EAGLES

Expert Advisory Group on
Language Engineering Standards

EAGLES

Computational Lexicons Working Group Reading Guide

EAGLES Document EAG–CLWG–FR–2

Version of May, 1996

Contents

1 Authors 1

2 Working group members 1

2.1 Core working group members . . . 1

2.2 Subgroup members . . . 2

2.2.1 Task 1: Architecture . . . 2

2.2.2 Task 2: Methodology . . . 2

2.2.3 Task 3: Morphosyntax . . . 3

2.2.4 Task 4: Syntax . . . 3

2.2.5 Task 5: Synthesis . . . 3

3 Introduction 4

4 Principles of EAGLES work on standards 4

4.1 Approaches to standardisation . . . 4

4.2 Phases of work towards standards for the lexicon . . . 5

4.3 Types of documentation . . . 5

4.4 Levels of recommendation . . . 6

4.5 List of the main documents produced by the CLWG . . . 7

4.6 Short summaries of the main reports . . . 8

4.6.1 Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages (Morphsyn) . . 8

4.6.2 Recommendations on syntactic subcategorisation (Synlex) . . . 8

4.6.3 The ELM reports . . . 8

4.6.4 The report on tagset and tagger interaction (Tags) . . . 9

4.6.5 The report on tagset mapping (TSM) . . . 9

5 The content of the CLWG work — a short overview 9

5.1 Levels of description covered . . . 10

5.2 Formats — Representation . . . 10

5.3 Standardisation by consensus: Requirements in the field of the lexicon . . . 10

5.4 Parameters of description of the standards proposals . . . 11

6 The working methodology of the CLWG **12**

7 Examples of results **13**

7.1 Inventory and first proposals phase 13

7.1.1 Formalisation of the proposals 14

7.1.2 Operational guidelines — The practical side of the proposed standards 14

7.2 Validation phase 14

7.2.1 The tagged reference corpora 14

7.2.2 The tagset mapping exercise 16

7.2.3 The tagging *vs.* tagset evaluation exercise 18

1 Authors

U. Heid	(CLWG Editor)
Simone Teufel	
Christine Stöckert	
Universität Stuttgart	
Azenbergstrasse 12	Tel: +49.711.1211373
HD-70174 Stuttgart	Fax.: +49.711.1211366
Germany	E-mail: uli@ims.uni-stuttgart.de

2 Working group members

2.1 Core working group members

N. Bel	
Universidad de Barcelona	
Fundacion Bosch Gimpera	
Barcelona	Tel: +34.3.2033597
E–08017	Fax: +34.3.2054656
Spain	E-mail: nuria@gilcub.es

N. Calzolari	
ILC–CNR	
Via della Faggiola 32	Tel: +39.50.560.481
I–56100 Pisa	Fax: +39.50.589.055
Italy	E-mail: glottolo@ilc.pi.cnr.it

D. Gibbon	
Universität Bielefeld	
Postfach 100131	Tel: +49.521.1063510
D–4800 Bielefeld	Tel: +49.521.1063510
Germany	E-mail: gibbon@spectrum.uni-bielefeld.de

U. Heid	
Universität Stuttgart	
Azenbergstrasse 12	Tel: +49.711.1211373
D–7000 Stuttgart	Fax.: +49.711.1211366
Germany	E-mail: uli@ims.uni-stuttgart.de

P. Hellwig	
Universität Heidelberg	
Karlstrasse 2	Tel: +49.6221.543245
D - Heidelberg	Fax: +49.6221.543242
Germany	E-mail: hellwig@novelll.gs.uni-heidelberg.de

G. Heyer	
Triumph Adler	
Fürtherstrasse 212	Tel: +49.911.9326047
D–8500 Nürenberg	Fax: +49.911.9326282
Germany	gerd@triumph-adler.de
H. Khatchadourian	
Cap Gemini Innovation	
86/90 rue Thiers	Tel: +33.1.49.10.51.50
F–92513 Boulogne-Billancourt	Fax: +33.1.49.10.06.15
France	E-mail: khatcha@capsogeti.fr
S. Sabbagh (Host)	
GSI-ERLI	
1, Place des Marseillais	Tel: +33.1.48.93.81.21
F–94227 Charenton-Le-Pont Cedex	Fax: +33.1.43.75.79.79
France	E-mail: simon.sabbagh@erli.gsi.fr
A. Sanfilippo	
Sharp Laboratories of Europe	
Oxford Science Park	Tel: +44.18.65.74.77.11
UK–OX4 4GA Oxford	Fax: +44.18.6574.77.17
United Kingdom	E-mail: antonio@sharp.co.uk
G. Thurmair	
Siemens-Nixdorf AG	
Otto-Hahn Ring 6	Tel: +49.89.63.64.41.93
D-8000 Munich 83	Fax: +49.89.63.64.96.46
Germany	E-mail: gregor.thurmair@mch.sni.de
A. Zaenen (Chairperson)	
Rank Xerox Research Center	
6, Chemin de Maupertuis	Tel: +33.76.61.50.73
F–38240 Meylan	Fax: +33.76.61.50.99
France	E-mail: Annie.Zaenen@xerox.fr

2.2 Subgroup members

2.2.1 Task 1: Architecture

Responsible: B. Menon
U. Heid (Universität Stuttgart, D)
D. Maret (SITE, F)
N. Modiano (Cap Gemini Innovation, F)

2.2.2 Task 2: Methodology

Responsible: N. Modiano
N. Bel (Universidad de Barcelona, E)
N. Calzolari (Istituto di Linguistica Computazionale, CNR, I)
D. Gibbon (Universität Bielefeld, D)
U. Heid (Universität Stuttgart, D)
P. Hellwig (Universität Heidelberg, D)

