The domain of spoken language technologies ranges from speech input and output systems to complex speech understanding and generation systems. **Spoken Language System and Corpus Design** provides the potential user of speech technologies as well as the developer of applications with the essential knowledge for

- the precise formulation of individual requirement profiles
- the competent and accurate comparison of available technologies
- the design, construction, transcription and use of spoken language corpora for speech research and development.

ISBN 3 11 015365 3

Editors
DAFYDD GIBBON · ROGER MOORE · RICHARD WINSKI

Spoken Language System and Corpus Design

Mouton de Gruyter

1 User's guide

1.1 Background

This handbook has been produced as a result of an initiative by the Commission of the European Union in February 1993 to launch, under the auspices of the DG XIII Linguistic Research and Engineering (LRE) Programme, a project aimed at accelerating the provision of common functional specifications for the development of large-scale speech and language resources within Europe. The initiative became known as "EAGLES" – the Expert Advisory Group on Language Engineering Standards – and it attracted a budget from the European Commission of approximately 1.25 MECU over a period covering almost three years.

1.1.1 EAGLES objectives

The overall strategic aims of the EAGLES initiative were as follows:

- to produce publicly defined and commonly agreed specifications and guidelines for specific areas of language engineering,
- to bring together industry and academia in an attempt to reconcile the often heterogeneous interests and approaches pursued by the two groups,
- to create focal points of expertise in Europe,
- to complement related European R&D projects falling under the LRE, ESPRIT and EUREKA programmes,
- to enhance the usability, portability and comparability of EU projects' results and thus maximise return on investment in the development of language products and services,
- to contribute to consensus building on an international scale by interacting with national and international standardisation initiatives, and
- to safeguard the multilingual dimension of Europe.

In addition, more specific objectives were defined as follows:

- to promote and accelerate cooperation and consensus building in specific areas of language engineering in Europe,
- to produce prenormative specifications and guidelines for the description and representation of linguistic knowledge and data, including methods for the assessment and evaluation of systems and components which make use of such information,
- to play an active role in the definition, demonstration, evaluation, validation, promotion and dissemination of said specifications, thus furthering harmonisation of the methods and formats used for the encoding and interchange of linguistic knowledge and data,
- to exploit and complement the results achieved in EAGLES members' own R&D activities,

- to actively seek cohesion with cooperative R&D efforts in Europe and worldwide,
- to produce a set of guidelines based on a broad consensus among participating organisations, which will be made publicly available and to which European and national R&D activities will be invited to adhere, and
- to provide input to national and European standardisation bodies active in relevant fields.

1.1.2 EAGLES organisational structure

The activities of the EAGLES initiative were distributed over five specialist technical working groups (see Figure 1.1). These working groups were set up in response to the most urgently felt requirements in advanced language engineering: common methodologies for the creation and interchange of electronic language resources such as text and speech corpora, computational lexicons and grammar formalisms, and the evaluation and quality assessment of language processing systems and components. There was also felt to be a need to attempt to reconcile the needs and practices of the speech and language R&D communities.

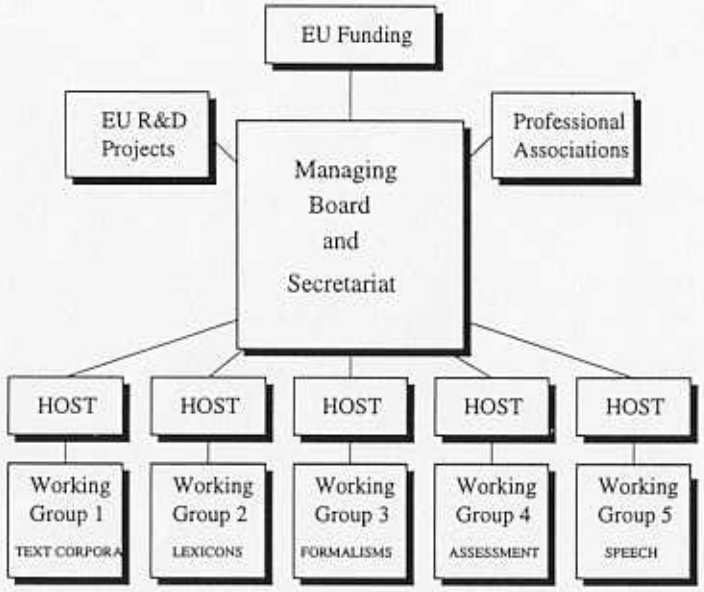

Figure 1.1: The main structure of the EAGLES Group

The EAGLES Management Board was constituted from member organisations representing European projects in natural language and speech (such as MULTILEX, PLUS, ACQUILEX, NERC, GENELEX, SAM-A, SUNDIAL, EUROLANG, TWB, ONOMASTICA and DELIS) and European associations and

coordinating bodies such as ELSNET, ESCA, FOLLI and the European Chapter of the ACL. The Board was chaired by Prof. Rohrer from the University of Stuttgart.

The project funding arrangements were such that, in the early stages, only travel and administrative expenses were covered by the EAGLES initiative. This meant that, whilst some cover was provided by other EU funded projects, the bulk of the costs incurred in the production of this handbook (and the other outputs from the EAGLES initiative) have been met by purely voluntary donations of time and effort on the part of the many contributors involved.

1.1.3 EAGLES workplan

Work towards the EAGLES objectives was conducted in accordance with two phases of activity each of roughly fifteen months duration (see Figure 1.2). Interim recommendations were released to the speech and language communities at the halfway point during 1994.

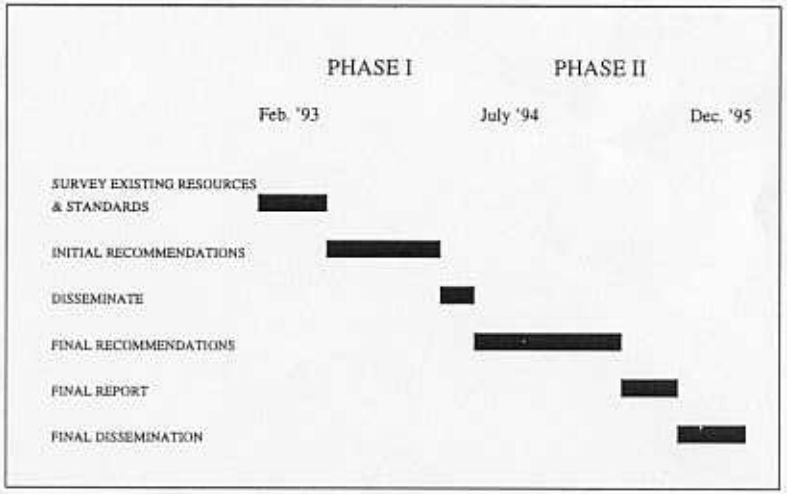

Figure 1.2: The EAGLES workplan

The initial activity was to conduct a survey of existing resources and standards. This was achieved by contacting people working in the field, identifying priority requirements, assembling working papers, reports and material, identifying relationships with other activities, scoping the EAGLES activities, producing an initial survey report and disseminating it within each Group and externally. From this survey, a set of initial recommendations was disseminated to the communities for comment.

