EAGLES

Expert Advisory Group on
Language Engineering Standards

EAGLES

Text Corpora Working Group
Reading Guide

EAGLES Document EAG–TCWG–FR–2

Version of May, 1996

Contents

1 Author

J. Llisterri
Instituto Cervantes
Libreros, 23
28801 Alcalá de Henares
E – Madrid Spain
Tel: +34.1.885.6192/6206
Fax: +34.1.883.5010
E-mail: eagles@cervantes.es

2 Working group members

2.1 Core working group members

S. Armstrong	
ISSCO	
54 route des Acacias	Tel.: +41.22.705.7113
CH – 1227 Geneva	Fax.: +41.22.300.1086
Switzerland	E-mail: susan@divsun.unige.ch
N. Ide	
Laboratoire Parole et Langage	
URA 216 CNRS	
29, Avenue Robert Schuman	Tel: +33.42.20.58.66
F – 13621 Aix-en-Provence Cedex 1	Fax: +33.42.20.59.05
France	E-mail: ide@univ-aix.fr
G. Leech	
Department of Linguistics and Modern English Language	
Lancaster University	
Lancaster	Tel.: +44.1524.593036
United Kingdom	Fax.: +44.1524.843085
LA1 4YT	E-mail: G.N.Leech@lancaster.ac.uk
J. Llisterri	(Group Editor & Host)
Instituto Cervantes	
Libreros, 23	
28801 Alcalá de Henares	Tel: +34.1.885.6192/6206
E – Madrid	Fax: +34.1.883.5010
Spain	E-mail: eagles@cervantes.es
F.A. Marcos Marín	
Laboratorio de Lingüística Informática	
Departamento de Lingüística	
Ap. 46348	Tel: +34.1.397.4529/4109/5250
E – 28080 Madrid	Fax:+34.1.535.2463
Spain	E-mail: marcos@vm1.sdi.uam.es

J. McH. Sinclair
Corpus Linguistics Group
School of English
Birmingham University
Westmere Mews
50 Edgbaston Park Road
Birmingham
United Kingdom
B15 2TT
Tel.: +44.121.414.6789
Fax.: +44.121.414.3288
E-mail: J.Sinclair@bham.ac.uk

J. Véronis
Laboratoire Parole et Langage
URA 216 CNRS
29, Avenue Robert Schuman
F – 13621 Aix-en-Provence Cedex 1
France
Tel: +33.42.20.58.66
Fax: +33.42.20.59.05
E-mail: veronis@univ-aix.fr

A. Zampolli (Chairperson)
ILC–CNR
Via della Faggiola 32
I – 56100 Pisa
Italy
Tel: +39.50.560.481
Fax: +39.50.589.055
E-mail: glottolo@ilc.pi.cnr.it

2.2 Subgroup members

2.2.1 Task 1: Typology

Responsible: J. Sinclair
W. Teubert (Institut für Deutsche Sprache, D)
P. van Sterkenberg (Instituut voor Nederlandse Lexicologie, NL)

2.2.2 Task 2: Corpus representation and encoding

Responsible: N. Ide
L. Romary (Centre de Recherche en Informatique de Nancy, F)
J. Véronis (Laboratoire Parole et Langage, CNRS, F)

2.2.3 Task 3: Linguistic annotation

Responsible: G. Leech
Morphosyntactic Annotation:
G. Arrarte (Instituto Cervantes, E)
N. Calzolari (Istituto di Linguistica Computazionale, CNR, I)
J-M. Langé (IBM Paris, F)
W. Paprotté (Universität Münster, D)
Syntactic Annotation:
H. van Halteren, (University of Nijmegen, NL)
J-M. Langé (IBM Paris, F)
S. Montemagni (Istituto di Linguistica Computazionale, CNR, I)
A. Voutilainen (University of Helsinki, FI)

2.2.4 Task 4: Documentation

Responsible: F. Marcos Marín

2.2.5 Task 5: Tools

Responsible: J. Véronis
E. Atwell (University of Leeds, UK)
T. McEnery (University of Lancaster, UK)
T. Erjavec (Jozef Stefan Institute, Slovenia)
G. Francopoulo (GSI-ERLI Paris, F)
H. van Halteren (University of Nijmegen, NL)
N. Ide (Vassar College, USA & CNRS, F)
P. Paroubek (INaLF, F)
V. Petkevic (Charles University, Czech Republic)
L. Romary (Centre de Recherche en Informatique de Nancy, F)