G. Heyer (TA Electronic Publishing, D)
D. Maret (SITE, F)
B. Menon (GSI-ERLI, F)

2.2.3 Task 3: Morphosyntax

Responsible: P. Hellwig
N. Bel (Universidad de Barcelona, E)
D. Bleiching (Universität Bielefeld, D)
N. Calzolari (Istituto di Linguistica Computazionale, CNR, I)
P. Guerriero (ILTEC, P)
U. Heid (Universität Stuttgart, D)
D. Maret (SITE, F)
M. Monachini (Istituto di Linguistica Computazionale, CNR, I)
A. Schiller (Universität Stuttgart, D)
C. Stöckert (Universität Stuttgart, D)
U. Von Rekowski (GSI-ERLI, F)
A. Zaenen (Rank Xerox Research Center, F)

2.2.4 Task 4: Syntax

Responsible: A. Sanfilippo
R. Barnett (University of Lancaster, UK)
N. Calzolari (Istituto di Linguistica Computazionale, CNR, I)
S. Flores (GSI-ERLI, F)
P. Hellwig (Universität Heidelberg, D)
P. Kahrel (University of Lancaster, UK)
G. Leech (University of Lancaster, UK)
M. Melero (GILCUB, Barcelona, E)
S. Montemagni (Istituto di Linguistica Computazionale, CNR, I)
J. Odijk (Philips, NL)
V. Pirrelli (Istituto di Linguistica Computazionale, CNR, I)
S. Teufel (Centre for Cognitive Science, Edinburgh, UK)
M. Villegas (GILCUB, Barcelona, E)
L. Zaysser (GSI-ERLI, F)

2.2.5 Task 5: Synthesis

Responsible: U. Heid
N. Bel (Universidad de Barcelona, E)
N. Calzolari (Istituto di Linguistica Computazionale, CNR, I)
B. Menon (GSI-ERLI, F)
A. Sanfilippo (Sharp Laboratories, UK)
A. Zaenen (Rank Xerox Research Center, F)

3 Introduction

This short document provides the reader with an overview of the work of the Eagles Computational Lexicons Working Group (CLWG), carried out in the timeframe October 1994 to April 1996. It describes the context in which the different documents produced in this phase should be seen, and their interrelationships.
The aim of this introduction is to give a short summary and a coherent picture of the work of the CLWG. We do so by giving the following information:

- A description of the principles underlying the EAGLES work in general and the work of the CLWG in particular (see section 4);
- A summary of the axes of the work of the CLWG, in terms of fragment coverage, language coverage, prioritary topics, etc. (see section 5);
- A brief account of the methodology developed in the CLWG for the creation of guidelines and proposals for standards in the field of computational lexicons and their applications (see section 6);
- A summary of the outcome of the main axes of work, in the fields of proposals for standards and of their validation (see section 7).

All parts are extensively cross-linked to the actual documents themselves. Thus, the reader can switch to the respective documents directly, whenever appropriate, or can first switch to the abstracts contained in section 4.6 of this document before (or as an alternative to) going to the documents themselves.

4 Principles of EAGLES work on standards

4.1 Approaches to standardisation

The task of the Expert Advisory Group on Language Engineering Standards (EAGLES) is to propose guidelines for both contents and form of certain tasks and resources in the field of Language Engineering. This implies taking stock of the current practice, finding out about best practice and thereafter proposing what seems to be the most advanced and generally acceptable and applicable way of carrying out the tasks in question, or the most useful yet generic way of organising and populating linguistic resources.
Standards may come about in different ways: de facto standards are often due to market forces (although the de facto standard product may or may not be optimal for the majority of tasks envisaged); formal, official Standards carry the stamp of a national or international standards body, such as ISO, for example; both cannot be expected to naturally arise in a very new field of research and development, as is Language Engineering. Paradoxically, it is however in just such a new field that the need for guidelines and for a clear statement of minimal common ground is felt most urgently. This is why EAGLES in general and with it the CLWG felt that there is a need for early standardisation — not imposed, but coming about through a consensus of the relevant actors in the field[1].
Standardisation by consensus is typically a medium to long term process characterised by a number of properties:

- Standardisation by consensus presupposes a broad overview of the field, based on prominent examples of current (best) practice;
- Standardisation by consensus must be based on a critical mass of experts from the field;
- Standardisation by consensus must prove (demonstrate) its relevance, both for practical applications and for leading edge technology development;
- Proposals for consensual standards need to be accessible and testable, such that users can easily find out about the relevance of these proposals for their own work.

[1]The CLWG has spent some time and effort on the definition of its own working procedures and methodology, in the 1993–94 timeframe; some of the principles of EAGLES and of the CLWG are described in EAGLES (1993a). For a quick summary, see section 6.

4.2 Phases of work towards standards for the lexicon

The consensual approach to standardisation has implications for the working methodology and for the phasing of the work leading to standards proposals. This work typically falls into the following (generic) conceptual phases:

Synopsis: Survey of the field, synopsis and comparative (qualitative) assessment of existing practice (leading to a survey and a best practice study);

Inventory and first proposals: Creation of an inventory of best practice components, on the basis of the outcome of the survey; organisation, structuring and formalization of such inventory, leading to proposals for standards, on different levels of formality;

Validation: Proof of applicability of the proposals produced, by means of application tests, validation exercises, etc.

These phases have been backed up by dissemination and awareness actions.
This model has been followed, and further developed, in the work of the CLWG. The procedure can be seen as an activity of filtering, condensing and testing (see below, section 6): a broad range of existing items is collected, screened and compared, to filter out the generalisable and uncontroversial parts. These are condensed and formalised, to make sure the condensate is communicable, reproducible and well-understood. Before making it available to the relevant expert public at large, the proposal is thoroughly tested and validated in applications.