The second phase of the project involved obtaining and evaluating the feedback to the initial recommendations, identifying areas of further development, producing and disseminating a set of final recommendations and obtaining feedback on the final results.

1.2 Spoken Language systems, standards and resources

1.2.1 Spoken Language systems

There is a wide range of technologies which fall under the general banner of "spoken language processing" (SLP) including:

- "automatic speech recognition" ASR (also known as "direct voice input" DVI, and "speech input" SI),
- "automatic speech generation" ASG (also referred to as "direct voice output" DVO, "speech synthesis" SS, and "text-to-speech" TTS),
- "speech input/output" SIO (which includes "speech understanding systems" SUS,
- "spoken dialogue systems" SDS, and "speech-to-speech translation systems" STS),
- "speech coding" (covering wide-band coding at over 4k bps, narrow-band secure voice between 1200 bps and 4k bps, and very-low data-rate speech communications at under 1200 bps),
- "speech analysis or paralinguistic processing" (which includes speaker identification/verification, language identification/verification and topic spotting),
- general speech processing applications such as "speech enhancement" and "voice conversion", and "speech systems technology" (which is concerned with the technology of database recording, corpus transcription, annotation, storage and distribution).

Many of these technologies rely heavily on the availability of substantial quantities of recorded speech material: first, as a source of data from which to derive the parameters of their constituent models (manually or automatically), and second, in order to assess their behaviour under controlled (repeatable) test conditions.

Of course very few spoken language processing applications involve standalone spoken language technology. Spoken language provides an essential component of the more general human–computer interface alongside other input/output modalities such as handwriting, typing, pointing, imaging and graphics (see Figure 1.3). This means that the actions and behaviours of the speech-specific components of a spoken language system inevitably have to be orchestrated with respect to the other modalities and to the application itself by some form of interactive dialogue process (simultaneously taking into account the wide range of human factors involved).

The complexity of the human–computer interface, and the subtle role of speech and language processing within it, has been (and continues to be)

a prime source of difficulty in deploying spoken language systems in "real" applications. Not only are field conditions very different to laboratory conditions, but there has been a serious lack of agreed protocols for testing such systems and for measuring their overall effectiveness.

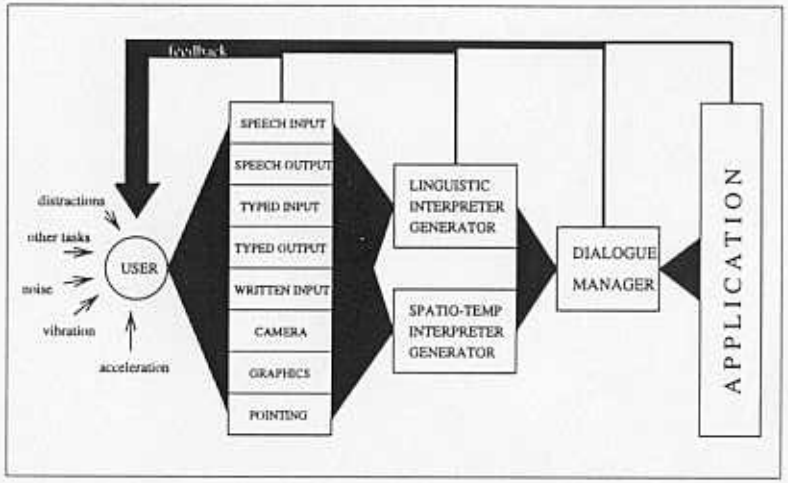

Figure 1.3: Multimodal human–computer interface (HCI) including speech/language input/output

1.2.2 Standards and resources for Spoken Language systems

As indicated above, the domain of the spoken language technologies ranges from speech input and output systems to complex understanding and generation systems, including multimodal systems of differing complexity (such as automatic dictation machines) and multilingual systems (with applications in different languages, but also, as in speech-to-speech translation systems, integrating processors for more than one language). 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 together with suitable measurement and evaluation tools. These requirements currently still determine considerable differences between spoken and written language in terms of paradigms and techniques of measurement and evaluation, which range from different practical and legal requirements for corpus construction to differences in experimental paradigms for the quality control of working systems. In these areas, the *de facto* standards are derived from the consensus within the spoken language community on evaluation methods and the resources required for these.

Of course, spoken language technology is still a relatively young area and thus the so-called standards that are discussed here represent only the first rung of the ladder towards the more formal standards which might emerge at a later date. The use of the term "standards" in the R&D community and in the context of this handbook is more usefully interpreted in terms of guidelines and recommended practices. The emergence of more prescriptive actions such as professional codes of conduct, quality marks and formal standards still lies very much in the future.

Nevertheless, the requirement for agreed standards and guidelines pervades all of the links in the spoken language system R&D chain starting from the research community (for algorithm development and benchmarking), to product developers (for performance optimisation), system integrators (for component selection), manufacturers (for quality assurance), sales staff (for marketing), customers (for product selection) and users (for service selection).

Of course, activity in the area of standards and resources for spoken language systems is not new; for many years, the majority of spoken language R&D groups have appreciated the value of sharing recorded speech material and the importance of establishing appropriate infrastructure in terms of standardised tools, research methodology, data formats, testing procedures etc. Indeed, the national research communities in a number of countries have put into place mechanisms for discussing and exchanging such information either as a result of an initiative on the part of the research community itself (for example, the Speech Technology Assessment Group – STAG – was set up in the UK under the auspices of the Institute of Acoustics in 1983 and the IEEE operated a similar working group in the USA over ten years ago) or mediated by a central agency (such as GRECO in France and DARPA in the USA). Also, several national standards organisations have become involved, notably the National Institute for Standards and Technology (NIST – formally the National Bureau of Standards) in the USA, the National Physical Laboratory (NPL) in the UK and AFNOR in France.

