EAGLES

Editors' Introduction

EAGLES Document EAG–EB–FR1

Version of June, 1996

Contents

1 Contributors

1.1 Authors

N. Calzolari	
Co-Chief Editor, EAGLES	
ILC–CNR	
Via della Faggiola 32	Tel.: +39.50.56.04.81
Pisa I-56126	Fax.: +39.50.58.90.55
Italy	E-mail: eagles@ilc.pi.cnr.it

J. M^{c}Naught	
Co-Chief Editor, EAGLES	
Centre for Computational Linguistics	
UMIST	
Manchester	Tel.: +44.161.200.3098
United Kingdom	Fax.: +44.200.3099
M60 1QD	E-mail: jock@ccl.umist.ac.uk

1.2 Editorial Board and EAGLES WWW server support

T. Ayazi	
EAGLES Secretariat	
ILC–CNR	
Via della Faggiola 32	Tel.: +39.50.56.04.81
Pisa I-56126	Fax.: +39.50.58.90.55
Italy	E-mail: eagles@ilc.pi.cnr.it

A. Enea	
EAGLES Webmaster	
ILC–CNR	
Via della Faggiola 32	Tel.: +39.50.56.04.81
Pisa I-56126	Fax.: +39.50.58.90.55
Italy	E-mail: enea@ilc.pi.cnr.it

T. Paskiewicz	
EAGLES Editorial Assistant	
Centre for Computational Linguistics	
UMIST	
Manchester	Tel.: +44.161.200.3097
United Kingdom	Fax.: +44.200.3099
M60 1QD	E-mail: teresap@ccl.umist.ac.uk

2 About this document

This is the Editors' Introduction to the results of the EC LRE EAGLES (Expert Advisory Group on Language Engineering Standards) project, which ran from February 1993 to May 1996.
As EAGLES was a highly distributed project involving numerous experts in the construction of many documents of various types, yielding different results, the main aims of this introduction are to:

- Enable the reader to understand the motivation for and organisation of EAGLES; and to
- Guide the reader by explaining the nature of EAGLES results and the relationships among what is a highly structured set of documents.

Taken together, the set of documents is referred to as the *EAGLES Guidelines*. They represent the efforts of a wide scientific and industrial community to make recommendations for best practice in a number of core areas of Language Engineering and, where this has been possible, to propose standards in areas where broad consensus has been achieved. We emphasise that the standards described here are the first step in a long-term process: they represent the consensual view of the community on several issues and are to be seen as providing a basis on which further work can proceed.
The Editors thank all who have contributed to the EAGLES initiative. There have been many kinds of contributions, all of which played an important part. All documents in the *EAGLES Guidelines* have been planned, composed, discussed, revised and edited by teams of experts over three years, managed by the EAGLES Working Groups and their Group Editors and coordinated by the Project Coordinator and the Editorial Board, thus it is invidious to mention individuals by name. Each document contains a list of those who contributed to it. Furthermore, the wider community has been involved in commenting on preliminary drafts of those documents that were ready at the end of the first phase of EAGLES activity.
This Introduction is structured as follows. Firstly, we give a broad overview of the EAGLES initiative, its motivations and its aims (section 3). We then briefly describe the EAGLES methodology of work (section 4.1) and interaction with other projects (section 4.2). The general structure of EAGLES is then outlined (section 5) in terms of its various groups (5.1.2), bodies (5.1) and officers (5.1.4), and the organisation of the work within EAGLES is briefly described. Next, a description of the overall structure of the *EAGLES Guidelines* is provided (section 6). This lists the different types of documents which all together constitute the *Guidelines*, classifying them according to their current status. Interested parties may contribute to the initiative by submitting comments and other types of feedback, as is explained in section 7. Such feedback from the wider community is crucial to the success of EAGLES, given its standardising mission which ultimately relies on widespread consensus. Lastly, we give details of how to obtain material from the EAGLES WWWW and ftp sites or in paper form (section 8).

3 Introduction to the EAGLES initiative

Every so often, in the history of ideas, a conjunction of circumstances causes a particular idea to gain prominence and to galvanise research and development in a number of related disciplines. Such ideas, which may have been of previously peripheral interest, are suddenly seen to be critical for future progress. They have an effect at many levels — research, training, industrial development — and rapidly produce, in technological fields especially, a telling impact on society.
Recently, many researchers, language engineers and technology planners have become aware of the idea of *reusability* (see section 3.2.2.2) and of its crucial role in facilitating the development of practical language technology products that respond to the needs of users.
However, reusability in the language technology field relies, as it does in other fields of technology, on the existence of common practices, guidelines, standards and compatible frameworks. Standards, whether these be *de facto* standards or national and international Standards, are the necessary key to true reusability.
With widely known and broadly accepted standards, interchangeability of language technology components becomes feasible; tools can be built to accept input or produce output in standard format; resources can be designed to a standard, or be mapped into a standard form; products of one type can be compared, if they adhere to relevant standards.

The *EAGLES Guidelines* are the result of a first attempt by a large community of language engineers to propose standards, guidelines and recommendations for good practice in several core areas of our field:

- Computational lexicons;
- Text corpora;
- Computational linguistic formalisms;
- Spoken language resources; and
- Assessment and evaluation.

We are thus concerned, more particularly, with the following aspects of Language Engineering:

- Resources (lexicons, corpora);
- Means to describe and manipulate knowledge and data (metalanguages, formalisms);
- Means to assess and evaluate resources, tools and products; and
- Generic and/or portable systems.

Furthermore, our efforts are focussed on both speech and natural language (NLP) processing as there is increasing integration between the speech and natural language communities.

3.1 About this introduction

In the following, we briefly characterise the EAGLES project, then present the background to EAGLES. We describe the motivation for the project and explain its underlying basic concepts, its objectives and foreseen impact. We also include a short section to place EAGLES in its historical perspective both to emphasise the long-term nature of this type of work and to acknowledge the invaluable efforts that have taken place over the years which have helped give rise to, and determine the nature and form of, the EAGLES project.

3.2 Background to EAGLES

3.2.1 What is EAGLES?

The *Expert Advisory Group on Language Engineering Standards* is an initiative of the European Commission (EC). It was launched, in February 1993, within EC DG XIII's *Linguistic Research and Engineering* programme, as project LRE–61–100. The aim of EAGLES is to accelerate the provision of standards for:

- Very large-scale language resources (such as text corpora, computational lexicons and speech corpora);
- Means of manipulating such knowledge, via computational linguistic formalisms, mark up languages and various software tools;
- Means of assessing and evaluating resources, tools and products.

Leading industrial and academic players in the Language Engineering field have actively participated in the definition of this initiative and have lent invaluable support to its execution. Moreover, the initiative is a direct result of a series of recommendations made to the EC over several years. Reports from recent EC Language Engineering strategy committees have strongly endorsed standardisation efforts in Language Engineering:

> *The importance of working towards standards and protocols is well recognised, with the establishment of the EAGLES project that brings together senior representatives of all the major speech and language development projects in Europe.*
>
> (Oakley, 1993, 71)

Moreover, there is a recognition that standardisation work is not only important, but is a necessary component of any strategic programme to create a coherent market, which demands sustained effort and investment:

> *The EAGLES initiative to encourage the development of standards and protocols for all aspects of language and speech engineering should be continued as a priority.*
>
> (Oakley, 1993, 73)

It is important to note that the work of EAGLES must be seen in a long-term perspective. This is especially true for any attempt aiming at standardisation in terms of International Standards. EAGLES will not and cannot result in such Standards: this is the preserve of the International Organization for Standardization (ISO). However, the road to standardisation is a long one. Moreover, successful Standards are those which respond to commonly perceived needs or aid in overcoming common problems. In terms of offering workable, compromise solutions, they must be based on some solid platform of accepted facts and acceptable practices.