4.3 Types of documentation

The activities of the CLWG could just be summarised in scientific papers. This has been done to some extent: cf. McNaught (1993), Calzolari (1994), Calzolari (1995), Calzolari & Zampolli (1994) and Calzolari (forthcoming); also, this work has been presented at a number of international events (workshops, conferences, summer schools, etc. However, given that much of the work of the CLWG deals with relatively large quantities of linguistic data or is intended to cover sizeable fragments, it makes sense to separate out the results of the three phases mentioned above, as far as their documentation goes.
The approach and phasing sketched out above typically produce different kinds of documentation which address different tasks and possibly different target groups. These types of documents are, in the case of the CLWG, the following:

- Formal guidelines (specifications) for the area concerned (from the inventory and first proposals phase). See the documents EAGLES (1996a), EAGLES (1996c), EAGLES (1996d) and EAGLES (1996b), for morphosyntax, as well as EAGLES (1996e) for syntactic subcategorisation.

- Operational guidelines for the practical application of the specifications (from the inventory and first proposals phase). These are included in EAGLES (1996a), EAGLES (1996c), EAGLES (1996d) and EAGLES (1996b), for morphosyntax, as well as in EAGLES (1996e) for syntactic subcategorisation.

- Consensus reports: surveys, descriptions of best practice and first proposals (from the survey phase). See EAGLES (1996g) for morphosyntax and chapter 3 ofEAGLES (1996e) for syntactic subcategorisation.

- Reports on validation exercises (from the validation phase). Validation has been done in particular for the morphosyntax area; the usability of the proposals for corpus tagging and the interrelationships between tagging methods, tagsets and tagging results have been assessed in EAGLES (1996h); the possibilities of mapping the EAGLES proposals onto existing corpus tagsets have been assessed in EAGLES (1996b).

- Background material on specific questions (from all different phases). A full list is found in section 4.5. For references, see the bibliography.

This above list is organised from the more formal types of documents (i.e. linguistic specifications, ready to be modelled in a computational linguistic formalism) to the more informal (narrative, textual) ones. It is important to note that not all the tasks initially identified were carried out — this was invariably due to judicious management of modest resources and the effects of working with largely voluntary labour. Also, certain tasks have given rise to a lower level of documentation than others, either due to labour and time problems or to a lack of achievable consensus in the current timeframe. Thus, only the main documentation is here presented.

4.4 Levels of recommendation

The proposals made by the CLWG are organised in such a way as to encompass a number of levels of recommendation. This can be seen most clearly in the morphosyntax work.

In the synopsis phase, the CLWG has (as far as can be seen, most likely for the first time ever) tried to come up with a proposal for a set of parallel inventories of descriptive devices, for the EU languages (EAGLES, 1996g). This parallelism is highly important: work on contrastive grammars and lexicons can greatly profit from a methodology based on the construction of parallel fragments, which use, whenever possible, the same descriptive devices and isomorphic classifications. The methodology of doing this has been developed and demonstrated, in a broad coverage scenario, in the morphosyntax work of the CLWG. To be able to account for language specificities, and for the differences in granularity relevant for different languages at different levels of morphosyntactic description, a four-layered system of recommendation types has been used.

These recommendation types distinguish *obligatory*, *recommended* and *optional* classifications. Table 1 below describes these levels in more detail:

Level	Information types	Recommendation types
L–0	part of speech	obligatory
L–1	morphosyntactic (agreement features)	recommended
L–2a	refined morphosyntactic features; non-morphosyntactic features; 3 or more languages	optional
L–2b	language-specific morphosyntactic refinement	

Table 1: Levels of recommendation in EAGLES (1996g).

4.5 List of the main documents produced by the CLWG

EAGLES document acronym	Title
Guide	Reading guide
Lexarch	Lexicon architecture
Method	Methodology
Morphsyn	Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages.
ELM–DE	ELM–DE: EAGLES specifications for German morphosyntax
TSM	Validation of tagset mapping (ELM-EN)
ELM–EN	ELM–EN: EAGLES specifications for English morphosyntax (Appendix to the TSM report)
ELM–IT	ELM–IT: EAGLES specifications for Italian morphosyntax
ELM–FR	ELM–FR: EAGLES specifications for French morphosyntax
Tags	Study of the relation between tagsets and taggers
Synlex	Recommendations on subcategorisation

4.6 Short summaries of the main reports

4.6.1 Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages (Morphsyn)

The Morphosyntax area was among the first in which EAGLES concentrated its efforts, because it is a more mature area, where a lot of work has already been done, many systems, approaches, and data exist for many languages: this constitutes a solid platform for making reasonable and acceptable proposals for standards.
The objective of the report is to propose a common core set of morphosyntactic distinctions to be encoded in lexicons of the European languages.
The report is organised as follows. After the Introduction, giving a concise motivation of the need for standards in the area of morphosyntax, section 2 highlights the methodology adopted by the group and the different phases of work, and mentions the interactions with some European projects. Section 3 deals with a number of general issues such as the relationship between lexicon specifications and corpus tagsets. Section 4 gives some general information about the language-specific applications. Sections 5 to 18 are the central part of the report, with the current status of the EAGLES morphosyntactic proposal, divided by POS and, for each POS, by language-specific applications. Section 19 contains the references.