1.2.2.1 Spoken Language standards and resources in Europe

The most significant activity on spoken language standards and resources in Europe has without doubt been the ESPRIT Speech Assessment Methods (SAM) project which ran from 1987 to 1993 (Fourcin 1993; Winski and Fourcin 1994). The SAM project arose out of the need to develop a common methodology and standards for the assessment of speech technology systems which could be applied within the framework of the different European languages. The definition of the project took place in the context of several ongoing national and international programmes of research including the UK Alvey programme, GRECO in France, COST in Europe and DARPA in the USA.

1.2.2.2 The ESPRIT SAM project

The SAM project was based on a collaboration between almost thirty laboratories in eight different countries: six countries within the EU and two from EFTA. Work was conducted in three interconnected areas:

- speech recognition assessment,
- speech synthesis assessment,
- enabling technology and research.

Within this structure SAM established a set of common tools which have become widely used in a large number of participating and non-participating speech research laboratories. These tools included a reference workstation, a recommended set of protocols for recording, storing, annotating and distributing speech data, and a standard machine readable phonetic alphabet. The SAM reference standard workstation (SESAM) was designed to provide a gateway between one European speech research laboratory and another. The minimum hardware requirements were an IBM PC-AT (or compatible) computer, an analogue interface board (OROS-AU21 or AU22), 1Mbyte of extended memory and a CD-ROM reader. SESAM hosted all SAM software products including EUROPEC, VERIPEC, PTS and ELSA for speech data collection and annotation, EURPAC and SAM_SCOR for measuring the performance of speech recognition systems, and SOAP for measuring the performance of speech synthesis systems.

The first SAM corpus – EUROM-0 – was distributed on a single CD-ROM and contained five hours of speech material. A second corpus – EUROM-1 – used the same standard format with sixty talkers in each of eight languages, speaking phonetically balanced CVC words, number sequences up to 9999 and situationally linked sentence.

1.2.2.3 Other EU projects

In parallel with (and subsequent to) SAM, a number of other EU funded projects have focused on spoken language standards and resources. For example, SQALE was concerned with the assessment of large-vocabulary automatic speech recognition systems across different EU languages and both SUNDIAL and SUNSTAR were directed towards the assessment of multimodal interactive systems.

Other 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 the German VERBMOBIL project. In particular, one of the single, most important achievements of the SPEECHDAT project has been to initiate the creation of the European Language Resources Association (ELRA).

1.2.2.4 The European Language Resources Association

The European Language Resources Association was established in Luxembourg in February, 1995, with the goal of creating an organisation to promote the creation, verification, and distribution of language resources in Europe. A non-profit organisation, ELRA aims to serve as a central focal point for information related to language resources in Europe. It is intended that it will help users and developers of European language resources, as well as government agencies and other interested parties, exploit language resources for a wide variety of uses. It will also oversee the distribution of language resources via CD-ROM and other means and promote standards for such resources. Eventually, ELRA will serve as the European repository for EU-funded language resources and interact with similar bodies in other parts of the world (such as the LDC – see below).
ELRA membership is open to any organisation, public or private. Full Membership, with voting rights, is available to organisations established in the EU or European Economic Area. Organisations based elsewhere may participate as subscribers. Purely for organisational purposes, members are classified by their chief interest (spoken, written, or terminological resources). The annual membership fee has been set at a level which would encourage broad participation.

1.2.2.5 Spoken Language standards and resources worldwide

At the international level, the NATO Research Study Group on Speech Processing (NATO/AC342/Panel III/RSG10) has, since the late 1970s, provided an effective mechanism for exchanging information on spoken language standards and resources between Canada, France, Germany, the Netherlands, the UK and the USA (Moore 1986). RSG10 was responsible for the first publicly available multilingual speech corpus, and has subsequently released on CD-ROM a database of noises from a range of selected military and civil environments (NOISE-ROM) and related experimental test data (NOISEX).
Also, at each IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) during the 1980s, Janet Baker of Dragon Systems regularly organised a side-meeting to discuss speech databases and opportunities to share such data between different laboratories.

1.2.2.6 COCOSDA

More recently, the International Committee for Collaboration in Speech Assessment and Databases – COCOSDA – was established in 1990 to encourage and promote international interaction and cooperation in the foundation areas of Spoken Language Processing (Moore 1991). COCOSDA provides a forum for international action and discussion and gives platforms for groups

of workers to exchange information and to set up collaborations in the field of Spoken Language Engineering. Very many of the world's leading workers are amongst it members and the group discussions are open and unconstrained by any special interests. Meetings take place annually as a satellite event to one of the major international conferences.

1.2.2.7 The Linguistic Data Consortium

In the US, the Linguistic Data Consortium (LDC) was founded in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Based at the University of Pennsylvania, the LDC is a broadly-based consortium that, in 1995, included about 65 companies, universities, and government agencies. An initial grant of $5 million from ARPA amplified the effect of contributions (both of money and of data) from the broad membership base, so that there is guaranteed to be far more data than any member could afford to produce individually. In addition to distributing previously-created databases, and funding or co-funding the development of new ones, the LDC has helped researchers in several countries to publish and distribute databases that would not otherwise have been released.
The operations of the LDC are closely tied to the evolving needs of the research and development community that it supports. Since research opportunities increasingly depend on access to the consortium's materials, membership fees have been set at affordable levels, and membership is open to research groups around the world. Although US government investment in LDC database development is continuing, a significant fraction of the consortium budget comes from membership fees. These fees are now adequate to support the central staff organisation, pay database publication costs and underwrite some database creation.

1.3 The EAGLES Spoken Language Working Group (WG5)

The Spoken Language Working Group (WG5) was constituted from nine senior members of the European spoken language R&D community. The selected individuals represented a range of industrial, academic and government interests and each had direct expertise in the area of spoken language standards and resources. The nine members of the Working Group were:

- Prof. Roger K. Moore (Chairman)
  DRA Speech Research Unit, Malvern, UK
- Dr. Richard Winski (Host)
  Vocalis, Cambridge, UK
- Prof. Dafydd Gibbon (Rapporteur)
  Fakultät für Linguistik und Literaturwissenschaft, Bielefeld, Germany
- Dr. John McNaught (Coordinator's Representative)
  Centre for Computational Linguistics, UMIST, Manchester, UK

- Dr. Giuseppe Castagneri
  CSELT, Turin, Italy
- Prof. Jean-Marc Dolmazon
  INPG-ICP, Grenoble, France
- Dr. Norman Fraser
  Vocalis, Cambridge, UK
- Prof. Louis Pols
  Institute of Phonetic Sciences, University of Amsterdam, Netherlands
- Prof. Hans Tillman
  Institut fur Phonetik und Sprachliche Kommunikation, Munich, Germany