With Language Engineering, we have a relatively immature field. It is not possible to propose Standards across the board in such a circumstance. It is not even appropriate to consider Standards for narrowly defined areas of Language Engineering, insofar as such areas involve interpretation of abstract knowledge about language: research is still confronted with hard problems about the formal nature of language in general and languages in particular, as used in many varied communicative situations, by people from different linguistic backgrounds, and has not arrived at any deep understanding or common approach (Bates and Weischedel, 1993). Equally, there are hard problems to be met at the computational level, when one tries to find flexible, expressive computational means to support discovery, formulation and formalisation of knowledge about language, means to constrain general purpose machines to interpret and apply such knowledge as language experts to language processing tasks and means for users of resultant systems to participate appropriately in these tasks. One should also not forget that, whereas in a text-based world, messages conviently appear in a recognisable and tractable (although not necessarily understandable) form, in a speech-based world there are severe problems in simply knowing what parts of a complex acoustic signal, possibly produced by many sources of air vibration, convey a message. The degree of variation in speech behaviour from individual to individual and even for one individual over time is moreover immense, which compounds the problem. Thus, looked at from the perspective of how close one is to achieving fully robust, flexible, accurate language processing technology, the field is indeed immature and not at all ready to propose Standards of the International Standard variety, except in areas associated largely with the physical handling of data.

However, the field is nevertheless highly active and has met with modest success in several areas. These successes have indeed led to a welcome transfer of knowledge to industry, for example through EC sponsored programmes such as ESPRIT, LRE, MLAP and LE. There is a growing number of companies specialising in natural language processing and/or speech processing.

However, although industry is capable of producing language engineering applications, the widespread development and adoption of such applications is threatened due to the lack of standards in the domain.

EAGLES was set up to ameliorate this situation, through bringing together representatives of major collaborative European R&D projects in relevant areas, to determine which aspects of our field are open to short-term *de facto* standardisation and to encourage the development of such standards for the benefit of consumers and producers of language technology. This work is being conducted with a view to providing the foundation for any future recommendations for International Standards that may be formulated under the aegis of ISO.

At the very least, EAGLES work towards *de facto* standards will allow the field to establish broad consensus on key issues, providing thus an opportunity for consolidation and a basis for technological advance and expansion of knowledge.

3.2.2 Basic concepts

3.2.2.1 Need for infrastructure linguistic resources

In the EU in particular, it is felt that progress in NLP and speech applications is hampered by a lack of generic technologies and of large-scale language resources, by a proliferation of different information formats, by variable linguistic specificity of existing information and by the high cost of development of resources.
In this section, we look briefly at why large-scale infrastructure resources are necessary to enable further progress in development and delivery of language engineering products.
Examples of such infrastructure resources include dictionaries, corpora, test suites and grammars. With the maturation of language technology, there come concerns about such matters as scaling-up, robustness, coverage, task suitability and accuracy.
It is generally true to say that language engineering technology was until very recently largely developed according to a methodology which involved:

- Extensive reliance on the intuitive knowledge of humans (linguists, lexicographers, phoneticians);
- Concentration on development of descriptions of either general language or of specialised subsets of language for some particular domain and purpose, with little commonality among any such descriptions;
- Limited reference to small-scale linguistic resources.

The many systems produced according to such a methodology tend to fall into one of two classes:

1. Good in a highly limited area, but unable to expand to any other;
2. Able to tackle any area, but with hugely variable results, often on the poor side.

Lengthy experimentation with systems of the second type can lead to the discovery of areas in which they may be of more use than others. Much depends on the purpose to which systems are put, e.g. one may be satisfied with poor performance where one is not overly concerned with high quality.
The development of systems based on linguistic theory has mitigated to some extent many of the problems that were caused by earlier systems whose *ad hoc* linguistic descriptions were a cause of much erratic behaviour. However, adoption of a theoretically based approach does not in itself guarantee a robust and accurate system, if the data one is modelling in one's theory are limited in extent and variety — or remain undifferentiated even though they reflect different language behaviour.
Today, it is accepted that language technology must be based on adequate linguistic knowledge about the types of messages to be processed. This implies deep and detailed knowledge about language structure, vocabulary, the role of context, collocation, discourse, stress, intonation, pronunciation, interaction in conversational situations and a host of other aspects of language behaviour, not forgetting those aspects concerned with multilingual communication.
In the last few years, therefore, there has been a rapidly increasing interest in building and sharing very large scale linguistic resources for both speech and natural language processing.
Resources such as corpora reflect actual usage of language in a variety of situations, for a variety of purposes. They provide documented evidence of language behaviour. As such, they are primary resources compared to, for example, dictionaries or grammars which are secondary, derived resources. Those working in speech have been accustomed to using corpora for system training purposes for some time now. In natural language processing, extensive use of corpora is of recent date. In lexicography, there is a long tradition of corpus-based work for a small number of languages. Only recently, however, has the need for very large-scale corpora come to the fore.

It is important to note that there is a pressing need not only for text corpora, but also for spoken language corpora. This latter type comes in many different varieties (as do text corpora). Given the amount of speech data — with associated derived annotations and analyses — required to support the development of a speech system even for some limited task, it is clear that the speech field, as the natural language field, is heavily reliant on corpus resources for successful progress.

Thus, no matter whether one is engaged in symbolic or statistically based approaches to Language Engineering, or is engaged in speech or natural language R&D (or both, as is increasingly common) large corpora are seen today as necessary resources for the acquisition of knowledge about language and language behaviour.

Computational lexicons represent another area where a pressing need is felt for the provision of large scale resources. Almost every language technology system requires a lexicon. Today's linguistic theories place great stress on the lexicon: much detailed information is required. It is expensive to produce such resources with the level of detail required. Attempts have been made to process existing publishers' machine readable dictionaries to extract the required information, however, as these are invariably oriented towards human needs and contain relatively unformalised information, success has been variable.

Language technology systems that are intended to process real messages in everyday working environments have a high requirement to have, insofar as feasible, a fully up-to-date dictionary that offers exhaustive coverage of the target domain in the depth required to support linguistic processing.

Thus, large scale lexicons are required, from which subsets of data may be selected for particular applications. There is an intimate link between corpus and lexicon: in order to have a reasonable lexicon, one must base it on a corpus. This does not preclude the augmentation of information through human intuition. The volumes of data and information involved are such that it is difficult for any one body to identify, collect, process, select, augment, disseminate and maintain lexical information to support language technology. It is quite impossible for the type of small to medium sized company that, in Europe, is the typical provider of language technology, to engage in endeavours related to linguistic resources on this scale.

For several years now, the lexicon has been recognised as the major bottleneck preventing widespread development of language technology. This is as much true for natural language processing as for speech processing.

Furthermore, as part of product development, producers must be able to assess and evaluate their language technology products. End-users have similar needs to determine how useful and cost-effective some product may be for their purposes. Such requirements thus lend further motivation to establishing corpora and other types of collections of data, for test purposes. A particular kind of testing utility is the test suite: an exhaustive systematic compilation of instances of language behaviour against which the performance of a system may be tested.

There is thus a widespread need for large scale infrastructure resources, for many reasons. This need is directly related to the work of EAGLES: without standard means of designing, constructing, manipulating, accessing and disseminating such resources, little progress will be made towards meeting this need. At the heart of this whole area lies the notion of reusability, to which we turn next.

3.2.2.2 Reusability

Interest in reusability arose due to the significant costs involved, for industry, in building the large-scale infrastructure resources that are necessary for the development, testing and maintenance of successful language technology products.

The cost and effort of developing large-scale language resources and the tools to manipulate them are high. It is uneconomic to build a resource for the development of one application only, or a sophisticated tool that is tied to a particular resource. There is thus great interest in the design of resources and tools such that these can be reused. A large lexicon, for example, richly annotated with linguistic information to support speech and natural language processing, is only worth building if it can be used to support a range of applications both now and in the future. If one has a small resource, one might prefer to acquire extra information from elsewhere, if possible, rather than duplicate existing work at perhaps greater cost. However, such acquisition can only take place if the information desired can be accessed and readily interpreted.

At the core of this problem lies the crucial notion of *reusability*. There are two interpretations of reusability which are highly relevant in Language Engineering:

1. The *accidental* property of being usable for a purpose other than that originally intended;
2. The property of being deliberately designed to be usable for multiple purposes.

An example of a reusable resource under the first interpretation is a publisher's machine readable dictionary, which is conceived originally for human use but whose contents are processed in order to extract information for the dictionary of a natural language processing system.
An example under the second interpretation is a lexicon designed to support theory-based language engineering applications, whose contents may be mapped into particular monolingual or multilingual application dictionaries expressed according to the principles of some particular theory. Such a dictionary is variously referred to as a *neutral dictionary* or a *polytheoretical dictionary*[1].
This dual nature of reusability was first articulated by J. M^{c}Naught at a meeting in Dublin in 1987 with S. Perschke and A. Zampolli, in relation to lexical resources. For further discussion of reusability, see Calzolari (1990) and Heid and M^{c}Naught (1991).
It is apparent that, given the volumes of lexical data to be processed for even a modest resource and the time and cost involved, reusability, in the second sense, is effectively a mandatory requirement for any future resource. Furthermore, the following is also apparent:

- Collaborative construction and sharing of major resources is highly desirable — no one body can hope to develop (and indeed thereafter maintain) very large-scale resources;
- The many small to medium sized firms (in the EU, especially) engaged in supplying language applications are typically unwilling and unable to supply the requisite lexical resources to underpin such applications — they however have an interest in acquiring data from elsewhere which could be reused in their applications;
- End-users of language technology are likewise interested in flexible acquisition of lexical data from possibly numerous resources.

