

THE OXFORD HANDBOOK OF

COMPUTATIONAL
LINGUISTICS

Edited by

RUSLAN MITKOV

OXFORD
UNIVERSITY PRESS

OXFORD

UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi São Paulo Shanghai Taipei Tokyo Toronto

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

© editorial matter and organization Ruslan Mitkov 2003
© chapters their several authors 2003

The moral rights of the authors have been asserted
Database right Oxford University Press (maker)

First published 2003

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organizations. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

ISBN 0-19-823882-7

1 3 5 7 9 10 8 6 4 2

Typeset in Minion
by Peter Kahrel, Lancaster
Printed in Great Britain
on acid-free paper by

Biddles Ltd., Guildford & King's Lynn

CONTENTS

<i>Preface</i>	ix
RUSLAN MITKOV	
<i>Abbreviations</i>	xi
<i>Introduction</i>	xvii
MARTIN KAY	

PART I FUNDAMENTALS

1. Phonology	3
STEVEN BIRD	
2. Morphology	25
HARALD TROST	
3. Lexicography	48
PATRICK HANKS	
4. Syntax	70
RONALD M. KAPLAN	
5. Semantics	91
SHALOM LAPPIN	
6. Discourse	112
ALLAN RAMSAY	
7. Pragmatics and Dialogue	136
GEOFFREY LEECH AND MARTIN WEISSER	

8. Formal Grammars and Languages 157
CARLOS MARTÍN-VIDE
9. Complexity 178
BOB CARPENTER

PART II PROCESSES, METHODS, AND RESOURCES

10. Text Segmentation 201
ANDREI MIKHEEV
11. Part-of-Speech Tagging 219
ATRO VOUTILAINEN
12. Parsing 233
JOHN CARROLL
13. Word–Sense Disambiguation 249
MARK STEVENSON AND YORICK WILKS
14. Anaphora Resolution 266
RUSLAN MITKOV
15. Natural Language Generation 284
JOHN BATEMAN AND MICHAEL ZOCK
16. Speech Recognition 305
LORI LAMEL AND JEAN-LUC GAUVAIN
17. Text-to-Speech Synthesis 323
THIERRY DUTOIT AND YANNIS STYLIANOU
18. Finite-State Technology 339
LAURI KARTTUNEN
19. Statistical Methods 358
CHRISTER SAMUELSSON

20. Machine Learning	376
RAYMOND J. MOONEY	
21. Lexical Knowledge Acquisition	395
YUJI MATSUMOTO	
22. Evaluation	414
LYNETTE HIRSCHMAN AND INDERJEET MANI	
23. Sublanguages and Controlled Languages	430
RICHARD I. KITTREDGE	
24. Corpus Linguistics	448
TONY MCENERY	
25. Ontologies	464
PIEK VOSSEN	
26. Tree-Adjoining Grammars	483
ARAVIND K. JOSHI	

PART III APPLICATIONS

27. Machine Translation: General Overview	501
JOHN HUTCHINS	
28. Machine Translation: Latest Developments	512
HAROLD SOMERS	
29. Information Retrieval	529
EVELYNE TZOUKERMANN, JUDITH L. KLAVANS, AND TOMEK STRZALKOWSKI	
30. Information Extraction	545
RALPH GRISHMAN	
31. Question Answering	560
SANDA HARABAGIU AND DAN MOLDOVAN	

32. Text Summarization	583
EDUARD HOVY	
33. Term Extraction and Automatic Indexing	599
CHRISTIAN JACQUEMIN AND DIDIER BOURIGAULT	
34. Text Data Mining	616
MARTI A. HEARST	
35. Natural Language Interaction	629
ION ANDROUTSOPOULOS AND MARIA ARETOULAKI	
36. Natural Language in Multimodal and Multimedia Systems	650
ELISABETH ANDRÉ	
37. Natural Language Processing in Computer-Assisted Language Learning	670
JOHN NERBONNE	
38. Multilingual On-Line Natural Language Processing	699
GREGORY GREFFENSTETTE AND FRÉDÉRIQUE SEGOND	
<i>Notes on Contributors</i>	717
<i>Glossary</i>	727
<i>Index of Authors</i>	761
<i>Subject Index</i>	775