From the outset of the EAGLES initiative, it was clear to the Spoken Language Working Group that very substantial resources already existed in regard to established spoken language corpora and widely accepted systems of data description. Also available were integrated systems of assessment and, in some areas, detailed methods of evaluation. In particular, the prior work of the ESPRIT Speech Assessment Methods (SAM) project had already laid down very substantial groundwork in these areas.
There was, however, an urgent need to provide a central focus for the consolidation and appropriate promulgation of these developments – a "handbook" – and the EAGLES initiative provided a unique framework for achieving this in a European context.
As a consequence, the overall objectives of the EAGLES Spoken Language Working Group (WG5) were set down as follows:

- to consult widely with the spoken language science, research, technology and application community,
- to provide a focus for liaison with other national and international bodies in the field,
- to evaluate existing resources and methodologies,
- to identify areas of consensus in respect of spoken language resources and standards,
- to facilitate interchange and cooperation between the speech and natural language communities, and
- to communicate the results in the form of a "handbook of standards and resources for spoken language systems".

The main technical topics addressed by the Spoken Language Working Group were spoken language resources, systems and terminology (Moore 1994b; Winski et al. 1995).
In the resources area, the Group considered that what was required was a review of contemporary and planned national and international corpora, a catalogue of existing speech data archives and a listing of existing distribution centres. The Group also felt that common protocols, formats and tools should be collated covering all of the important aspects of the design,

specification, collection, representation, storage and distribution of spoken language corpora.
In the spoken language systems area, the Group identified two important subtopics: system specification (including requirements definitions, interface standards, dialogue engineering and multimedia) and system assessment (covering methodologies for spoken language input assessment, spoken language output assessment, interactive systems and other speech technologies such as speaker verification).
For spoken language terminology, the Group defined the following activities as being required: the creation of an initial word list, it's extension and/or reduction by members of the Group, the definition of each headword and the addition of glosses in major European languages. However, the Group noted that the funding resources available within the EAGLES initiative were unlikely to be sufficient to complete this important task.

1.3.1 Subgroups of the EAGLES Spoken Language Working Group

Organisationally, the work of the Spoken Language Working Group was performed by a number of Subgroups each focussing on specific technical aspects of the area (see Figure 1.4) and each drawing in expertise from outside the main Working Group.
The structure of the Subgroups was designed to parallel the contents of the planned handbook as it was felt that this would not only provide an effective working structure, but would also simplify the management of the overall activity. In general, each Subgroup had three members, one of whom was a member of the main Spoken Language Working Group.

1.3.2 Relationships with the other EAGLES Working Groups

Within the overall EAGLES initiative, there were obvious technical overlaps between the activities of the five Working Groups and, more specifically, between the Spoken Language Working Group and all four other Working Groups (see Figure 1.1). The exchange of information between groups was encouraged, therefore in the first instance, selected members of the Spoken Language Working Group also participated in other Groups. Also, in view of the very specific commonalities that existed between the Spoken Language Working Group and the Lexicon and Corpus Working Groups, common "Cross-Groups" were established.

1.3.3 Workshops

The objectives of the Spoken Language Working Group were progressed through a series of meetings and workshops which took place during the course of the project. Each event involved members of the main Group and selected Subgroups. The primary workshops took place in London (24th–25th February 1993 and 5th July 1993), Cambridge (1st–2nd November

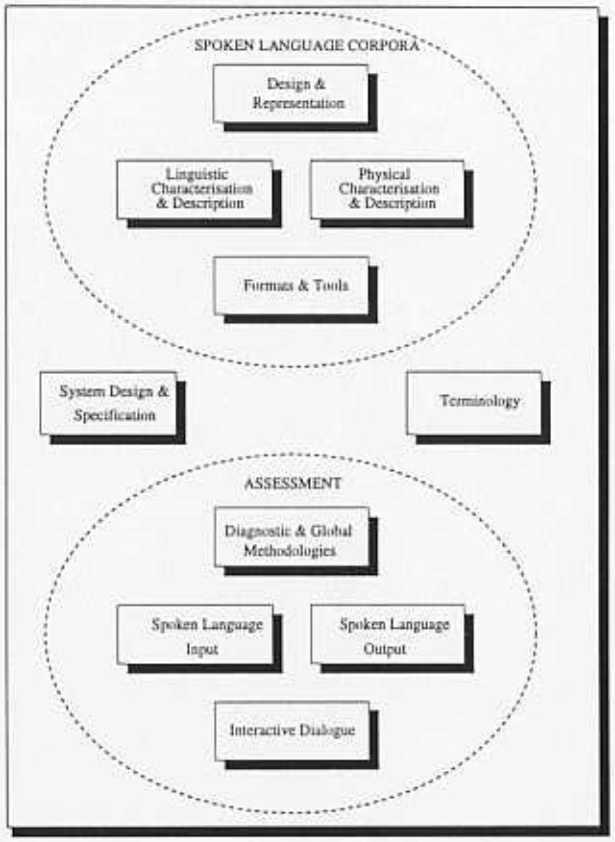

Figure 1.4: Working structure of the EAGLES Spoken Language Working Group

1993), Paris (19th–20th January 1995), London (9th–10th March 1995), Lisbon (24th–25th May 1995) and London (4th–5th December 1995).
The initial workshops focused on the development of a detailed structural outline of the proposed handbook and activities were initiated in all of the identified technical areas in parallel. Subsequent workshops reflected the adoption of a more pipelined approach and this enabled the more well developed technical areas to be progressed to completion at an earlier stage. The penultimate workshop in Lisbon was concerned with finalising the overall style and content of the handbook and involved over thirty participants (two-thirds from university research laboratories and one-third from industrial R&D establishments).

1.3.4 Production of the handbook

At the start of the EAGLES initiative, it was intended that all authoring activities would involve voluntary effort on the part of a large number of individual contributors. However within the Spoken Language Working Group, it had become apparent by the halfway stage in the project that this strategy would not guarantee that the agreed comprehensive plan for the handbook would in fact be realised. Therefore, in July 1994, a change in the funding arrangements was negotiated such that identified individuals could be paid a modest fee for acting as technical authors for the main chapters of the handbook.
These revised arrangements were a great success in unblocking the authoring log-jam and an initial draft of the handbook was completed in October 1994. This "interim" handbook covered 50% of the planned topics and was circulated widely throughout the international spoken language R&D community (see below).
The remaining topics were addressed by the Working Group during the second half of the project and a first draft of the full handbook was completed in September 1995. This was presented to the community for the first time at EUROSPEECH'95 (and at the following COCOSDA meeting).