4.6.2 Recommendations on syntactic subcategorisation (Synlex)

The report (EAGLES, 1996e) is organised into chapters. The introduction describes the background, objectives and working procedures of the subgroup on subcategorisation.
Chapter 3 provides a review and comparison of different approaches to subcategorisation adopted in linguistic theories, practical NLP lexicons and bracketed corpora. In chapter 4, the results of the comparisons drawn in the preceding chapter are used to compile frame-based representations of subcategorisation structures. After a general description of the frame structure, specific details are given concerning the encoding of categorial and morphosyntactic properties, grammatical functions, order, verbal diatheses, control, thematic roles and semantic class. In chapter 5, a formalisation of the encoding scheme presented is given using an object-oriented specification written in SGML (ISO 8879). In the final chapter, a set of macros is given which synthesises the formal frame descriptions presented in chapter 4; using these macros, a list of examples is compiled for each of the languages (Spanish, German and Italian) represented within the group.

4.6.3 The ELM reports

There are specification reports for German (EAGLES, 1996a), French (EAGLES, 1996c), Italian (EAGLES, 1996d) and — slightly less detailed — for English (EAGLES, 1996b) (an appendix of TSM (EAGLES, 1996h)).
The reports give a detailed morphosyntactic description of word forms of these languages, specifying the relevant distinctions at the levels of category agreement features and distributional features.
All reports are parallel in structure. A usage guide explains the descriptive devices and their practical use. Subsequent chapters are devoted to the individual word classes (noun, verb, adjective, etc.) and are again parallel in structure. They contain

1. A label inventory for the category in question;
2. A type hierarchy table showing the relevant subclassifications;
3. A description of the core definition of each class;
4. Part of speech delimitation tables and feature delimitation tables, allowing operational reproduction of the distinctions proposed;
5. An example list carrying full annotations according to the EAGLES proposals.

The types (1)–(3) of the documents can be used as a definition of the morphosyntactic specifications, and (3) and (4) support their use as operational guidelines.

The example lists (5) and an alphabetical index of all word forms discussed and annotated in each document allow for an easy and comfortable use as reference works.

4.6.4 The report on tagset and tagger interaction (Tags)

The report (EAGLES, 1996f) describes the use of a tagset (STTS) derived from, the German ELM specification in a series of tests for the evaluation of the impact of:

- Different tagging methods on the tagging results, if the same tagsets are used;
- Tagset modifications on the tagging results, if the same tagger is used,
- Perceived linguistic differences between training text and test text.

The emphasis is on the design of these tests, and on directions for the interpretation of their results, rather than on the concrete individual results themselves. The document shows the applicability of the ELM specification as a tagset for corpus annotation, as well as the impact of certain much discussed tagset modifications on the results (e.g. giving up the separate class of "proper names" and merging it with "common nouns").
The report first states background and objectives (section 3) and briefly describes the resources and tools used for the tests in question (section 4). It goes on to make proposals for test setups for the three types of problems described above, including proposals for the kinds of statistics which need to be extracted from the results and compared (section 5). These setups and interpretation guidelines are then applied to concrete tests (German material), and the results are displayed and interpretations proposed.
The appendices contain detailed descriptions of the tested phenomena, suggestions for further tests on the impact of tagset modifications and an overview of the tagset used.

4.6.5 The report on tagset mapping (TSM)

This report describes the experiments carried out in view of a mapping between the English version of the EAGLES morphosyntax proposals and existing corpus tagsets (EAGLES, 1996h). The tagsets chosen as examples are:

- The British National Corpus (BNC) tagset, CLAWS5, and
- the UPenn tagset (used in the morphosyntactic annotation of texts prepared for the University of Pennsylvania Treebank project).

The first part contains a concise description of the objectives of the mapping exercise, of the preparation of an ELM specification for English (cf. the description of the ELM reports, above, in section 4.6.3), the technical realisation of the mapping procedures and the use and output of the mapping from the LIQUY tool.
The second part contains a draft ELM specification for English, which mainly follows the organisation of the other ELM reports. It is however less worked-out, because it has not been used for manual tagging. Thus, the parts which support the operational use (delimitation tables) are not available in as much detail as they are for German, French and Italian.
The annex contains the Prolog specification of ELM–EN, the compiled feature hierarchy and lists of the BNC and UPenn tagsets.
The mapping is exemplified by a list of all possible queries in ELM–EN and all results produced on BNC and on UPenn.

5 The content of the CLWG work — a short overview

This section describes the context of the work of the CLWG in a general way. More details on the content of the work can be found in section 7, and in the relevant papers.

5.1 Levels of description covered

Lexical descriptions typically tend to be broken up into packages, according to the traditional view of levels of linguistic description:

- Morphology and morphosyntax;
- Syntax (e.g. subcategorisation);
- The syntax–semantics interface;
- Semantics;
- Pragmatics, discourse-related information, etc.

The CLWG has concentrated on morphology/morphosyntax (key reports are EAGLES (1996g) and the series of ELM reports, i.e. EAGLES (1996a), EAGLES (1996c), EAGLES (1996d) and EAGLES (1996b)) and syntactic subcategorisation (The key report is EAGLES (1996e)). These fields are most consolidated and thus most immediately ripe for preliminary standardisation.

5.2 Formats — Representation

Work on lexical description typically deals with two aspects of the lexicon:

1. The linguistic description of lexical items ('content');
2. The (formal or formalisable) representation of lexical descriptions ('container').

As far as standardisation proposals are concerned, the CLWG has concentrated on the first issue, leaving the work towards proposals for the second to the Computational Linguistic Formalisms Working Group. This does however not mean that the CLWG disregards the evident need for a formal expression of the contents-oriented proposals put forward: on the contrary, the ELM documents series (ELM stands for *Eagles/Lexicon/Morphosyntax Subgroup*), as well as the subcategorisation report (EAGLES, 1996e) have been produced to satisfy the requirement of creating formal specifications which can be used in an implementation.