3.2.2.3 Standardisation

Language Engineering is a difficult area which is, moreover, in a state of transition, as designers move from largely *ad hoc* systems to systems based on the results of research into linguistic theory. We are thus faced with the need to represent language data and knowledge about language at various levels of abstraction, as well as with the need to interpret information within some theoretical framework. Substantial work has already been done concerning standardisation in relation to physical representation. Indeed, standardisation efforts promoted over the last decade by the IT industries and by international and European standards bodies, such as ISO, CCITT and CEN–CENELEC, have led to stable means for the encoding of data. Widely used norms exist for the representation of textual data, covering multilingual character sets up to powerful metalanguages for document structuring and description. The US–European Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 1994) has made considerable progress in tailoring such metalanguages to the many types of electronic text.
What is far more problematic, especially when faced with competing theories, is how we agree on a core set of basic notions, and then how one *interprets* the meaning of some physical representation which ultimately refers to some abstract entity: for example, a linguistic label, whose name has been agreed upon — different theories have different views as to the meaning to be attached to this label. We are thus faced with a dual problem when attempting to set up standards in Language Engineering: we have to deal with physical format as well as with interpretation of the format. Moreover, we have to deal with various types of information. Linguistic knowledge is traditionally expressed according to linguistic level: phonetic, phonological, morphological, syntactic, semantic, pragmatic and so on. Each level has many theories or approaches associated with it. Our knowledge of each level differs in terms of completeness and adequacy and our ability to formalise what we do know. No level can be said to be completely understood and formalisable. In general,

[1]There is some disagreement in the field over terminology here as may be expected in a new, volatile subject area. The force of the concept is however not affected.

more confidence is felt regarding our ability to handle 'lower' levels such as phonetics and morphology, whereas, for the 'higher' levels such as semantics and pragmatics, there is a feeling that much research still needs to be done to provide even minimally workable solutions. This is a gross characterisation: the detailed picture is much too complex to discuss in depth here, especially when one thinks of approaches which cut across traditional linguistic level boundaries and theories which take a holistic view of language.

There are also needs to integrate products in present or future language technology environments, to interface them with various types of information system, to expand them to different domains and languages and also to maintain and, indeed, further develop them.

3.2.3 Recent history

Concerns about standardisation and reusability came to prominence in the mid 1980s, particularly among those working on lexical issues in centres carrying out research in natural language processing. Contemporary trends in linguistic theory made it clear that the lexicon would be the keystone of linguistic description for the foreseeable future. As confidence grew in being able to build large computational grammatical descriptions of language, concern also grew over their associated, much larger, lexical descriptions: if one were to change, for example, one's grammatical theory, how could one cope with the knock-on effects on the lexicon? As the lexicon is expressed in the same theoretical framework, a reworking of potentially thousands of lexical entries would be, for most projects, prohibitively expensive. This is all the more true in a European context where we have to deal with numerous languages and the mappings between them. In a grammar-oriented view of the time, formally, a lexicon was but a convenient place to gather data about the terminal elements of a grammar: in reality, there is no separation, formally, between a grammar and its terminal elements.

In the early 1980s, work had begun on extraction of information from publishers' machine readable dictionaries. Although the bulk of this work was undertaken for natural language processing purposes, there was also much interest in using such extracted information for psycholinguistic studies, for information retrieval purposes and for other tasks where lexical knowledge is of importance. Much of this extraction work was conducted for *ad hoc* purposes and there was little need felt for cooperation or for common frameworks. Such processing of publishers' machine readable dictionaries is still very much with us in the 1990s, although it has evolved strong concerns with reusability of the results.

However, it is only since the mid 1980s that standardisation and reusability issues have taken on increasing importance, and not just in the field of the lexicon. Nevertheless, it was the lexicon field which provided the impetus for such issues to be addressed in other fields.

The seminal event which set in motion a number of efforts towards standardisation and reusability took place in 1986: a workshop on *Automating the Lexicon: Research and Practice in a Multilingual Environment*, otherwise known in the field as *The Grosseto Workshop*, organised by N. Calzolari, LṘolling, J.C. Sager, D. Walker and A. Zampolli (Walker, Zampolli and Calzolari, 1995). The serious nature of this workshop is evident from the list of its sponsoring bodies: the European Commission (EC), Consiglio Nazionale delle Ricerche (CNR), University of Pisa, Association for Computational Linguistics (ACL), Association for Literary and Linguistic Computing (ALLC), Association Internationale de Linguistique Appliquée (AILA) and European Association for Lexicography (EURALEX). Participation was by invitation and was organised to provide a mix of linguists, lexicographers, lexicologists, computational linguists, artificial intelligence specialists, cognitive scientists, publishers, lexical software producers, translators, funding agency representatives and representatives of professional associations, drawn from several European countries, North America and Japan. The topic of the *Computational Lexicon* was thus tackled from a broad perspective. Numerous recommendations arose from this workshop (cf. Walker et al. (1995) and Zampolli (1991)). These in turn led to a common statement that high priority should be assigned to the design and development of *large, reusable, multifunctional, precompetitive, multilingual linguistic resources*.

An explosion of follow-up meetings took place between 1986 and 1988, including:

- Workshop on the Lexical Entry (New York, 1986);
- Panel on *The Lexicon in a Multilingual Environment*, COLING–86;

- Conference on *Standardization in Lexicography*, sponsored by the European Science Foundation, Saarbrücken, 1986;
- Conference on *Advances in Lexicography*, Center for the New OED, Waterloo (Canada), 1986;
- Session on *Words and World Representations*, Theoretical Issues in Natural Language Processing, Las Cruces (New Mexico), 1987;
- Workshop on *The Lexicon in Theoretical and Computational Perspectives*, Summer Linguistic Institute, Stanford, 1987;
- Summer School on *Computational Lexicology and Lexicography*, Pisa, 1988.

Similar events have taken place in increasing numbers since then.
In addition, as a direct result of the Grosseto Workshop, several groups came into being:

1. *Ad hoc* Working Group on the feasibility of a neutral or polytheoretical lexicon (the 'Pisa group');
2. Specialist Working Group on *Dictionaries and the Computer*, sponsored by EURALEX;
3. Special Interest Group on the Lexicon (SIGLEX), sponsored by the Association for Computational Linguistics.

The first of these is especially important, as it brought together for the first time representatives of the then major contemporary schools of linguistics, to investigate in detail the possibility of representing the linguistic information frequently used in parsers and generators (i.e. the major syntactic categories, subcategorisation and complementation, verb classes, nominal taxonomies, etc.) in such a way that it could be used and reused in various theoretical frameworks. Not only did this group look at different linguistic syntactic theories and their points of convergence and divergence, but it also looked into how such theories could handle phenomena in several different European languages.
The Grosseto recommendation to give priority to the design and development of large, reusable, multifunctional, precompetitive, multilingual linguistic resources has not, as yet, yielded any such resource. This is not surprising, given the degree of cooperation and level of funding needed for such enterprises. To some extent, however, there has been hesitation over the need to develop very large scale reusable resources. There are practical organisational questions to be answered regarding funding, running, access modalities, acquisition policy and so forth. There are fundamental questions to be answered regarding choice of representation for data, interpretation of representation, which levels of linguistic knowledge to recognise and how these relate within and across language descriptions. Associated questions arise regarding the nature of tools for acquisition of and access to data. The strong requirement for reusability has affected in some way or another every discussion about all such issues. Reducing the problem to a simple one of money, no-one will spend a great deal of money on building a very large scale resource if it cannot be guaranteed to be reusable, even if the case for the existence of such large resources is made and accepted. This is especially true in a European context, where little money has been available for language engineering in comparison to what is spent on e.g. certain aspects of agriculture, defence or other areas of information technology. Large-scale infrastructure projects tend to be less favoured, financially, than applications-oriented projects, especially when there is a demand for rapid development of applications. However, most language engineering applications can only be successfully developed and delivered if there is an adequate linguistic resource infrastructure in place. This fact, well-known to the community, has also been recognised for many years by the EC and by its advisory bodies.
One of the most recent confirmations of the key role of infrastructural linguistic resources is to be found in the mid term review carried out for the EC's Telematics programme:

> *Language Engineering requires an infrastructure of a type that is unusual in other subjects [...] Without a major attack on these infrastructural needs, applications engineering will make little progress in the language engineering field.*

(Oakley, 1993, 70)

Bates and Weischedel (1993) also place great emphasis on the role of infrastructure for language processing. It is to the credit of the EC that it took a strong interest in encouraging developments in the area of reusable linguistic resources, from an early date. Given the nature of EC programmes and the internal organisation of the EC, four projects were initiated in various frameworks at the beginning of the 1990s, looking at aspects of the reusability area. Almost immediately, these projects began to discuss ways and means of cooperating to reach common objectives, under the aegis of the EC.