1.3.5 Consultation with the R&D Community

From the beginning, the Spoken Language Working Group placed great emphasis on the need to reflect the wide (and possibly disparate) views of the spoken language R&D community at large. Therefore, the activities of the Group were made public at every opportunity (for example, at the annual meetings organised by the International Coordinating Committee on Speech Databases and Assessment – COCOSDA). Also, specific consultation periods were established between November 1993 and July 1994, and from October 1994 to May 1995 in which draft documentation was made available publicly on the Internet.
An additional consultation was coordinated at the international level by the central EAGLES administration. Over forty institutions worldwide were sent copies of the interim handbook for formal review. These included European industries such as Philips, Hewlett-Packard, Daimler-Benz, Siemens, GEC, Vecsys, Alcatel, Dragon UK, MATRA and ENSIGMA; European PTTs such as Telefonica, British Telecom, Telia, Jutland Telephone, CNET and the Dutch PTT; European research institutions such as ENST, University of Valencia, IDIAP, KTH, University of Essex, University of Patras, University of Catalunya, University of Leeds and Cambridge University; non-European industries such as Dragon Systems, ETL, IBM, VPC, Entropic and Apple; non-European PTTs such as NTT, ATR, Nynex, Bellcore and Bell Labs; non-European research institutions such as OGI, University of Sydney, Uni-

versity of Tsukuba, Australian National University, University of Berkeley and MIT; and important spoken language resource centres such as the Linguistic Data Consortium (LDC) and the US National Institute of Science and Technology (NIST).
The feedback received from this formal review process was incorporated into the activities of the Group and thence into the handbook itself.

1.4 Overview of the handbook

Previous work on standards and evaluation within the spoken language community have lead to an initial documentation of existing practice which is relatively comprehensive but in many respects heterogeneous and widely dispersed. The purpose of this handbook, therefore, is to collect and catalogue this material within a single document. That is not to say that the handbook is recommending or defining a single European standard, rather, it points to contemporary working practices and *de facto* standards where they already exist.
The handbook has been realised as a series of necessarily interrelated chapters, where each chapter provides some introductory background (including definitions of basic terminology) and concise summaries of common approaches, including alternatives, where these exist. Factors pertaining to recommended approaches are outlined, and preferred methods are identified wherever possible.
The overall style of the handbook is to focus, wherever possible, on clear straightforward "recommendations" supported by appropriate overviews, justifications, exemplifications and reference material. Also, each chapter is intended to be somewhat independent of the others, so that the handbook can appear in its final published form not as a single library volume, but as a set of practical paperbacks and, for convenience in reference to specific points, a fully linked hypertext version.
Clearly, in an exercise of this magnitude, harmonisation all of the key concepts cannot be guaranteed. Not only will the reader come across occasional terminological inconsistencies but it is also possible that some recommendations may be in direct conflict with each other. Such circumstances are probably not errors, but a direct consequence of the lack of concordance that has been attempted so far in the spoken language technology R&D community. Subsequent revisions of the handbook will attempt to resolve these issues.

1.4.1 Intended readership

It is intended that the handbook should provide an essential reference work useful to a wide range of laboratories which are concerned with almost any aspect of spoken language technology. In particular, in addressing the production of the handbook, the Working Group 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, and
- corporate end-users of spoken language technology, who need to specify, procure or integrate system components, and who require guidance related to system specification and assessment.

The handbook is not intended to be a textbook about state-of-the-art algorithms and techniques in spoken language technology. However, there are dangers involved in the simplistic use of a handbook without a good understanding of the methods and principles involved. A reader who wishes to find out more along these lines can find appropriate tutorial material in a number of relevant books (Ainsworth 1988; Allerhand 1987; Bloothooft et al. 1995; Bristow 1984, 1986; Holmes 1988; House 1988; Lea 1980; Linggard 1985; Mariani 1989; Roe and Wilpon 1994; Rowden 1992; Witten 1982), the proceedings of the major conferences on spoken language processing such as the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), the European Conference on Speech Communication and Technology (EUROSPEECH), the International Conference on Spoken Language Processing (ICSLP), the International Congress of Phonetic Sciences (ICPhS), and journals such as Speech Communication (North-Holland), Computer Speech and Language (Academic Press), the Journal of the Acoustical Society of America, and Natural Language Engineering (Cambridge University Press).
Likewise, although it is intended that the handbook should provide support to non-experts, it should be noted that there is a limit to what can be sensibly achieved on a "DIY" (do-it-yourself) basis. For example, some of the areas touched on in the handbook (room acoustics, for example) are huge fields in themselves. Hence, if detailed guidance is required, it is important to realise that there are a number of individuals and companies who already provide technical consultancy and services in such areas.

1.4.2 Scope

The scope of the handbook fundamentally addresses the resources required for specifying, developing and evaluating spoken language 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, collection, representation, characterisation, storage and distribution of speech corpora, as well as upon assessment methodologies for the component technologies and integrated systems.
The handbook is essentially divided into four main parts. The first part (Chapters 2 to 5) is concerned with the design of spoken language systems and addresses spoken language resources (the design, collection, characterisation and annotation of corpora). The second part (Chapters II:2 to II:4) is concerned with spoken language characterisation (spoken language lexicon design, language models, and physical characterisation). The third part (Chapters III:2 to III:6) covers assessment methods (for recognition, synthesis, verification and interactive systems). The fourth part is a substantial body of reference material.

1.4.2.1 Spoken Language system design and specification

One of the difficulties which arises from the complexity of the human–computer interface (HCI) and the position of spoken language within it, is that people concerned with implementing applications are unable to select appropriate HCI components (such as automatic speech recognisers, for example). This arises not just from a lack of standardised evaluation criteria for system components but also from a lack of clear understanding of the implications on overall performance of the performance of each system component.
One possible model for understanding the relationship between spoken language system applications and the corresponding technology is illustrated in Figure 1.5. The key notion which sets it apart from previous models developed by the spoken language R&D community is that it not only focuses on the fact that there are many factors which influence the performance of spoken language systems and that it is necessary to distinguish between "capabilities" and "requirements", but it also emphasises that the purpose of introducing spoken language technology into an application is to achieve the appropriate operational benefits. It is only when all of these features become properly integrated into agreed methods for spoken language system assessment that it will be possible to arrive at a meaningful (and comprehensive) definition of the "suitability" of particular technologies for particular applications.
The model shown in Figure 1.5 indicates clearly that successful implementation of spoken language systems depends only indirectly on the technical features of the system components and on the operational benefits being sought in the applications themselves. What is more important is to develop a process for converting technical features into technical and operational capabilities, and for converting operational benefits into operational and technical requirements. These processes were felt by the Working Group to be so important to the system design process (and hence to the success of

