4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

EAGLES SPOKEN LANGUAGE WORKING GROUP: OVERVIEW AND RESULTS

Richard Winski
email: richard@vocalis.com
Vocalis Ltd
Chaston House, Mill Court
Great Shelford, Cambridge
CB2 5LD - UK

Roger Moore
email: moore@signal.dra.hmg.gb
Defence Research Agency
St. Andrews Road
Malvern, Worcs
WR14 3PS - UK

Dafydd Gibbon
email: gibbon@spectrum.uni-bielefeld.de
Fakultät für Linguistik und
Litteraturwissenschaft
Universität Bielefeld, Postfach 100131
D-4800 Bielefeld 1 - Germany

Abstract

In this paper a brief overview is provided of the EAGLES project specifically with reference to progress achieved in the Spoken Language Working Group. The goals and achievements are presented primarily with respect to the production of a handbook documenting existing working practices and guidelines for spoken language resource creation and description in Europe. Future prospects include further extension and development of the handbook material to cover present activities more adequately, to extend the language coverage and to create a more widely representative consultation base.

1. Introduction

In this paper a brief overview is provided of the EAGLES project [1] specifically with reference to progress achieved within the Spoken Language Working Group (SLWG). The goals, working structures, methods, and achievements are first briefly summarised. We then outline the major achievement of the project, the handbook of Spoken Language working practices and guidelines, with some discussion of important liaisons developing with other projects and bodies. The paper concludes with an overview of current plans and prospects for further extension and development of these activities.

The domain of spoken language technologies ranges from speech input and output systems to complex understanding and generation systems, including multi-modal systems of widely differing complexity (such as automatic dictation machines) and multilingual systems (including for example translation systems). The definition of *de facto* standards and evaluation methodologies for such systems involves the specification and development of highly specific spoken language corpus and lexicon resources, and measurement and evaluation tools.

In these areas the *de facto* standards are derived from the consensus within the spoken language community previously established in a number of European [3] and national projects, with reference to important initiatives in the US and Japan. Primary among these have been the SAM projects (centred on component technology assessment and corpus creation), SQALE (for large vocabulary systems assessment) and both SUNDIAL and SUNSTAR (for multi-modal systems.) Past and present projects with significant outputs in the domain of assessment and resources include ARS, RELATOR, ONOMASTICA and SPEECHDAT, as well as major national projects and programmes of research such as VERBMOBIL in Germany. This has led to an initial documentation of existing practice which is relatively comprehensive but in many respects heterogeneous and widely dispersed. The Spoken Language Working Group of the EAGLES project has addressed the task of collecting and unifying this existing body of material to provide an up-to-date baseline reference documentation to serve current and immediate future needs.

2. EAGLES Project overview

2.1 Overall objectives and structures

The lack of generic technologies and resources and the wide diversity of formats and specifications has hindered the effective reutilisation of existing resources. In 1993 the EAGLES initiative was launched within the framework of the CEU's DGXIII Linguistic Research and Engineering (LRE) Programme, to accelerate the provision of standards for developing, exploiting and evaluating large-scale language resources. The project is now in its final phase of activities and is currently planning to publish its results as a handbook of agreed working practices and recommendations.

EAGLES (Expert Advisory Group on Language Engineering Standards) consists of five Working Groups hosted by designated R&D centres; a Management Board charged with overall co-ordination and supervision and a central support team based at the project co-ordinator site at CPR, Pisa. The Working Groups are made up of experts from European industry and academia, and are concerned with the following five areas: Text Corpora, Computational Lexicons, Grammar Formalisms, Evaluation, Spoken Language.

The aims of EAGLES are:

- to produce public, commonly agreed specifications and guidelines for specific areas of language engineering
- to complement European R&D projects
- to promote adoption of EAGLES results in future R&D ventures; and
- to feed results to national and international standardisation initiatives

2.2 Spoken language objectives in EAGLES

The specific aims within the Spoken Language Working Group (SLWG) have been:

- to consult widely with the SL science, research, technology and application community
- to evaluate existing resources and methodologies
- to identify areas of consensus in respect of resources and standards
- to provide a handbook of recommendations and guidelines to be disseminated widely
- to facilitate interchange and co-operation between the speech and NL communities
- to provide a focus for liaison with other national and international bodies in the field

The strategy adopted to achieve these objectives has been firstly to consult with EAGLES members and other co-opted experts, through a series of technical workshops involving researchers and industrialists active in the field. One of the central objectives has been to define and subsequently commission the production of a handbook using recognised technical experts in each domain. The project has aimed to take full account of the major developments in this field, and wherever possible to provide a set of practical, working recommendations which encapsulate currently perceived best practice for resource creation and description. Initial drafts have been circulated within the SLWG itself and also to associated projects and individuals to ensure that the EU community viewpoint is adequately represented as far as possible prior to wider dissemination and consultation internationally.