As it was clear that one could not immediately undertake the development of very large scale reusable linguistic resources, these projects were all concerned with preparatory or enabling aspects:

ESPRIT BRA ACQUILEX: Basic Research Action on acquisition of lexical knowledge from machine readable dictionaries and its representation in a common framework;

EUROTRA–7 Study: Feasibility of designing large scale reusable lexical and terminological resources;

ESPRIT MULTILEX: Industry-led research and development project aiming at providing specifications of standards for multilingual lexicons.

MLAP NERC: Study for the design of a Network of European Reference Corpora.

At roughly the same time, a major project was launched within the EUREKA framework, namely GENELEX. EUREKA projects are collaborative projects involving predominantly industrial consortia, partially funded by European governments. GENELEX was set up to establish a convergence path for various lexical assets owned by the consortium into a common conceptual structure. It has been concerned with many of the issues regarding reusability that exercised EUROTRA-7 and MULTILEX, although it was not initially conceived as a standards-related project.

It can be appreciated that ACQUILEX (Calzolari and Briscoe, 1995) was concerned primarily with reusability in the first sense we have identified, whereas EUROTRA–7 and MULTILEX were concerned with reusability in the second sense, and GENELEX is concerned with both.

EUROTRA–7 (Heid and McNaught, 1991) in particular deserves mention. Its goal was to assess the feasibility of designing large scale, reusable lexical and terminological resources. A broad range of different possible sources of lexical material, and in particular of different applications requiring a lexical component, was investigated, with attention focussing on different theoretical frameworks, different needs with respect to granularity of information, depth and coverage of description, and so on, to provide an account of a variety of diverging and converging needs.

An important observation of EUROTRA–7, which led to a methodological recommendation for future actions towards developing specifications for reusable linguistic resources, was — stated in very simplistic terms — that different theories describe essentially the same facts, but make different generalisations and use different descriptive devices. The methodological claim deriving from this was, on the one hand, that one should go back to the most fine-grained, observable differences and phenomena, i.e. to reach an extreme atomisation of linguistic observations, and, on the other hand, that one should aim for complete explicitness of descriptive devices and should provide explicit and reproducible criteria for each observable difference taken into account.

This shared layer of granular observations would then constitute a common data pool, accounting for atomic facts, subclassified according to explicit criteria, and represented in some problem-oriented high-level formalism (a typed feature based language was favoured). The global *reuse scenario* defined within EUROTRA–7 foresaw such a common data pool being at the heart of a model composed of three major areas: acquisition, representation and application. Within this model, lexical data are represented in terms of the common data pool. There is an interface, through specific acquisition tools, to sources of lexical information, and another interface, through a compiler/interpreter level which takes care of *recombining* specific descriptive devices and of filtering relevant information, to individual application lexicons which need application specific lexical descriptions.

EUROTRA–7 was thus important in terms of carrying out work towards initial specification of a model for a reusable lexicon. In addition, this study made numerous recommendations regarding the need for work towards standards, the importance of bringing industry and academia together in collobarative ventures and the importance of setting up a forum to share knowledge and to tackle issues of community-wide relevance together.

Overlapping with EUROTRA–7, the ESPRIT MULTILEX project (MULTILEX, 1993) was concerned with proposing standards for multilingual lexicons for natural language processing. Emphasis was placed on means of exchanging and sharing lexical data. This led to work not only on architectures for lexicons, but also to work on the various possibilities for exchange, together with their various implications, for example:

- Exchange via some central lexicon (common data pool);
- Exchange via a standard exchange format, independent of some central lexicon.

These can be seen as variations on a theme. The crucial area came to be seen as the development of standard ways of encoding and interpreting lexical data. The difficult aspect has always been obtaining agreement on the interpretation of some encoding: we may agree on a name for some object, but disagree (if indeed we realise there is disagreement) over the interpretation of that named object.

MULTILEX was able to carry forward work started during EUROTRA–7. In particular, it went into greater detail regarding levels of linguistic knowledge. A traditional division of the lexicon was followed, i.e. into levels ranging from orthography and phonology through morphosyntax and syntax to semantics. An additional pragmatic level was incorporated, running orthogonal to the other levels (i.e. pragmatic information was essentially distributed over all other levels). Each level was investigated to determine whether it was ripe for standardisation — and if so to what extent standardisation could take place, in what expected timeframe, etc.; if not, to determine what other efforts might be envisaged to push towards standardisation.

The results of MULTILEX served to confirm what had earlier been suspected (via EUROTRA–7 and the work of the *Ad Hoc Working Group on the Neutral Lexicon*): there was room for agreement at the levels of morphosyntax and syntax, however at higher levels (semantics) the field was neither mature enough nor rich enough, in terms of workable, generally acceptable, widely used semantic models, to allow practical standardisation. Despite its short timespan, the standards-related phase of MULTILEX was able to propose a reasonable inventory of standard labels for morphosyntax, together with a preliminary design for an architecture capable of supporting a multilingual lexicon. Although MULTILEX provided definitions for the inventory of labels and examples of their use within a typed feature formalism, this work resulted in only a partial solution to the reusability problem: what was missing was a full lexical specification, in that there was no link from the inventory to any underlying pool of subclassified linguistic phenomena at the atomic level of minimal observable difference.

However, the positive results of MULTILEX, related to standards, were that it helped to advance our knowledge; it enabled motivated decisions to be taken about which areas were worth pursuing in terms of standardisation; it provided a preliminary design for a reusable multilingual lexicon; it allowed collaboration to be established between industry, academia and appropriate funding bodies; and it offered an inventory of labels as a starting point for future work. It should be pointed out that MULTILEX, in a subsequent phase, experimented with the design and implementation of tools and lexical databases appealing to the proposed lexical standards for various purposes, work which has provided valuable insights for later projects.

GENELEX (Antoni-Lay, Francopoulo and Zaysser, 1994) has taken a slightly different view of the reusability issue, by aiming at developing a generic, application-independent model of a lexicon, thoroughly documenting it and providing explicit guidelines on its use. The GENELEX lexicon is 'theory-welcoming' in that it can adapt to competing theories, via its associated formalism. Various tools are being developed to aid in management of GENELEX dictionaries and large-scale dictionaries are being built.

GENELEX, MULTILEX and EUROTRA–7 therefore had much in common. Importantly, there was also a degree of overlap among these projects in terms of participants. Given their common interests and the growing importance of the need for standards, they came together to form a Coordination Group, under the aegis of the EC.

NERC (Network of European Reference Corpora) (Calzolari, Baker and Kruyt, 1995) was launched in 1992 with the objectives of making recommendations to the EC about the future of language corpus provision in Europe, seeking to establish Europe-wide cooperative actions in relation to large-scale resources, designing a distributed repository of such resources in the European Union, experimenting with means of collecting and disseminating such resources via networks and appropriate media, and specifying an infrastructural network of large-scale resources on the basis of existing EU networks and organisations. Thus, NERC has become a focal point for corpus-related work in the Union.

Although there have been many projects in the area of natural language processing that have been foundational or enabling projects with respect to the present initiative, there has also been highly significant input from the speech community, in the form of the ESPRIT SAM project (Speech Input/Output Assessment Methodology and Standardisation) (SAM, 1992) This significant project covered many aspects of speech and resulted in a wealth of documents on standards related aspects (housed at UCL, London) and numerous standard tools for speech (notably the SAM Workstation), among other valuable contributions.