The work of the CLWG uses different representation formalisms to express its proposals, at different levels. In the morphosyntax area, proposals have been represented in different formats, ranging from a tabular presentation of the first proposals in EAGLES (1996g) over a set of hierarchical tables inspired by a typed feature logic representation (in the ELM reports series) to a specification implemented in Prolog (in the tagset mapping exercise documented in EAGLES (1996h). The results of the work on syntactic subcategorisation are represented in a feature structure format inspired by HPSG, and a Document Type Definition (DTD) has been constructed which can accommodate the syntactic lexicon entries proposed. The different parts of EAGLES (1996e) distinguish the different representations.

5.3 Standardisation by consensus: Requirements in the field of the lexicon

Above, a few properties of the consensual approach to standardisation have been sketched out; throughout its work, the CLWG has aimed at satisfying the requirements related with the approach of standardisation by consensus.

- *Broad overview:*
  - The survey phase was based on input from both the fields of computational lexicography and of corpus linguistics (corpus processing is seen as the most prominent application of both morphosyntactic and syntactic classification of lexical material. In this context, we would like to particularly thank the EAGLES Subgroups on morphosyntactic and on syntactic corpus annotation for the valuable input and the amount of effort offered as a contribution to the work summarised here).

  - The morphosyntax work, although highly language-specific, covers all of the European Union languages, plus several languages of Eastern and Central European countries[2].

- *Critical mass:*
  - Broad language coverage has been ensured (see above); contributions on the different languages have been provided by both industry and academia; the first proposals, documented in EAGLES (1996g), have been discussed in many LRE and LE projects, whose feedback was then integrated.
  - The comparison of individual specifications for the individual languages (i.e. EAGLES (1996a), EAGLES (1996b), EAGLES (1996d), EAGLES (1996c)) is eased by the fact that all specifications are strictly parallel (based on the same descriptive devices, the same presentation, etc.).
  - The availability of detailed documentation allows for easy-to-reproduce descriptions. This is enhanced by the fact that an integrated package is being prepared which will contain the morphosyntax specifications, the practical guidelines for their use in corpus tagging, and tagged sample texts of 50–60,000 wordforms, for both German and Italian (see below).

- *Proof of relevance, for practical applications and for leading edge technology development:*
  - The breadth of the fragment coverage and the depth of the description makes the proposals applicable outside toy situations.
  - In addition, more sophisticated applications can hook up their own, specific descriptions to the basic ones proposed by the CLWG.
  - The formalisation employed makes the proposals reinterpretable (the reinterpretation being even automatically verifiable) and thus more easily reusable than proposals for which such possibilities do not exist.

- *Accessibility and testability:* a pilot resource has been produced (see section 7.2.1), to be distributed via the European Linguistic Resources Association (ELRA), containing 50–60,000 wordform mini-corpora of Italian and German newspaper text, annotated according to the proposals of the CLWG. The preparation of these corpora was carried out jointly with the ELSNET Language and Speech Resources Task Group. We most gratefully acknowledge the financial support provided by ELSNET for this task.

5.4 Parameters of description of the standards proposals

From the above, the following parameters can be derived:

- Depth of description: which lexical information types?
- Breadth of description: fragment coverage
  (single exemplary word classes vs. broad coverage of all classes).
- Language coverage.
- Degree of formalisation (from less to more formal):
  - Inventory of labels (list of descriptive devices);
  - Definition of a vocabulary (as in formal languages);
  - Control of consistency within the vocabulary, e.g. through typing and constraints on appropriateness;
  - Full specification defining also the interaction of the descriptive devices for individual classes of lexical objects.

The work on morphosyntax and on subcategorisation can be summarised, along these parameters, in table 2, below:

[2] The latter extensions have been carried out, in part, in the projects MULTEXT and Multext goes East

Parameter	Morphosyntax	Subcategorisation
References	(several reports)	(Synlex report)
Information types	category morphosyntactic features	phrase types, grammatical functions, complement–argument linking
Word classes	all	verbs
Languages	EU languages	generic, EU languages
Formalisation: - Inventory of labels - Vocabulary definitions - Consistency control - Full specifications	 + (Morphsyn report) + (Morphsyn report) + (ELM reports) + (ELM reports)	 + + +

Table 2: Morphosyntax and syntactic subcategorisation work

6 The working methodology of the CLWG

This section summarises the working methodology of the CLWG; this methodology is conditioned by the constraints related with the standardisation-by-consensus approach briefly described above in sections 4 and 5.3.

Recall the phases of the work and their output, as introduced above in sections 4.2 and 4.3. For convenience, we here summarise these in table 3. The documents cited are from the subgroup on morphosyntax. The syntax group managed to condense its output into one major report, which, in its internal structure, reflects the relevant phases as well.

Phase	Step	Output
Synopsis	Survey of best practice	Survey report (Morphsyn report)
	Comparative assessment	Survey report (Morphsyn rep.)
Inventory	List of labels, first proposals	Survey and proposals report (Morphsyn report)
Proposals	Formal specification (Descriptive guidelines)	ELM reports series
	Proposals for use in lexical classification (Operational guidelines)	ELM reports series
Validation	Test: automatic mapping with existing approaches	TSM report
	Pilot resources: medium scale application experiment; Further testing: tagset development	EAGLES/ELSNET IT and DE corpora (Tags report)
Maintenance	(ongoing task)	

Table 3: Phases, steps and output of the work of CLWG

In the 1995–96 timeframe, emphasis has been on the inventory/proposals phase and on the validation phase[3]. The maintenance work is not really a phase of the standards development cycle, but is a very important element of the life cycle of standardisation proposals as such. It is an ongoing task relevant beyond the phase described here.