The structure of EAGLES has foreseen the need for interaction between the different working groups in NL and SL, and specific cross-groups have been created to foster liaison between these in the areas of lexica, evaluation and corpora. It is envisaged that a greater degree of interaction will be possible as the results of the different working groups become available later in the course of the project, and in particular within future funded phases of EAGLES activities.

Liaison with related national and international initiatives has primarily been advanced on the basis of regular updates via ELSNET newsletters and COCOSDA bulletins, and at ESCA and ESPRIT programme meetings. Contacts have been established with several EU projects in the ESPRIT and LRE RTD programmes. In particular there is now an emerging relationship with ELRA, the newly established organisation concerned with creation, validation and dissemination of European language resources.

3. SL handbook: objectives, scope and readership

The central objective of the SL handbook is to collect and catalogue the existing but often widely dispersed sources of information regarding spoken language resources within a single document. It is intended that the handbook will provide an essential reference work useful to a wide range of laboratories which are concerned with any aspect of speech technology. In addressing the production of the handbook, the project has kept in mind that the potential readership should include

- research workers and system developers who require convenient access to an organised body of specific reference material
- workers in other countries who require access to well-documented common practice in central Europe
- newcomers to the field who require introductory material, primarily research workers in related disciplines and students
- corporate end-users of speech technology, who need to specify, procure or integrate system components, and who require guidance related to system specification and assessment

The scope of the handbook fundamentally addresses the resources required for specifying, developing and evaluating speech technology components, including automatic speech recognition, speaker recognition and speech synthesis, which themselves are integrated to form interactive systems such as spoken dialogue systems. There is an emphasis upon the design and characterisation of speech corpora, the primary resource essential to both speech science and related technology developments, as well as upon assessment methodologies for the component technologies and integrated systems. It was specifically decided however not to attempt to address issues related directly to technology developments *per se*, such as methodologies and techniques for designing specific system components.

4. Achievements

The handbook has been realised as a series of necessarily inter-related chapters, where each chapter provides some introductory background, including definitions of basic terminology, and then provides concise summaries of common approaches, including alternatives, where these exist. Factors pertaining to recommended approaches are outlined, and preferred methods and recommendations are identified wherever possible. A chapter on tools catalogues the software and hardware tools that are available to support resource creation. A selected bibliography is included as well as useful reading lists of a tutorial nature and an index. The current chapter plan addresses:

1. Introduction
2. System Design and Specification
3. Corpus Design
4. Corpus Collection
5. Corpus Representation
6. Lexica
7. Language Modelling
8. Dialogue
9. Physical Characterisation and Description
10. Assessment Methodologies and Experimental Design
11. Assessment of Recognition Systems
12. Assessment of Synthesis Systems
13. Assessment of Speaker Verification Systems
14. Assessment of Interactive Systems
15. Tools
16. Terminology & Glossary

A number of appendices provide valuable reference material on various topics, including:

A. Computer readable phonetic alphabets
B. SAMPA description
C. Speech file formats (SAM, NIST, Verbmobil)
D. Recording protocols (studio, telephone)
E. Compendium of public domain SL Corpora
F. EUROM databases overview
G. Speechdat and Polyphone databases overview
H. Current list of document servers
I. Directory of speech agencies (ESCA, ELRA, ELSNET, LDC)

The current handbook cannot be considered a final or complete statement of guidelines and recommendations as agreed by the EU SL community. Nevertheless it is expected that the present work substantially reflects the community position on a large range of relevant topics, and will prove to be an important interim working document for the provision of commonly agreed working standards and ultimately, where appropriate, may support progression of these *de facto* conventions and practices towards formal representation.

In its present incarnation the handbook already reflects the results of fruitful co-operation between the EAGLES project and the LRE project SPEECHDAT, which itself is concerned with creating an infrastructure and implementation model for the creation of commonly required spoken language resources. This co-operative basis of resource specification and description, and close identification with previous speech technology projects, seeks to ensure that the current set of EAGLES recommendations are relevant and closely related to present-day requirements in both industry and research.

One of the single, most important achievements of the SPEECHDAT project to date has been initiating the creation of an association - the European Language Resource Association (ELRA) - to oversee the creation, validation, marketing and distribution of the growing body of specifically European language resources, both text and spoken. ELRA will provide the executive structures required to implement the strategies for speech resource creation, validation and distribution initially formulated within the SPEECHDAT project. It is foreseen that the co-operation fostered between EAGLES and ELRA will continue to develop as a closely interlinked relationship, much as between a legislature and an executive body.