On the basis of SAM alone, there is every justification for claiming that the achievement of the spoken language community in the EU in developing solid codes of working practice for multinational, multilingual projects is unparallelled.

The situation sketched above created the appropriate conditions for the launching, firstly, of a project definition study in 1992 and then, in February 1993, of the EAGLES project, aiming specifically at defining standards or preparing the ground for future standard provision.

The areas of concern to EAGLES are text corpora, computational lexicons, grammar formalisms, evaluation and assessment, and spoken language. For each area a core working group (WG) has been established, where leading experts of both the research and the industrial communities are represented, combining their efforts towards the development of a common basic European infrastructure and agreed linguistic specifications.

4 EAGLES working strategy

4.1 EAGLES methodology

The basic idea behind EAGLES work is for the group to act as a catalyst in order to pool concrete results coming from current major European projects. Relevant common practices or upcoming standards are being used where appropriate as input to EAGLES work, particularly in the areas of lexicons, text encoding and speech. Numerous theories are being taken into account, where appropriate, as any recommendation for harmonisation must take into account the needs and nature of the different major contemporary theories. EAGLES is also drawing strong inspiration from the results of major projects whose results have contributed to advancing our understanding of harmonisation issues.

The major efforts in EAGLES concentrate on the following types of activities, which, as seen in the following, show how, along very general lines, the work is organised in the five working groups.

- Detecting those areas ripe for short-term standardisation *vs.* areas still in need of basic research and development;
- Assessing and discovering areas where there is a consensus across existing linguistic resources, formalisms, speech systems and common practices;
- Surveying and assessing available proposals or contributed specifications in order to evaluate the potential for harmonisation and convergence and for emergence of standards;
- Proposing common specifications for core sets of basic phenomena, recommendations for good practice, for standard methodologies, etc., on which a consensus can be found;
- Setting up guidelines for representation of core sets of basic features, for representation of resources, etc.;
- Feasibility studies for less mature areas;
- Suggesting actions to be taken for a stepwise procedure leading to the creation of multilingual reusable resources, elaboration of evaluation methodologies and tools, etc.

4.2 Interaction of EAGLES and other projects: impact of EAGLES recommendations

By its very nature, EAGLES must interact closely with the scientific and industrial R&D community not only in Europe, but world-wide.

Since EAGLES involves many bodies active in European NLP and speech projects, close collaboration with these projects is assured and, significantly, in many cases, free manpower has been contributed by the projects, which is a sign of both their commitment and of the crucial importance they place on reusability issues. Procedures have been established allowing EAGLES to access relevant material developed by EAGLES participants working in other projects.

The most relevant impact of and feedback to EAGLES work comes from those large projects which are applying the EAGLES guidelines, or comparing and mapping EAGLES recommendations with their own specifications. Feedback from these sources comprises the most important *evaluation* of EAGLES results, since the real impact of EAGLES activity must be measured by its acceptance and use within projects more than by individuals.

A *large number* of European and national projects has already made extensive use of EAGLES results, actively interacting with on-going EAGLES work in a reciprocal and cyclical process of proposing, testing, evaluating, providing feedback and revising. As an example, the EAGLES morphosyntax proposals (EAGLES, 1996a), based on extensive comparison of approaches (EAGLES, 1996b) are already being applied — and consequently tested and evaluated — in a number of national and European projects, such as LRE DELIS, RENOS, CRATER, MECOLB, MULTEXT, COPERNICUS MULTEXT–East and TELRI, MLAP–PAROLE, ESPRIT–ELSNET, French GRACE, German Textcorpora und Erschliessungswerkzeuge, LE–SPARKLE, ELRA, EUROWORDNET and PAROLE. This interaction — in both directions — between EAGLES, which is an infrastructural type of action, and a large number of topic-specific R&D projects and applications, is an essential component of the EAGLES overall work style and has been very fruitful. This feature will be enforced and strengthened in future EAGLES work.

Standards must emerge from state-of-the-art developments, and as such they are not to be imposed. Consolidation of a standards proposal must be viewed, by necessity, as a slow process and, by definition, as a non-innovative action, comprising, in our case, after the phase of putting forward proposals, a cyclical phase involving EAGLES external groups and projects with:

- Careful evaluation and testing by the scientific community of recommendations in concrete applications;
- Application, if appropriate, to a large number of European languages
- Feedback on and readjustment of the proposals until a stable platform is reached, upon which a real consensus — acquiring its meaning by real usage — is arrived at;
- Dissemination and promotion of consensual proposals.

What can be defined as *new* in this long process is the highlighting of the areas for consensus (or of the areas in which consensus could be reached) and the gradual consciousness of the stability that evolves within the communities involved.

A first benefit is the possibility, for those entering the field, of focussing their attention on as yet unsolved problems without losing time in rediscovering and reimplementing what many others have already worked on. This is the only way our discipline can really move forward, but this entire process is not without dangers or possible objections.

4.3 Overall cost-effectiveness of EAGLES

Work that otherwise would have been expensively distributed and duplicated over several projects (and that probably would not have thus been adequately accomplished) has been done just once within EAGLES in a collaborative manner, at a reduced cost. A great deal of effort has thereby been provided for a relatively small investment.

Projects currently have a framework, with EAGLES, that allows consensus to be built with broad participation, and that can sensibly handle feedback from the wider LE community. EAGLES has been able to synthesise proposals and to provide recommendations, methodologies that can subsequently be tested out and applied in RTD projects and has ensured the beginning of a convergence of various projects in using common guidelines, so that their results are or will be reusable and exchanges may be accomplished, all with large-scale savings.

5 EAGLES organisation

The structure of EAGLES has resulted from recommendations made by leading industrial and academic centres, and by the EC's Language Engineering strategy committees.
A very large number of EU research centres, industrial organisations, professional associations and networks across the EU are providing labour, for free, towards the common effort: *more than 100 sites* are involved in the different EAGLES groups or subgroups of which approximately 40% from industry distributed throughout all the Working Groups. However, it was found that, especially in areas where material had to be gathered, compared and synthesised, where feedback had to be assimilated and integrated, and where major revision of proposals had subsequently to be carried out to achieve wider consensus, the work was both time-consuming and labour-intensive and therefore had to be paid for.
In human terms, there is now a large community available that has learned how to work together towards standardisation objectives, is highly motivated to do so and has good industrial participation in key areas of spoken and written LR.
EAGLES has engendered a broad sense of cooperation, and has offered a forum in which *both* users *and* developers can together arrive at workable recommendations in Language Engineering. Users and developers have contacted EAGLES to get early information on guidelines, to offer advice and collaboration.
With financial support from the EC's LRE Programme, coordination is being carried out by the Consorzio Pisa Ricerche (Italy), while five working groups execute the detailed work.

5.1 EAGLES bodies

EAGLES consists of

- A management board;
- Five working groups, hosted by designated R&D centres; and
- A central support team.

5.1.1 Management board

The management board is charged with overall coordination and administration; it provides a forum for scientific exchange among EAGLES organisations, supervises the working groups and endorses and promotes their results.
At present the constitution of the Board reflects a cross-section of interested European companies and academic centres, including Cap Gemini Innovation (F), Siemens Nixdorf (D), GSI ERLI (F), EUROLANG (F), Electronic Publishing Partners (D), Vocalis Ltd (UK) and the universities of Edinburgh (UK), Stuttgart (D) and Pisa (I), representing RTD projects, and also coordination bodies such as ELSNET, EACL, ESCA and FOLLI.
Prof. Ch. Rohrer (University of Stuttgart) is the current chairman and Dr. J. Peckham (Vocalis Ltd, Great Shelford) vice-chairman.
The various projects represented in the management board have agreed to provide:

- Background material as input for EAGLES in terms of results arrived at in their projects; and
- Expertise through individual members in the working groups.

5.1.2 Working groups

The work towards common specifications is carried out by the five working groups. These are concerned with firstly establishing common methodologies for the five areas of concern within EAGLES, in order to subsequently arrive at *de facto* standards. The table below gives information on the working group hosts, chairpersons and WG editors editors.