The phases build one on the other, and they implement a stepwise process (from smaller to larger, from dispersed to concentrated, from consensual to tested, from informal to formally represented and implemented (runnable specifications) proposals). This can be symbolised as in figure 1, below. This working methodol-

[3]Most of the survey work was carried out in the first phase (1993–94) of the EAGLES project.

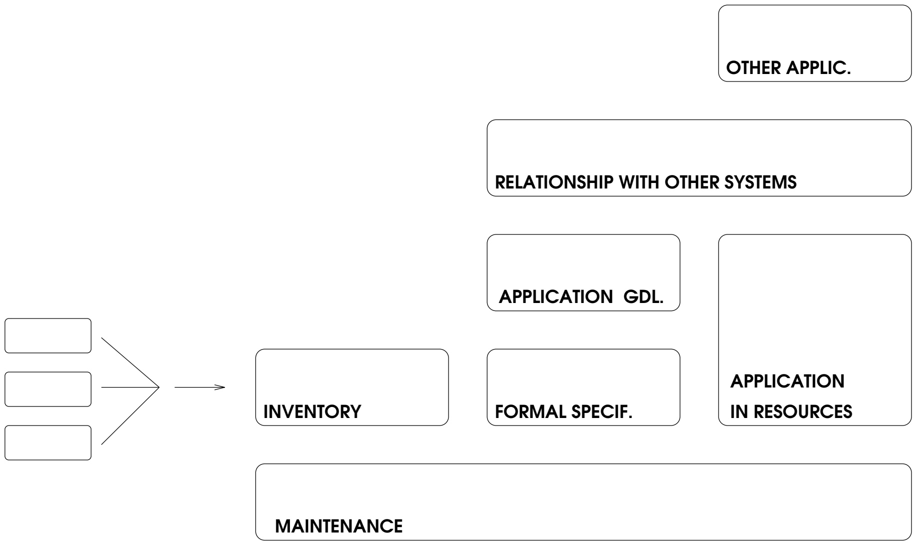

Figure 1: Steps towards proposals for standards for lexical description

ogy has proven applicable in the fields of morphosyntax and syntactic subcategorisation. It may be tried as well in other tasks addressed with the same objective.

7 Examples of results

This section serves to very briefly summarise and exemplify, in tables and figures, the main outcome of the work of the CLWG, in the 1994–96 timeframe.

7.1 Inventory and first proposals phase

The main results of the inventory and first proposals phase are summarised in table 4:

Area	Output	Documents
Morphology	Paradigm proposals for different languages	background reports (EAGLES, 1993c; EAGLES, 1993b; EAGLES, 1994)
Morphosyntax	4-layered general proposals; inventory of labels	Morphsyn report (lower part of tables)
	Formalisation in typed representation (tables)	the ELM series
	Formalisation in Prolog	ELM_EN report (EN and IT)
Syntax: subcategorisation	Operational guidelines for lexical classification	the ELM series
	Frame-oriented description of subcategorisation information	Synlex report (section 4)
	Formalisation in E–R diagrams	Synlex report (section 5.1)
	Formalisation in a DTD	Synlex report (section 5.2)

Table 4: Summary of the main outcome of the Inventory and first Proposals Phase

7.1.1 Formalisation of the proposals

The formalisations have been made available with a number of objectives in mind:

- To provide interpretation guidelines for the tables containing the inventory of descriptive devices and of lexical classes distinguished with their help;
- To provide as precise definitions as possible of the classifications of interest, in order to base operational guidelines on these definitions (see below);
- To support automatic applications (such as the tagset mapping system relating ELM–EN with the UPenn and BNC tagsets, cf. section 7.2.2, below, and (EAGLES, 1996h));

7.1.2 Operational guidelines — The practical side of the proposed standards

Along with the inventories and their declarative description (in the different specification and formalisation documents), *operational guidelines* for the practical application of the proposed standard in concrete classification tasks have been given. This is most clearly visible in the ELM series (i.e. EAGLES (1996a), EAGLES (1996b), EAGLES (1996d) and EAGLES (1996c)), where *delimitation tables* for part-of-speech classes and for attributes have been given, whenever manual annotators of text corpora or linguists carrying out a manual lexical classification job had problems in distinguishing among the classes proposed by the CLWG[4].
An example of such a delimitation table is given in figure 5, below: it serves to distinguish between proper names and common nouns, in German.

7.2 Validation phase

Due to the timescales of the EAGLES project, validation work was carried out only in the morphosyntax area. The main results are summarised in table 6.

7.2.1 The tagged reference corpora

Italian and German newspaper texts have been morphosyntactically annotated according to the EAGLES Italian (EAGLES, 1996d) and Italian (EAGLES, 1996a) specifications. Table 7 summarises the distribution of texts in the German corpus; the Italian corpus[5] is structured in a very similar way.
The texts were prepared as follows:

- The texts were automatically tagged and then manually corrected (For German: automatically pre-tagged for level (b), fully manually tagged on level(c) — see below for levels);
- The overlap in the manual work (two or more linguists working on the same text without knowing it) is some 10–20% of the textual material;
- The material was cross-checked by extraction of all wordforms tagged the same way and by plausibility checks against morphology systems, as well as by manual checking on closed class items.

The material will be made available in the following forms:

Level (b) — STTS tags (DE)

Level (b) — ELM–DE attribute/value pair annotation (DE)

Level (c) — ELM–DE attribute/value pair annotation (DE)

Level (c) — ILC–Pisa tagset (IT)

[4]Input for this has been provided by, among others, the manual annotators who produced the tagged reference corpora (see below, section 7.2.1).