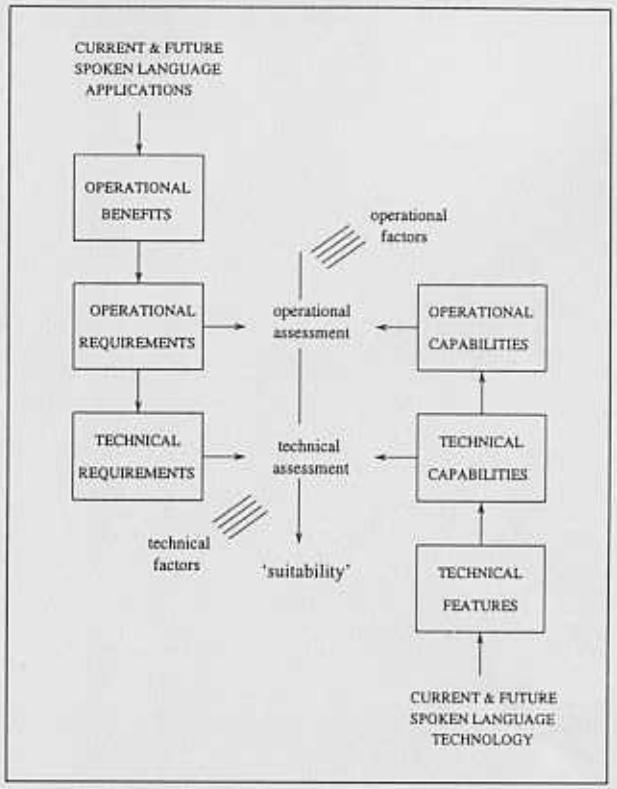

Figure 1.5: A model of the relationship between the applications of spoken language systems and the underlying technology

the technology in the market place), that a chapter outlining design issues should take pole position at the front of the handbook (Chapter 2).

1.4.2.2 Spoken Language resources

Broadly speaking, the spoken language R&D community can be partitioned into two main interest groups: those concerned with "speech science" and those concerned with "spoken language technology". In the main, speech science is the domain of phonetics, linguistics and psychology, and spoken language technology is the domain of engineering, computer science, mathematics and AI. Both areas have a strong need for significant quantities of both transcribed speech data (orthographic, phonetic, prosodic etc.) and digitised acoustic speech recordings (together with the means for accessing selected subsets of the material using the relevant transcriptions and annotation).
Three types of recorded speech are typically of interest (Moore 1992b):

1. analytic–diagnostic material which is of primary importance to progress in basic speech science and which is specifically designed to illuminate specific phonetic and linguistic behaviour (for example, lists of all consonant–vowel–consonant syllables in a given language);
2. general purpose material which includes vocabularies which are either common or which are typical of a wide range of speech technology applications (for example, alpha-numeric words or standard control terms);
3. task-specific material which reflects different levels of formalised spoken monologue/dialogue within constrained discourse domains (for example, train timetable enquiries).

Clearly general purpose speech corpora are easy to collect and are useful in a general sense but, of course, they have only limited practical value. On the other hand, although task-specific corpora can be time-consuming to collect and are only relevant to a specific domain, they are obviously directly useful for the purposes of commercial exploitation. Diagnostic corpora are time consuming to design, but they are extremely useful for research purposes and, in the long term, could prove to be the most valuable resource for spoken language R&D.
At the current time there is a growing requirement for recorded speech which is in some sense more "natural" than the so-called "lab-speech" that has been normally collected and studied up to now. This is true for all three types of material identified above. In this context a range of different speaking styles are now of interest: read speech – including talkers with different amounts of formal training and familiarity with the subject matter, spontaneous speech arising from a directed monologue, spontaneous speech arising from a dialogue between human interlocutors, spontaneous speech arising from simulated human–computer interaction – using the so-called "Wizard of Oz" protocol, and spontaneous speech arising from "real" human–computer interaction.
These issues (and the technology required for acquiring spoken language data) are presented in depth in the handbook chapters on spoken language corpus design (Chapter 3) and collection (Chapter 4).
Of course, recorded spoken language data is, in itself, of limited value; the raw acoustic signal needs to be associated with the appropriate phonetic and linguistic transcripts. This is achieved by "annotating" the data with markers which make such relationships explicit and which provides the means by which the data can be accessed, thereby facilitating the organised study of the data and both automatic parameter estimation and assessment for spoken language systems. These issues are dealt with in the chapter on spoken language corpus representation (Chapter 5).
An important linguistic component of any spoken language corpus, and a key feature of a spoken language system, is the set of words that are employed and their associated properties (such as information about pronunciation, grammatical and semantic features) – the "lexicon". This area is treated in the chapter on spoken language lexica (Chapter II:2).
Another key linguistic aspect of spoken language which has particular relevance in spoken language technology systems, is concerned with "language modelling" (Chapter II:3).
As well as the linguistic characterisation of spoken language corpora described above, there is also a need to be able to characterise such data from an acoustical and electrical perspective. All aspects of the recording chain become important, from the nature of the recording environment, through the types of microphones or headphones that might be used, to issues such as methods for calculating the signal-to-noise ratio. These factors are presented in the chapter on the physical characterisation of spoken language corpora (Chapter II:4).

1.4.2.3 Assessment of Spoken Language systems

In the assessment of spoken language systems it is possible to distinguish three main methodologies: live "field" trials, laboratory-based tests and system modelling paradigms (Moore 1992a). The first of these of course is likely to provide the most representative results but, from a scientific point of view, there are likely to be a number of uncontrolled conditions and this limits the degree of generalisation that can be made from application to application. Field trials also tend to be rather costly operations to mount. Laboratory testing is per force more controlled and can be relatively inexpensive, but the main problem is that such tests may be unrepresentative of some (possibly unknown) key field conditions and give rise to the observed large difference between performance in the laboratory and performance in the field. The third possibility, which is itself still the subject of research, is to model the system (and its components) parametrically. In principle, this approach could provide for a controlled, representative and inexpensive methodology for assessment but, as yet, this area is not sufficiently well developed to be useful.
Also, the term "assessment" covers a range of different activities. For example, a suitable taxonomy of assessment activities should include:

- "calibration" (does the system perform as it should),
- "diagnosis" (how well does the system perform under parametrically controlled conditions),
- "characterisation" (how well does the system perform over a range of diagnostic conditions),
- "prediction" (how well will the system perform under different conditions) and
- "evaluation" (how well does the system perform overall).