WG	Host	Chair	WG Editors
Text Corpora	Instituto Cervantes Madrid, Spain	A. Zampolli (I)	J Llisterri (E)
Computational Lexicons	GSI-ERLI Paris, France	A. Zaenen (F)	U. Heid (D)
Grammar Formalisms	DFKI Saarbrucken, Germany	H. Uszkoreit (D)	J. Capstick (D) G. Erbach (D)
Evaluation and Assessment	Center for Sprogteknologi Copenhagen, Denmark	M . King (CH)	C. Mazzi (DK)
Spoken Language	Vocalis Ltd Great Shelford, UK	R. Moore (UK)	D. Gibbon (D)

Each working group (WG), in carrying out its work, has set up a number of small subgroups (SGs), each corresponding to a different task. The number, size and lifespan of SGs vary according to the number and nature of the tasks.
In addition, there is a number of *cross-groups* to ensure consistency of approach between different working groups and the proper exploration of common topics.
The officers of the WGs are the chairperson, the task responsibles of SGs and the WG editors.
Details of working group membership and of all those who contributed in various ways to the creation of EAGLES material are to be found in the 'Contributors' section of individual documents.
The main functions of the chairpersons are to:

- Prepare/propose a programme of work, task and resource allocation (all of which are kept under constant review);
- Give scientific guidance;
- Monitor the document flow and production process.

The main functions of the task responsibles are to:

- Coordinate the work of their subgroup;
- Schedule contributions from subgroup members and prepare technical reports as specified in the workplans;
- Provide the WG chairperson with a schedule of dates by which the subgroup will deliver intermediate and final reports.

The main functions of the internal editors are to:

- Collate and integrate the various contributions stemming from the task responsibles;
- Compile front and back matter such as introductions and bibliographical references;
- Interact closely with the Chief Editors.

5.1.3 Hosting organisations

A host organisation is associated with each working group, to provide infrastructure support to the chairpersons, editors and WG members in fulfilling their tasks and, specifically, to:

- Manage the funds allocated to the group;
- Contribute to the group's work; by providing logistic and organisational support;
- Monitor the document flow;
- Gather and maintain material disclosed to the group by participating organisations or third parties.

5.1.4 Co-ordinator

The central support team (Director: A. Zampolli, CPR, Pisa) is responsible for ensuring cohesion of the distributed activities, for mediating between the participating bodies and for supplying administrative/technical support.
The central editorial board is composed of two co-Chief Editors: N. Calzolari (ILC, Pisa) and J. M^c^Naught (CCL, UMIST). The editorial board ensures overall coordination and harmonisation of documentation and also provides scientific and technical guidance where appropriate to ensure the efforts of the working groups remain focussed on EAGLES objectives.

5.2 International cooperation and dissemination

EAGLES has established links with other international initiatives whose aims and orientations are similar or closely connected to EAGLES, such as

- The Text Encoding Initiative (TEI), promoted by the three major professional associations in natural language processing (ACL, ALLC and Association for Computing in the Humanities (ACH));
- ELRA, which will make use of EAGLES recommendations and proposals for its validation actions.

A cooperative effort with the NSF has produced an international survey of the state-of-the-art in speech and natural language processing R&D, which is under publication (Cole et al., 1996) and already accessible on the Web.
The European Network of Centres of Excellence in Speech and Language (ELSNET) has been entrusted by the management board with the provision of dissemination facilities to EAGLES. The regularly appearing newssheet of ELSNET, *ELSNews*, carries mini-reports of EAGLES activities and plans.

6 Structure, contents and status of the *EAGLES Guidelines*

6.1 Status label definitions

Even though we are formally at the end of a phase of EAGLES activity, it must be clear that:

- The results stemming from this initiative are not at the same level of completion since some Working Groups began work in 1993 and their results have therefore undergone a wide process of testing, application, feedback and revision, whereas others began work only in 1995 and are therefore in a phase of first proposals; and
- Not all the Working Groups, and the Subgroups within them, had the objective of reaching a normalisation proposal, some of them having considered as possible – within the timeframe and person months effort available – only preparatory work in view of future normalisation activity proper.

We attach therefore to each document (or sometimes to parts of a document) a label describing its current status. The labels used are given in table 1.

Label	Description
R	Recommendations
P	Preliminary recommendations or first proposals
F	Formal specifications and explicit guidelines
V	Validation document
B	Background document
D	Data produced according to EAGLES recommendations
L	Links to related project documents

Table 1: Status labels of EAGLES documents

No labels are attached to summaries, introductions, conclusions, syntheses, etc. The definitions of these labels are, in more detail:

R: Recommendations — Applies to those parts where work has reached a stage allowing the formulation of recommendations concerning common specifications, definitions, terminology, etc. Usually these recommendations have already been widely publicised in the community, have been tested, applied, evaluated and accordingly revised.

P: Preliminary recommendations — Applies to those parts where preliminary recommendations —- sometimes rather general — could be formulated, but these are being made available outside EAGLES only now and therefore did not have the chance of being commented on, evaluated, tested, etc. We expect to have major feedback on these from now onwards and thus envisage revisions to them.

F: Formal specifications and explicit guidelines — Applies to those parts which provide formal, explicit guidelines and specifications for implementing the above recommendations. This typically involves taking broad recommendations and formalising them to the point of demonstrating how to instantiate them for specific languages.

V: Validation document — Applies to documents produced in testing EAGLES recommendations within EAGLES internal validation tasks, especially in relation to application-specific areas.

B: Background document — Applies to those parts which are at an earlier stage of development and where only survey or other preparatory work has been done, or where, in certain cases, individual views have been put forward as a trigger to consensus building.

D: Data produced according to EAGLES recommendations — Applies to linguistic data which have been produced, analysed or encoded according to EAGLES recommendations, outside EAGLES itself, by affiliated projects or other bodies collaborating with EAGLES.

L: Links to related project documents — Applies i) to links to Web or ftp sites of related projects, having contributed to EAGLES either with input or in the testing, application, validation phases, and ii) to particular documents of such projects of particular relevance to EAGLES.

6.2 Looking for feedback

Because of the very nature of EAGLES, which aims at developing consensual and broadly agreed common specifications, a large number of groups active in Europe in the different areas of concern were involved in building on the results of the major European projects to develop preliminary specifications or to establish common ground more precisely. Many of these groups were or are engaged in testing and validating our preliminary proposals or in elaborating further background material intended to provide a basis for additional proposals in a second phase of EAGLES.

From the outset, in EAGLES, there has been a commitment to obtaining feedback (within the time constraints of the project) from the wider community on all proposals agreed on by the group. For this reason, preliminary proposals of the various EAGLES Working Groups were made public at the end of 1994.

In addition to the main objective of disseminating and making available the EAGLES results for wide consultation and application, we remain committed to the purpose of soliciting feedback, comments, suggestions and criticism (negative and positive) on a body of material which represents work that is:

- In some cases still preliminary or in progress;
- In different states of development;
- Not always coherent or harmonised;
- Not complete with respect to known problematic areas.

The reader is therefore requested to take particular note of the nature of this body of material and to make due allowance. We certainly do not claim that all the material reaches what is normally thought to be an acceptable level for general publication. Certain material, for example, has been only lightly edited or revised, in our concern to place it before the community in timely fashion for feedback. We are thus very much

aware of inconsistencies, repetitions, lack of cross-references, differences in style and presentation, etc. The reader should accord low priority to commenting on such matters.
We are also aware that sometimes different views of some issue are put forward by different working groups (or indeed by different members of one subgroup). We see this as a normal part of the search for consensus. If the reader is able to contribute in part to helping resolve controversial issues, or can provide telling reasons to opt for one proposed preference rather than some other, this would be most welcome.
The reader is thus invited to offer feedback on *content-related issues* as a first priority. We are keen to hear the reader's views on:

- The feasibility and applicability of our current (sometimes preliminary) proposals;
- The relevance of proposals to your own area of interest and to the needs and interests of your organisation;
- Whether, in your view, an adequate level of consensus has been reached already on some issue, as a result of our proposals;
- What further work needs to be addressed to arrive at a consensus, if our proposals are deemed inadequate or incomplete;
- Whether there is any undue bias and, if so, suggestions as to how this might be corrected;
- Whether any major or widely-used approaches, theories, projects, resources, etc., that have not been taken into account so far, should be, *only if you deem these to be critical to any harmonisation effort* (we have not been exhaustive in our descriptions or assessments of projects, etc., where these latter merely added weight to our deliberations rather than any aspect crucial to overall harmonisation — moreover, we do not have the capacity or mission to effect exhaustive surveys);
- Whether you prefer some proposed approach to another (there are cases where, at least within EAGLES, there appears to be no strong preference for adopting one proposal rather than another);
- Any relevant standardisation or harmonisation activities that should be taken into account but have not been;
- Possible terminological infelicities or conceptual misconceptions;
- What issues you would like to see addressed by some future EU standardisation effort.