[5]Similar in proportions, the Italian material amounts to slightly more than 50,000 wordforms.

type=com		**type=prop**	
Description	**Example**	**Description**	**Example**
Composita with common noun head	VW–**Vorstand**/`com`	Names of companies	der Vorstand von **Porsche**/`prop`
Names of products where used as concrete nouns	ein **Mercedes**/`com`, eine **Cola**/`com`	Names of products where used as names	**Glenfiddich**/`prop`
Proper names derived from common nouns	die **Grünen**/`com`	Real proper names	**CDU**/`prop`, **Susanne**/`prop`, Frau **Müller**/`prop`
Complex name of state	die **Bundesrepublik**/`com`	Names of a specific country	**Schweiz**/`prop`
Complex name of state	die Vereinigten **Staaten**/`com` von Amerika	Abbreviated name of state	die **USA**/`prop`, die **BRD**/`prop`
Product name	Er fuhr einen **Porsche**/`com`	Company name	Er fuhr zu **Porsche**/`prop` zur Arbeit
Product name	Er aß ein **Mars**/`com`	Planet name	Der **Mars**/`prop` ging heute morgen neben dem Mond auf
Names of party, spelled out	**Partei**/`pos` des Demokratischen Sozialismus, die **Grünen**/`com`, die **Demokraten**/`com`	Name of party, abbreviated	**CDU**/`prop`, **PDS**/`prop`

Table 5: A part-of-speech delimitation table from EAGLES (1996a)

Area	Output	Documents, Resources
Morphosyntax	Application in tagset for corpus tagging (IT, DE)	EAGLES / ELSNET reference corpora (available 7/96)
	Exercise in tagset mapping (ELM-EN *vs.* UPenn *vs.* BNC)	TSM report (tool, rules)
	Tests on tagset development (impact of tagset changes on tagging results)	Tags report

Table 6: Main outcome of the validation phase

Economy	ca.	17,000 wordforms
Politics	ca.	14,000 wordforms
Culture	ca.	18,000 wordforms
Sports	ca.	9,000 wordforms
Local Events	ca.	8,500 wordforms
Total	ca.	66,500 wordforms

Table 7: EAGLES/ELSNET reference corpus for German

Level (c) — ELM–IT attribute/value pair annotation (IT)

The following is a sample of the German text, in fully-fledged ELM feature structure annotation (cf. level (c) above):

```
          Mexiko    [ pos=noun & type=prop ]
               :    [ pos=resid & type=punct & punct-t=c-final ]
             Die    [ pos=art ]
               "    [ pos=resid & type=punct & punct-t=non-c-final ]
 Praesidentenmache  [ pos=noun & type=com ]
               "    [ pos=resid & type=punct & punct-t=non-c-final ]
               .    [ pos=resid & type=punct & punct-t=c-final ]
    Mexikanische    [ pos=adj & use=attr ]
         Politik    [ pos=noun & type=com ]
             ist    [ pos=verb & type=aux & fin=fin & ( vm-f=ind | vm-f=konj )]
        Almachie    [ pos=noun & type=com ]
             und    [ pos=conj & type=coord ]
       Tradition    [ pos=noun & type=com ]
               .    [ pos=resid & type=punct & punct-t=c-final ]
```

7.2.2 The tagset mapping exercise

The tagset mapping exercise was carried out with the following objectives:

- To make ELM fully automatically testable by using it as a formal specification;
- To compare ELM–EN with two of the major proprietary tagsets (BNC's CLAWS 5 tagset and the morphosyntactic tagset used by the University of Pennsylvania Treebank Project), to identify possible gaps and/or inconsistencies;
- To gain insight into automatic tagset mapping;
- To use ELM–EN as a common query language for BNC and the UPenn corpora, at least experimentally.

Mapping rules deal with 1:1, 1:n, n:1 and n:m cases. A small scale exception lexicon was created for idiosyncratic cases (e.g. *that* as 'IN' (= preposition) in the UPenn tagset).
The following are a few sample mapping rules, for the BNC:

```
  cqp_name(upenn, 'BNC').
[pos = 'AJ0']    =>      [adj & po].
[pos = 'AJC']    =>      [adj & comp].
[pos = 'AJS']    =>      [adj & sup].
[pos = 'AT0']    =>      [art].
[pos = 'AV0']    =>      [adv & (general '|' degree)].
[pos = 'CJC']    =>      [conj & coord].
[pos = 'CJS']    =>      [conj & subord].
[pos = 'CRD']    =>      [numeral & card].
[pos = 'DT0']    =>      [det & (indf '|' dem) '|' pron & dem].
[pos = 'DTQ']    =>      [det & wh].
[pos = 'EX0']    =>      [unique & existential].
[pos = 'NN0']    =>      [noun & com].
```

The following are a few sample entries from the exception lexicon:

```
[no]       << [pos = 'AT0'] >> [det & indf].
[that]     << [pos = 'CJT'] >> [conj & subord '|' pron & rel].
[of]       << [pos = 'PRF'] >> [conj & subord].
['bound'] << [pos = 'VVN'] >> [verb & s_aux].
['going'] << [pos = 'VVG'] >> [verb & s_aux].
```

If used with the LIQUY tagset mapping tool, the following output is generated; the queries serve to retrieve BNC evidence by using ELM–EN as a query language:

- Simple sample queries:

```
| ?- test(bnc).
[noun]
--------------------------------------------------------------
[(pos = "NN0" | pos = "NN1" | pos = "NN2" | pos = "NP0")]


[com & noun]
--------------------------------------------------------------
[(pos = "NN0" | pos = "NN1" | pos = "NN2")]
```