Of all these, the last – evaluation – has received a the bulk of the attention in spoken language systems assessment.

Given the complexity of the human–computer interface discussed above, it is clear that assessment protocols are required which address a large number of different types of spoken language system. For example, such systems range from laboratory prototypes to commercial off-the-shelf products, from on-line to off-line systems, from stand-alone to embedded systems, from subsystems to whole systems and from spoken language systems to spoken language based HCI systems.
The majority of research in the area of spoken language system assessment has concentrated on evaluating system components (such as measuring the word recognition accuracy for an automatic speech recogniser, for example) rather than overall (operational) effectiveness measures of complete HCI systems. Since the publication of the NBS guidelines in 1985, there have been considerable developments at the international level. In Europe, the ESPRIT SAM project established a standard test harness for both recognisers and synthesisers and in the US a very efficient assessment paradigm has been funded by the Advanced Projects Research Agency (ARPA) which included an efficient production line of "hub and spoke"-style experiments involving the coordinated design, production and verification of data, distribution through the LDC, and with NIST responsible for the design and administration of tests and the collation and analysis of the results.
These activities point strongly to the importance of establishing appropriate "benchmarks", either through the implementation of standard tests, or by reference to human performance or to reference algorithms.
Throughout these issues, it is vitally important that the relevant practitioners are fully competent in the process of experimental design and in the understanding of key issues such as statistical significance. For these reasons, the handbook specifically includes a chapter on this (Chapter III:2) at the front of the chapters on assessment.
The chapter on experimental design is followed by chapters which cover the assessment of the three main component technologies: automatic speech recognition (Chapter III:3), speaker verification (Chapter III:4) and speech synthesis (Chapter III:5). These are followed by a chapter concerned with the assessment of interactive spoken language systems (Chapter III:6).

1.4.2.4 The reference material

The handbook is structured such that the supporting material for each chapter has been separated from the main text and collated to form a substantial body of reference material spanning all aspects of spoken language standards and resources. The main reference materials covered are:

- Character codes and computer readable alphabets
- SAMPA computer readable phonetic alphabet
- SAM file formats
- SAM recording protocols
- SAM software tools
- EUROPEC recording tool
- Database management system (DBMS) guide
- Speech standards review
- EUROM-1 database overview
- POLYPHONE project overview
- European speech resources
- Transcription and documentation conventions for SPEECHDAT
- The Bavarian archive for speech signals

1.4.3 The main chapters of the handbook

For general orientation purposes, the reader is recommended to refer to the chapters on system design and specification (Chapter 2) and assessment methodologies and experimental design (Chapter III:2).

1.4.3.1 System design

The chapter on "system design" is specifically aimed at potential users of spoken language technology (such as system designers or technology procurers) who need to know how to relate the technical features of the technology to the operational benefits they are seeking to achieve. It is intended that this chapter should be able to help such users to communicate effectively with the technologists and technology suppliers, to give guidance as to what questions they should ask, and to provide a means for specifying their requirements in a way which is meaningful to themselves and to the technologists.
The chapter starts with an introduction to the difference between a system's "capability profile" and the requirements of a given application. This is followed by an enumeration of the many and varied factors which influence the performance of the types of spoken language systems covered by the rest of the handbook. Automatic speech recognition systems are treated first, and over twenty factors are presented which range from aspects such as variability in the fluency of the speaker through to variability in the characteristics of telephone handsets. This is followed by a discussion of the different configurational possibilities for speaker verification/identification systems and a brief description of the key facts of speech synthesis systems. Interactive voice systems are introduced and the importance of error recovery strategies is identified.
The chapter goes on to outline key issues associated with the software and/or hardware aspects of the system platform, and highlights the possibilities for system simulation and prototyping, as well as a variety of practical matters ranging from the physical interface between a spoken language system and the host application to dealing with multilinguality.

1.4.3.2 Spoken Language corpus design

The chapter on "spoken language corpus design" is targeted not only at users of speech corpora within the domain of spoken language technology but also to use in other areas such as sociolinguistics, language learning and pathology. It starts with a discussion of the most important differences between written and spoken language data, and then presents examples of the many application areas which require access to spoken language corpora. The second half of the chapter describes how to specify a spoken language corpus, first in terms of the required linguistic content and, second, in terms of the number and types of speakers involved. The latter issue is dealt with in some detail, and relevant speaker characteristics are covered which, among many other things, include the age and sex of each speaker, their smoking and drinking habits, and whether or not they have received any professional speech training.

1.4.3.3 Spoken Language corpus collection

The chapter on "spoken language corpus collection" concentrates on the practical aspects of collecting spoken language material. In the first part, the dimensions of data collection are described which cover different recording scenarios such as studio versus location recording, or interviews versus read material, for example. It is also pointed out that important data about spoken language may be collected from sensors other than a microphone, for example by means of multi-channel recordings of signals derived from laryngography, electropalatography or NMR (Nuclear Magnetic Resonance) imaging.
The second part of this chapter contains recommendations for the actual collection of spoken language data covering the necessary equipment and the data management protocols needed. The legal aspects of recording arbitrary spoken language material are also discussed and appropriate recommendations given. It is the intention that the recommendations contained within this chapter should enable any reasonably competent person to establish a suitable recording environment that will deliver data in a controlled manner and to an acceptable level of technical quality.

1.4.3.4 Spoken Language corpus representation

The chapter on "spoken language corpus representation" describes how, to be of value, a set of "raw" speech recordings needs to be augmented with symbolic annotation covering a range of phonetic and linguistic levels of description. The transcription of spoken language data is discussed (including problems which arise with spontaneous speech or overlapping speech in dialogues), and mechanisms for segmenting and labelling the data are described. This is followed by an extensive presentation of the many possible representational structures ranging from simple orthography, through detailed low-level acoustic-phonetic analysis, to prosodic transcription and other non-linguistic phenomena (such as hesitations or acoustic non-speech events, for example).

1.4.3.5 Spoken Language lexica

The chapter on "spoken language lexica" provides a framework for relating concepts such as the creation of lexica for specific applications, the transfer of lexical resources from one application to another and the automation of these processes. The chapter covers topics such as the basic features of spoken language lexica, the types of information contained within a spoken language lexicon (such as surface, morphological, grammatical, semantic and pragmatic information), lexicon structure (including appropriate formalisms), lexical access and lexical knowledge acquisition (from dictionaries, for example).