2.2.6 Task 6: Multilingual corpora

Responsible: S. Armstrong

2.3 Cross-group on spoken texts

A cross-group has been set up with members drawn from the Corpora WG and from the Spoken Language WG to address the specific problems of spoken texts:
Responsible: J. Llisterri
W. Barry (Universität des Saarlandes, D) from the SLWG
J. Sinclair (University of Birmingham, UK) from the CWG
H. Tillmann (Universität München) from the SLWG

3 What is a corpus?

In the EAGLES recommendations on corpus typology (EAGLES, 1996e), a *corpus* is defined as:

> *Corpus: A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.*

Words such as 'collection' and 'archive' refer to sets of texts that do not need to be selected, or do not need to be ordered, or the selection and/or ordering do not need to be on *linguistic criteria*, They are therefore quite unlike corpora.
Linguistic criteria to be applied to the selection and ordering may be:

External — in that they concern the participants, the occasion, the social setting or the communicative function of the pieces of language;

Internal — in that they concern the recurrence of language patterns within the pieces of language.

These criteria are reviewed in more detail in the recommendations on corpus typology (EAGLES, 1996e) where a classification of different types of corpora can also be found.
Since this document is devoted to *computer corpora*, it is appropriate to start by the definition also proposed in the above document:

> *Computer corpus: a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks.*

4 Written data, spoken data and speech data

Language corpora can comprise only written, only spoken, or both written and spoken data.
Spoken data can be collected for a variety of R&D tasks. We distinguish here between two different approaches:

1. The approach of the R&D speech community: the central issue is the speech signal itself, and its acoustic and articulatory properties. The symbolic representation is usually made by means of a *phonetic alphabet*. We refer to data of this type as *speech data*, and we call the deliberately designed collection of speech data *speech corpora* or *speech databases*.

2. The approach of corpus linguists and the NLP community: the central issue is the use of languages and the analysis of the data at various linguistic levels of description, i.e. lexical, morphosyntactic, syntactic, etc. The representation is usually made by means of a *phonological alphabet*, or of an *enriched orthographical system*. We use the term *spoken corpora* to refer to such corpora.

The document on spoken texts (EAGLES, 1996f) points out in detail the differences and similarities between the two approaches.
In EAGLES, speech corpora are dealt with by the Spoken Language Systems Working Group.
The EAGLES Text Corpora Working Group aims at providing guidelines for encoding both written and spoken data: usually, in fact, large language corpora comprise both types of data.
However, recognising the fact that spoken data increasingly represent a point of convergence of the interests of both the NLP and speech communities, the provision of guidelines for encoding of spoken data has been assigned to a subgroup for spoken corpora, formed by members of the two Working Groups (TCWG and SLWG).

5 Organisation of work

The activities of the Text Corpora Working Group have been organised into different tasks. A core group member is responsible for each task and, if necessary, for the constitution of a subgroup to perform it. The documents in this report are the result of the work of task responsibles and of the contribution of the corresponding subgroup. They also incorporate the results of the feedback received after initial dissemination.
A final workshop took place in Madrid on January 18–20, 1996, to which 40 european corpus experts were invited. Sessions were organised for each of the Working Group tasks, in which preliminary versions of the final documents were presented. The results of the discussions and the feedback received after the workshop have also been a useful input to the actual contents of the final report.

6 About the documents of the Text Corpora Working Group

The present set of reports from the EAGLES Text Corpora Working Group represents part of the work done by the group from February, 1993 to May, 1996.
The documents included here are of very different nature, reflecting in some cases the maturity and the previous work in the domains addressed; some of them contain either preliminary or final recommendations, while others provide general design principles or methodology proposals, or include significant background information in the form of extensive specificational materials or surveys that point to other sources of information.
Reports have been written for the tasks for which the state-of-the-art allowed the subgroup to provide at least preliminary recommendations:

- Corpus and text typology;
- Corpus encoding;
- Linguistic annotation;

- Tools; and
- Spoken texts.

6.1 Corpus and text typology

Two complementary reports are presented for this task: one concerned with corpus typology (EAGLES, 1996e) and a second one dealing with text typology (EAGLES, 1996g).
The document on *Corpus typology* starts with definitions that aim at distinguishing what is a corpus from what should not be considered as such. Then several characteristics are considered as a basis for a distinction between a corpus and a *special corpus*: quantity, quality, simplicity and documentation. Finally, specific types of collections like spoken corpora, samples, sublanguage corpora, reference corpora, monitor corpora, parallel and comparable corpora are defined.
The document on *Text typology*, to be read in conjunction with the previous report, distinguishes internal and external criteria for the classification of texts included in a corpus, and discusses in detail notions such as *topic* and *genre*. In view of both types of criteria, a typology of texts is proposed. Topic and *style* are also extensively discussed as part linguistic internal criteria. Spoken language corpora are also addressed. Finally, an Appendix is included which shows the common features of different classification systems in European corpora. This Appendix should be read together with the main document.