INTRODUCTION

MARTIN KAY

Computational Linguistics is about as robust a field of intellectual endeavour as one could find, with its books, journals, conferences, professorial chairs, societies, associations and the like. But, of course, it was not always so. Computational Linguistics crept into existence shyly, almost furtively. When shall we say it all began? Perhaps in 1949, when Warren Weaver wrote his famous memorandum suggesting that translation by machine might be possible. The first conference on machine translation took place at MIT in 1952 and the first journal, *Mechanical Translation*, began in 1954. However, the phrase ‘Computational Linguistics’ started to appear only in the mid-1960s. The journal changed its name to *Mechanical Translation and Computational Linguistics* in 1965 but the words ‘and Computational Linguistics’ appeared in very small type. This change coincided with the adoption of the journal by the Association for Machine Translation and Computational Linguistics, which was formed in 1962.

The term ‘Computational Linguistics’ was probably coined by David Hays during the time that he was a member of the Automatic Language Processing Advisory Committee of the National Academy of Sciences. The publication of this committee’s final report, generally known as the ALPAC report, certainly constituted one of the most dramatic moments in the history of the field—proposing, as it did, that machine translation be abandoned as a short-term engineering goal in favour of more fundamental scientific research in language and language processing. Hays saw this coming and realized that, if the money that had been flowing into machine translation could be diverted into a new field of enquiry, the most pressing requirement was for the field to be given a name. The name took hold. Redirection of the funds did not.

Progression from machine translation to Computational Linguistics occurred in 1974 when *Machine Translation and Computational Linguistics* was replaced by the *American Journal of Computational Linguistics*, which appeared initially only in microfiche form. In 1980, this became *Computational Linguistics*, which is still alive and vigorous today.

By the 1980s, machine translation began to look practical again, at least to some people and for some purposes and, in 1986, the circle was completed with the publication of the first issue of *Computers and Translation*, renamed *Machine Translation* in 1988. The *International Journal of Machine Translation* followed in 1991.

Warren Weaver's vision of machine translation came from his war-time experience as a cryptographer and he considered the problem to be one of treating textual material, by fundamentally statistical techniques. But the founders of Computational Linguistics were mostly linguists, not statisticians, and they saw the potential of the computer less in the possibility of deriving a characterization of the translation relation from emergent properties of parallel corpora, than in carrying out exactly, and with great speed, the minutely specified rules that they would write. Chomsky's *Syntactic Structures* (1957) served to solidify the notion of grammar as a deductive system which therefore seemed eminently suited to computer applications. The fact that Chomsky himself saw little value in such an enterprise, or that the particular scheme of axioms and rules that he advocated was ill suited to the automatic analysis of text, did nothing to diminish the attractiveness of the general idea.

Computational Linguistics thus came to be an exercise in creating and implementing the formal systems that were increasingly seen as constituting the core of linguistic theory. If any single event marks the birth of the field, it is surely the proposal by John Cocke in 1960 of the scheme for deriving all analyses of a string with a grammar of binary context-free rules that we now know as the Cocke-Kasami-Younger algorithm. It soon became clear that more powerful formalisms would be required to meet the specific needs of human language, and more general chart parsers, augmented transition networks, unification grammars, and many other formal and computational devices were created.

There were two principal motivations for this activity. One was theoretical and came from the growing perception that the pursuit of computational goals could give rise to important advances in linguistic theory. Requiring that a formal system be implementable helped to ensure its internal consistency and revealed its formal complexity properties. The results are to be seen most clearly in syntactic formalisms such as Generalized Phrase Structure Grammar, Lexical Functional Grammar, and Head Driven Phrase Structure as well as in application of finite-state methods to phonology and morphology.

The second motivation, which had existed from the beginning, came from the desire to create a technology, based on sound scientific principles, to support a large and expanding list of practical requirements for translation, information extraction, summarization, grammar checking, and the like. In none of these enterprises is success achievable by linguistic methods alone. To varying extents, each involves language not just as a formal system, but as a means of encoding and conveying information about something outside, something which, for want of a better term, we may loosely call 'the world'. Much of the robustness of language comes from the imprecision and ambiguity which allow people to use it in a casual manner. But this works only because people are able to restore missing information and resolve ambiguities on the basis of what makes sense in a larger context provided not only by the surrounding words but by the world outside. If there is any field that should be responsible for the construction of comprehensive, general models of the world, it

is presumably artificial intelligence, but the task is clearly a great deal more daunting even than building comprehensive linguistic models, and success has been limited.

