

LINGUISTICA
COMPUTAZIONALE

VOLUME XI

TOWARDS A NETWORK
OF EUROPEAN
REFERENCE CORPORA

REPORT OF THE NERC CONSORTIUM
FEASIBILITY STUDY

COORDINATED BY ANTONIO ZAMPOLLI

EDITED BY:
NICOLETTA CALZOLARI, MONA BAKER,
JOHANNA G. KRUYT



GIARDINI EDITORI
E STAMPATORI
IN PISA

Linguistica Computazionale is published twice (two separate issues) per year.
Subscriptions should be sent to the Publisher: *Giardini editori e stampatori in Pisa*,
via delle Sorgenti 23, I-56010 Agnano Pisano (Pisa), Italy.
Tel. +39-50 934242 · Fax +39-50 934200.
Postal current account n. 12777561.

Subscription rates (per year): Lit. 50,000 individuals (US \$ 43 abroad) · Lit. 70,000
institutions (US \$ 57 abroad).

Direttore responsabile: Antonio Zampolli
Registrazione presso il Tribunale di Pisa n. 1 dell'11/03/1981

Direttore responsabile: Antonio Zampolli
Registrazione presso il Tribunale di Pisa n. 1 dell'11/03/1981

TABLE OF CONTENTS

Introduction	XI
1. Corpus linguistics, the use of computer in the humanities, computational linguistics	XI
2. The emergence and the evolution of the concept of reusable Linguistic Resources	XVIII
3. The NERC Consortium and the NERC feasibility study	XXII
4. Conclusion	XXX
5. Acknowledgements	XXXI
CHAPTER 0 – IMPLEMENTATION PLAN	1
1. Introduction	1
2. Proposals: Corpora	2
3. General Principles for Organization of the Work	11
4. NERC Implementation Plan: Organizational and Managerial Aspects	12
5. Proposals: Computational Lexicons	27
6. Conclusion	28
APPENDIX 1	30
APPENDIX 2	32
CHAPTER 1 – USER NEEDS	37
1. Introduction	37
2. A Description of Information Sources	40
3. Recommendations	54
CHAPTER 2 – CORPUS DESIGN CRITERIA	57
1. The problem	57
2. Investigations	58
3. Main results	59
4. Conclusions	61
5. Recommendations	62

CHAPTER 3 – TEXT REPRESENTATION: WRITTEN LANGUAGE/SPOKEN LANGUAGE	73
A. Written Language	73
1. Introduction	73
2. Why the SGML standard?	74
3. The Text Encoding Initiative	75
4. Definition of a minimal level of text representation for European Corpora	77
5. Conclusion and Recommendations	83
B. Spoken Language	85
1. Introduction	85
2. Organisation of the Chapter	86
3. Speech Community	87
4. The Text Encoding Initiative	88
5. Transcription Conventions	90
APPENDIX A	92
APPENDIX B	96
CHAPTER 4 – TEXT ACQUISITION AND REUSABILITY ACCESS AND MANAGEMENT SOFTWARE TOOLS	107
A. Text Acquisition and Reusability	107
1. Introduction	107
2. SGML Retroconversion Techniques	108
3. Main Acquisition Methods	110
4. Conclusion and Recommendations	114
B. Access and Management Software Tools	116
1. Introduction	116
2. Basic Access Software	118
3. Corpus Maintenance, Development and Availability	122

1. Introduction	12
2. Phonetic/Phonemic and Prosodic Annotation	13
3. Morphosyntactic Annotation	13
4. Syntactic Annotation	15
5. Annotation beyond the Syntactic	17
CHAPTER 6 – CORPUS ANNOTATION TOOLS	17
1. Introduction	17
2. Current software tools: Parsers etc.	17
3. Current software tools: Applications	18
4. New corpus software tools: grammatical analysis	18
4.1. Lemmatisation	18
4.2. Tagging	18
4.3. Parsing	18
5. Lexical tools	18
6. Lexicogrammar	18
7. Multilingual software	18
CHAPTER 7 – KNOWLEDGE EXTRACTION	19
1. Introduction	19
1.1. The relation between corpus-based and rule-based work in NLP	19
1.2. Linguistic knowledge acquisition: a major bottleneck in NLP	19
1.3. On the impact of corpus-based work in Linguistics and NLP applications	19
1.4. Three knowledge acquisition models	19
1.5. Structure of the chapter	19
2. Methodologies of Linguistic Knowledge Extraction	19
3. Applications	20
GENERAL NERC BIBLIOGRAPHY	22