5. Future Prospects

The objectives presented here for the EAGLES SLWG represent one step in the enormous task of documenting current resources and methods employed within the entire field of spoken language science and technology developments. It is of course inconceivable that such a task can be thoroughly and comprehensively accomplished within such a limited period of time and within the very modest resources devoted to this project. Initial feedback from the spoken language community however has confirmed the value of this activity. Discussions are now under way to define the nature and extent of further work in this important area, and especially to provide an effective mechanism for regular updating of the handbook material. The main areas of activity for the near future are considered to be:

(i) Survey of existing practice. Industrial participation has so far been considerable, but the coverage of opinion within the field needs to be extended on a broader basis than has so far been possible: first, a further in-depth survey should be made of the requirements of industrial developers and users; second, a survey of resources and needs in Eastern Europe and the Newly Independent States formerly in the Soviet Union is required. Equally important is coverage of results of Fourth Framework Programme projects.

(ii) Extension of language base. Existing documentation covers the main languages of the European Union, and definition of standard representation techniques for transcription and signal annotation of other languages is urgently required. Of increasing interest in this respect are the languages of Eastern Europe.

(iii) Revision and completion of existing documentation. The presently available documentation is still incomplete and requires fuller consultation on some of the more recently produced material. Several areas, including corpus collection and lexical database techniques and tools as well as the evaluation methodology for complex systems, require updating and additions in the light of recent developments. More precise user targeting is required, with an explicit distinction in information granularity between management/planning and laboratory/project user levels.

(iv) Publication and dissemination. The available documentation requires new dissemination and publication concepts in line with recent developments in the use of new media and broad-band networks. Efficient development and production techniques for different modes of publication and dissemination of complex documents in conventional and hypertext form are required. Legal aspects of accessibility of resources and documentation need to be addressed.

(v) Co-ordination with other bodies. The relation between European standardisation and evaluation work and European associations such as ELRA, as well as with national spoken language archives and validation centres, requires further study and negotiation. Equally important is continued interaction with related international endeavours, primarily via the COCOSDA initiative. Finally, language projects initiated in the Fourth Framework Programme are expected to provide a continued source of fruitful interactions with future EAGLES activities.

(vi) Co-ordination with written language standardisation and evaluation groups. Some of the results of core work in spoken language which is of secondary value to written language work, such as pronunciation transcriptions for lexica and dialogue corpora, is available as a service to written language groups. However, in addition to the separate consolidation of work in the two complementary areas, joint consultation will be required in the foreseeable future on complex systems such as automatic dictation systems or speech to speech translation systems.

References

1. Moore, R. *"The EAGLES Working Group in Spoken Language."* in *"Advanced Speech Applications. European Research on Speech Technology." (Research Reports ESPRIT Volume 1).* Ed K.C. Varghese, S. Pfleger and J.P. Lefevre. Springer-Verlag. 1994.

2. Winski, R and Fourcin, A. *"A Common European Approach to Assessment, Corpora and Standards."* in *"Advanced Speech Applications. European Research on Speech Technology." (Research Reports ESPRIT Volume 1).* Ed K.C. Varghese, S. Pfleger and J.P. Lefevre. Springer-Verlag. 1994.

3. *"Advanced Speech Applications. European Research on Speech Technology." (Research Reports ESPRIT Volume 1).* Ed K.C. Varghese, S. Pfleger and J.P. Lefevre. Springer-Verlag. 1994.

4. *"Handbook on Spoken Language Resources."* EAGLES SLWG publication (to be published September 1995)

The EAGLES draft handbooks are available as postscript documents which can be accessed from the central EAGLES ftp server (nicolet.ilc.pi.cnr.it with user name and password: eagles). An EAGLES world wide web information page is accessible at http://www.ilc.pi.cnr.it/EAGLES/home.html.
Publication of the printed handbook is in preparation. The central project editors may be contacted at ceditor@tnos.ilc.pi.it for more details.

Acknowledgements

The EAGLES project is fundamentally conceived as a community undertaking. This breadth of participation is reflected in the following list of active members, technical authors and contributors whose corporate efforts have cumulatively resulted in the EAGLES SLWG handbook. Many individuals have contributed and we apologise if we have inadvertently omitted to include any individual in the list below.

SLWG members:
J-M. Dolmazon, G. Castagneri, N. Fraser, D. Gibbon, R.K. Moore, L. Pols, H. Tillmann, R. Winski

Principal technical authors:
F. Bimbot, L. Boves, G. Chollet, K. Choukri, E. den Os, C. Draxler, N. Fraser, D. Gibbon, P. Howell, L. Knohl, V. Kraft, H. Ney, R. van Bezooijen, V. van Heuven, D. van Leeuwen

Contributors:
W. Barry, C. Benôit, J. Blauert, M. Cartier, P. Dalsgaard, C. Delogu, A. Fourcin, V. Hazan, U. Jekosch, D. Johnston, H. Klaus, L. Lamel, J. Llisterri, F. Néel, G. Pérennou, H. Steeneken, I. Trancoso, J. Wells, B. Williams, D. van Bergem, J. Zeiliger

Editorial Support:
D. Gibbon, I. Mertins, J. McNaught