1.4.3.6 Language models

The chapter on "language models" is different from the other chapters in that it is more concerned with details of techniques and algorithms. This is because of the central role language modelling plays in spoken language systems and in characterising a spoken language corpus. The chapter covers the different formalisms involved, the definition of the key concept of "perplexity" and a range of practical schemes for developing high quality language models.

1.4.3.7 Physical characterisation and description

The chapter on "physical characterisation and description" is essentially concerned with the non-linguistic aspects of a spoken language corpus. This includes such features as the characteristics of talkers and listeners, the recording environment, the transducer(s) and any communications channel. It also deals with "reproducibility assurance procedures"; that is, recommendations for ensuring the integrity of the data (for example, calibration techniques and the use of reference signals).

1.4.3.8 Assessment methodologies and experimental design

The chapter on "assessment methodologies and experimental design" is intended to provide general guidance to all practitioners in the field in matters relating to formal methods for designing and executing statistically significant experiments and for the meaningful interpretation of experimental results. This relates both to the design of representative spoken language corpora and to the evaluation of spoken language systems.

1.4.3.9 Assessment of recognition systems

The chapter on the "assessment of recognition systems" presents information on the substantial amount of work that has been done in this area over the past years. The chapter starts with a classification of different recognition systems and then introduces various performance measures. A taxonomy of different assessment methodologies is described ranging from the straightforward use of spoken language corpora, to more diagnostic methods and artificial test signals. This is followed by a discussion of the parameters which affect performance including those which affect the speaker (such as workload stress or noise) and those which affect the recogniser (such as noise).
The second half of the chapter provides recommendations on testing procedures for two main classes of speech recognition system: the small-vocabulary isolated-word recogniser and the large-vocabulary continuous speech recogniser. In both cases, attention is given to the training of the system, the test procedures and scoring and analysing the results.

1.4.3.10 Assessment of speaker verification systems

The chapter on the "assessment of speaker verification systems" opens by presenting a taxonomy of system types in which the difference between identification and verification is made clear, and issues such as text-dependency are illuminated. This is followed by an analysis of the factors which influence the performance of speaker recognition systems and the set of recommended scoring procedures which should be used. The chapter concludes with some specific points concerning the forensic use of speaker recognition systems.

1.4.3.11 Assessment of synthesis systems

The chapter on the "assessment of synthesis systems" starts with a taxonomy of assessment task and techniques, distinguishing, for example, between laboratory and field assessment, and between human judgements and automatic testing. A methodology is then presented covering the choice of subjects for listening experiments, the required test procedures and suitable benchmarks and reference conditions. Recommendations are made for "black box" testing of overall output quality, and for "glass box" testing at many detailed levels of analysis. A taste is also given to future developments in synthesis evaluation.

1.4.3.12 Assessment of interactive systems

The chapter on the "assessment of interactive systems" presents recommendations for the specification, design and assessment of interactive systems in which spoken language dialogue plays a major part. After defining different types of dialogue system, the chapter describes in some detail the "Wizard of Oz" paradigm for system simulation and the central role it plays in the design and assessment of interactive systems. The chapter goes on to address methods for characterising dialogue systems, tasks and users, and presents an assessment framework which includes high-level metrics such as correction rate and transaction success.

1.5 The current state of play

The current handbook cannot be considered a final or complete statement of guidelines and recommendations as agreed by the EU spoken language technology community for the following reasons:

- it may have serious omissions,
- some chapters have not been consulted as fully as others,
- it does not cover the full range of spoken language technologies identified earlier, and
- it may go out of date quite quickly due to the speed of technological progress.

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.

1.6 Possible future actions

1.6.1 Revision and completion of existing documentation

The presently available documentation on spoken language resources, standards and evaluation methodology contains gaps and required 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.

1.6.2 Extended 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.

1.6.3 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.

1.6.4 Terminology

Although part of the original plan, it was realised that very little attention could be given to this vitally important area simply due to the lack of available resources to fund the detailed work that would need to be undertaken. Some groundwork has been done, but a significant effort is needed to bring it to completion.

1.6.5 Move to prescriptive recommendations

Most of the recommendations put forward in the current handbook are based on *de facto* standards and simply describe the current working practices in spoken language technology R&D. In any future activity, it would be possible to move towards a more prescriptive framework, in which serious consideration is given to recommending particular methods and techniques over and above some others. Clearly this would require a continued commitment to the process of community consultation and feedback.

1.6.6 Publication and dissemination

The available documentation requires new dissemination and publication concepts in line with recent developments in the use of new media and broadband 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.

1.6.7 Coordination 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.
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, are 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.
It is also the case that the mere existence of the current handbook presents a worthwhile starting point for the negotiation of more formal standards through the mechanisms of national and international standards authorities.

1.7 Contact points

An initiative of this nature can only succeed with the full backing of the R&D community. Already, many individuals have committed themselves to making a contribution to the activities of the Group. Inputs are still welcome from all corners of the field. Potential contributors can either contact the following individuals, or access the central EAGLES team (see below).

Chairman: Roger K. Moore
DRA Speech Research Unit
St. Andrews Road
Malvern, Worcs, WR14 3PS
United Kingdom
Tel/Fax: +44 1684 89 4091/5103
email: moore@hermes.mod.uk

Rapporteur: Dafydd Gibbon
Universität Bielefeld
Fakultät für Linguistik und Literaturwissenschaft
P 10 01 31
33501 Bielefeld
Germany
Tel/Fax: +49 521 106 3510/2996
email: gibbon@spectrum.uni-bielefeld.de

Host: Richard Winski
Vocalis Ltd.
Chaston House
Mill Court
Gt. Shelford
Cambridge
United Kingdom
Tel/Fax: +44 1223 84 6177/6178
email: richard@vocalis.com

1.8 Acknowledgements

The EAGLES project has been 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 handbook of standards and resources for spoken language systems.
As well as the chairman/host/rapporteur management team and the members of the Working Group mentioned earlier, the following individuals have made substantial contributions to the construction of this handbook:

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. Benoît, D. van Bergem, J. Blauert, M. Cartier, P. Dalsgaard, C. Delogu, J. Esling, K. Fellbaum, A. Fourcin, M. Grice, V. Hazan, U. Jekosch, D. Johnston, H. Klaus, K. Kohler, L. Lamel, J. Llisterri, F. Neel, G. Pérennou, J. van Santen, H. Steeneken, A. Syrdal, I. Trancoso, J. Wells, B. Williams, J. Zeiliger

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

Part I

Spoken language system and corpus design