6.2 Corpus encoding

The document on preliminary recommendations for corpus encoding (EAGLES, 1996a) is a first version developed in collaboration with the LRE MULTEXT project. The work is aimed at providing a set of encoding standards for corpus-based work in natural language processing applications, and is compliant with the Text Encoding Initiative. It involves the specification of

> *a minimal encoding level that corpora must achieve to be considered standardised in terms of descriptive representation (marking of structural and typographic information) as well as general architecture (so as to be maximally suited for use in a text database.*

General principles and recommendations common to all documents constitute the first two parts of this report. Special attention is paid to the corpus encoding header in part 3, and part 4 offers a full definition of three levels of conformance in the encoding of primary data.
Provisions for encoding of linguistic annotation and a data architecture for linguistic corpora are also found in this document.
A full set of Appendices contains useful complementary information.

6.3 Linguistic annotation

Morphosyntactic annotation is addressed as part of the linguistic annotation task (EAGLES, 1996c). This report constitutes a proposal which results from intensive collaboration with the EAGLES Computational Lexicons Working Group.
Recommendations for three different degrees of standardisation are discussed and recommendations for morphosyntactic categories are presented. An intermediate tagset with obligatory, recommended and optional attributes or values is described in the body of the report. An Appendix containing an English and an Italian tagset mapped onto the intermediate tagset for morphosyntactic annotation is included.
The document on syntactic annotation (EAGLES, 1996d) is structured in 5 sections plus an Appendix in which the annotation scheme proposed is illustrated with small text samples from Dutch, English, Spanish and German.
After an introductory section in which definitions, goals and several conceptual distinctions are discussed, section 3 addresses the different layers of annotation. The guidelines are found in section 4, where obligatory and recommended annotations and optional categories are proposed; this is complemented by the discussion of some of the issues in the practical application of the scheme in section 5.

6.4 Tools

Work on corpus tools in EAGLES has been jointly developed with the LRE MULTEXT project and is presented in two documents, one discussing the reusability of linguistic software (EAGLES, 1996h), the other giving guidelines for linguistic software development (EAGLES, 1996b).
In the first document, benefits derived from the standardisation of linguistic software are presented in the first place. A discussion of the main aspects involved in software reusability leads to the establishment of the general principles for the definition of a coherent environment for the development of linguistic software. This environment is presented in the second document. In its present version, principles and guidelines are summarised, and certain parts of the environment are already developed: character sets, representation of textual data and of the linguistic annotation, programming in C and consistency in the command line interface.

6.5 Spoken texts

The document on spoken texts (EAGLES, 1996f) is devoted to the specific problems that arise when spoken language is represented in written form, and is closely related to several parts of the report from the EAGLES Spoken Language Working GroupWorking Group.
The report starts by discussing and comparing the needs of the speech community and the corpus linguistics community in terms of materials and transcription. Existing transcription and representation practices for spoken texts are reviewed, with special attention to the proposals of the Text Encoding Initiative (TEI) and the Network of European Reference Corpora (NERC). Levels of transcription are also discussed.
The other sections are devoted to recommendations based on the previous review concerning data acquisition, events to be encoded in a spoken text, orthographic representation and segmental and suprasegmental transcription.

7 Conclusion

Taken together, these documents represent the present state of the work carried out by the EAGLES Text Corpora Working Group. However, many issues are still unresolved and need further discussions within the whole of the corpus linguistics community. It is hoped that the public dissemination of this report will create new opportunities for collaboration and will contribute, in the long term, to the creation of language resources that can be shared and used for the advancement of our field.

8 References

EAGLES (1996a) Recommendations on corpus encoding. EAG–TCWG–CES/R-F. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.
EAGLES (1996b) Guidelines for linguistic software development. EAG–TCWG–LSD2/B. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.
EAGLES (1996c) Recommendations for the morphosyntactic annotation of corpora. EAG–TCWG–MAC/R. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.
EAGLES (1996d) Preliminary recommendations for the syntactic annotation of corpora. EAG–TCWG–SASG1/P-B. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.
EAGLES (1996e) Preliminary recommendations on corpus typology. EAG–TCWG–CTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.
EAGLES (1996f) Preliminary recommendations on spoken texts. EAG–TCWG–SPT/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.
EAGLES (1996g) Preliminary recommendations on text typology. EAG–TCWG–TTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.
EAGLES (1996h) Reusability of linguistic software. EAG–TCWG–LSD1/B. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.