6.3 Structure of the *EAGLES Guidelines*

The logical structure of the EAGLES Guidelines is conveyed in visual form in the model diagram of figure 1 (here representing the Computational Lexicons WG). The Guidelines are hierarchically structured according to the different status of their parts, and divided by Working Group (WG): Text Corpora WG, Computational Lexicons WG, Evaluation WG, Computational FormalismsWG, Spoken Language WG.

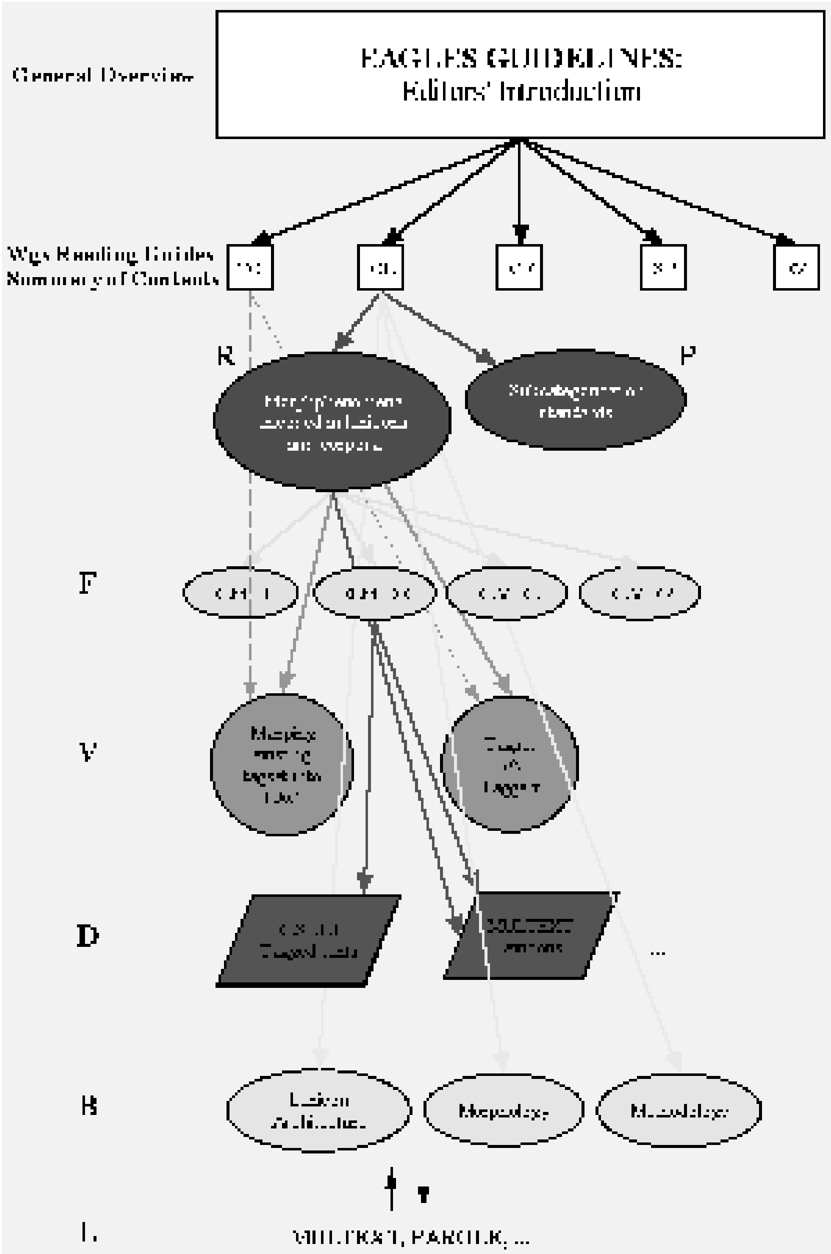

Figure 1: Logical structure of *EAGLES Guidelines*

Each Working Group has adopted a slightly different structure for the set of documents produced. The differences are determined by a number of factors, such as dissimilarities in the field covered, partially differing strategies of working, partially differing types of target readers/users.
All of the WG report(s) are headed by an *Introduction*, written by the WG Editor or Chairperson, which provides a Reader's Guide to the documents of the Working Group, similar to what we present here for the *EAGLES Guidelines* as a whole. The remaining set of documents is structured into a number of varying levels, tagged with one of the status labels described in section 6.1 above. The lists of documents for each WG are given in the following sections.

6.4 Reports of the Text Corpora Working Group

Status	Title
	Reading guide
P	*Preliminary recommendations on corpus typology*
P	*Preliminary recommendations on text typology*
R/F	*Recommendations on Corpus encoding*
R	*Recommendations for the morphosyntactic annotation of corpora*
P/B	*Preliminary recommendations for the syntactic annotation of corpora*
P	*Preliminary recommendations on spoken texts*
B	*Reusability of linguistic software*
B	*Guidelines for linguistic software development*

6.5 Reports of the Computational Lexicons Working Group

Status	Title
	Reading guide
R	*Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages*
P	*Preliminary recommendations on subcategorization*
F	*ELM-DE: EAGLES specifications for German morphosyntax*
V **F**	*EAGLES validation task on tagset mapping* *ELM-EN: EAGLES specifications for English morphosyntax (Appendix)*
F	*ELM-IT: EAGLES specifications for Italian morphosyntax*
F	*ELM-FR: EAGLES specifications for French morphosyntax*
V	*Study of the relation between tagsets and taggers*
B	*Lexicon architecture*
B	*Computational lexicons methodology task*

6.6 EAGLES Evaluation of Natural Language Processing Systems report

Status	Part	Title
	Chapter 1	Reading Guide
P	Chapter 2	The Framework Model
B/D	Chapter 3	Related Work
	Chapter 4	Conclusions and Directions for Future Work
	References	
	A.	ISO Terms and Guidelines
	B.	Requirements Analysis for Linguistic Engineering Evaluation
	C.	Evaluation of Writers' Aids
	D.	Evaluation of Translators' Aids
	E.	User Profiles
	F.	Translation Tools
	G.	Feature Checklist Examples
	H.	A Practical Evaluation of Writers' Aids
	I.	Evaluation of Knowledge Management Systems
	J.	SdT: A Case Study
	Appendices	
	Bibliography	

6.7 Reports of the Computational Linguistic Formalisms Working Group

Status	Part	Title
	Chapter 1	Reading Guide
B	Chapter 2	State of affairs at the beginning of EAGLES
B	Chapter 3	New developments since the start of EAGLES
B	Chapter 4	Trends towards convergence
B	Chapter 5	Reuse of grammars
P	Chapter 6	Needs for industry and research
P	Chapter 7	Recommendations concerning formalisms for new NL projects
P/F	Chapter 8	Exchange formats for grammar formalisms
	Chapter 9	Conclusion and planned activities
	Bibliography	
B		*Report of the EAGLES Workshop on Implemented Grammar Formalisms*
B		*Report of the EAGLES Workshop on Linguistic Adequacy of Linguistic Formalisms for NLP*
B		*Report of the EAGLES Workshop on the Connection between CLP and Linguistic Formalisms*

6.8 EAGLES Handbook on Spoken Language Systems

Status	Part	Title
	Chapter 1	Reading guide
P	Chapter 2	System design
P	Chapter 3	SL Corpus design
P	Chapter 4	SL Corpus collection
P	Chapter 5	SL Corpus representation
P	Chapter 6	SL lexica
P	Chapter 7	Language models
P	Chapter 8	Physical characterisation and description
P	Chapter 9	Assessment methodologies and experimental design
P	Chapter 10	Assessment of recognition systems
P	Chapter 11	Assessment of speaker verification systems
P	Chapter 12	Assessment of synthesis systems
P	Chapter 13	Assessment of interactive systems
	Appendices	
	A.	Character codes and computer readable alphabets
	B.	SAMPA computer readable phonetic alphabet
	C.	SAM file formats
	D.	SAM recording protocols
	E.	SAM software tools
	F.	EUROPEC recording tool
	G.	Digital storage media
	H.	Database Management Systems (DBMSs)
	I.	Speech standards
	J.	EUROM-1 database overview
	K.	Polyphone project overview
	L.	European speech resources
	M.	Transcription and documentation conventions for Speechdat
	N.	The Bavarian Archive for Speech Signals

7 Comments and feedback

Comments and reactions to the above listed reports will be taken into account during the upcoming phase of EAGLES. We recommend that comments be sent using the *Comments Cover Form*. Comments should be sent — if possible by e-mail — to the Chief Editors and to the Working Group Editors according to the type of comments:

General comments concerning the EAGLES project should be sent to the Chief Editors;

Specific technical comments related to work done within one of the five working groups should be sent both to the Chief Editors and to the WG Editor of the Working Group concerned who will in turn deliver it to the appropriate member(s) of the WG.