- Queries leading to inexact matches

```
[noun & com & pl]
--------------------------------------------------------------
%%%%%%% Warning:  Constraint [pl] ignored
%%%%%%%           due to 1:n case in [com&noun]
%%%%%%%
[(pos = "NN0" | pos = "NN2")]


[verb & fin & aux & 2]
--------------------------------------------------------------
%%%%%%% Warning:  Noise to be expected:
%%%%%%%           [aux&fin&pres&verb&1](Due to tag "[pos=VHB]")!

%%%%%%% Warning:  Noise to be expected:
%%%%%%%           [aux&fin&pres&verb&1](Due to tag "[pos=VDB]")!

%%%%%%% Warning:  Noise to be expected:
%%%%%%%           [aux&fin&pres&verb&1](Due to tag "[pos=VBB]")!


%%%%%%% Warning:  Constraint [2] ignored
%%%%%%%           due to 1:n case in [aux&fin&past&verb]

%%%%%%% Warning:  Constraint [2] ignored
%%%%%%%           due to 1:n case in [aux&fin&pres&verb&pl]

%%%%%%% Warning:  Constraint [2] ignored
%%%%%%%           due to 1:n case in [aux&fin&impr&pres&verb]
%%%%%%%
```

```
[(pos = "VBB" | pos = "VBD" | pos = "VDB" |
  pos = "VDD" | pos = "VHB" l pos = "VHD")]
```

The latter example shows the treatment of noise and silence in non-1:1 situations.

7.2.3 The tagging *vs.* tagset evaluation exercise

The proposals made in EAGLES (1996g) are intended to be applicable to different European languages and to be independent from a particular NLP application. The ELM series contains typed specifications again not geared towards a specific single application.

In the tests on the interaction between tagging methods, tagsets and tagging results, it was shown that the EAGLES-based ELM–DE specifications indeed allow tagset to be derived which can be practically used for the tagging of German and which leads to acceptable results.

Moreover, part of the history of the tagset could be followed, and the impact of the modifications introduced could be evaluated.

The following tests have been run, all with tagsets derived from ELM–DE:

Tagger evaluation — Tests allowing the impact of different statistical tagging methods on the results to be assessed, by comparing the performance of different taggers on the same training and test data, using the same tagset;

Tagset evaluation — Tests allowing the impact of tagset modifications on the results to be assessed, by using different versions of a given tagset on the same texts; differences between the versions of the tagset were documented and classified, and the impact of each modification was tested;

Text type evaluation — Tests allowing the impact of perceived linguistic differences between training texts and test (or: application) texts on the results to be assessed, by using texts from different text types in training and testing, tagsets and taggers being unchanged otherwise.

References

Calzolari, N. (1994). European efforts towards standardizing language resources, *in* P. Steffens (ed.), *Machine translation and lexicon*, Springer, Heidelberg.

Calzolari, N. (1995). Large lexical resources for language engineering, *Proceedings of the ESSLI workshop on the computational lexicon*, UNED, Barcelona.

Calzolari, N. (forthcoming). La standardizzazione di risorse linguistiche in europa: Il progetto LRE–EAGLES, *Quaderni MURST* .

Calzolari, N. & Zampolli, A. (1994). Standards to make natural language resources shareable resources, *Proceedings of the international workshop on shareable natural language resources*.

EAGLES (1993a). Computational lexicons methodology task, *Technical Report EAG–CLWG–Method/B*, ILC–CNR, Pisa.

EAGLES (1993b). German morphology in the framework of the GENELEX model, *Technical Report EAG–LSG–T4.2*, ILC–CNR, Pisa.

EAGLES (1993c). Working group on computational lexicons: Task 4 — Morphosyntax, *Technical Report EAG–LWG–T4–1*, ILC–CNR, Pisa.

EAGLES (1994). Guidelines, *Technical Report EAG–LWG–T4–2*, Pisa, Pisa.

EAGLES (1996a). ELM-DE: EAGLES specifications for German morphosyntax. Lexical specification and classification guidelines, *Technical Report EAG-CLWG-ELM_DE/F*, ILC–CNR, Pisa.

EAGLES (1996b). ELM-EN: EAGLES specifications for English morphosyntax. Lexical specification and classification guidelines, *Technical Report EAG–CLWG–ELM_EN/F*, ILC–CNR, Pisa.

EAGLES (1996c). ELM-FR: EAGLES specifications for French morphosyntax. Lexical specification and classification guidelines, *Technical Report EAG–CLWG–ELM_FR/F*, ILC–CNR, Pisa.

EAGLES (1996d). ELM-IT: EAGLES specifications for Italian morphosyntax. Lexical specification and classification guidelines, *Technical Report EAG–CLWG–ELM_IT/F*, ILC–CNR, Pisa.

EAGLES (1996e). Preliminary recommendations on subcategorisation, *Technical Report EAG–CLWG–Synlex/P*, ILC–CNR, Pisa.

EAGLES (1996f). Study of the relation between tagsets and taggers, *Technical Report EAG-CLWG-Tags/V*, ILC–CNR, Pisa.

EAGLES (1996g). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages, *Technical Report EAG–CLWG–Morphsyn/R*, ILC–CNR, Pisa.

EAGLES (1996h). Validation of tagset mapping, *Technical Report EAG–CLWG–TSM/F-V*, ILC–CNR, Pisa.

McNaught, J. (1993). The EAGLES standards-setting initiative, *Proceedings of the SALT Club Workshop on the Unified Lexicon*, DTI/University of Durham, Durham.