Svendsen
Knut Kvale

Sweden
KTH
Bjørn Granstrøm
Mats Blomberg
Lennart Nord
Kjell Elenius
Roger Lindell

Televerket
Fred Lundin
Gunnar Hult
Jaan Kaja
Bj"rn Lindstr"m

UK
UCL - Prime Contractor
Adrian Fourcin
Evelyn Abberton
Bill Barry
Val Hazan
Martine Grice
Paul Howard-Jones
Stephen Nevard
Jane Espinasse (Project Officer initially)
Kate Jones (Project Officer finally)
Georgie Harland (responsible for all the initial phase administration)

Logica
Richard Winski
Kamran Kordi

National Physical Laboratory (NPL)
Hilary Fuller
Mike Goldsmith
Tim Sherwood

Royal Signals and Radar
Establishment (RSRE)
Roger Moore

Mike Tomlinson

EC Administration

Initiation, Definition and Extension (1541) Phases
Jan Roukens, Didier Bouis

Main Phase (2589)
David Talbot, Patrick Van Hove

SAM-A (6819)
David Talbot, Aniyan Varghese

SpeechDat (LRE 66314)
Jan Roukens, Jose Soler