Final versions of those documents that have been revised in the light of feedback will become steadily available from now on. Those documents that have not been widely commented on will be revised as feedback comes in and final versions will thus appear at a later date. Interested readers should check the EAGLES ftp or WWW servers to obtain news of progress regarding availability of final versions.

8 How to get EAGLES material

EAGLES material may be acquired by three main modes:

1. Connect to the ftp site at Pisa:

Address	:	ftp.ilc.pi.cnr.it (under construction)
(IP number	:	131.114.202.11)
Username	:	anonymous
Password	:	your E-mail address
Directory	:	eagles/reports

   Instructions on how to proceed will be given by the ftp server.

2. Connect to the WWW site at Pisa:

   http://www.ilc.pi.cnr.it/EAGLES96/home.html (under construction)

3. Send a message to:

   EAGLES Secretariat
   Istituto di Linguistica Computazionale
   Via della Faggiola 32
   Pisa
   I–56126 Fax: +39.50.589055
   Italy E-mail: eagles@ilc.pi.cnr.it

References

Antoni-Lay, M.-H., Francopoulo, G. and Zaysser, L. (1994). A generic model for reusable lexicons: The GENELEX project, *Literary and Linguistic Computing* **9**(1): 47–54.

Bates, M. and Weischedel, R. (eds) (1993). *Challenges in natural language processing*, Cambridge University Press, Cambridge.

Calzolari, N. (1990). Lexical databases and textual corpora: perspectives of integration for a lexical knowledge base, *in* U. Zernik (ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon*, Laurence Erlbaum, Hillsdale, NJ, pp. 191–208.

Calzolari, N., Baker, M. and Kruyt, T. (eds) (1995). *Towards a network of european reference corpora*, Vol. XI–XII of *Linguistica Computazionale*, Giardini Editori e Stampatori in Pisa, Pisa.

Calzolari, N. and Briscoe, E. (1995). ACQUILEX-I and -II: Acquisition of lexical knowledge from machine readable dictionaries and text corpora, *Cahiers de Lexicologie* **67**(2): 95–114.

EAGLES (1996a). Recommendations for the morphosyntactic annotation of corpora, *Eag–csg/ir–tr.1*, ILC–CNR, Pisa.

EAGLES (1996b). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora, *Eag-lwg-morphsyn*, ILC–CNR, Pisa. A common proposal and applications to european languages.

Heid, U. and McNaught, J. (eds) (1991). *EUROTRA-7 Study: Feasibility and project definition study of the reusability of lexical and terminological resources in computerised applications*, IMS, University of Stuttgart, Stuttgart.

MULTILEX (ed.) (1993). *Standards for a multifunctional lexicon*, CAP GEMINI INNOVATION for the MULTILEX Consortium, Paris.

Oakley, B. (ed.) (1993). *Telematics Programme: Mid-term review*, Commission of the European Communities, Luxembourg.

SAM (ed.) (1992). *Multilingual speech input/output: Assessment, methodology and standardisation. Final report, year three, 1 March 1991 – 28 February 1992*, number SAM-UCL-G004, ESPRIT PROJECT 2589 (SAM), London.

Sperberg-McQueen, C. and Burnard, L. (1994). *Guidelines for electronic text encoding and interchange: TEI-P3*, ACH-ACL-ALLC Text Encoding Initiative, Chicago and Oxford.

Walker, D., Zampolli, A. and Calzolari, N. (eds) (1995). *Automating the lexicon: Research and practice in a multilingual environment*, Oxord University Press, Oxford.

Zampolli, A. (1991). Technology and linguistic resources, *in* M. Katzen (ed.), *Scholarship and technology in the humanities*, British Library Research, London.

A EAGLES Comments Cover Form

```
EAGLES COMMENTS --- COVER FORM

Please fill out Parts A and B if you are MAKING COMMENTS.
Please attach whole form at head of comments.

A. COMMENTS MADE BY:

Name:                                  Date:
Affiliation:                                 (1-31)/(1-12)/199_
E-mail address:
Telephone:
Fax:
Postal address:


B. COMMENTS FOLLOW ON:
                                                 (delete as appropriate)

1. Text Corpora WG                                     YES     NO
   (SEND COMMENTS TO: tcwgeditor@tnos.ilc.pi.cnr.it)

2. Computational Lexicons WG                           YES     NO
   (SEND COMMENTS TO: clwgeditor@tnos.ilc.pi.cnr.it)

3. Computational Linguistics Formalisms WG             YES     NO
   (SEND COMMENTS TO: clfwgeditor@tnos.ilc.pi.cnr.it)

4. Evaluation of NLP Systems WG                        YES     NO
   (SEND COMMENTS TO: ewgeditor@tnos.ilc.pi.cnr.it)

5. Spoken Language Systems WG                          YES     NO
   (SEND COMMENTS TO: slwgeditor@tnos.ilc.pi.cnr.it)

6. General aspects of EAGLES                           YES     NO
   (SEND COMMENTS TO: ceditor@tnos.ilc.pi.cnr.it)

THANK YOU FOR YOUR TIME AND VALUABLE CONTRIBUTION TO EAGLES


C. EAGLES ACTIONS ON COMMENTS

1. Acknowledgement sent on:           by:

2. Action on comments:

   a. Dealt with by Chief Eeditors directly on:

   b. Dealt with by WG editor directly on:
```

```
c. Distributed for discussion by core WG on:

   Core WG decision:

d. Sent to responsible of subgroup:               on:

   Subgroup decision:
```

= COMMENTS FOLLOW = COMMENTS FOLLOW = COMMENTS FOLLOW =

B Names and addresses of editors

B.1 Chief Editors

N. Calzolari	
Co-Chief Editor, EAGLES	
ILC–CNR	
Via della Faggiola 32	Tel.: +39.50.56.04.81
Pisa I-56126	Fax.: +39.50.58.90.55
Italy	E-mail: ceditor@tnos.ilc.pi.cnr.it

J. McNaught	
Co-Chief Editor, EAGLES	
Centre for Computational Linguistics	
UMIST	
Manchester	Tel.: +44.161.200.3098
United Kingdom	Fax.: +44.200.3099
M60 1QD	E-mail: ceditor@tnos.ilc.pi.cnr.it

B.2 Working Group Editors

B.2.1 Text corpora

J. Llisterri	
Instituto Cervantes	
Libreros, 23	
28801 Alcalá de Henares	Tel: +34.1.885.6192/6206
E – Madrid	Fax: +34.1.883.5010
Spain	E-mail: tcwgeditor@tnos.ilc.pi.cnr.it

B.2.2 Computational lexicons

U. Heid	
Institut für Maschinelle Sprachverarbeitung	Computerlinguistik
Universität Stuttgart	
Azenbergstrasse 12	Tel: +49.711.1211373
D–7000 Stuttgart	Fax.: +49.711.1211366
Germany	E-mail: clwgeditor@tnos.ilc.pi.cnr.it

B.2.3 Formalisms

J. Capstick	and G. Erbach
DFKI	
Stuhlsatzenhausweg 3	Tel: +49.681.302.5288
D–66123 Saarbrücken 11	Fax: +49.681.302.5341
Germany	E-mail: clfwgeditor@tnos.ilc.pi.cnr.it

B.2.4 Evaluation

A. Neville	
Centre for Language Technology	
Njalsgade 80	Tel: +45.35.329.065
DK–2300 Copenhagen	Fax: +45.35.329.089
Denmark	E-mail: ewgeditor@tnos.ilc.pi.cnr.it

B.2.5 Spoken Language

D. Gibbon
Fakultät fur Linguistik und Literaturwissenschaft
Universität Bielefeld
Postfach 100131
D–4800 Bielefeld 1
Germany

Tel: +49.521.106.3510/09/18
Fax: +49.521.106.2996
E-mail: slwgeditor@tnos.ilc.pi.cnr.it