As a result, Computational Linguistics has gained a reputation for not measuring up to the challenges of technology, and this in turn has given rise to much frustration and misunderstanding both within and outside the community of computational linguists. There is, of course, much that still remains to be done by computational linguists, but very little of the responsibility for the apparently poor showing of the field belongs to them. As I have said, a significant reason for this is the lack of a broader technological environment in which Computational Linguistics can thrive. Lacking an artificial intelligence in which to embed their technology, linguists have been forced to seek a surrogate, however imperfect, and many think they have found it in what is generally known as 'statistical natural language processing'.

Roughly speaking, statistical NLP associates probabilities with the alternatives encountered in the course of analysing an utterance or a text and accepts the most probable outcome as the correct one. In 'the boy saw the girl with the telescope', the phrase 'with the telescope' is more likely to modify 'saw' than 'the girl', let us say, because 'telescope' has often been observed in situations which, like this one, represent it as an instrument for seeing. This is an undeniable fact about seeing and telescopes, but it is not a fact about English. Not surprisingly, words that name phenomena that are closely related in the world, or our perception of it, frequently occur close to one another so that crisp facts about the world are reflected in somewhat fuzzier facts about texts.

There is much room for debate in this view. The more fundamentalist of its proponents claim that the only hope for constructing useful systems for processing natural language is to learn them entirely from primary data as children do. If the analogy is good, and if Chomsky is right, this implies that the systems must be strongly predisposed towards certain kinds of languages because the primary data provides no negative examples and the information that it contains occurs, in any case, in too weak dilution to support the construction of sufficiently robust models without strong initial constraints.

If, as I have suggested, text processing depends on knowledge of the world as well as knowledge of language, then the proponents of radical statistical NLP face a stronger challenge than Chomsky's language learner because they must also construct this knowledge of the world entirely on the basis of what they read about it, and in no way on the basis of direct experience. The question that remains wide open is: Just how much of the knowledge of these two kinds that is required for NLP is derivable, even in principle, from emergent properties of text? The work done over the next few years should do much to clarify the issue and thus to suggest the direction that the field will follow thereafter.

This book stands on its own in the sense that it will not only bring people working in the field up to date on what is going on in parallel specialities to their own, but also introduce outsiders to the aims, methods, and achievements of computational

linguists. The chapters of Part I have the same titles that one might expect to find in an introductory text on general linguistics. With the exception of the last, they correspond to the various levels of abstraction on which linguists work, from individual sounds to structures that span whole texts or dialogues, to the interface between meaning and the objective world, and the making of dictionaries. The difference, of course, is that they concentrate on the opportunities for computational exploration that each of these domains opens up, and on the problems that must be solved in each of them before they can contribute to the creation of linguistic technology.

I have suggested that requiring a formal system to be implementable led linguists to attend to the formal complexity properties of their theories. The last chapter of Part I provides an introduction to the mathematical notion of complexity and explores the crucial role that it plays in Computational Linguistics.

Part II of the book gives a chapter to each of the areas that have turned out to be the principal centres of activity in the field. For these purposes, Computational Linguistics is construed very broadly. On the one hand, it treats speech recognition and text-to-speech synthesis, the fundamentals of which are more often studied in departments of electrical engineering than linguistics and on the other, it contains a chapter entitled 'Corpora', an activity in which students of language use large collections of text or recorded speech as sources of evidence in their investigations. Part III is devoted to applications—starting, as is only fitting, with a pair of chapters on machine translation followed by a discussion of some topics that are at the centre of attention in the field at the present.

It is clear from the table of contents alone that, during the half century in which the field, if not the name, of Computational Linguistics has existed, it has come to cover a very wide territory, enriching virtually every part of theoretical linguistics with a computational and a technological component. However, it has been only poorly supplied with textbooks or comprehensive reference works. This book should go a long way towards meeting the second need.